BCCWJ の接続詞の品詞情報の解析精度について

馬場 俊臣（北海道教育大学教育学部）

On the Precision of the POS Information: Focusing on the Conjunctions in the BCCWJ

Toshiomi Baba (Hokkaido University of Education, Sapporo Campus)

要旨

接続詞を扱った研究において BCCWJ の品詞情報を利用する際の留意点を示すために、BCCWJ で「接続詞」の品詞情報が付与された語（長単位）の解析精度の調査を行い、以下の結果を得た。(1)サンプル調査（非コアデータ各 100 件）の結果、品詞情報「接続詞」の使用頻度上位 20 語の適合率は 63.0%～100.0%の範囲にあり、特に「で」「唯」「又」の適合率が低い。(2)「又」の詳細調査（非コアデータ 1000 件）の結果、適合率は 85.8%であり、レジスター別では「特定目的・ブログ」42.4%が特に低い。(3)「で」の詳細調査（非コアデータ 1000 件）の結果（ただし 200 件の途中経過）、適合率は 62.5%であり、レジスター別では「特定目的・知恵袋」44.1%が特に低い。なお、本研究は、品詞情報付与に関する解析器改良のための参考資料を提供するものでもある。

１．はじめに

『現代日本語書き言葉均衡コーパス』（BCCWJ）を利用した接続詞研究の問題点と可能性に関する基礎的研究の一環として、本稿では、BCCWJ の接続詞に関する品詞情報の信頼性を見るために、品詞情報「接続詞」[1]の解析精度に関する調査結果を報告する。

BCCWJ の解析精度は、「長単位・短単位とも、データ全体に対して人手修正を行ったコアデータは 99%以上、データの一部に対して人手修正を行ったコアデータ以外のデータは 98%以上」（小椋、冨士池(2011):39）とされるが、品詞によって解析精度は若干異なると予想される。また、同じく接続詞であっても語により解析精度が異なると予想される。

BCCWJ を利用した重要な研究の一つに、品詞比率に基づいた文章・文体研究がある[2]。こうした巨視的な研究では、品詞の違いによる解析精度の若干の異なりは、分析結果に殆ど影響を与えず何ら問題は生じない。しかし、例えば特定の品詞に限定して、その品詞に属するいくつかの語（ないし語群）の比率を問題にする場合は対象とする語の解析精度の違いが分析結果に影響を及ぼす可能性がある。特に接続詞は、属する語の種類（異なり語）が少なく、一つ一つの語の解析精度の違いが場合によっては分析結果に大きな影響を及ぼす恐れがある。

BCCWJ を利用する際の基本としては、利用マニュアル[3]や小木曽(2014)に示されているように「解析誤り」「形態素解析の弱点」があることを前提として、研究目的・研究対象

---

1 品詞情報として「接続詞」が付与されていることを、以下「品詞情報「接続詞」」又は単に括弧を付けて「接続詞」と略記する。他の品詞についても同様である。

2 品詞比率とジャンル（レジスター）等の文体・文章構造の違いとの関連を分析した研究として、冨士池他(2011)、鯨井(2011)などの研究がある。なお、左記の二つの研究では、誤解析に対する人手修正を施したコアデータ（長単位）を使用している。

3 国立国語研究所コーパス開発センター(2011)、国立国語研究所コーパス開発センター(2013)。

に応じて人手による点検が必要になる。こうした点検を行うことによって、語による解析精度の違いの問題を避けることができる。

しかし、検索結果をそのまま利用する場合などでは特に、一つ一つの語の解析精度の違いがどの程度有りうるのかという知見を予め知っておくことが重要である。

本稿では、このような問題意識に基づいて、BCCWJ の「接続詞」の品詞情報の信頼性を見るために、「接続詞」の用例の解析精度に関する調査を行い、その結果を報告する。調査内容は次の通りである。

(1) 「接続詞」の使用頻度上位 20 語（長単位）についてサンプル調査（非コアデータ各 100 件）を行い、語ごとの適合率[4]を明らかにする。（3 節）

(2) 適合率が低い「又」（使用頻度第 1 位）について、サンプル数を増やした詳細調査（「接続詞」「副詞」各 1000 件）を行い、「接続詞」及び「副詞」の適合率を明らかにし、さらに、レジスター別での違いも明らかにする。（4 節）

