『日本語歴史コーパス』の 文脈化単語埋め込みに基づく意味空間

浅原 正幸（国立国語研究所 コーパス開発センター）
加藤 祥（目白大学 外国語学部）

内省が効かない古典語について研究を進めるにあたり，統語・語義的に類似用例を提示する技術が求められている．近年，自然言語処理の分野で単語埋め込みの研究が盛んになり，単語の出現毎に異なるベクトルを付与することにより統語・語義的類似度を計量する「文脈化単語埋め込み」の技術が確立した．本研究では 220 億語規模の現代語の『国語研日本語ウェブコーパス』の事前学習モデルを語彙素に基づき構築し，共通の語彙素が付与されている『日本語歴史コーパス』に文脈化単語埋め込みを付与した．本稿では，文脈化単語埋め込みに基づく意味空間により，古典語に対してどのような研究ができるかについて検討する．

Distributional Semantics for “Corpus of Historical Japanese” Based on Contextual Word Embeddings

ASAHARA Masayuki (National Institute for Japanese Language and Linguistics, Japan)
KATO Sachi (Mejiro University)

Because introspection is not effective for the analysis of ancient languages, a technique to syntactically and semantically present the word similarities is required. Recently, researches on word embeddings have been conducted in the field of natural language processing, and the technique of “contextual word embeddings” has been established to assign a different word vector for each word token. The contextual word embeddings enable us to calculate the cosine between two word (or sentence) tokens that define syntactic and semantic similarities. We developed a pre-training model of BERT based on lexemes from the 22 billion token “NINJAL Web Japanese Corpus” and assigned contextual word vectors on the “Corpus of Historical Japanese” using common lexeme standards. This study explored the effect of contextual word embeddings on historical linguistic studies.

１．はじめに

内省が効かない古典語について研究を進めるにあたり，電子化されたコーパスに基づいて類似用例を比較検討することが重要である．通時的にコーパスを検討するにあたり，各語の統語的ふるまいや語義の変化に対応するために，単純な文字・単語・形態に基づく検索だけでなく，統語的ふるまいや語義に基づく類似用例を検索する技術が求められている．

分布意味論の分野において，単語を低次元の実数ベクトルで表現する単語埋め込み技術が研究されてきた．文脈化単語埋め込み[1]は，既存の単語埋め込み技術と異なり，単語の出現毎に異なるベクトルを割り当てるために，語義のあいまい性解消にも有効である．自然言語処理の分野では，事前学習モデルの研究が盛んであるが，BERT [2] (https://github.com/google-research/bert) も単語単位と文単位の文脈化単語埋め込みが出力できる．

日本語でも BERT のモデルが多数整備されている．その中で『NWJC-BERT』[3]は，形態素解析辞書 UniDic の語彙素表記を語彙として，『国語研日本語ウェブコーパス』(https://masayu-a.github.io/NWJC/, 以下 NWJC) [4]から訓練した，言語研究向けのモデルである．NWJC-BERT は，自然言語処理の応用を目指したモデルの構築ではなく，形態素解析用辞書 UniDic (https://unidic.ninjal.ac.jp/) の語彙素表記に基づき統語的ふるまいや語義の違いをベクトルで表現することができ，定量的な評価ができるモデルを構築した．本研究では NWJC-BERT を用いて『日本語歴史コーパス』[5]に文脈化単語埋め込みを悉皆付与した．これにより，統語的ふるまいや語義に基づく類似用例の検索ができるようになる．さらに『現代日本語書き言葉均衡コーパス』に対する文脈化単語埋め込み情報 BERTed-BCCWJ [6] と比較することにより，語義の通時的な変遷について，数値化できる．

本稿では，モデルやデータの構築方法および利

用した言語資源について示すとともに，文脈化単語埋め込みを用いた研究事例についても示す．また，同データのオープン化の可能性について検討を行う．

なお，本発表は，日本語学会 2020 年度春季大会の発表で『日本語歴史コーパス』の一部（「竹取物語」「土左日記」「徒然草」「方丈記」ほか）について検討を行ったもの[7]をコーパス全体に拡張し，新たにデータの仕様や可視化について詳細に論述したものである．

２．利用する言語資源

２．１．『日本語歴史コーパス』

国立国語研究所で整備している『日本語歴史コーパス』は，奈良時代から明治・大正時代にかけてのさまざまな資料を電子化したうえで，形態論情報が付与された大規模コーパスである．形態論情報は形態素解析用辞書 UniDic に基づいた読み・品詞・語彙素などの情報が付与されている．

