読み時間と統語・意味分類
Between Reading Time and Syntactic/Semantic Categories

浅原 正幸†，加藤 祥†
Masayuki Asahara, Sachi Kato
† 人間文化研究機構 国立国語研究所
National institute for Japanese Language and Linguitics, Japan
masayu-a@ninjal.ac.jp

Abstract

This article presents the contrastive analysis between reading time and syntactic/semantic categories in Japanese We overlaid the reading time annotation BCCWJ-EyeTrack and a syntactic/semantic category information annotation on the 'Balanced Corpus of Contemporary Written Japanese'. Statistical analysis based on a mixed linear model showed that verbal phrase tends to be shorter reading time than adjective, adverbial phrases or nominal phrases. The results suggest that the preceding phrases associated with the presenting phrases promotes the reading process to shorten the gazing time.

Keywords — Reading Time, Syntactic Category, Semantic Category, Readability

1. はじめに

従来の文処理研究は，仮説を立てたうえで適切な作例を作成し，作例に対する被験者の読み時間を検証する確認的データ分析 (Confirmatory Data Analysis) により進められてきた。以下では『現代日本語書き言葉均衡コーパス』[10](以下 BCCWJ) に対する読み時間アノテーション BCCWJ-EyeTrack [3] と分類語彙表番号アノテーション [16] を重ね合わせ，探索的データ分析 (Exploratory Data Analysis) により，統語・意味分類が読み時間に与える影響について検討を試みる。

前者の読み時間アノテーションは，BCCWJ 新聞記事コアデータ 21 記事を刺激として，日本語母語話者 24 人分の読み時間を収集したものである。自己ペース読文法に基づく SELF データと，視線走査法を単語出現順に集計しなおした FFT(First Fixation)・FPT(First-Pass)・SPT(Second-Pass)・RPT(Regresion Path)・TOTAL データの 6 種類からなる。

後者の分類語彙表番号アノテーションは，BCCWJ の短単位と長単位に対して語義の曖昧性を人手で解消しながら国立国語研究所で整備されている分類語彙表 [18, 19] の分類番号を付与したものを，文節単位に写像（文節最右要素もしくは文節に含まれる要素）したうえで分析する。

この 2 つのデータの重ね合わせを行い，被験者と呈示サンプルをランダム効果とした，線形混合モデルによる対照比較を行った。

以下 2 節では利用するデータである BCCWJ-EyeTrack と BCCWJ に対する分類語彙表アノテーションについて説明する。3 節では統計分析手法について説明する。4 節では結果と考察を示す。5 節にまとめと今後の研究の方向性について示す。

2. 利用するデータ

2.1 BCCWJ

利用するデータは『現代日本語書き言葉均衡コーパス』(BCCWJ)[10] とそれに対する各種アノテーションである。ここでは BCCWJ について説明する。BCCWJ は，現代日本語の書き言葉を適切なサンプリング手法で集積した均衡コーパスである。このうちコアデータは人手による形態論情報 (短単位・長単位・文節境界) が付与されている。

本研究ではコアデータのうちの生産実態に基づきサンプリングされた新聞記事サンプル (PN_core) を用いる。新聞記事サンプルの一部には，以下に述べる読み時間の情報アノテーション (BCCWJ-EyeTrack) と分類語彙表番号アノテーションが付与されている。

2.2 BCCWJ-EyeTrack

BCCWJ-EyeTrack [3](表 1) は， BCCWJ の新聞記事サンプルに読み時間データを自己ペース読文法と視線走査法により，日本語母語話者 24 人分の読み時間を付与したものである。以下，データの詳細について説明する。

表 1　データ形式

列名	データ型	摘要
surface	factor	出現書字形
time	int	読み時間
logtime	num	読み時間 (常用対数)
measure	factor	読み時間の種類
sample	factor	サンプル名
article	factor	記事情報
metadata_orig	factor	文書構造タグ
metadata	factor	メタデータ
length	int	文字数
space	factor	文節境界空白の有無
subj	factor	実験協力者 ID
setorder	factor	文節境界空白の呈示順
dependent	int	係り受け関係
sessionN	int	セッション順
articleN	int	記事呈示順
screenN	int	画面呈示順
lineN	int	行呈示順
segmentN	int	文節呈示順
is_first	factor	最左要素
is_last	factor	最右要素
is_second_last	factor	右から 2 つ目の要素