(3) 適合率が最も低い「で」について、サンプル数を増やした詳細調査（「接続詞」「格助詞」「助動詞」各 1000 件）を行い、「接続詞」及び「格助詞」「助動詞」の適合率を明らかにし、さらに、レジスター別での違いも明らかにする。（5 節）

なお、本研究は、BCCWJ を利用した今後の接続詞研究[5]に対して重要な基礎的知見を提供するとともに、品詞情報付与に関する解析器の改良のための参考資料を提供するものでもある。

２．BCCWJ 全体の品詞情報の解析精度について

調査結果を示すに先立って、公表されている BCCWJ 全体の品詞情報の解析精度を示す。本稿の調査は、BCCWJ において、「接続詞」の品詞情報が付与された長単位[6]の語彙素を対象とする。検索ツールとして、品詞情報を用いた検索ができる「中納言」を利用する。

BCCWJ の形態論情報の付与では、「短単位解析には解析エンジン MeCab と形態素解析用辞書 UniDic を、長単位解析には短単位解析結果から長単位を自動構成する解析器」（小椋、冨士池(2011):44）を用いており[7]、また（短単位全体の）「1 億語のうち約 100 万語（コアデータ）については、自動解析後に人手修正を行い、解析精度 99%以上の高精度なデータとし、形態素解析システムの学習用データとして用いた」（同:64）とのことである。

接続詞に関しては、UniDic における接続詞（短単位）は 30 語であり（UniDic-mecab version 1.3.12 の接続詞辞書（Conjunction.csv）による）、さらに、長単位では 32 の「連語」（従って、そうして、其れとも、では等）が接続詞として扱われている（同:69）。

BCCWJ の形態論情報の解析精度は、コアデータは 99%以上、コアデータ以外のデータは 98%以上（同:39）とのことである。レジスター別では、「白書、書籍（文学）、書籍

---

4 本稿では解析精度として「適合率」を用いた。「適合率」は「正しく品詞情報を付与された長単位数／当該品詞情報を付与された長単位数」で求めた。本稿の調査では「再現率」は調査しておらず、従って「F 値」も求めていない。脚注 8 も参照。

5 接続詞研究においても BCCWJ を利用した研究が増えている。ただし、検索ツールや検索方法の詳細、また、検索結果に対する人手による点検の有無の詳細が示されていないものがある。コーパスを用いた研究の特徴の一つに追試可能性が挙げられる。それを保証するためには、検索及び用例確定の方法を明示することが必須となろう。

6 多くの接続詞研究において接続詞として扱われる語の単位は、「長単位」にほぼ相当する。

7 本稿での指摘は MeCab+UniDic により付与された品詞情報の問題点でもある。

（文学以外）、新聞、Web（Y!知恵袋）」の各レジスターの「品詞」の解析精度（F 値）[8]は、それぞれ 0.995693、0.9866095、0.989596、0.989116、0.984112 となっており、98%以上を実現している（同:45）。BCCWJ の利用マニュアルに記載されている解析精度は F 値のみであり、適合率及び再現率は示されていない。小木曽他(2010)では、「新聞」（毎日新聞 2007 年度版）・「文学作品」（新潮文庫の 100 冊）・「ブログ」（Yahoo!ブログ）を用いて UniDic-mecab と他の解析器との精度比較を行い「UniDic-mecab 1.3.12」での適合率、再現率、F 値を示している。新聞、文学作品、ブログの順にそれぞれ「品詞」の適合率は 0.9879、0.9772、0.9756 であり 98%前後以上である。

３．高頻度接続詞の適合率

３.１　調査の目的と方法

本節では、品詞情報「接続詞」の語のうち、使用頻度上位 20 語（長単位）（以下、「高頻度接続詞」と呼ぶ）について、サンプル調査（非コアデータ各 100 件）を行い、語ごとの適合率を明らかにする。

まず、高頻度接続詞を取り出すために、「中納言」長単位検索で「品詞　大分類　接続詞」を指定し、全レジスター対象に検索[9]を行った[10]。検索総件数は 668,836 件である。語彙素を単位として集計し、頻度合計上位 20 位までの語を選定した（表 1 参照）[11]。

次に、各接続詞からサンプルを抽出した。コアデータについては自動解析後に人手による修正を行っているため、サンプル調査の対象は非コアデータのみとする。「中納言」長単位検索で「語彙素」「品詞　大分類　接続詞」を指定し検索[12]を行い、検索結果画面上