本研究では 2019.3 データを利用した．分析対象の形態素数・文数（文相当単位数）を表 1 に示す．

表 1　『日本語歴史コーパス』の形態素数・文数
Table 1. The number of morphemes and sentences in “Corpus of Historical Japanese” by periods/eras

	形態素数	文相当単位数
奈良時代編	99,194	4,809
平安時代編	1,013,024	39,136
鎌倉時代編	972,674	49,304
室町時代編	415,573	30,899
江戸時代編	624,411	53,922
明治・大正編	15,226,278	1,122,005
和歌集	268,457	17,159

２．２．『NWJC-BERT』

自然言語処理の分野において，単語埋め込みの研究が進められている．その中で ELMo [1] (https://allennlp.org/elmo) は双方向 LSTM を複数層重ねたモデルで，各隠れ層の線形和を単語埋め込みとして得る．得られた単語埋め込みは，各語の出現毎に文脈に応じてその統語的ふるまいや語義の特徴を捉えた異なるベクトル（文脈化単語埋め込み）を出力できる．

BERT [2] は，大量のテキストデータから，隣接文推定や単語穴埋めのタスクを生成し，双方向 Transformer を複数層重ねたモデルで解析する，事前学習モデルである．転移学習やファインチューニングなどの技術を用いて，自然言語処理のその他のタスクに転用することにより，既存手法を上回る性能を達成している．この BERT のモデルの Embeddings 層を取り出すことで，各語の出現毎のベクトルを出力するほか，文頭に位置する [CLS] の最終層に割り当てられるベクトルを用いて，文単位のベクトルも出力できる．また，未知語が出現した場合にも，制御語 [UNK] を割り当てたうえで，その文脈に基づくベクトルを付与できる．

『NWJC-BERT』は NWJC の 6 単語以上の文 12.8 億文 226 億語により訓練した BERT のモデルである．語彙の分析に特化するため，形態素解析用辞書 UniDic に登録されている語彙素表記に基づいて訓練を行った．NWJC は 1 文単位のコーパスのため，ランダムに 1 文を分割することで「隣接単語列推定」を生成して訓練した．単語リスト (vocab.txt) は，UniDic の機能語すべて 154 語彙素表記と UniDic 分類語彙表番号対応表[8]に出現する 48,790 語彙素表記（一部前述の機能語と同じ表記あり）と制御語 5 種からなる．語彙素に基づいて訓練を行っているため，表層形に展開すると UniDic の 872,831 表層形中 54.6% の 468,460 表層形を被覆する．なお，『国語研日本語ウェブコーパス』は 1 文単位のコーパスであるために，訓練時にはランダムで 1 文を 2 つに分割する処理を行った．global_step 2,000,000 回のモデルを言語資源協会から公開している (https://www.gsk.or.jp/catalog/gsk2020-e/).

３．データの構築方法

本節では，今回構成した『日本語歴史コーパス』に対する文脈化単語埋め込みデータの構築方法と仕様について説明する．

『日本語歴史コーパス』 2019.3 データの語彙素表記を抽出し，文単位にしたうえで，『NWJC-BERT』のモデルに入力し，文単位のベクトルと形態素単位のベクトルを出力した．768 次元 12 層のベクトルを出力するが，最終層の-1 層をデータ化した．なお，空白など一部の形態素は除外した．データ形式はタブ区切りで以下の通り．

文ベクトルデータの形式：
1 列目：corpusName コーパス名
2 列目：pSampleID サンプル ID
3 列目：pStart 開始位置
4 列目：768 次元ベクトル

形態素ベクトルデータの形式：
1 列目：corpusName コーパス名
2 列目：pSampleID サンプル ID
3 列目：pStart 開始位置
4 列目：boundary 文境界情報
5 列目：orthToken 形態素表層形
6 列目：lexeme 語彙素表記
7 列目：veclexeme BERT が認識した語彙素表記
8 列目：768 次元ベクトル

