既存コーパスの修正とその有効性
ー「YNU 書き言葉コーパス」を例にー

長谷川　守寿

1. 目的

本稿の目的は、既存のコーパスに修正を加えることの有効性を検証することである。実際には、形態素解析を行い、品詞数が正しく得られる形に「YNU 書き言葉コーパス」(金澤編 2014)のデータを修正し、修正前のデータと修正後のデータを比較することで、どのような違いが見られるのかを明らかにし、修正の有効性を考察する。なお、今後出典が明示されていない引用・例文は全て金澤編(2014)からである。必要な部分のみ提示し、中略した場合のみ“(略)”を入れる。

現行の「YNU 書き言葉コーパス」は、「取り返えし(p. 121)」や「自分のからた(p. 128)」のように、「パソコンで入力すれば書けたであろう(p. 17)」という基準から修正されたことになっているが、実際には修正されていない例が少なからず見られる。この状態では量的調査に用いにくく、結果を応用する際にも問題が生ずる。そのため、量的調査に堪えうるコーパスとするには修正を行う必要がある。

そこで本稿では、既存のコーパスに修正の必要がある場合、どのように問題の箇所を探し出すかを紹介する。そして実際にどのような箇所で修正が行われたかを明らかにし、修正前・修正後の品詞数の比較を通じて、修正の有効性について考察する。

2. 「YNU 書き言葉コーパス」について

「YNU 書き言葉コーパス」は、金澤編(2014)によると「12 種類のタスクによる作文を手書きで書」(p. 14)いたものである。さらに指示は母語訳で与えられ、被調査者が作文を書く際には、時間の制限がなく、自分のペースで書き進めることができるが、辞書の使用は認められていないという特徴がある。

そして、この手書きのデータは電子情報としてテキスト化されている。テキスト化の際には二種類のテキストが作成され、一つは「オリジナルデータ」と呼ばれ、「被調査者が書いた通りに、できるだけ再現したもの」(p. 15)である。もう一つは「補正データ」と呼ばれ、「その後のデータベースでの検索の利便性を主な目的として、『オリジナルデータ』を補正したもの」(p. 16)である。補正のポイントは、具体例も挙げて 3.1 で詳述する。

上述の「取り返えし」や「自分のからた」のように、筆者は「補正データ」の補正が不十分である点が問題だと考える。次章で修正（本稿では「補正」を修正と呼ぶ）が必要な箇所の検出方法を示し、修正の方針について述べる。

3. 方法

3.1 で「YNU 書き言葉コーパス」の「オリジナルデータ」に対して行われた修正のポ

イントを紹介し、本研究での修正の方針と、不足していると考える修正の詳細について述べる。さらに3.2で、問題となる箇所の発見方法について説明する。なおデータ修正に関する先行研究は、管見の限り見当たらなかった。データの修正については、バージョンアップの際にデータ作成者から述べられることが多く[*1]、修正自体を扱った研究はないものと思われる。

3.1. 修正の方針

金澤他(2014)に示されている「YNU 書き言葉コーパス」のオリジナルデータに対する修正は、「誤漢字と送り仮名の修正(p. 16)」「全てひらがな書きで読みにくいものの漢字への修正(p. 16)」「文脈から推測できるものの修正(p. 17)」の3点にまとめられる。なお形式の面では、全て一文を一行に変更し、不要な改行・空欄は削除されており、これらには問題はないと考える。

基本的な判断基準として挙げられている「パソコンを使用したとすれば書けたであろう」というのは、この3点に関わってくると思われるが、実例を挙げて詳細に説明し、さらに本稿で新たに加える修正の方針についても説明する。

3.1.1. 誤漢字と送り仮名の修正

「誤漢字と送り仮名については適宜修正した」(p. 16)とあり、例えば「持たせていただきます」を「待たせていただきます」に、「卒校」を「卒業」に修正していることから、誤漢字の修正には、字形が似ているものの修正だけではなく、意味的に推測できるものも修正していると判断した。

さらに誤漢字の基本的な判断基準を「パソコンを使用したとすれば入力の誤りに気がついて正しく書けたであろう」(下線筆者、以下同)と拡張し、補正する。例えば、「とうしょうかん」は現在の補正データではこのまま入力されているが、入力した場合、変換されるのは(筆者のパソコンの場合)「東証刊」なので、入力の誤りに気づき「としょかん」と正しく入力できると考え、「図書館」に修正する。