---

8 適合率（精度）、再現率、F 値は分類の評価指標として用いられる。適合率は付与された品詞がどのくらい正しいかを表す指標である。再現率は実際にある品詞であるものをどれくらいカバーして付与できているかを表す指標である。F 値は適合率と再現率の調和平均である。接続詞を例にすると、次の式で求められる。

（適合率）＝（品詞情報「接続詞」を付与されて正しく接続詞であった件数）／（品詞情報「接続詞」を付与された件数）×100[%]

（再現率）＝（品詞情報「接続詞」を付与されて正しく接続詞であった件数）／（調査対象全体で実際に接続詞である件数）×100[%]

（F 値）＝2×（適合率）×（再現率）／（（適合率）＋（再現率））

9 検索条件式は、「キー: 品詞 LIKE "接続詞%" WITH OPTIONS unit="2" AND tglWords="10" AND limitToSelfSentence="0" AND endOfLine="CRLF" AND tglKugiri="" AND encoding="UTF-8" AND tglFixVariable="2"」である。なお、「中納言」では 10 万件以上の一括ダウンロードができないため、いくつかのレジスターごとに分割してダウンロードを行った。

10 本稿での「中納言」検索結果は、高頻度接続詞及び「又」の詳細調査に関しては 2013 年 11 月～2014 年 2 月、「で」の詳細調査に関しては 2014 年 12 月～2015 年 1 月の期間で得られた結果である。

11 「『現代日本語書き言葉均衡コーパス』長単位語彙表 ver1.0」（DVD データに基づく語彙表）では、「だから」「だが」「所が」の頻度合計はそれぞれ 21,010、17,871、11,394 であり、本調査と比べいずれも非コアデータの頻度が 2 件、1 件、6 件低くなっている。理由は不明である。

12 検索条件式（例として「又」を挙げる）は次の通りである。

キー: (語彙素 = "又" AND 品詞 LIKE "接続詞%") IN (registerName="出版・新聞" AND core="false") OR (registerName="出版・雑誌" AND core="false") OR (registerName="出版・書籍" AND core="false") OR (registerName="図書館・書籍" AND core="false") OR (registerName="特定目的・白書" AND core="false") OR (registerName="特定目的・ベストセラー" AND core="false") OR (registerName="特定目的・知恵袋" AND core="false") OR (registerName="特定目的・ブログ" AND core="fals

で表示された 500 件の内、最初の 100 件を調査対象とした。検索結果の画面表示については、「検索ヒット数が 500 件を超える場合、検索結果からランダムで選ばれた 500 件が表示されます。」（中納言オンライン「マニュアル」更新日：2014-04-02 版）とのことであり、無作為抽出とみなした。

得られた各接続詞の用例 100 件の品詞を、前後の文脈を読み取りながら人手により確認した。副詞など接続詞以外の品詞との判別が特に問題となるものについては、次のような置き換え可能性を目安にして判断した。また、コアデータでの品詞判定も参考にした。判定に迷う場合は接続詞とした。

「又」[13]：「並びに、その上に、又は」に置き換えられるかどうか。「再び、同様に、一方、一体全体・まったく」に置き換えられる場合は副詞。

「更に」：「その上に、それに加えて」に置き換えられるかどうか。「ますます、もっと、少しも（～ない）」に置き換えられる場合は副詞。

「其れから」：「そして」に置き換えられるかどうか。「その時から」に置き換えられる場合は「代名詞＋格助詞」、両方可能な場合は接続詞扱い。

「唯」：「ただし」に置き換えられるかどうか。「単に」に置き換えられる場合は副詞。

「猶」：言い添える内容が続くかどうか。「相変わらず、やはり、一層、ちょうど（のごとし）」に置き換えられる場合は副詞。

「で」：「それで」に置き換えられるかどうか。

「其れでも」：「でも」に置き換えられるかどうか。「でも」に置き換えられず「それで」に置き換えられる場合は「それ」は代名詞。

３.２　高頻度接続詞の適合率の調査結果（語彙素別）

調査結果は、表 1 の通りである。

調査対象 20 語全体の適合率は 93.8%であり、非コアデータ全体の F 値 98%以上よりは低いが、高い適合率になっている。ただし、語ごとに見ると、適合率 90%未満の語が「又」82.0%、「更に」89.0%、「其れから」87.0%、「唯」76.0%、「猶」89.0%、「で」63.0%の 6 語ある。「又、更に、唯、猶」は副詞の誤判定[14]が目立つ。この 4 語には副詞の同形の語彙素がある。「其れから」は代名詞「其れ」との誤解析が目立つ。「で」の適合率は特に低く格助詞及び助動詞の誤判定が目立つ。

