BCCWJ2 構築に向けた漢字カタカナ文抽出ツールの開発

平林 照雄[1] 呉 寧真[1] 山崎 誠[1] 小木曽 智信[1]
[1] 国立国語研究所
{teru-hirabayashi, gons, yamazaki, togiso}@ninjal.ac.jp

概要

2024 年度より、「現代日本語書き言葉均衡コーパス」（BCCWJ）の拡張として BCCWJ2 の構築が進められている。BCCWJ2 は現代日本語を中心とする一方、英文や、古文・漢文資料からの引用、擬古文、方言表現など、現代日本語辞書での形態素解析で誤りが生じやすい文を含む。漢字・カタカナおよび記号のみからなる文（漢字カタカナ文）もその一つである。本論文では、漢字カタカナ文を対象にタグ付与を支援する抽出ツールを提案し、抽出性能と実運用での有効性を示す。

1 はじめに

国立国語研究所では、2024 年度より、「現代日本語書き言葉均衡コーパス」（BCCWJ）の拡張として BCCWJ2 の構築が進められており、一部が 2025 年度末に公開される。その対象語数は、2006 年〜2010 年の書籍サンプルから空白及び記号類を除いた総語数約 2300 万語で、表 1 の件数となる。

表 1 2006 年〜2010 年 BCCWJ2 ジャンル別語数

NDC	空白と記号類を除いた総語数	総語数
0 総記	524,629	622,687
1 哲学	1,603,993	1,871,052
2 歴史	1,778,286	2,066,435
3 社会科学	5,893,725	6,835,671
4 自然科学	2,494,104	2,968,632
5 技術	2,198,568	2,607,977
6 産業	1,118,247	1,304,970
7 芸術	1,857,837	2,187,939
8 言語	388,746	481,899
9 文学	5,532,644	6,590,638
計	23,390,779	27,537,900

これらの書籍は現代日本語で記述されており、古文や漢文のみで構成された作品は、原則として収録していない。しかし、古文・漢文資料からの引用等の形で、古文的・漢文的表現が文中に含まれる場合がある。また、方言的表現や、英文やプログラムのコード、URL 等が含まれる文も収録されている。このような文に対して形態素解析（以下、解析と呼ぶ）を行う場合、現代日本語向けに整備された辞書で解析したり、文処理を施さずに解析を実行したりすると、正しい解析結果を得られない。

そこで BCCWJ2 では、以下のように XML でタグ付けを行い、その属性に応じて文処理や辞書選択を行うことで、解析精度の向上を図っている。

```
<proc dic="Kindai-bungo" norm="kata2hira">
人倫ノ重キヲ思ヒ。
</proc>
```

上記のタグが付与された文では、カタカナをひらがなに変換した後、近代文語 UniDic[1)] で解析を行う。その解析結果は、表 2 となる。一方、文処理や辞書選択を行わず、現代書き言葉 UniDic[1)] で解析した場合の結果を、表 3 に示す。表 2 と表 3 を比較すると、語の分割及び品詞判定が原文の用法に即したものになっており、解析結果が改善している。

このように、タグ付けと変換処理を適切に行うことで解析性能の向上が確認されている。一方で、現状ではこれらのタグの大半は手作業によりコアデータ[2)] にのみ付与されており、非コアデータへの適用は十分に進んでいない。しかし、非コアデータは規模が大きく、コアデータと同様の手作業による確認やタグ付与を行うことは現実的ではない。そのため、非コアデータに対しても適用可能な、効率的なタグ付与支援手法の整備が求められている。

そこで本研究では、BCCWJ2 に含まれる漢字カタカナ文を対象として、当該文を自動的に抽出し、手作業によるタグ付与を支援するツールを提案する。提案手法は、非コアデータを含む大規模データに対しても適用可能であることを重視し、実際のコーパス構築作業を想定した設計とする。本論文では、提案ツールの抽出性能を評価するとともに、BCCWJ2

1） https://clrd.ninjal.ac.jp/unidic/download.html
2） コアデータとは、自動解析後に全件について作業者が確認・修正を行ったデータを指す。非コアデータとは、自動解析を基本とし、必要最小限の手作業のみを補佐的に加えたデータを指す [1]