3.1.2. ひらがなの漢字への修正

「全てひらがな書きで読みにくいものは漢字に変換して修正」(p. 16)とあるが、この場合の「もの」の解釈が問題となる。オリジナルデータには(1)のように、(文の途中までの)一行が全てひらがなで書かれたものはあるが、「文」全てがひらがなで書かれたものはなかった。そこで「もの」は「語」と解釈する。この修正方針により、「としょかん」を「図書館」に修正し、「たのみたいこと」を「頼みたいこと」に修正したものと判断する。

(1) あなたのまわりにあなたのことをしんぱいしてるみんなが (task_05_K015)

[*1] 例えば台湾・東呉大学で作られた「LARP at SCU」コーパス内の「00-第2版ご案内 .pdf」など

3.1.3. 文脈から推測できる語句の修正

(2)のように、オリジナルデータでは再現が不可能な文字として「●」で表されている語でも、推測できるものについては再現し、この場合「『困っています』に入力し直した」(p. 17)とある。同様に、文脈から推測できる語句の修正として、母語話者としての判断で推測できるものは修正する。これはタスクが決まっており、文脈が明確なため可能となるが、推測の可能性が複数ある場合は、後述の通り修正は行わない。

(2) たのみたいことがあって(略)としょかんではないようで●っています。
(task_01_K029)

3.1.4. 語の品詞抽出を目的とした修正

形態素解析をして正しい品詞に区切ることができるような修正を加える。ここが本稿のオリジナルである。なお修正を徹底しても、語として抽出できるとは限らない場合がある。そこで、修正の際には形態素解析の結果を考慮し、日本語として正しく表記されていても解析結果に問題があれば、正しく解析できることを優先する。

例えば、表記の問題がない(3)を形態素解析にかけると、「長ねぎとたまご」は「長:形容詞」「ね:助詞」「ぎとたまご:名詞」と解析されてしまう。そこで、この場合は「長ネギと玉子」と表記を変更する(「たまご」は「玉子」でも「卵」でも可能であるが、統一した)。同様に(4)の「○○町」は、「○(記号)/○(名詞-数詞:ゼロ)/町(名詞)」と3語に解析されるので、「○○ 町」のように、間に全角空白文字を入れることで「○(記号)/○(記号)」となるように修正した[*2]。

このように、修正には形態素解析用辞書に依存する側面もあるが、本調査で使用するMeCab[*3]とIPA辞書[*4]で形態素解析できるようになれば、(語の区切り方は異なるが)UniDic[*5]でも解析できるようになる。ただし(5)の「おそれ」のような語の場合、そのままでは「お(接頭詞)/それ(名詞・代名詞)」になり、修正が必要となる。意味を考えると「恐れ・畏れ」などではなく「虞」にしなくてならなくなり、被調査者が書いていない漢字を使用していいのかという問題も生ずる。しかしこのような修正を加えることにより、語として正しく区切られ、被調査者の意図に沿ったデータとなる。

次にどこまで正しいものを目指すかであるが、品詞の大分類が一致すればよいとし、小分類や読みまでの一致は考えない。このような修正を加えることで、品詞数を対象とした調査ができる頑強なデータができあがると考える。

(3) 少しあとに長ねぎとたまごを入れます。 (R_task_09_K004)

---

*2 詳しくは長谷川・西尾(2017)を参照

*3 様々な形態素解析用辞書を使用することができる形態素解析エンジン

*4 「現代日本語書き言葉均衡コーパス」に利用されている形態素解析用辞書

*5 形態素解析エンジンChaSenとともに使われることが多い形態素解析用辞書

(4) 私は、○○町に住む町民の１人です。 (R_task_06_J008)
(5) 子どものアイデンティティに影響が出るおそれがあるため (R_task_10_C061)

3.1.5. 修正しない箇所

修正の方針について述べてきたが、修正しないものとその理由を説明する。

本研究は、正しく語に区切り、品詞数の正確な検出を行うことを目標とするため、方針に沿わない、動詞の自他の誤り(6)、助詞の誤り(7)、語の誤り等は修正しない。これを(8)を用いて説明する。「もめ豆腐」は後述の特殊拍に関する誤りとして「もめん豆腐」に修正するが、動詞・助詞の誤りである「入って」「をつかめやすい」を「あって」「が染みやすい」に修正することはしない。