このように、語ごとに見た場合、適合率が特に低い語があり、注意が必要である。

---

e") OR (registerName="特定目的・法律" AND core="false") OR (registerName="特定目的・国会会議録" AND core="false") OR (registerName="特定目的・広報誌" AND core="false") OR (registerName="特定目的・教科書" AND core="false") OR (registerName="特定目的・韻文" AND core="false") WITH OPTIONS unit="2" AND tglWords="200" AND limitToSelfSentence="0" AND endOfLine="CRLF" AND tglKugiri="" AND encoding="UTF-8" AND tglFixVariable="2"

13 「又」の接続詞と副詞の判別の詳細については、4 節参照。

14 本稿では、品詞分類の誤りを「誤判定」と呼び、それ以外の形態素境界の誤りや長単位の構成に関する誤りなどを「誤解析」と呼び、便宜的に呼び分ける。

表 1　高頻度接続詞（サンプル調査）の適合率（語彙素別）[15]

順位	語彙素	コアデータ頻度	非コアデータ頻度	頻度合計	調査件数	接続詞	他品詞等	適合率	他品詞等内訳
1	又	899	85,543	86,442	100	82	18	82.0%	副詞 13、誤解析「又は」5
2	然し	561	68,041	68,602	100	100	0	100.0%	
3	そして	426	62,269	62,695	100	100	0	100.0%	
4	及び	660	48,295	48,955	100	99	1	99.0%	動詞 1
5	でも*	307	36,397	36,704	100	100	0	100.0%	
6	又は*	151	29,560	29,711	100	100	0	100.0%	
7	或いは	106	26,490	26,596	100	98	2	98.0%	副詞 2
8	だから*	172	20,840	21,012	100	100	0	100.0%	
9	更に	275	18,614	18,889	100	89	11	89.0%	副詞 11
10	だが*	177	17,695	17,872	100	100	0	100.0%	
11	其れから*	54	16,570	16,624	100	87	13	87.0%	誤解析(代名詞+格助詞)13
12	唯	159	16,388	16,547	100	76	24	76.0%	副詞 23、誤解析「只松」1
13	然も	106	14,570	14,676	100	100	0	100.0%	
14	猶	89	12,272	12,361	100	89	11	89.0%	副詞 10、誤解析「尚穆王」1
15	但し	80	11,667	11,747	100	99	1	99.0%	誤解析「但一人」1
16	所が*	105	11,295	11,400	100	100	0	100.0%	
17	で	74	10,866	10,940	100	63	37	63.0%	格助詞 18、助動詞 3、誤解析（助動詞）9、誤解析(その他)5、(「て」の)誤字 2
18	即ち	38	10,717	10,755	100	100	0	100.0%	
19	従って*	36	9,900	9,936	100	100	0	100.0%	
20	其れでも*	91	9,807	9,898	100	93	7	93.0%	誤解析(代名詞+格助詞+係助詞)7
	計				2,000	1,875	125	93.8%	

3.3　高頻度接続詞の適合率の調査結果（レジスター別）

同じ調査データを用いレジスター別の適合率を集計した。表 2 に、20 語全体の数値と適合率の低い「又、唯、で」の 3 語の数値を示した。

表 2　高頻度接続詞（非コアデータ、サンプル調査）の適合率（レジスター別）

レジスター	20 語全体		又		唯		で	
	調査件数	適合率	調査件数	適合率	調査件数	適合率	調査件数	適合率
出版・書籍	589	95.4%	32	81.3%	24	75.0%	11	36.4%
出版・雑誌	76	96.1%	1	100.0%	3	66.7%	5	100.0%
出版・新聞	13	100.0%	0	0.0%	1	100.0%	0	0.0%
図書館・書籍	610	94.4%	19	89.5%	25	64.0%	21	76.2%
特定目的・白書	106	94.3%	22	77.3%	0	0.0%	0	0.0%
特定目的・教科書	11	100.0%	3	100.0%	0	0.0%	0	0.0%
特定目的・広報誌	35	97.1%	3	66.7%	0	0.0%	0	0.0%
特定目的・ベストセラー	60	93.3%	0	0.0%	3	66.7%	3	66.7%
特定目的・知恵袋	146	84.9%	6	100.0%	24	83.3%	20	45.0%
特定目的・ブログ	149	86.6%	10	60.0%	9	88.9%	38	68.4%
特定目的・韻文	2	50.0%	0	0.0%	1	0.0%	0	0.0%
特定目的・法律	60	96.7%	0	0.0%	0	0.0%	0	0.0%
特定目的・国会会議録	143	96.5%	4	100.0%	10	90.0%	2	50.0%
計	2000	93.8%	100	82.0%	100	76.0%	100	63.0%

