『日本経済新聞記事オープンコーパス』と『日本語話し言葉コーパス』語義と読みの対応表の作成

大井 恵奈 [1]　古宮 嘉那子 [1]　柏野 和佳子 [2]　浅原 正幸 [2]
[1] 東京農工大学生物システム応用化学府　[2] 国立国語研究所・総合研究大学院大学
s231186v@st.go.tuat.ac.jp　kkomiya@go.tuat.ac.jp　{waka,masayu-a}@ninjal.ac.jp

概要

本研究では，日本経済新聞記事オープンコーパス，日本語話し言葉コーパスを対象とし，読みと意味の対応表作成のための情報付与を行い，データ整備を行う．日本経済新聞記事オープンコーパスには，語義情報が付与されているものの，読み情報が付与されていなかったため，専門家による読みのアノテーションを行った．日本語話し言葉コーパスは，話し言葉の書き起こしであるため，読み情報は正確であったが，語義情報はついていないため，語義曖昧性解消システムを利用して語義データを付与した．以上のデータから，読みと意味の対応表を作成し，関係性について調査する．

1　はじめに

日本語では漢字の読みによって意味が変わってしまう場合がある．例えば，「最中」は「もなか」と「さいちゅう」では意味が全く異なる一方で，「今日」の読みは「きょう」も「こんにち」も似た意味でありながら，微細な意味の違いがある．このような微妙な意味の違いを整理するためには意味と読みの両方が付与されたデータが必要である．

そこで本研究では，『日本経済新聞記事オープンコーパス』[1]，『日本語話し言葉コーパス』[2] を対象に，形態素レベルでの読み仮名付与の実態を明らかにすることを目的とする。『日本経済新聞記事オープンコーパス』に対しては，33346 形態素に対して人手による読み仮名付与を実施した．ある漢字に複数の読み仮名がある場合，その文脈で複数の読み仮名が可能である場合はすべての読み仮名を付与し，いずれかの読み仮名でしか適用できない場合は，その単一の読み仮名を付与した．これらの読み仮名の選択にはレジスタや語義情報が密接な関係があると考えられる．特に，読み仮名の曖昧性には，レジスタの特性や多義語における語義選択の影響が考えられる．本研究では，この曖昧性の実態を解明するため，分類語彙表の語義と読み仮名の分布を対照して検討を行った．『日本経済新聞記事オープンコーパス』には，『分類語彙表』[3] に基づく語義情報が付与されている．さらに『日本語話し言葉コーパス』に自動解析技術で語義情報を付与したデータを利用することにより，読み仮名の選択と語義情報との関係性を明らかにすることを目指す．

2　コーパス

『日本経済新聞記事オープンコーパス』とは，日本経済新聞の朝夕刊（2013 年 1〜2 月）から選択した 96 記事を元に作成した日本語の書き言葉コーパスである．本コーパスには語義情報を付与したデータが存在するが読み情報は付与されていない．また，『日本語話し言葉コーパス (Corpus of Spontaneous Japanese)』(以下 CSJ とする) とは，日本語の自発音声を大量にあつめて多くの研究用情報を付加した話し言葉研究用のデータベースである．本コーパスは音声データの書き起こしであるため，正確な読み情報が得られるが，語義情報はまだ公開されていない．本研究ではこれら二つのコーパスを利用し，意味と読みの関係を整理する．

3　人手による日本経済新聞記事オープンコーパスへの読み仮名付与作業

本節では，『日本経済新聞記事オープンコーパス』における読み仮名付与作業について説明する．本作業は，短単位で形態素解析されているデータを対象に，人手で実施したものである．

まず，短単位形態素解析用辞書である『UniDic』

表 1　複数の読み仮名

表層形	読み仮名
その	ソノ
後	ア ト@ゴ@ノチ
に	ニ
税率	ゼイリツ
が	ガ

表 2　略語の読み仮名

表層形	読み仮名
日本	ニッポン@ニホン
の	ノ
ＤＦ	ディフェンダー
を	ヲ
抜き去り	ヌキサリ

表 3　数値表現の読み仮名

表層形	読み仮名
1	センキュウヒャクゴジュウ
9	↓
5	↓
0	↓
年	ネン
大会	タイカイ

表 4　外国人名の読み仮名

表層形	読み仮名
許	キョ@シュー
其亮	キ リ ョ ウ@チーリャン

を用い，各形態素に対して可能なすべての読み仮名を展開した．その上で，文脈に基づき曖昧性を解消するための作業を人手で行った．

複数の読み仮名がその文脈で可能な場合には，それらすべての読み仮名を付与する方針を採用した．表 1 の例では「後」の読み仮名が「アト」「ゴ」「ノチ」のすべてが可能であるために「@」をデリミタとして併記した．略語については，一般的に使用される読み仮名を復元し，これを付与した．表 2 の例では「ＤＦ」に対して「ディフェンダー」を付与した．