また推測できるものの候補が複数ある場合、どちらか分からないものは修正しない。(9)の「おかけてください」は「かけてください」と「おかけください」の可能性があり、(10)の「苦らく」は「つらく（辛く）」と「くらく（暗く）」の可能性が考えられる。一意に決定できないものについては、修正しないこととする。

(6) 計画的にできるように経済面の支援を増えるべきと考えて (R_task_04_C008)
(7) 留学生といっても優秀な留学生が来てほしいと思います。 (R_task_04_C003)
(8) もめ豆腐には小さなすきまがいっぱい入って醤汁をつかめやすい (R_task_09_C008)
(9) ご不明な点があれば、気軽く、声をおかけてください。 (R_task_07_C045)
(10) その当時はとても苦らく長い時間が過ぎていくのですが、 (R_task_09_C008)

3.2. 修正箇所の発見方法

本稿では、形態素解析の結果が文構成・語構成の面から正しくない品詞の連続となっている部分を「問題の箇所」とし、その中で本稿の方針に従って何らかの変更を加える箇所を「修正箇所」と呼ぶ。一部には必ずしも正しいわけではないが、望ましい結果を得るための変更もこれに含める。

本稿では、正しく語として解析できることを目的とし（特に品詞）、必要となる修正箇所の発見は形態素解析の結果を目視することで行う。形態素解析用辞書にIPA辞書、形態素解析エンジンとしてMeCabを選択するが、それは「日本語の形態素解析エンジンの中で最もよく使われている」（末吉2019, p23）からである。また「MeCabではIPA辞書(mecab-ipadic)がデフォルトのシステム辞書として使われている（同書）」とあり、使用者の多いKH Coderでも、使用する辞書はIPA辞書がデフォルトであることを考えれば、形態素解析用辞書としてIPA辞書の選択が妥当であると判断した。

また文のチェックに形態素解析を使う理由は、「幼稚園の配布文書コーパス」の作成の際、表記の揺れや誤入力の発見に有効だと判明したからである（長谷川・西尾2017）。例えば、「授かった」の送り仮名が異なる「授った・授ずかった」は、(11)(12)のようになり、語構成の面から正しくない品詞の連続となっているため、目視で誤り

が発見しやすい。本稿でもその方法を用い、形態素解析した結果を目視で確認し、後述の方針に従い修正する。そして、この作業を二回行うことでデータの精度を上げる。

なお、目視による事前調査では見落とされた「成積・観迎」（→成績・歓迎）等も、本方法を用いることで発見でき、形態素解析を誤りの検出に使用することは、学習者の作文でも有効であると思われる（“→”は正しい形で修正後を示す）。

(11)　授【名詞】／っ【動詞】／た【助動詞】
(12)　授【名詞】／ず【助動詞】／かっ【動詞】／た【助動詞】

正しいかどうかの判定には、「境界」「品詞」「語彙素」の段階が考えられる（小木曽2014）。「境界」は語の区切りが正しいかどうかで、「品詞」は「境界」に加えて語の品詞が正しいかどうか、「語彙素」はいわゆる辞書の見出しに相当し、語の境界・品詞に加えて語の同定が正しく行われたかどうかで、この順に厳しくなる。(13)(14)(15)が「境界」「品詞」「語彙素」のレベルで正しく行われていない例である。(13)は語の区切りが正しく出来ていない例であり、「長ねぎ／と／たまご」と区切られるべきところである。(14)は、語としては正しく区切られているが、「動詞」と誤った判定がされている。(15)は区切りも品詞も正しいが、「おりる」の活用形と判定されている。なお、“／”は語の区切り、“【】”は形態素解析で割り出された品詞、“《》”は語彙素である。

本稿では「品詞」レベルでの修正を行う。「語彙素」レベルで正しいものに修正するには、IPA 辞書では表記の多様性に対応できず、望むことができない。具体的には、(15)の場合、「布を織りました」のように表記を書き換える必要があるため、今後の課題になる*6。

(13)　長【形容詞】／ね【助詞】／ぎとたまご【名詞】　(R_task_09_K004)
(14)　私／は／ある【動詞】／日　(R_task_05_K026)
(15)　布／を／おり《おりる:降りる》／まし／た　(R_task_12_J027)