15 「*」を付けた語彙素は、長単位で「連語」の接続詞となる語彙素である。

20 語全体では、調査件数が少ない「特定目的・韻文」を除けば、「特定目的・知恵袋」84.9%及び「特定目的・ブログ」86.6%の適合率が若干低くなってはいるが、全体的にレジスター間で大きな違いは見られない。しかし、（調査件数が少ないレジスターを除くと）「又」では「特定目的・白書」77.3%、「特定目的・ブログ」60.0%、「唯」では「図書館・書籍」64.0%、「で」では「出版・書籍」36.4%、「特定目的・知恵袋」45.0%が特に低くなっており、レジスターの違いによる適合率の大きな違いが見られる。

３.４　詳細な調査の必要性

高頻度接続詞の適合率の調査によって、調査対象 20 語全体の適合率は高いが、語ごとでは適合率の低い語があること、また、20 語全体ではレジスターの違いによる適合率の違いはほぼ見られないが、適合率の低い「又」「唯」「で」ではレジスターによる適合率の違いが見られることが明らかになった。

本節では高頻度接続詞について各 100 語を対象として調査を行ったが、サンプル数が少ないという問題点がある。サンプル数を増やしてより詳細な調査を行う必要がある。本稿では、適合率の低い語のうち「接続詞」使用頻度第 1 位の「又」及び適合率の最も低い「で」について詳細な調査を行う。

４.「又」の詳細調査

４.１　調査の目的と方法

「接続詞」使用頻度第 1 位の「又」に関してより厳密な適合率を明らかにするため、またレジスターによる適合率の違いを詳細に分析するため、「接続詞」及び「副詞」の品詞情報が付与された「又」について調査（以下、「詳細調査」と呼ぶ）を行った。

詳細調査の前に、念のために、形態素解析システムの学習用データとして用いた人手による修正済みのコアデータについて適合率を確認する調査を行った。「中納言」長単位検索で品詞情報を「接続詞」及び「副詞」と指定しコアデータ対象に検索[16]を行い、得られた用例の品詞を前後の文脈を読み取りながら人手により確認した[17]。その結果、「接続

16 検索条件式は次の通りである。「副詞」の検索では「接続詞」の箇所を「副詞」に置き換えた。

キー: (語彙素 = "又" AND 品詞 LIKE "接続詞%") IN (registerName="出版・新聞" AND core="true") OR (registerName="出版・雑誌" AND core="true") OR (registerName="出版・書籍" AND core="true") OR (registerName="特定目的・白書" AND core="true") OR (registerName="特定目的・知恵袋" AND core="true") OR (registerName="特定目的・ブログ" AND core="true") WITH OPTIONS unit="2" AND tglWords="300" AND limitToSelfSentence="0" AND endOfLine="CRLF" AND tglKugiri="" AND encoding="UTF-8" AND tglFixVariable="2"

17 「並びに、その上に、又は」（接続詞）、「再び、同様に（「～もまた」等）、一方（「秋はまた収穫の季節でもある」等）、一体全体・まったく（「どうしてまたそんなことをしたのだ」「またなんときれいな花だ」等）」（副詞）への置き換えを目安に品詞判定を行った。また、コアデータでの品詞判定も参考にした。接続詞と副詞の両方に解釈可能な用例など判定が難しい用例は、付与された品詞情報を正解として処理した。なお、「又貸し」「又聞き」等は全体で名詞とした。「又の名」「又の日」も全体で名詞（小椋、小磯、冨士池、宮内、小西、原(2011)「資料　要注意語」p.20 参照）とした。また、「山また山」「一人また一人」のような同じ名詞を繋ぐ用法は辞書により扱いが異なる。コアデータでは「一羽また一羽と死んでいきました」は接続詞としているが、詳細調査対象の非コアデータでは「足音が一歩、また一歩と大きくなった」「人また人でぎっしり埋まる」は「副詞」と判定されている。今回の調査ではコアデータに従い接続詞として扱う。