自己ペース読文法は，他の文節をマスクしたうえで 1 文節単位を逐次的に呈示する読み時間測定手法である。読み戻しができないため，文節単位の読み時間がそのままデータとなる。このデータを SELF と呼ぶ。

視線走査法で取得したオリジナルのデータから文字の半角単位に Start Fixation Time （注視開始時刻）と End Fixation Time （注視終了時刻）と Fixation Time （注視時間）を得る。このデータを国語研文節単位でグループ化しなおした注視順データを集計して，テキスト生起順データに加工する。テキスト生起順データは以下の 5 種類からなる。

- First Fixation Time (FFT)
- First-Pass Time (FPT)
- Regression Path Time (RPT)
- Second-Pass Time (SPT)
- Total Time (TOTAL)

説明のために図 1 の例を用いる。

First Fixation Time (FFT) は注視範囲に視線が 1 回目に停留した注視時間である。例中の「初年度決算も」の FFT は 5 の注視時間となる。

First-Pass Time (FPT) は，注視範囲に視線が 1 回目に停留し注視範囲から出るまでの総注視時間である。出る方向は右方向でも左方向でも構わない。例中の「初年度決算も」の FPT は 5, 6 の注視時間の合計である。

Regression Path Time (RPT) は，注視範囲に視線が 1 回目に停留し，注視範囲に再度停留して次に右切片から出るまでの総注視時間である。左側に戻る場合には再度注視範囲に戻るまで合算する。例中の「初年度決算も」の RPT は 5, 6, 7, 8, 9 の注視時間の合計である。左側に戻っても再度注視範囲に停留しない場合は合算しない。例中「上回り,」の RPT は 4 の注視時間である。

表 2　分類番号の構造 「この」(分類番号: 3.1010)

類	部門	中項目	分類項目
相 (3)	関係 (.1)	真偽 (.10)	こそあど (.1010)

Second-Pass Time (SPT) は，注視範囲に 1 回視線が停留し，注視範囲から出たあと，2 回目以降に注視範囲に停留する総注視時間である。例中の「初年度決算も」の RPT は 9, 11 の注視時間の合計である。尚，FPT + SPT が次に説明する Total Time になる。

Total Time (TOTAL) は注視範囲に視線が停留する総注視時間である。例中「初年度決算も」の RPT は 5, 6, 9, 11 の注視時間の合計である。

テキスト生起順データにおいて，サッケードの時間は集計しない。

これらの読み時間情報 (time, logtime) に対して，出現書字形 (surface)・記事情報 (sample, article)・文書構造 (metadata_orig, metadata) のほか，出現書字形文字数 (length)，文節単位の空白の有無 (space)，実験協力者 ID (subj)，係る文節数 (dependent)，実験協力者ごとの呈示順序 (sessionN, setorder, articleN, screenN, lineN, segmentN) ，画面水平方向の位置 (is_first,is_last,is_second_first) を付与したデータを分析に用いる。係る文節数は BCCWJ-DepPara [1] のものを用いる。

2.3　BCCWJ に対する分類語彙表番号アノテーション

『分類語彙表』[18] は「語を意味によって分類・整理したシソーラス（類義語集）」である。初版はおよそ 33,000 語を収録していたが,『分類語彙表–増補改訂版–』[19] は区切り文字を含めて 101,070 件からなる。本研究では分類語彙表増補改訂版の CSV データ [1] を用いる。

『分類語彙表』は表 2 に示す分類番号を用いて，単語の分類項目の体系的位置づけを行う。分類番号は，1-4 の最初の 1 桁が「類」と呼ばれ，品詞 (統語的分類) を表す。1 が名詞の仲間である体の類を，2 が動詞の仲間である用の類を，3 が形容詞・形容動詞・副詞・