表 2　文処理や辞書選択を行った場合の漢字カタカナ文の解析結果例

表層形	語彙素	語彙素読み	品詞	活用型	活用形
人倫	人倫	ジンリン	名詞-普通名詞-一般		
の	の	ノ	助詞-格助詞		
重き	重い	オモイ	形容詞-一般	文語-形容詞-ク	連体形-一般
を	を	ヲ	助詞-格助詞		
思ひ	思う	オモウ	動詞-一般	文語四段-ハ行	連用形-一般
。	。		補助記号-句点		

表 3　文処理や辞書選択を行わない場合の漢字カタカナ文の解析結果例

表層形	語彙素	語彙素読み	品詞	活用型	活用形
人倫	人倫	ジンリン	名詞-普通名詞-一般		
ノ	の	ノ	助詞-格助詞		
重	重	ジュウ	名詞-普通名詞-助数詞可能		
キ	機	キ	名詞-普通名詞-助数詞可能		
ヲ	を	ヲ	助詞-格助詞		
思	思う	オモウ	動詞-一般	五段-ワア行	連用形-省略
ヒ	ひい	ヒイ	感動詞-一般		
。	。		補助記号-句点		

における実運用への適用事例を示し、タグ付与作業の効率化および品質向上への有効性を検証する。

2　漢字カタカナ文を判定する手法

事前調査により、漢字カタカナ文を現代書き言葉 UniDic で形態素解析すると、漢字カタカナ文で頻出する「ノ」や「ヲ」などの助詞は正しく解析されることが多い一方、「ニ」や「ハ」などの助詞は誤解析が多い傾向にある。また、一つの文章中に漢字カタカナ文が複数回出現する傾向もみられた。

以上の傾向より、漢字・カタカナまたは記号のみで構成される文すべてを対象文とし、対象文集合から以下の条件を満たす文を漢字カタカナ文候補として自動抽出する。これらの条件は、事前調査で観察された解析誤りの傾向に基づき、コーパス構築時の実用性を重視して経験的に設定したものである。

1. 対象文に含まれるカタカナ語について、表層形が以下の助詞または助詞複合形と完全一致する語を抽出し、それらのうち少なくとも 1 つが、文頭以外または記号類の直後以外に出現する場合、その対象文を漢字カタカナ文と判定する。(ハ, ガ, ヲ, ニ, ヘ, ト, デ, モ, ヤ, ノ, バ, テ, ニテ, ニテハ, ヨリ, マデ, ホド, コソ, サエ, ダニ, ナド, トカ)[3)]
2. 1 により漢字カタカナ文と判定された対象文を含む同一文章内の対象文を、漢字カタカナ文と判定する。
3. 対象文を現代書き言葉 UniDic で形態素解析した結果、助詞と判定されるカタカナ語が 1 つ以上含まれ、かつそれらのうち少なくとも 1 つが記号類の直後以外または文頭以外に出現する場合、その対象文を漢字カタカナ文と判定する。
4. 対象文に含まれるカタカナについて、カタカナをひらがなに変換した場合と変換しない場合とで、それぞれ現代書き言葉 UniDic による形態素解析を行い、変換前に「漢字のみの一語」に分割された語が、変換後に「漢字＋ひらがなを含む一語」へと分割結果が変化する語を含む場合、当該対象文を漢字カタカナ文と判定する。

3) 助詞「カ」は「一カ月」などの誤抽出が多いため抽出対象外とした。

3　漢字カタカナ文判定ツールの適用実験

3.1　実験設定

漢字カタカナ文判定ツールの性能および実運用上の有効性を評価することを目的として、BCCWJ2 収録予定データのうち、2006 年の非コアデータを対象に実験を行った。収録全文から、漢字、カタカナまたは記号のみで構成される文を抽出し、8,647 文を対象とした。これらの文に対して、共著者 1 名が手作業で漢字カタカナ文かを判定し、<proc>タグの付与を実際に行った。その際、タグ付与作業に要した時間を、以下の二つの方法でそれぞれ計測した。

(i) 8,647 文を参照しつつ判断を行い、元文章の該当箇所に<proc>タグを付与
(ii) 8,647 文に漢字カタカナ文判定ツールを適用し、抽出された候補文について判断を行った上で、元文章の該当箇所に<proc>タグを付与