詞」の「又」899 件のうち 889 件が接続詞であり適合率 98.9%であった。また、「副詞」の「又」247 件のうち 241 件が副詞であり適合率 97.6%であった。コアデータに関しては 98%前後以上の高い適合率であることが確認された。

非コアデータを対象とした「又」の詳細調査の手順・方法を示す。まず、コアデータと同様に品詞情報を指定し非コアデータ対象に検索[18]を行い、「接続詞」の「又」の用例 85,543 件、「副詞」の「又」の用例 28,756 件を得た。これらの用例に対して、それぞれ層別無作為抽出（レジスターの 1 層）を行い、「接続詞」「副詞」各 1000 例を調査対象の用例として、前後の文脈を読み取りながら人手により品詞を確認した。なお、「接続詞」及び「副詞」の用例の抽出率は、それぞれ 1.17%、3.48%である。

４.２　「又」詳細調査での適合率の結果及び誤判定の要因

「又」の詳細調査による品詞判定の結果を表 3 に示す。

表 3　「又」詳細調査（非コアデータ）での適合率

品詞情報	人手による品詞判定				計	適合率
	接続詞	副詞	誤解析	誤字		
「接続詞」	858	117	25	0	1000	85.8%
「副詞」	160	828	11	1	1000	82.8%
計	1018	945	36	1		

「接続詞」の「又」1000 件のうち 858 件が接続詞であり適合率 85.8%であった。接続詞以外は、副詞の誤判定 117 件、誤解析 25 件（「又は」23 件、「またぐ」「三ツ又沼」各 1 件）であった。「副詞」の「又」1000 件のうち 828 件が副詞であり適合率 82.8%であった。副詞以外は、接続詞の誤判定 160 件、誤解析 11 件（「又の名」3 件、「俟つ」2 件、「尾亦、胡亦堂、興復、又七郎、又左、股」各 1 件）、誤字「復雑」（複雑）1 件であった。

「接続詞」の「又」に関しては 3 節での 100 例サンプル調査での適合率 82.0%に比べると若干高くなってはいるが、それでも 90%を下回っている。品詞情報を利用する際に十分留意する必要がある。

ただし、「接続詞」の「又」の正解 858 件と「副詞」の「又」のうち接続詞の用例 160 件とを合わせると 1,018 件となる。少なくとも「又」は、仮に「接続詞」1000 件の数値をそのまま利用したとしても大きな違いが生じないという見方もできるかもしれない[19]。

誤判定の起こる要因は断定できないが、読点（「、」及び「，」）の直後の「又」の誤判定が目立った。直前 1 文字別の適合率（調査件数 6 件以上のみ）を表 4 に示す。

表 4 の通り、「接続詞」「副詞」各 1000 件の用例のうち、ともにほぼ 4 分の 1 の用例が読点の直後の用例である。「、」の直後の「接続詞」の適合率は 73.1%であり、「，」及び「、」の直後の「副詞」の適合率はそれぞれ 21.1%、58.7%であり極めて低い。また、「接続詞」全体の副詞の誤判定 117 件のうち読点の直後の用例は 55 件(47.0%)であり、「副詞」全体の接続詞の誤判定 160 件のうち読点の直後の用例は 114 件(71.3%)であり、誤

[18] 検索条件式は、非コアデータを指定した以外は、注 13 と同様である。

[19] ただし、4.3 に示すようにレジスター別では大きな違いが生じる場合がある。特に「特定目的・ブログ」では、「接続詞」の「又」には副詞が 5 割以上含まれるのに対し「副詞」の「又」には接続詞が 144 例中 3 例あるのみであり、「接続詞」の「又」の使用頻度をそのまま用いるのは危険である。

判定の多くは読点の直後である。このように、読点の直後での誤判定の多さが、全体の適合率を下げる一つの大きな要因となっていると見られる[20]。

表 4　「又」詳細調査（非コアデータ）での直前 1 文字別適合率(調査件数 6 件以上のみ)

「接続詞」		
直前 1 文字	調査件数	適合率
、	208	73.1%
は	23	87.0%
(全角スペース)	190	90.0%
て	12	91.7%
.	19	94.7%
。	412	97.3%
,	41	97.6%
？	28	100.0%
全体	1000	85.8%