[1] http://pj.ninjal.ac.jp/corpus_center/archive.html#bunruidb

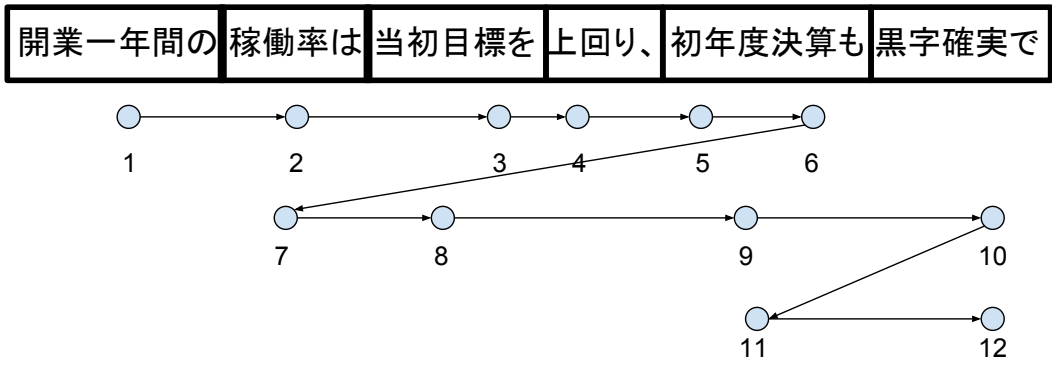

図 1　読み時間の集計方法

連体詞などの仲間である相の類を，4 が接続詞・感動詞などのその他の類を表す。ピリオドをはさんで 4 桁からなる数値が意味分類を表す。意味分類のうち 1 桁目は「部門」と呼ばれ，.1 が抽象的関係 (関係) を，.2 が人間活動の主体 (主体) を，.3 が精神および行為 (活動) を，.4 が生産物および用具 (生産物) を，.5 が自然物および自然現象 (自然) を表す。また，意味分類のうち 2 桁目までを「中項目」と呼び，4 桁目までを「分類項目」と呼ぶ。さらに，増補改訂版より分類項目の下位分類として,「段落」が定義されている。

BCCWJ のコアデータの一部に対して『分類語彙表』の分類番号を付与する作業が進められている [16]。分類語彙表を人手で UniDic の語彙素番号に対応させたデータ [17] により同データの短単位と長単位の両方について，可能な分類番号を枚挙し，人手で語義の曖昧性解消を行うとともに，未定義の部分に追加して分類番号を付与する。

本分析には BCCWJ に対する分類語彙表番号アノテーションデータの長単位データに基づき，文節内再右自立語の分類番号を分析対象とする。統語分類として「類」(WLSPLUWA と呼ぶ) を用い，意味分類として「部門」(WLSPLUWB と呼ぶ) を用いて，統計分析を行う。

3.　統計処理

まず，対象は BCCWJ-EyeTrack の全データとする。データの前処理として，`metadata` が {`authorsData`, `caption`, `listItem`, `profile`, `titleBlock`} のものを除外した。さらに視線走査実験結果の 0 (fixation がない対象) のデータポイントを除外した。

分析は常用対数時間に対して線形混合モデルに基づいて行う [4]。モデリングには R の lme4 パッケージを用いた。最初に一度モデル化したうえで，標準偏差 ± 3.0 を超えるデータポイントを除外した。`subj` と `article` をランダム切片として，次のような式に基づき分析を行った。なお，ランダム切片に対する係数の組み合わせによるモデル選択は行っていない。

```
logtime ~ space * sessionN + length + dependent
  + is_first + is_last + is_second_last
  + articleN + screenN + lineN + segmentN
  + WLSPLUWA + WLSPLUWB
  + (1 | subj) + (1 | article)
```

4.　結果と考察

4.1　結果

表 3 に結果を示す。

まず，分類語彙表番号以外の情報を確認する。

視線走査法においては FFT 以外のものについて，空白ありのほうが読み時間が短くなる。単純に読み時間を短くするという観点でリーダビリティを上げるには，文節間に空白を入れたほうがよい。文節長は FFT 以外について，長くなればなるほど読み時間が長くなる。これは，文節長が長くなればなるほど，表示面積が大きくなり，視線が停留する確率が線形に高くなるためだと考える。係り受けでは FFT 以外について，多くの係り受けがある文節ほど読み時間が短くなる。これは後に述べる Anti-locality 現象 [7] の追認である。レイアウト情報 (is_first, is_last, is_second_last) は，折り返しの視線移動に基づく影響を勘案するものである。 最左要素 (is_first) に関しては SPT 以外で読