4.　修正箇所の特徴

上述のように、データを形態素解析し、その結果を確認しながら本文を修正し、さらにもう一度同様の作業を行った。その中での修正を全て列挙することは、紙幅を越えるので、修正の徹底が必要なもの、特徴的なものを挙げる（内容の重複もある）。

4.1.　送り仮名の修正

送り仮名の間違いは、「見かって・残こして（→見つかって・残して）」のような動詞だけではなく、(16)(17)のように形容詞「早れば（→早ければ）」・形容動詞「気

*6 表記の多様性に配慮し、形態素解析用辞書にUniDicを使用するという選択もあるが、語を細かく分割しすぎるという別の問題が発生してしまう

軽るに（→気軽に）」で見られた。

(16) 私も外国語の教育は早れば早いほど始めること (R_task_10_C010)
(17) 今や北京や上海などどの都市でも気軽るに食べられます。 (R_task_09_C049)

4.2. 漢字の修正

(18)のように、音（オン）は正しいが形が正しくないものなどは、パソコンを使用すれば適切なものに変換できると考えられる。他にも「間単・救求車（→簡単・救急車）」など様々なものが見られた。量的な結果には反映しないが、結果としてこの種類の修正が多かった。

他にパソコンを使えば書けるであろう漢字には、(19)の「期問（期間）」のように部首が同じである似た漢字や、「分析・生義・経騒（→分析・生姜・経験）」のように字形の似た漢字が該当する。こんな形だったかと思って書いているものと推測するが、どのような被調査者か調べたところ、韓国語母語話者と日本語母語話者のみで、中国語母語話者はいなかった。

(18) ちょうど就職活動の際中に、肺炎にかかっちゃって (R_task_05_J009)
(19) でも長い期問おいておいて (R_task_09_K032)

4.3. 特殊拍に関連する修正

特殊拍は長音・撥音・促音とする（窪薗 1999）。特殊拍が挿入・追加されているものには(20)(21)、特殊拍が追加されているものには(22)、欠落しているものには(23)(24)(25)、長音の位置が違うものには(26)のような例が見られた。

(20) そして、オプショーンになるものは (R_task_09_K020)
(21) おれが後にいるからお前は前だけみてさっさっと行け。 (R_task_05_K020)
(22) その後韓国ほんばんの味も感じてみてください！ (R_task_09_K019)
(23) 2005年までには線形でだんだ下がって (R_task_03_K032)
(24) いろいろな野菜、ピマン、人参、たまねぎ、などを (R_task_09_K011)
(25) 「二人にお仕置きしなきゃ」、とおしゃった。 (R_task_12_C059)
(26) グロバール時代にふさわしい、時代に遅れない人に (R_task_10_K034)

4.4. 清音・濁音に関連する修正

濁音が清音になっているもの(27)、清音が濁音になっているもの(28)、濁音の位置が間違っているもの(29)が見られた。学習者の書き言葉コーパスには様々なものがあるが、これらは他のコーパスにはあまり見られない特徴である。

最後に特殊拍と清濁の要因が複数見られる例があり、「大道市（だいどうし）→大都市（だいとし）」のように濁音化と長音が追加された(30)が挙げられる。

(27) 今一番重要なことは自分のからたに大事にして、 (R_task_05_C025)
(28) 情けない～だぶん始まった時間が遅くなったせいかも。 (R_task_11_C006)
(29) 皆も時間があれば、せび中国本場の作り方で餃子を (R_task_09_C047)
(30) 日本の大道市、たとえば東京とか横浜と似ているようだが (R_task_07_K019)

4.5. ひらがな書きされた語の修正

ある語が、全てひらがな書きされているものには、「つまづく（躓く）」「ひげ（髭）」「かささぎ（鵲）」のように難しい漢字が書けず、ひらがなになったと思われるものもあるが、読み手への配慮でひらがな書きされていると思われるものもある。

例えばタスク 12 は「あなたは、小学校新聞の昔話コーナーで、あなたの国の昔話を書いてほしいと頼まれました。新聞の発行が７月なので『七夕伝説』のストーリーを書くことにしました。小学生にもわかるように、どのような話か詳しく書いてください」(p. 34) である。日本語母語話者の作文には、(31) のように「織姫・彦星」と漢字表記されているものもあれば、読み手が小学生であることを配慮してか、(32) のように「おりひめ・ひこぼし」とひらがな表記しているものもある。問題となる例は他に「おせち・しんで・おたがい・はなればなれ・むこ・むすめ・おつかれさま」など多数見つかった。これらは漢字（一部漢字）に変換した。