「副詞」		
直前 1 文字	調査件数	適合率
,	19	21.1%
、	242	58.7%
に	43	81.4%
て	31	83.9%
の	7	85.7%
ら	40	87.5%
を	17	88.2%
は	126	92.1%
「	17	94.1%
で	41	95.1%
が	59	96.6%
も	202	98.5%
。	9	100.0%
と	18	100.0%
ば	7	100.0%
れ	51	100.0%
全体	1000	82.8%

４.３　「又」詳細調査での適合率の結果（レジスター別）

同じ調査データを用いレジスター別の適合率を集計した（表 5 参照）。

表 5　「又」詳細調査（非コアデータ）での適合率（レジスター別）

レジスター	「接続詞」		「副詞」	
	調査件数	適合率	調査件数	適合率
出版・書籍	274	91.2%	257	77.8%
出版・雑誌	27	96.3%	25	92.0%
出版・新聞	5	100.0%	3	66.7%
図書館・書籍	236	83.9%	356	82.9%
特定目的・白書	161	86.3%	4	25.0%
特定目的・教科書	17	94.1%	3	33.3%
特定目的・広報誌	36	97.2%	3	66.7%
特定目的・ベストセラー	22	77.3%	51	86.3%
特定目的・知恵袋	86	89.5%	67	92.5%
特定目的・ブログ	66	42.4%	144	97.2%
特定目的・韻文	1	0.0%	4	100.0%
特定目的・法律	0		0	
特定目的・国会会議録	69	97.1%	83	65.1%
計	1000	85.8%	1000	82.8%

[20] コアデータの読点の直後の用例のみを取り出してみると、「接続詞」全 120 件中 4 件が副詞であり（適合率 96.7%）、「副詞」全 14 件中 1 件が誤解析（名詞「又の名」）であった（適合率 92.9%）。

レジスター別（調査件数 10 以下のレジスターは除く）に見ると、「接続詞」の「又」では、「特定目的・ブログ」42.4%（特に「、」の直後全 14 件の適合率 14.3%）、「特定目的・ベストセラー」77.3%（特に「、」の直後全 9 件の適合率 44.4%）が特に適合率が低い。「副詞」の「又」では、「特定目的・国会会議録」65.1%（特に「、」の直後全 20 件の適合率 10.0%）が特に適合率が低い[21]。

レジスター別の使用頻度に基づいた接続詞の分析を行う際には、適合率が低いレジスターがあることを十分に考慮する必要がある。

５．「で」の詳細調査

サンプル調査で適合率が最も低かった「で」に関しても、「又」と同様の方法で詳細調査（「接続詞」「格助詞」「助動詞」各 1000 件）を行っている途中である（表 6 参照）[22]。現段階（各 200 件の途中経過）では、「接続詞」に関しては、適合率が 62.5%と低く、レジスター別では「特定目的・知恵袋」44.1%が特に低くなっている。また、格助詞や助動詞の誤判定や誤解析は「で」の直前が空白（全角スペース）や記号類（、）」等）の場合、数式などを削除している場合、文頭の「であるから、でないから」等の場合に目立つ。

表 6　「で」詳細調査（非コアデータ）での適合率（途中経過）

品詞情報	人手による品詞判定						計	適合率
	接続詞	格助詞	助動詞	接続助詞	誤解析	誤字		
「接続詞」	125	31	13	1	29	1	200	62.5%
「格助詞」	0	182	15	0	3	0	200	91.0%
「助動詞」	0	46	139	0	15	0	200	69.5%
計	125	259	167	1	47	1		

６．まとめ

BCCWJ を利用した接続詞研究が増えている。接続詞研究において BCCWJ の品詞情報を利用する際の留意点を示すために、本稿では、BCCWJ で「接続詞」の品詞情報が付与された語（長単位）の解析精度の調査（非コアデータ対象）を行い、以下の結果を報告した。

① 高頻度接続詞 20 語全体の適合率は 93.8%であり、非コアデータ全体（全品詞）に比べると低いが、高い適合率になっている。しかし、語ごとに見ると、適合率は 63.0%～100.0%の範囲にあり適合率の低い語がある。適合率 90%未満の語は、「又」82.0%、「更に」89.0%、「其れから」87.0%、「唯」76.0%、「猶」89.0%、「で」63.0%の 6 語である。「又、更に、唯、猶」は副詞の誤判定が目立つ。