数値表現など，短単位では適切な読み仮名が付与できない場合には，読み仮名の曖昧性が解消されるより長い単位に対して読み仮名を付与した．『日本語話し言葉コーパス』は転記時に位取り記数法を用いるために，このような問題は生じない．『現代日本語書き言葉均衡コーパス』[4] は表現を位取り記数法を用いて正規化する工程 (NumTrans) があり，このような問題は生じない．表 3 の例は，「１９５０」の「０」には読みが付与されないほか，「１」はその後置する数字の桁数から「セン」という読み仮名が付与されるために，複数単位に対して読み仮名を付与した．

外国人名は，わかる範囲で「日本語読み」と「現地読み」を併記した．表 4 の例では，中国人名「許其亮」に対して，日本語読み「キョ キリョウ」と現地読み「シュー チーリャン」を併記した．

なお，記号など発音されないものについては，読み仮名を付与しなかった．

4　NIKKEI-WLSP: 語義情報データ

本節では，先行研究加藤ら [5] の『日本経済新聞記事オープンコーパス』の語義アノテーションデータについて説明する．本研究では，『日本経済新聞記事オープンコーパス』に収められた 96 記事・33,346 語に対し，語義情報として『分類語彙表』の分類番号を人手で付与した．この分類番号は，各語の意味に対応する適切なカテゴリーを示すものであり，特に多義語については，その文脈に基づいて適切な語義を選択して付与した．

付録の表 9 に，語義情報（類・部門：分類番号の小数点以下一桁まで）の統計情報を示す．各分類に属する語の頻度と割合を確認することができる．このデータは，人手により語義の曖昧性解消がなされていることから，語義情報からの読み仮名の対応を検討することができる．

5　自動解析による CSJ への語義付与

本節では，CSJ の語義アノテーションデータについて説明する．本研究では，浅田ら [6] により，日本語 BERT の Fine-tuning を利用して，コーパス内の単語全てを対象とする all-words WSD を CSJ に行ったデータを使用した．付録の表 10 に，付与された語義情報（類・部門：分類番号の小数点以下一桁まで）の統計情報を示す．

6　語義と読みの分析

本節では，読みと語義番号についての対応を調査した結果を示す．

まず，4 節および 5 節で付与したデータに基づいて作成したデータから，読みと語義番号がともに付与されている各単語について，語義番号と読みの一対一のペアを抽出した．この際，複数の読みが可能な場合については，複数の読みに対してそれぞれ語義のペアを作成した．なお，読みと語義番号がともにひとつしかないものは分析の対象外とした．対象としたコーパスごとに，それぞれ読みと語義の組が出現したかどうかの表を作成した．日本経済新聞オープン記事コーパスからは 102 件，CSJ からは 539 件の語について，このような表を作成した．その結果，コーパスごとに単語と読みのペアの出現の傾向が異なることが分かった．例として表 5 に日本経済新聞記事オープンコーパスの「後」の読みと語義の組が出現したかどうかの表を示す．表中の ○

表 5　『日本経済新聞オープン記事コーパス』における「後」の読みと語義の出現

「後」	1.1670 体 時間的前後	1.1650 体 時間的順序	3.1670 相 時間的前後
ノチ	×	○	○
ゴ	×	○	○
アト	○	×	○

表 6　CSJ における「後」の読みと語義の出現

「後」	1.1740 空間	3.1670 時間前後	1.1643 未来	1.1650 時間順序	1.1670 時間前後
コウ	×	×	×	×	○
ゴ	×	×	×	○	○
アト	○	○	○	○	○
ノチ	×	○	×	○	○

は出現したことを示し，× は出現しなかったことを示す．また表 6 に CSJ についての同様の表を示す．表中の番号は分類語彙表の分類番号である．これらの区分と例を付録の表 11 に示す．これらの表から，「後」には必ず全ての読みと語義のペアが存在するわけではないことが分かる．ただし，CSJ のアノテーションは自動的なものであり，その正解率は 91.8%程度であることに注意されたい．

表 5 から，日本経済新聞オープン記事コーパスでは「後」を「アト」と読む場合，分類語彙表の 1.1670（体・関係・時間・時間的前後）もしくは 3.1670（相・関係・時間・時間的前後）が割り振られ，分類語彙表の 1.1650（体・関係・時間・時間的順序）の意味を持つ例が出現しないことが分かる．つまり，日本経済新聞オープン記事コーパスにおいては，時間的順序を表す「後」は「アト」と読んだ例が出現しなかったということである．一方で，表 6 から，CSJ では「アト」という読むときも 1.1650 の意味を持つ例が出現していることが分かる．さらには，1.1740（体・関係・空間・左右・前後・たてよこ）や 1.1643（体・関係・時間・未来）の意味を持つ例があることも分かった．これらの違いは日本経済新聞は書き言葉であり，CSJ は話し言葉であることに起因している可能性が高い．ただし，日本経済新聞より CSJ の方がコーパスサイズが大きいことが原因である可能性もある．

表 7, 8 にそれぞれ日本経済新聞オープン記事コーパスと CSJ における語義と読みの数の対応を示す．語義がひとつのときと複数であるときに，読みがひとつであるか，複数であるのかをまとめたものである．これらの表から，一つの語義に対して複数の読みを持つ単語よりも，一つの読みに対して複数の語義を持つ単語の割合が大きいことが読み取れる．な