(31) 織姫と彦星は出会って一目でお互いを好きになり (R_task_12_J015)
(32) ひこぼしは牛にえさをあげることができず、おりひめははたをおることができなくなってしまいました。 (R_task_12_J001)

4.6. 形態素解析の結果を考慮した修正

これまでの分類に当てはまらない修正について述べる。(33) の「所謂」のように、漢字表記のままでは正しく語に区切れないため（所【名詞・固有名詞】謂【名詞・一般】）、表記の変更が必要になるものもある。さらに地名・商品名・料理名等も、一語として解析されるように適宜変更が必要である（紙幅の都合で詳細は省略する）。

なお書き忘れか理由は不明だが、「お土産」の一部が欠けている (34) のような例も多数見られたが、これらもパソコンを使用していれば正しく書けると推測する。

(33) 日本より収入が少なく、所謂発展途上国であることも (R_task_04_K009)
(34) 大連は港町であり、海産物が(略)有名です。お産としていい (R_task_07_C010)

4.7. 交ぜ書きの修正

4.5 のように全てひらがな書きではなく、一部がひらがなになっているものも見受けられた。(35) (36) はパソコンがあれば「準備・お年寄り」と書けたものと判断する。他に「仕ごと（→仕事）・出し（→出汁）・沸とう（→沸騰）」等が見つかった。

(35) 就職する前の準びと卒業論文の材料の収集もできる (R_task_05_C036)
(36) 高齢化が進む中、お年よりの方々はリハビリセンターまで (R_task_06_C049)

4.8. 文脈から分かる修正

4.8.1 のように音から正しい形の推測が容易な修正から、4.8.2 のように内容を読み込まないと判断できない修正まで様々である。いくつか出現数が多かったものを挙げる。

4.8.1. 表現・語順の逆転した語の修正

「補正データ」には(37)のように表現・語順の一部が逆転しているものが見られ、他には「介紹（→紹介）・分部（→部分）・値価（→価値）・故事（→事故）・響影（→影響）・したかがなく（→しかたがなく）・ことろ（→ところ）」などがあり、学習者だけでなく(38)のように母語話者の作文にも存在した。

(37) 以上の理由で私は市民総合病院の続存を請求します。 (R_task_06_C010)
(38) ○○もいはま不安でいっぱいだと思う。 (R_task_05_J019)

4.8.2. 読みと文脈から推測できる語の修正

読みと文脈の類推から分かるものとして、(39)の「約間（若干）、(40)の「道歩（徒歩）」などがある。また(41)「少ない」と「小さい」の混同の例も見られた。また●については(42)の「震災復興」のように文脈から推測が可能な語がいくつか見られた。

(39) 2008年には約間上昇し、2009年にはもとの9割まで回復して (R_task_03_K005)
(40) ミョンドンを中心にして道歩で6時間 (R_task_07_K021)
(41) 外国語は少さい時習うのが一番だということかな。 (R_task_06_C043)
(42) 日本の震●復興の力になれるように頑張っています。 (R_task_04_C008)

4.9. 忠実な再現が原因と思われる語の修正

ここからは特徴的といえる点を挙げる。まず観察すると、手書きされたものをなるべく忠実に入力しようとした事が分かる例がある。例えば、(43)の「教育」は「育」に点がなかったものをそのまま再現したと考えられる（なお後続文では「育」と正しく書けている）。(44)は母語話者の例であるが、手書きされた「よりよい」の「い」が「り」に近かったのではないかと推測される。忠実な再現を目指すのか、書き手の意図に沿った表記にするのか、事前の方針決定が重要であると思われる。

(43) 早期英語教育の意識調査(略)以上の英語教育を受けました。 (R_task_10_K010)
(44) 学生のよりより学校生活のためにご検討の程、 (R_task_04_J030)