表 3　線形混合モデルに基づく分析結果

	Dependent variable:					
	logtime					
	SELF	FFT	FPT	SPT	RPT	TOTAL
space=True	−0.001	−0.006	−0.017***	−0.039***	−0.018***	−0.029***
空白あり	(0.002)	(0.004)	(0.005)	(0.009)	(0.006)	(0.005)
length	0.086***	−0.003	0.135***	0.022***	0.115***	0.130***
文節長	(0.001)	(0.002)	(0.003)	(0.005)	(0.003)	(0.003)
dependent	−0.008***	−0.003	−0.016***	−0.016***	−0.012***	−0.018***
係り受け	(0.002)	(0.002)	(0.003)	(0.006)	(0.004)	(0.003)
is_first	0.052***	0.019***	0.090***	−0.027**	0.030***	0.069***
最左要素	(0.004)	(0.006)	(0.008)	(0.013)	(0.009)	(0.008)
is_last	0.033***	−0.009	0.014*	−0.052***	0.088***	−0.007
最右要素	(0.004)	(0.006)	(0.008)	(0.016)	(0.010)	(0.008)
is_second_last	−0.010***	−0.001	0.034***	−0.005	0.045***	0.034***
右から 2 番目の要素	(0.004)	(0.006)	(0.007)	(0.012)	(0.008)	(0.007)
sessionN	−0.022	−0.022	−0.041*	−0.036**	−0.049*	−0.047*
セッション順	(0.021)	(0.016)	(0.024)	(0.018)	(0.025)	(0.024)
articleN	−0.028***	−0.004	−0.005	−0.002	−0.007	−0.001
記事順	(0.005)	(0.004)	(0.007)	(0.007)	(0.007)	(0.008)
screenN	−0.029***	−0.004	−0.018***	−0.015***	−0.017***	−0.025***
画面順	(0.002)	(0.003)	(0.003)	(0.006)	(0.004)	(0.003)
lineN	−0.010***	−0.010***	−0.018***	−0.018***	−0.007**	−0.018***
行番号	(0.001)	(0.002)	(0.003)	(0.005)	(0.003)	(0.003)
segmentN	−0.004***	0.003***	−0.005***	−0.009***	−0.013***	−0.012***
セグメント番号	(0.001)	(0.001)	(0.001)	(0.002)	(0.002)	(0.001)
WLSPLUWA2	−0.047***	−0.038***	−0.096***	−0.029**	−0.088***	−0.101***
用の類	(0.004)	(0.006)	(0.007)	(0.014)	(0.009)	(0.008)
WLSPLUWA3	−0.036***	−0.003	−0.056***	−0.034*	−0.054***	−0.071***
相の類	(0.005)	(0.008)	(0.010)	(0.020)	(0.012)	(0.010)
WLSPLUWA4	−0.031*	−0.020	−0.127***	−0.238**	−0.137***	−0.189***
その他の類	(0.018)	(0.033)	(0.040)	(0.100)	(0.049)	(0.042)
WLSPLUWAFALSE	−0.030	0.020	−0.075	−0.031	−0.109	−0.160**
分類語彙表未登録語	(0.019)	(0.061)	(0.076)	(0.299)	(0.092)	(0.079)
WLSPLUWB.2	0.001	0.014**	0.018**	0.011	0.005	0.018**
主体	(0.004)	(0.006)	(0.007)	(0.013)	(0.009)	(0.008)
WLSPLUWB.3	−0.007**	0.015***	0.024***	0.012	0.021***	0.023***
活動	(0.003)	(0.005)	(0.006)	(0.011)	(0.007)	(0.006)
WLSPLUWB.4	0.017***	0.005	0.022*	0.009	0.018	0.037***
生産物	(0.007)	(0.010)	(0.013)	(0.021)	(0.015)	(0.013)
WLSPLUWB.5	0.014	0.034**	0.017	0.054	0.024	0.040**
自然	(0.010)	(0.015)	(0.019)	(0.034)	(0.023)	(0.020)
space1:sessionN	−0.016	0.044	0.059	0.060*	0.061	0.061
	(0.042)	(0.031)	(0.049)	(0.035)	(0.050)	(0.048)
Constant	2.790***	2.299***	2.532***	2.456***	2.603***	2.672***
	(0.022)	(0.017)	(0.026)	(0.023)	(0.027)	(0.026)
Observations	17,628	13,232	13,232	4,769	13,232	13,232
Log Likelihood	7,177.714	1,336.057	−1,506.614	−1,020.231	−4,082.304	−2,123.398
Akaike Inf. Crit.	−14,307.430	−2,624.114	3,061.228	2,088.462	8,212.607	4,294.796
Bayesian Inf. Crit.	−14,120.770	−2,444.344	3,240.998	2,243.739	8,392.377	4,474.565