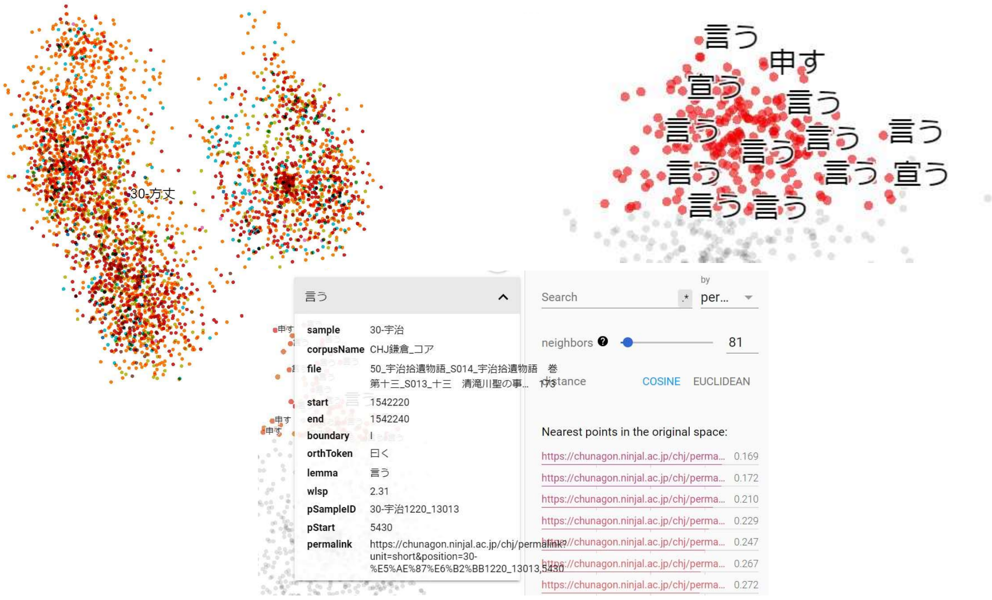

図 1 分類語彙表番号 2.3100 用-活動-言語-言語活動の可視化
Figure 1. Visualization of WLSP number 2.3100 Verbal-Action-Language-Linguistic Activity

サンプル ID と開始位置は検索系「中納言」上の位置情報で，位置検索などにより用例を確認できるほか，短単位語数表 (https://pj.ninjal.ac.jp/corpus_center/chj/chj-wc.html)に含まれる当該データのメタデータ情報が確認できる．

「BERT が認識した語彙素表記」(veclexeme) には，NWJC-BERT の語彙リスト vocab.txt に含まれる場合には「語彙素表記」(lexeme) が記載されるが，そうでない場合には未登録語である [UNK] が記載される．BERT は，未知語に対しても前後文脈からベクトル情報を付与できる．但し，空白や未知語が続くためにベクトルが不定になる場合には，当該箇所を除外した．

[コーパス中の語彙素表記] と[NWJC-BERT 上の語彙素表記] の違いは，後者は NWJC-BERT の vocab.txt に含まれない語彙素表記であった場合に，未知語を意味する [UNK] が記載される点である．なお，外部に公開する場合には，表層形や語彙素表記を伏せたうえで，「中納言」上の位置情報（サンプル ID と開始位置）に 768 次元ベクトルを付与した RDF などの 3 つ組データを公開する．

４．分析事例

文脈化単語埋め込みを用いたコーパスの分析について 3 つ紹介する．

４．１．分類語彙表番号に基づく分析

1 つ目の分析事例は，分類語彙表番号との対照である．我々は並行して『日本語歴史コーパス』に対する分類語彙表番号アノテーション[9]を進めており，現在までに『竹取物語』『土左日記』『徒然草』『方丈記』『宇治拾遺物語』『十訓抄』『今昔物語集』（一部）『虎明本狂言集』（一部）8 作品 459,129 語のアノテーションが完了している．形態素単位（短単位）のベクトルと分類語彙表番号を評価することで，ベクトルに対する語義の割り当てが可能になる．我々の以前の発表 [7] では，時間にまつわる単語（.16 以下，.1641 〈現在〉，.1642〈過去〉，.1643〈未来〉）について検討したが，今回は 2.3100 (古典対照分類語彙表 [10] においては 23100) （用−活動−言語−言語活動）3819 用例について検討する．可視化には Embedding Projector (https://projector.tensorflow.org/) を用いた．

図 1 に可視化の例を示す. 図 1 左上はベクトル空間を主成分分析した結果の上位 2 次元をプロットしたものである（寄与率 22.8%）．図 1 右上のように領域を選択することにより，元の語彙素が表示できる．また，図 1 下のように各点を選択することによって，表層形・語彙素表記の情報のほか，「中納言」へのリンク情報が表示される．このリンクをたどることにより，元の用例を確認できる． さらに他の要素を選択することにより cosine もしくは Euclidean 距離に基づく近傍の要素と，その要素の「中納言」へのリンク情報を表示できるため，類似用例の検索にも利用可能である．