方法 (i) の結果、132 文を漢字カタカナ文と認定した。2006 年の非コアデータのジャンル別総文数は表 4 に、対象となった漢字、カタカナまたは記号のみで構成される文及び、漢字カタカナ文のジャンル別割合は表 5 に示す。

表 4 2006 年非コアデータジャンル別文数

NDC	総文数
0 総記	7,515
1 哲学	12,561
2 歴史	18,421
3 社会科学	59,150
4 自然科学	23,344
5 技術	23,512
6 産業	10,200
7 芸術	16,033
8 言語	7,728
9 文学	49,517
計	228,073

表 5 2006 年非コア対象文数及び漢字カタカナ文のジャンル別文数

NDC	対象文数	漢字カタカナ文数
0 総記	274	4
1 哲学	314	3
2 歴史	891	54
3 社会科学	1,601	16
4 自然科学	837	7
5 技術	1,999	10
6 産業	556	8
7 芸術	1,208	0
8 言語	370	1
9 文学	597	29
計	8,647	132

続けて、ツールの性能の評価は、以下の二つの観点から行った。

第一に、提案ツールが漢字カタカナ文を候補としてどの程度適切に抽出できるかを確認するため、候補抽出性能について Precision、Recall、F1 を用いて評価した。この評価では、各文が漢字カタカナ文の候補として抽出されたかどうかのみを対象とし、人手による辞書属性の選択は評価対象に含めない。

第二に、コーパス構築における最終的なタグ付与済みデータに至るまでの工程において、提案ツール、人手による <proc> タグの付与・非付与の判断、および人手による辞書属性の選択といった各処理要素を段階的に適用した場合の処理結果を比較した。具体的には、各候補文について、原文のまま処理した場合、提案ツールのみを適用した場合、さらに人手によるタグ付与・非付与の判断を加えた場合、および人手による辞書属性の選択まで含めて処理を行った場合という各条件の下で処理を行い、各段階において最終的なタグ付与済みデータと一致した文数を比較した。

3.2 実験結果

漢字カタカナ文判定ツールの運用上の性能として、ツール適用による作業時間の変化を表 6 に示す。方法 (ii) での作業時間は、対象文からツールでの絞り込みにかかった時間 6 秒を含む。

表 6 方法による作業時間の変化

Methods		Time
(i)	(ツール使用無し)	1 時間 10 分 30 秒
(ii)	(ツール使用有り)	27 分 41 秒

ツールの抽出性能は表 7 である。

表 7 漢字カタカナ文判定ツールの精度

Recall	Precision	F1
0.924 (122/132)	0.371 (122/329)	0.529

次に、提案ツールによって抽出された漢字カタカナ文候補 329 文を対象として、タグ付与工程をどの段階まで進めた場合に、最終的なタグ付与済みデータと同じ状態に到達する文がどの程度得られるかを確認した。表 8 に、各処理条件の下で処理を行った結果を示す。表 8 では、原文のまま処理した場合、提案ツールを適用した場合、さらに提案ツールを適用後、人手によって <proc> タグを付与するか否かの判断を行った場合、および <proc> タグの付与が確定した文に対して、人手による辞書属性の選択までを行った場合という、段階的な処理条件ごとに結果を集計している。各段階について、その時点で得られた処理結果が、最終的なタグ付与済みデータと同一であった文の数を示している。

表 8 処理段階ごとに最終的なタグ付与済みデータと一致した文数

処理段階	一致した文数
原文ママ	197
ツール適用	16
＋人手判断（<proc>付与可否）	213
＋人手判断（辞書選択）	329

4 考察

本実験では、漢字カタカナ文判定ツールの適用による作業効率の変化、候補抽出性能、および辞書選択を含む最終的なタグ付与結果との関係について検討した。

まず、作業時間に着目すると、表 6 に示したように、ツールを用いた場合、用いない場合と比べ約 6

割の時間で作業を完了している。この結果は、提案ツールが大規模データに対する人手作業を支援する上で、実運用上有効であることを示している。

次に、候補抽出性能について考察する。表 7 に示したように、本ツールの候補抽出性能は、Recall が 0.924（122/132）と十分な値を示す一方で、Precision は 0.371（122/329）にとどまった。これは、本ツールが漢字カタカナ文を自動的に確定することを目的とするのではなく、漢字カタカナ文を漏れなく候補として抽出することを重視して設計されていることに起因する。そのため、候補文数が増加し Precision が低下することは想定された挙動であるが、その一方で、確認対象文数は 8,647 文から 329 文へと大きく削減されており、候補抽出段階としては実用上十分な絞り込みが達成されている。