Note: *p<0.1; **p<0.05; ***p<0.01

み時間が長くなり，最右要素やその隣の要素 (is_last, is_second_last) では FPT, RPT, Total などで読み時間が長くなる傾向にある。呈示順 (sessionN, articleN, screenN, lineN, segmentN) は，基本的に進めば進むほど読み時間が短くなる。これは実験協力者が実験に慣れてきた効果であると考える。

次に分類語彙表の類 (統語分類) について確認する。全ての読み時間指標について，用の類 (WLSPLUWA2) は体の類 (WLSPLUWA1) に対して，有意に読み時間が短くなる傾向がみられた。また相の類 (WLSPLUWA3) は FFT 以外で体の類に対して有意に読み時間が短く，SPT 以外で用の類に対して有意に読み時間が長い傾向がみられた。

最後に分類語彙表の部門 (意味分類) について確認する。抽象的関係 (WLSPLUWB.1) に対して主体 (WLSPLUWB.2)・活動 (WLSPLUWB.3)・生産物 (WLSPLUWB.4) が FFT, TOTAL に関して読み時間が長い傾向がみられた。

4.2 考察

ここでは主に読み時間が短くなる部分について検討を行う。

Anti-locality は先行文脈に係り元文節（単語）が多い要素ほど読み時間が短くなるという現象であり [7]，最初にドイツ語の二重目的語構文で報告され [8]，さらに日本語においても再検証された [13]。このような読み時間の短縮は，主辞後置言語において，係り元要素が多い要素を読むのに負荷がかかるという予測 [5] や，後続する主辞の処理コストは先行文脈の数の影響を受けないという予測 [11] などの，ワーキングメモリモデルによって説明できない現象であった。この現象はサプライザル [6, 9] の説明と親和性があるが，二重目的語構文における動詞述語の読み時間についてのみ報告されてきた。直接目的語と間接目的語の二つの係り元要素が先行する動詞述語のほうが，直接目的語のみ係り元要素が先行する動詞述語の読み時間より読み時間が短い。しかし，この結果は anti-locality を示す必要条件であるが，十分条件ではない。

文献 [3] では，均衡コーパスと係り受けアノテーションを用いることにより，より一般化した設定で anti-locality 現象を調査した。BCCWJ-DepPara [2] と読み時間データの重ね合わせから，係る文節数が多い文節ほど読み時間が短くなることを報告した。

本研究では，統語分類において，体の類＞相の類＞用の類の順で読み時間が短くなる傾向が確認された。体の類は一般に「モノ」などを表す名詞の仲間で，動詞や形容詞などの述語の項になりうる一方，修飾詞や項を取りうる句は限定的である。また判定詞などを取り名詞述語文を構成するが，相の類は形容詞・形容動詞・副詞・連体詞の仲間で，ガ格を取り述語になるものと修飾詞にのみなりうるものがある。用の類は動詞の仲間で，一般に節末などに出現し，項を取ることが多い。

係り受けと統語分類の結果は，先行要素記憶の負荷に基づくワーキングメモリモデルよりも，先行要素が後置要素を予測するモデルのほうが妥当であることを示唆している。言い換えると，修飾詞になったり，項を取ったりする用の類のほうが，予測されやすいということを反映している。係り受けを重ねたうえでも，この差が出ているのは，日本語において項が省略されていることによるものと考える。

また，意味分類においては，抽象的関係が他の要素よりも読み時間の短縮される傾向がみられた。関係は二項を取りうる要素であること，特に関係名詞の項は陽に表層にあまり出現しないことを反映している。

文献 [15] では，読み時間と節境界の対照を行い，節末で読み時間が短くなる傾向を報告している。一般に，項や修飾詞をとる用の類は節末に出現するため，今回の結果と親和性がある。文献 [14] では，読み時間と情報構造の対照比較を行い，ブリッジングなどにより想定可能情報や共有情報と比べて，情報の受容者側にとっての非共有情報の読み時間が長くなることを報告している。項や修飾詞を取りうる句は想定可能情報や共有情報になりやすいことと相関があると考える。

5. おわりに

本研究では，テキストの読み時間の傾向を分析するために，均衡コーパスに対する読み時間情報と分類語彙表番号アノテーションの対照比較を行った。結果，統語分類に対しては，体の類＞相の類＞用の類の順に読み時間が短くなる傾向がみられた。また意味分類（部門）に対しては，抽象的関係で読み時間が長くなる傾向がみられた。