分類語彙表番号が 2.3100 (23100) である 3819

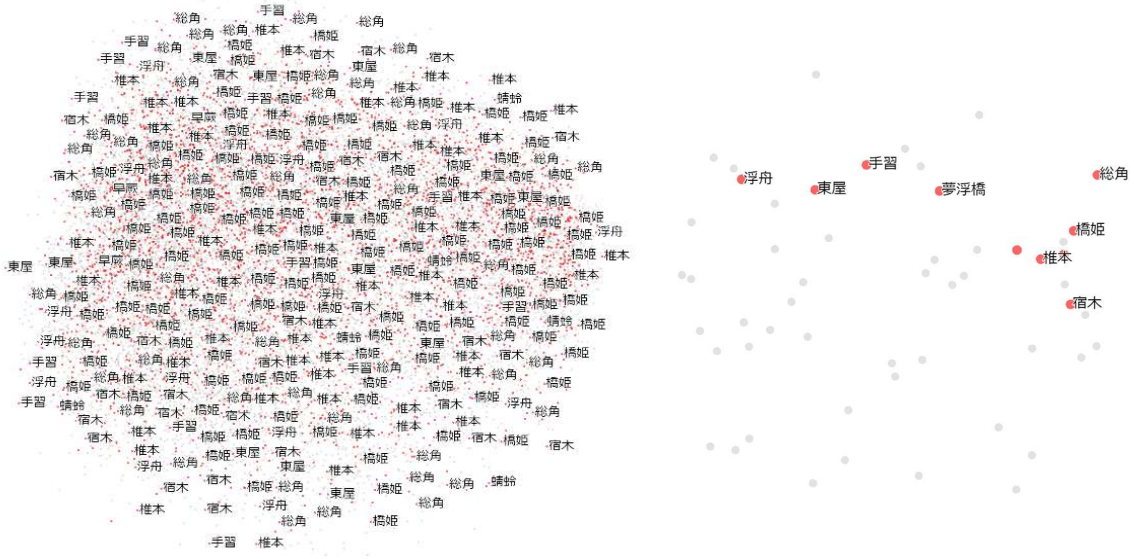

図 2 源氏物語の文ベクトル帖別の分布（宇治十帖のみラベル付与）
Figure 2 Distribution of Sentence Vectors for “Genji Monogatari”

用例中 2584 用例が「言う」，742 用例が「申す」であった．これらの 2 用例は言語活動を表現する語彙素表記の 87%を被覆する．各語彙素表記は主に同じ活用形に基づいてクラスタを作成する．

頻度 3 位に「宣う」(185 例)，頻度 4 位に「仰せる」(139 例)が確認できた．これらの事例は近接したところに出現した．例えば，以下の 2 事例のベクトル類似度が特に近かった．

公達のもの	おほ 仰せ	らるるに、さしらへするやうやはある。

CHJ: 30-十訓 1252_01052,1880
(日本語歴史コーパス中の位置情報，以下同様)

火を高くともして、隠れ居るかと見よ」と	のたまひ	ければ、法師ばら、「をかしくも仰せらるるかな」とて、

CHJ: 30-宇治 1220_14002,7380

このように，文脈化単語埋め込みを構成することで，用例の統語・語義的類似用例を提示できることが確認できた．

４．２．文単位のベクトルの利用

2 つ目の分析事例は，文単位のベクトルの分析である．『日本語歴史コーパス』の語数表に，サンプル ID に対応する書誌情報が含まれており，その情報に基づいて，文単位のベクトルを可視化できる．

以下では源氏物語 54 帖のうち，宇治十帖（「橋姫」以降）が，それ以外の箇所と文体が違うという説について検証する．安本[11]は，文長・和歌・直喩・擬声語・心理描写・色彩語・名詞・用言・助詞・助動詞の出現頻度・品詞の長さの 12 項目から，文長と助動詞の出現頻度以外の 10 項目について差異があることを示している．本稿では，このような違いがベクトル空間上にどのように表現されるかを確認する．

図 2 左は 16,613 文の各ベクトル表現を主成分分析により 2 次元のみ取り出したものである（次元 1:15.9%，次元 2:5.8%）．宇治十帖のみラベル付きの点で示してある．この図から,宇治十帖はそれ以外のベクトルと混在していることがわかる。以上から,宇治十帖とそれ以外の箇所の間に，特段の違いは確認できなかった．