さらに、辞書選択を含む最終的なタグ付与について考察する。表 8 に示したように、提案ツールによって抽出された 329 文の内、原文のままで既に 197 文が適切であり、これらは追加の処理を行わなくても正しく扱われていた文である。これに対し、提案ツールを適用することで、16 文に適切に処理をされ、人手によるタグの付与・非付与の判断のみの確認を行うことで、合計 213 文が正しく処理される結果となった。一方、辞書選択まで含めた処理を行った場合には、さらに 116 文が新たにタグ付与され、最終的に 329 文すべてが正しくタグ付与された。この結果は、適切なタグ付与には辞書選択が大きな割合を示していることを示している。

以上の結果から、提案ツールは、漢字カタカナ文を自動的に確定する処理ではなく、人手による確認および辞書選択を前提としたタグ付与工程において、候補提示を担う前段処理として位置づけられることが分かる。しかし、最終的なタグ付与結果には辞書選択が大きく寄与しており、この工程をどのように支援するかが今後の課題である。

最後に、本ツールを BCCWJ1 に適用して得られた漢字カタカナ文候補について、BCCWJ1 の解析結果を提供する『中納言』[4] 上での解析結果と、カタカナをひらがなに変換して解析した結果を比較した。

例として、以下の BCCWJ1 のサンプル ID: PB39_00419 内の 2 文を用いる。

―聞コエヌノカ。前方カラダレカ来ル。

『中納言』上での解析結果を表 9 に、カタカナをひらがなに変換し解析を行った結果を表 10 に示す。

表 9　『中納言』上での漢字カタカナ文の解析結果例

書字形出現形	語彙素	品詞
―	―	補助記号-一般
聞	聞き	名詞-普通名詞-一般
コエヌノカ		未知語
。	。	補助記号-句点
前方		カタカナ文
カラダレカ		カタカナ文
来		カタカナ文
ル		カタカナ文
。		カタカナ文

表 10　カタカナをひらがなに変換した場合の漢字カタカナ文の解析結果例

表層形	語彙素	品詞
―	―	補助記号-一般
聞こえ	聞こえる	動詞-一般
ぬ	ず	助動詞
の	の	助詞-準体助詞
か	か	助詞-終助詞
。	。	補助記号-句点
前方	前方	名詞-普通名詞-一般
から	から	助詞-格助詞
だれ	だれ	代名詞
か	か	助詞-副助詞
来る	来る	動詞-非自立可能
。	。	補助記号-句点

このように、漢字カタカナ文では、文中のカタカナ表記により、語分割が不安定となる事例が BCCWJ1 においても確認され、本手法は、そのような事例の検出に有効であることが確認された。

5　おわりに

本研究では、BCCWJ2 に含まれる漢字カタカナ文を対象として、手作業によるタグ付与を支援することを目的とした抽出ツールを提案し、その抽出性能および実運用上の有効性を検証した。提案ツールにより、大規模な非コアデータに対しても、人手による確認対象を効率的に絞り込めることを示した。

一方で、最終的なタグ付与結果には辞書選択が大きく関与しており、候補抽出後の処理をどのように支援するかが重要であることも明らかとなった。

今後は、漢字カタカナ文の時代区分の精緻化や古文的表現の自動抽出などへ対象を拡張し、BCCWJ2 に適切なタグ付けを行うための実用的な手法の検討を進めていく予定である。

4）　https://chunagon.ninjal.ac.jp/

謝辞

本研究は文化庁委託事業「信頼できる言語資源としての現代日本語の保存・活用のためのデジタル基盤整備事業」による成果の一部です。
https://www2.ninjal.ac.jp/BCCWJ2/

参考文献

[1] 国立国語研究所コーパス開発センター. 『現代日本語書き言葉均衡コーパス』利用の手引 第 1.1 版. p. 58, 2021.