今回の分析は，頻度主義的な統計分析手法である一般化線形モデルを用いた。今後，読み時間のモデリング手法として階層ベイズモデル [12] を用いて，分析を行いたい。

謝辞

本研究は JSPS 科研費 JP25284083,JP17H00917 の助成を受けたものです。また，本研究の一部は国立国語研究所コーパス開発センター共同研究プロジェクトによるものです。

参考文献

[1] Masayuki Asahara and Yuji Matsumoto. BCCWJ-DepPara: A Syntactic Annotation Treebank on the 'Balanced Corpus of Contemporary Written Japanese'. In *Proceedings of the 12th Workshop on Asian Langauge Resources (ALR12)*, pp. 49–58, 2016.

[2] Masayuki Asahara and Yuji Matsumoto. Bccwj-deppara: A syntactic annotation treebank on the 'balanced corpus of contemporary written japanese'. In *Proceedings of the 12th Workshop on Asian Language Resources (ALR12)*, pp. 49–58, December 2016.

[3] Masayuki Asahara, Hajime Ono, and Edson T. Miyamoto. Reading-Time Annotations for 'Balanced Corpus of Contemporary Written Japanese'. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pp. 684–694, 2016.

[4] R. H. Baayen. *Analyzing Linguistic Data: A practical Introduction to Statistics using R*. Cambridge University Press, 2008.

[5] E. Gibson. Linguistic complexity: Locality of syntactic dependencies. *Cognition*, Vol. 68, pp. 1–76, 2008.

[6] J. Hale. A probabilistic earley parser as a psycholinguistic model. In *Proceedings of the second conference of the North American chapter of the association for computational linguistics*, Vol. 2, pp. 159–166, 2001.

[7] L. Konieczny. Locality and parsing complexity. *Journal of Psycholinguistic Research*, Vol. 29, No. 6, 2000.

[8] L. Konieczny and P. Döring. Anticipation of clause-final heads. evidence from eye-tracking and srns. In *Proceedings of the 4th International Conference on Cognitive Science*, 2003.

[9] R. Levy and E. Gibson. Surprisal, the pdc, and the primary locus of processing difficulty in relative clauses. *Frontiers in Psychology*, Vol. 4, No. 229, 2013.

[10] Kikuo Maekawa, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, and Yasuharu Den. Balanced Corpus of Contemporary Written Japanese. *Language Resources and Evaluation*, Vol. 48, pp. 345–371, 2014.

[11] K. Nakatani and E. Gibson. An on-line study of japanese nesting complexity. *Cognitive Science*, Vol. 34, No. 1, pp. 94–112, 2010.

[12] Tanner Sorensen, Sven Hohenstein, and Shravan Vasishth. Bayesian linear mixed models using stan: A tutorial for psychologists, linguists, and cognitive scientists. *Quantitative Methods for Psychology*, Vol. 12, pp. 175–200, 2016.

[13] S. Uchida, E. T. Miyamoto, Y. Hirose, Y. Kobayashi, and T. Ito. An erp study of parsing and memory load in japanese sentence processing – a comparison between left-corner parsing and the dependency locality theory –. In *Proceedings of the Thought and Language/the Mental Architecture of Processing and Learning of Language 2014*, 2014.

[14] 浅原正幸. 読み時間と情報構造について (ちょっとみじかめ). 第 23 回言語処理学会年次大会発表論文集, pp. 911–915, 2017.

[15] 浅原正幸. 読み時間と節境界について. 第 154 回言語学会年次大会発表論文集, pp. 46–51, 2017.

[16] 加藤祥, 浅原正幸, 山崎誠. 『現代日本語書き言葉均衡コーパス』に対する分類語彙表番号アノテーション. 第 23 回言語処理学会年次大会発表論文集, pp. 306–309, 2017.

[17] 近藤明日子, 田中牧郎. 分類語彙表・unidic 見出し対応表の構築 – コーパスへの網羅的・系統的な語義情報付与を目指して –. 第 23 回言語処理学会年次大会発表論文集, pp. 90–93, 2017.

[18] 国立国語研究所（編）. 分類語彙表. 秀英出版, 1964.

[19] 国立国語研究所（編）. 分類語彙表 –増補改訂版–. 大日本図書, 2004.