4.10. 異なる文字コードで書かれたファイルの修正

「YNU 書き言葉コーパス」には、異なる文字コード、実際には S-JIS と UTF-16 が用いられているファイルが存在していることが分かった。「補正データ」の説明の中で「中国語母語話者の場合にも、漢字の誤りや中国の簡体字で記したため、日本語での入力が難しいものも見られ（略）正しいと考えられる漢字を推測して、それを入力し直した」(p. 16)とあるが、実際には(45)のように簡体字で入力されているものがある。これは、地元の名所を紹介する際に、S-JIS では「灵」が表示できないため、文字コードを変更して忠実に文字を表したものと思われる。なお、ある被験者のデータは全て UTF-16 であるなど、入力の際の方針は不明である。UTF-16 で入力されたファイルは 29 あったが、一括処理の際に問題となるため、全て S-JIS に統一し、S-JIS で該当する漢字（(45)の場合は「霊」）への変更を行った。

(45) 歴史のあるところと言えば、灵隐寺というお寺は (R_task_07_C003)

4.11. 入力時の誤りの修正

文脈から考えて明らかにデータ入力時の誤りだと思われる例がある。例えば(46)は「糸こんにゃく」、(47)は「身につける」の誤った例だと思われる。また(48)の「爽快」はその後に出現する「総会」の誤りだと考えられ、同様の誤りは他にも見られた。なお(48)のような誤りは形態素解析だけでは見つからず、目視が必要な例である。

(46) そして、野菜や糸こんなｙく や豆腐などを丸く順番に (R_task_09_J027)
(47) しっかり正しい日本語をに見つけるべきだと思う。 (R_task_11_J030)
(48) 先日の学生爽快で出た(略)。先日の学生総会におきまして、 (R_task_04_J030)

4.12. 混入している他言語の修正

中国語や英語の語・表現が使われている例も入っている。(49)(50)は語として抽出するため修正を行ったが、(51)(52)はそのままとした。このようなものをどのように扱うか、研究を行う際に計量の方針を決定する必要がある例である。

(49) 例えばにら、猪肉、えびで中身を作る場合 (R_task_09_C010)
(50) 今度、居民の声を反映します。 (R_task_06_C026)
(51) 中国語の中には(略)ことわざがあって(略)「吃在中国」と (R_task_09_C033)
(52) Have a good time in Dalian (R_task_07_C026)

5. 修正前後の語数の比較

上記のように修正を行ってきた。漢字の修正などは一語一語異なるので、数値での比較に向かない。しかし有効性について考察するため、修正の結果が有意な差として

出るのはどこかを明らかにする。手順としては以下の3種類の調査を探索的に行う。品詞数の計量には、RMeCab（石田 2017）を用い、検定にはカイ二乗検定を用いる。

調査1．修正前の全データと修正後の全データの品詞数はどこが違うか
調査2．母語話者別では、修正前データと修正後データの品詞数はどこが違うか
調査3．タスク別では、修正前データと修正後データの品詞数はどこが違うか

具体的に述べると、調査1は修正前の全てのファイル（1,080）と修正後の全てのファイル（1,080）の品詞数の比較を行う。

調査2は、修正前の日本語母語話者、中国語母語話者、韓国語母語話者のそれぞれのファイル（360）と修正後のファイル（360）の品詞数の比較を行う。

調査3は、12 のタスクそれぞれについて、話者の属性は考慮に入れず、修正前のファイル（90）と修正後のファイル（90）の品詞数の比較を行う。その結果、有意な差が見られたもののみ、ファイルをさらに母語話者別に分割し、比較を行う。

表1. 修正前後の品詞頻度（*<.05、**<.01）

順位	品詞	修正前	修正後
1	名詞	74382	73252
2	助詞	61613*	61509*
3	動詞	30940**	29887**
4	助動詞	22563	22500
5	副詞	5407	5385
6	形容詞	3776	3753
7	接続詞	2209	2228
8	連体詞	1983	2011
9	接頭辞	1494**	1312**
10	感動詞	523	505
11	フィラー	63	33
12	その他	2	0
	合計	204955	202375

5.1. 全データの比較

全データを比較した結果が表1である。全データをまとめた場合、助詞の使用度数は、修正前・修正後において有意水準 5% で差があった（$\chi^2(1)=5.30$、$p<.05$）。さらに動詞（$\chi^2(1)=8.59$、$p<.01$））、接頭辞（$\chi^2(1)=9.56$、$p<.01$）については、有意水準1%で差が見られた。