表 7　『日本経済新聞オープン記事コーパス』の読みと語義の対応

	読みが一つ	**読みが複数**
語義が一つ	-	11
語義が複数	76	15

表 8　CSJ の読みと語義の対応

	読みが一つ	**読みが複数**
語義が一つ	-	61
語義が複数	408	70

お，本調査では，『日本経済新聞記事オープンコーパス』内の「金」という単語を除き，読みと語義を共に複数持つ全ての単語に対し，その単語の可能な語義候補の 5 割以上の数の語義を取りうる読みが存在した．つまり，ひとつの読みでひとつの語義と決まっているわけではなく，読みが固定されたとしても，取りうる語義のバリエーションは半分以下に減らないものが大半であった．以上の結果から，読みと語義の関係を考慮する場合，複数の語義を持つ読みに着目することで，読み仮名の選択の正確性がより向上する可能性があると考えられる．

7　おわりに

本研究では，日本経済新聞記事オープンコーパス，日本語話し言葉コーパス (CSJ) を対象とし，読みと意味の対応表作成のための情報付与を行い，対応表を作成した．日本経済新聞オープン記事コーパスからは 102 件，CSJ からは 539 件の語について，読みと意味のペアの出現の表を作成した．

また，二つのコーパスの分析から，コーパスごとに単語と読みのペアの出現の傾向が異なることを示した．さらに，ひとつの読みでひとつの語義と決まっているわけではなく，読みが固定されたとしても，取りうる語義のバリエーションは半分以下に減らないものが大半であること，そのため，読みと語義の関係を考慮する場合，複数の語義を持つ読みに着目することで，読み仮名の選択の正確性がより向上する可能性について示した．

謝辞

本研究は JSPS 科研費 JP22K12145，及び国立国語研究所共同研究プロジェクト「アノテーションデータを用いた実証的計算心理言語学」と「テキストの読み上げのための読みの曖昧性の分類と読み推定タスクのデータセットの構築」の助成を受けたものです．

参考文献

[1] 日本経済新聞社. 日本経済新聞記事オープンコーパス, 2023. Available at: https://nkbb.nikkei.co.jp/alternative/corpus/.

[2] 国立国語研究所. 日本語話し言葉コーパスの構築法, 2006. Available at: https://doi.org/10.15084/00001357.

[3] 国立国語研究所. 分類語彙表増補改訂版データベース（ver.1.0.1）, 2018. Available at: https://github.com/masayu-a/WLSP.

[4] 前川喜久雄 (監修), 山崎誠 (編). 書き言葉コーパス—設計と構築—. 講座日本語コーパス 2. 朝倉書店, 2014.

[5] 加藤祥, 浅原正幸. 『日本経済新聞記事オープンコーパス』に対するメタデータと語義情報付与. 言語処理学会第 30 回年次大会発表論文集, 2024.

[6] 浅田宗磨, 古宮嘉那子, 浅原正幸. 『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悉皆付与. 言語処理学会第 30 回年次大会発表論文集, pp. 2767–2772, 2024.

表 9　NIKKEI-WLSP の語義情報の分布 (類・部門)

語義情報	頻度	(割合)
1.1 体-関係	6750	20.20%
1.2 体-主体	2963	8.90%
1.3 体-活動	4120	12.40%
1.4 体-生産物	506	1.50%
1.5 体-自然	312	0.90%
2.1 用-関係	1352	4.10%
2.3 用-活動	1565	4.70%
2.5 用-自然	14	0.00%
3.1 相-関係	1138	3.40%
3.2 相-主体	2	0.00%
3.3 相-活動	174	0.50%
3.5 相-自然	14	0.00%
4. 他	82	0.20%
ラベルなし	14354	43.00%
合計	33346	100.00%

表 10　CSJ の語義情報の分布 (類・部門)

語義情報	頻度	(割合)
1.1 体-関係	151560	30.15%
1.2 体-主体	4767	0.95%
1.3 体-活動	34308	6.83%
1.4 体-生産物	3126	0.62%
1.5 体-自然	5541	1.10%
2.1 用-関係	22603	4.50%
2.3 用-活動	38050	7.57%
2.5 用-自然	42	0.01%
3.1 相-関係	30976	6.16%
3.3 相-活動	1265	0.25%
3.5 相-自然	222	0.04%
4. 他	7294	1.45%
ラベルなし	202888	40.36%
合計	502642	100.00%

表 11　「後」の持ちうる分類語彙表の番号の意味と用例

分類番号	類・部門・中項目・分類番号	用例
1.1643	体・関係・時間・未来	先，先週，今後，明日
1.1650	体・関係・時間・時間的順序	次，初，順，優先
1.1670	体・関係・時間・時間的	前, 後
1.1740	体・関係・空間・左右・前後・たてよこ	前，後，横，左側
3.1670	相 (形容詞的)・関係・時間・時間的前後	まだ，以後，もう