② 高頻度接続詞 20 語全体では、レジスターの違いによる適合率の違いはほぼ見られない。しかし、適合率の低い「又」「唯」「で」では、レジスターによる適合率の違いが見られる。

③ 「又」の詳細調査の結果、適合率は「接続詞」85.8%、「副詞」82.8%である。レ

21 「特定目的・ブログ」「特定目的・国会会議録」で適合率が特に低くなったのは、行動の叙述（時間的）、並列的な事柄の提示（非時間的）というそれぞれの内容的な特徴も関わっていると思われる。

22 「で」のコアデータの適合率は「接続詞」90.5%、「格助詞」97.0%、「助動詞」99.0%である。「接続詞」は全 74 件、「格助詞」「助動詞」は検索結果画面に表示された最初の各 100 件を対象とした。

ジスター別では「接続詞」の「特定目的・ブログ」42.4%、「副詞」の「特定目的・国会会議録」65.1%が特に低い。読点の直後の「又」の誤判定が多く、全体の適合率を下げる大きな要因となっていると見られる。

④　「で」の詳細調査の結果（ただし途中経過）、「接続詞」の適合率は 62.5%であり、レジスター別では「特定目的・知恵袋」44.1%が特に低い。

接続詞研究では、従来、コーパス検索の際、多くは文字列検索が行なわれ、また、効率的に検索するために、文頭に限定したり読点が後続する場合に限定したりすることも多かった。今後の研究において、BCCWJ での品詞情報が利用できることは極めて有益なことである。接続詞全体での品詞情報の解析精度はコーパス全体（全品詞）よりも若干劣るが、接続詞全体として他品詞と比較する場合には大きな問題は生じないであろう。しかし、異なり語の少ない接続詞内部で個々の語（語群）を分析する場合には、品詞情報の解析精度の違いが問題となる。もちろん、BCCWJ の品詞情報を利用する際には、研究の目的や方法に応じて人手による点検が不可欠である。しかし、検索結果をそのまま利用する場合では、特に分析対象とする語の解析精度の違いを十分把握しておく必要がある。

今後は、誤判定、誤解析の要因を明らかにし解析精度の向上を図ることが期待される。本稿の結果は品詞情報付与に関する解析器改良のための参考資料を提供するものでもある。

文　献

小木曽智信(2014)「第 5 章　形態素解析」山崎誠(編)『講座日本語コーパス　2. 書き言葉コーパス―設計と構築―』朝倉書店, pp.89-115.

小木曽智信、小椋秀樹、小磯花絵、宮内佐夜香、渡部涼子、伝康晴(2010)「形態素解析辞書のベンチマークテスト―IPAdic・NAIST-jdic・UniDic のジャンル別精度比較―」, 言語処理学会第 16 回年次大会発表論文集, pp.326-329.

小椋秀樹、小磯花絵、冨士池優美、宮内佐夜香、小西光、原裕(2011)『『現代日本語書き言葉均衡コーパス』　形態論情報規程集　第 4 版　（下）』国立国語研究所.

小椋秀樹、冨士池優美(2011)「第 4 章　形態論情報」, 『現代日本語書き言葉均衡コーパス』利用の手引　第 1.0 版, pp.39-73.

鯨井綾希(2011)「主成分分析を用いた文章構造の特徴抽出――品詞構成の変動に注目した分析――」, 文芸研究, 172, pp.59-48.

国立国語研究所コーパス開発センター(2011)『『現代日本語書き言葉均衡コーパス』利用の手引　第 1.0 版』国立国語研究所コーパス開発センター.

冨士池優美、小西光、小椋秀樹、小木曽智信、小磯花絵(2011)「長単位に基づく『現代日本語書き言葉均衡コーパス』の品詞比率に関する分析」, 言語処理学会第 17 回年次大会発表論文集, pp.663-666.

関連 URL

国立国語研究所コーパス開発センター(2013)『『現代日本語書き言葉均衡コーパス』マニュアル　第 1.1 版（Web 公開用）』国立国語研究所コーパス開発センター.　http://www.ninjal.ac.jp/corpus_center/bccwj/doc/manual/BCCWJ_Manual.zip

「現代日本語書き言葉均衡コーパス 中納言 1.1.0」　https://chunagon.ninjal.ac.jp/

「『現代日本語書き言葉均衡コーパス』長単位語彙表 ver1.0」　http://www.ninjal.ac.jp/corpus_center/bccwj/freq-list.html