どの話者のデータの修正に現れるのか、どのタスクに見られるのか、さらに観察する。

5.2. 母語話者別の比較

母語話者別の修正前・修正後データを比較した結果を示す。日本語母語話者の場合、頻度に有意差が見られたのは、助詞（修正前頻度 19,698、修正前全体頻度 65,174、修正後頻度 19,655、修正後全体頻度 63,955、$\chi^2(1)=3.92$, $p<.05$）、動詞（修正前頻度 10,573、全体頻度 65,174、修正後頻度 9,764、修正後全体頻度 63,955、$\chi^2(1)=22.149$, $p<.01$）であった。

日本語学習者は、ともに接頭辞で有意差が見られ、それぞれ中国語母語話者（修正前頻度 554、修正前全体頻度 72,346、修正後頻度 480、修正後全体頻度 71,559、$\chi^2(1)=4.41$, $p<.05$）、韓国語母語話者（修正前頻度 450、修正前全体頻度 67,435、修正後頻度 377、修正後全体頻度 66,861、$\chi^2(1)=5.70$, $p<.05$）であった。

次にどのようなタスクのどのような修正として現れるのか、タスク別に見ていく。

5.3. タスク別の比較

修正前のタスクと修正後のタスクで有意差が見られたのは、12タスク中2つ、タスク6とタスク12であった（頻度10以上のもの）。

具体的には、タスク6の接頭辞（修正前頻度86、修正前全体頻度19,553、修正後頻度59、修正後全体頻度19,372、$\chi^2(1)=4.44$, $p<.05$）、タスク12の名詞（修正前頻度11,994、修正前全体頻度36,588、修正後頻度12,125、修正後全体頻度35,888、$\chi^2(1)=8.19$, $p<.01$）、動詞（修正前頻度6,534, 修正前全体頻度36,588, 修正後頻度5,741, 修正後全体頻度35,888, $\chi^2(1)=44.485$, $p<.01$）に有意差が見られた。上記の母語話者別の結果はこれに対応すると思われ、接頭辞は学習者データ、名詞・動詞については母語話者データが関係すると思われる。そこで、実際にどのような修正が行われたのか、探索的に調査を行う。

5.4. タスク別・母語話者別の比較

タスク6の学習者データで接頭辞を確認した。その結果が表2で、修正前と修正後のデータでは、「再」の減少が目立つ。確認したところ、これは全て韓国語母語話者の例で、(53)(54)のように「再活・再活治療」という表現を使っており、これらの語は形態素解析用辞書にないため、「再（接頭辞）・活」のように解析されたものと思われる。

表2. 修正前後の接頭辞（頻度3以上）

	修正前		修正後	
順位	語	頻度	語	頻度
1	再	28	お	15
2	お	14	再	9
3	ご	3	第	3
4	総	3		
5	第	3		

(53) そのゆえ、再活治療は多くの人々が必要としています。 (R_task_06_K003)
(54) ふういん科、再活治療科などもなくなるので (R_task_06_K012)

これらの表現は延べ27回、14人の作文に現れた。韓国語母語話者に確認したところ、「韓国語をそのまま（日本語の）漢字にしたという印象」（丸括弧内は筆者による）で、直訳と言える表現だそうである。このように、修正の効果が、あるタスクに局在する可能性もある。

タスク6では韓国語母語話者の修正が目立ったが、中国語母語話者の場合、接頭辞の修正はどうなのか、5.2のデータに対し、接頭辞に着目して確認した。その結果、修正前と修正後の接頭辞で出現に差が見られたのは、「旧（前6、後2）・故（前6、後0）・無（前6、後0）・急（前5、後0）・老（前5、後1）」であった。「旧」は6人に見られたものであるが、「無」は1人が6回使っており、使用の傾向が異なる。「故響（→故郷）」のような誤り(55)や、地名(56)、「故里・急救車」のように中国語をそのまま使っていると考えられる例(57)(58)が多い。

(55)　おすすめしたいのはもちろん私の故郷—湖北省です。　(R_task_05_C046)
(56)　無錫は中国の東にあるので、天候は暖かく、　(R_task_07_C043)
(57)　故里は山が多くて、森は良く見えます。　(R_task_07_C038)
(58)　それで、急救車も呼ばれたの！　(R_task_08_C003)