図 2 右は 54 帖単位（「雲隠」は未定義，「若菜」を「若菜上」「若菜下」に分割）に文ベクトルの平均をとったうえで主成分分析を行ったものである（次元 1:24.3%，次元 2:13.7%）．文単位に評価したものと異なり，ある程度のまとまり（各次元において近い値を有する）ことが確認できており，宇治十帖とそれ以外のものとの言語的な特徴を捉えられていると考える．

これらのことから，文単位のベクトルを構成しても宇治十帖とそれ以外との差異が明確に確認できず，主成分分析の 2 次までの寄与率も 21.7%であったが，帖単位で平均をとることによって差異がみえてくる（主成分分析の 2 次までの寄与率 38%）傾向が確認できた．安本においても,分析は帖単位の出現頻度に基づいており，こういったものが可視化できたと考える．今後，各次元の要素が何を表したものなのかといった詳細な検討（プロービング）が必要であろう．これについては今後の課題としたい．

４．３．ベクトルの分散の利用

3 つ目は，語彙素表記ごとの平均ベクトルと分散の評価である．欧ら[12]は，BERT に基づく単語埋め込み表現の分散値（分散ベクトルのノルム）

が，語義の広がりを表すことを想定し，単義語と多義語の分散値の差異を検討したが，明らかな差異は見られなかったとしている．そこで，『日本語歴史コーパス』の時代区分ごとに平均ベクトルと分散ベクトルのノルムを確認することで，語義の変遷が定量的に評価できるかを試みる．さらに本研究では分散共分散行列のフロベニウスノルムについて検討した．

以下では，語彙素表記「赤い」について検討する．「赤い」は調整頻度（100 万語あたりの頻度）が，奈良時代を除いて，おおよそ 30 後半から 50 後半くらいの調整頻度であった．表 2 に「赤い」の時代・ジャンル別の平均ベクトルのノルム・分散ベクトルのノルム・分散共分散行列のノルムを評価したものを示す．

表 2　「赤い」の時代別統計情報
Table 2 The statistics of “Akai” by periods/eras

時代・ジャンル ※明大=明治・大正	平均 ノルム	分散 ノルム	分散 共分散 ノルム
平安-仮名文学	20.3	13.1	87.6
和歌集	24.6	9.8	147.5
鎌倉-日記・紀行	23.8	11.0	185.5
鎌倉-説話・随筆	21.5	10.9	58.7
室町-キリシタン	25.2	7.5	117.7
室町-狂言	21.0	12.8	120.4
江戸-洒落本	22.4	10.0	84.8
江戸-人情本	21.3	11.9	96.4
明大-初期口語	22.1	12.0	129.6
明大-雑誌	21.0	11.4	45.9
明大-教科書	21.5	10.8	53.4

平均ベクトルノルム・分散ベクトルノルムにおいては，時代・ジャンルごとの差が小さいものの，分散共分散行列のフロベニウスノルムにおいては差異が確認できた．基本的に「モノが赤い」「顔を赤らむ」の用例が多く出現した．

和歌集・鎌倉-日記・紀行においては和歌の独特の連接をモデル化しているためノルムが大きくなったと考える．

かゝみ山やまかきくもりしくるれともみち	あかく	そ秋はみえける

CHJ:20 後撰 0955_07007,1433

室町時代においては，「赤い」が格要素に出現する表現が確認できた．

身共があかミじやうごをしりながら、かほの	赤 あかひ	がおかしひか

CHJ: 40-虎明 1642_02018,22590

一女ハ天照太神宮、山田が原に神とゞまりまし／＼て、	赤 あかき	をハ人とさづけ、くろきハぎうばと定め、一切衆生をりやくせんが

CHJ:40-虎明 1642_03021,1858,

明治・大正期の初期口語資料は，言文一致の過程における多様な表現が確認できた．

とかいふ旦那でかみの毛がちゞれて	あか 赤い	とはいへ日本ことばもよくわかる

CHJ:60C 口語 1872_02305,18410

西洋風に模擬て正室には花毛穀を布き高机倚子などを駢べ荷蘭書の	あか 赤き	小口の立派なる古本を土場店にて買ひ

CHJ: 60C 口語 1872_03102,1130

何も黒日には日輪が黒く光り半黒の日は黄色に光り白日には	あか 赤く	おてらしなさるといふ差別はござり升まい

CHJ:60C 口語 1872_07203,28610

欧らの元論文においては，ベクトルの集合の分散のノルムを評価することで語義の広がりを評価することを試みた．本研究では分散・共分散行列のノルムを評価することを試みた．統計的には分散共分散行列は主成分分析と関連が深い．主成分分析は分散共分散行列を対角化する固有値問題にほかならないために，分散ベクトルではなく分散共分散行列を検討することが妥当であると考える．

５．おわりに

本研究では『日本語歴史コーパス』に対する文脈化単語埋め込みの悉皆付与について解説した．UniDic 辞書・コーパス・ BERT モデルを UniDic 語彙素により統一的に扱うことにより，内省が効かない語彙のその文脈的特徴を評価・可視化できることを確認した．

分析事例として，3 つの分析を示した．1 つ目の分析は分類語彙表番号アノテーションに基づく分析で，2.3100 （古典対照分類語彙表においては 23100 : 用-活動-言語-言語活動）である 3819 用例について Embeddings Projector による可視化を試みた．メタデータに「中納言」への位置情報を埋め込むことで，統語・意味的に近い用例を効率的に探索できることを示した．2 つ目の分析は文ベクトルの利用である．源氏物語の文体分析において，文単位のベクトルと帖単位の平均ベクトルについて主成分分析を行い可視化した．文単位には宇治十帖とそれ以外の差異が確認できなかったが，帖単位の平均ベクトルにおいては安本が示すような文体の差異が可視化できた．3 つ目の分析は分散の利用である．欧らの元論文においては分散ベクトルのノルムを検討していたが，分散共分散行列のフロベニウスノルムについても検討した．結果，時代・ジャンルごとの表現の広がりを数値化できることを確認した．

今後の展開として，いかに一般の利用に供するかを検討する必要がある．データサイズが 300GB 程度になるため，現実的にはそのまま配布することは困難である．Linked Open Data としての公開を想定するが，適切な次元圧縮が必要であろう．

BERT を訓練する際に最初から少ない次元で訓練することも考えられる．また，データがベクトル表現であるために，そこに距離空間が定義される．SPARQL のようなリンクの有無による問い合わせではなく，リンクによって定義される距離空間に基づく近接性による検索系の設計を行いたい．

謝辞

本研究は科研費 17H00917, 18H05521, 19K00591, 19K00655 および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成果物です．

参考文献

[1] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer. “Deep Contextualized Word Representations”, Proc. of NAACL-2018, 2018, p.2227-2237.
[2] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”, Proc. of NAACL-2019, 2019, p.4171-4186.
[3] 浅原正幸, 西内沙恵, 加藤祥, NWJC-BERT: 多義語に対するヒトと文脈化単語埋め込みの類似性判断の対照分析, 言語処理学会第 26 回年次大会発表論文集, 2020, p.961-964.
[4] M. Asahara, K. Maekawa, M. Imada, S. Kato, H.Konishi, “Archiving and Analysing Techniques of the Ultra-large-scale Web-based Corpus Project of NINJAL, Japan”, Alexandria: The Journal of National and International Library and Information Issues, 2020, Vol. 25, No. 1-2, p.129-148.
[5] 国立国語研究所,『日本語歴史コーパス』(バージョン 2019.3), 2019.
https://chunagon.ninjal.ac.jp/chj/
[6] 浅原正幸, 加藤祥, BERTed-BCCWJ: 多層文脈化単語埋め込み情報を付与した『現代日本語書き言葉均衡コーパス』データ, 言語処理学会第 26 回年次大会発表論文集, 2020, p.161-164.
[7] 浅原正幸, 加藤祥,『日本語歴史コーパス』に対する文脈化単語埋め込み情報付与, 日本語学会 2020 年度春季大会, 2020.
[8] 近藤明日子, 田中牧郎, 「分類語彙表番号-UniDic 語彙素番号対応表」の構築, 国立国語研究所論集, No. 18, 2020, p.77-91.
[9] 浅原正幸, 加藤祥, 鈴木泰, 池上尚,『日本語歴史コーパス』4 作品に対する分類語彙表番号付与とその分析, 日本語学会 2018 年度秋季大会, 2018.
[10] 宮島達夫, 石井久雄, 安部清哉, 鈴木泰, 日本古典対照分類語彙表, 笠間書院, 2014.
[11] 安本美典, 宇治十帖の作者―文章心理学による作者推定, 文学・語学第 4 号, 1957.
[12] 欧陽恵子, 曹鋭, 白静, 馬ブン, 新納浩幸, BERT による単語埋め込み表現の分散値を用いた語義の広がりの分布, 言語資源活用ワークショップ 2020 (国立国語研究所), 2020.