タスク別には、中国人学習者はタスク 7 (老・無)、タスク 8 (救)、タスク 9・12 (旧) と、遍在する傾向が見られた。

次に、日本語母語話者のタスク 12 における動詞と名詞の修正前と修正後のデータを確認する (表 3)。

表3　タスク 12 における修正前後の動詞上位 20 語

	修正前		修正後			修正前		修正後	
順位	語	頻度	語	頻度	順位	語	頻度	語	頻度
1	する	288	する	282	11	ひむ	71	思う	37
2	なる	198	なる	193	12	織る	57	言う	36
3	ひめる	132	いる	124	13	働く	56	できる	31
4	こぼす	127	しまう	93	14	いる	48	見る	31
5	ひる	127	会う	76	15	せる	46	あげる	25
6	いる	123	働く	57	16	会える	37	作る	25
7	おる	112	織る	54	17	思う	37	見つける	24
8	しまう	93	いる	48	18	言う	36	怒る	23
9	おりる	87	せる	45	19	できる	31	遊ぶ	23
10	会う	73	会える	37	20	見る	31	探す	20

網掛けをした「ひめる・こぼす・ひる・おる・おりる・ひむ」は、修正前のみに現れる動詞で、これらを除くと修正前動詞 20 位の語と、修正後動詞 14 位までは一致していることが分かり、これらが適切に修正されたものと思われる。具体的には、「おりひめとひこぼしが」のような表現は、修正前はひらがな書きのままだったので、(59) のようになるが、漢字表記に修正することによって(60)のようになる。

(59)　おり【動詞】ひめ【動詞】と【助詞-格助詞】ひ【動詞】こぼし【動詞】が【助詞-接続助詞】
(60)　織姫【名詞】と【助詞-並立助詞】彦星【名詞】が【助詞-格助詞】

このように適切な修正により、修正後のデータでは、動詞が減り、名詞が増えたと考えられる。実際に名詞を確認すると、表4のような結果となり、修正後には「織姫・彦星」が加わっていることが分かる。

しかし有意差のある修正となった全データの助詞・母語話者の助詞については、個別のタスクを調べてみても、顕著な増減は現れなかった。これは、全データの助詞は幅広いタスクで修正されたため、個別のタスクでは現れなかった可能性があり、母語話者の助詞は修正後の全体の語数の減少の影響が考えられる。この点についてはさらに詳細な分析の必要がある。

表4. 修正前後の名詞（上位5語）

	修正前		修正後	
順位	語	頻度	語	頻度
1	人	252	織姫	262
2	天	150	人	252
3	2	141	彦星	156
4	仕事	109	天	150
5	二	101	2	141

6. まとめ

「YNU 書き言葉コーパス」を対象とした場合、語数を正確に計量するには幾多の問題があることが分かった。ひらがなや漢字表記の修正など問題の多くは、一件ずつの個別のものであり、量的な調査に向かない性質を持つ。しかし、全体としてどのような変化が見られたかを明らかにするため、修正箇所を探し修正を加えたデータに対して、品詞という観点から修正前と修正後の変化を調べ、有効性を検証した。その結果、名詞・動詞のように特定のタスクに存在する修正と、接頭辞等のように複数のタスクに遍在する修正が見られた。また、助詞のように特定のタスクに現れず広く行われる修正も見られ、品詞・タスクによっては有効な修正だったと考えられる。以上のように、正確な結果を得るため、調査対象とする表現・品詞によっては、適宜修正して使用することが求められる。

7. 今後の課題

本研究では、今後必要な修正を加えた品詞数計量のためのバージョンを作成・公開したい。作成の方針としては、全体を変更したバージョンにすることも考えられる。例えば、アジア圏の英語学習者コーパスである ICNALE には、提出された作文をそのまま収録したものだけでなく、学習者の意図が反映された“Edited Essay Module”（Ishikawa2018）のように、全体を変更したバージョンも含まれている。

これに対し I-JAS では、「T:ポーズ、長音の訂正」「G:発音や活用の誤り」「K:PC 入力時の変換ミス」等のように、「形態素解析の自動処理の精度を向上させたり、データの利用者に有益な情報を提供することが可能になるよう」（細井・八木 2020）、付属的な情報を付与するタグセットを開発し、本文にタグを付けるという方針も考えられる。

さらに本文とは別に、特記事項を保有するという方針もある。「日本・韓国・台湾の大学生による日本語意見文データベース」では、母語話者データの漢字や文法の誤用に関しては、本文とは別の部分に、「『誤楽』原文のまま」・「『窓』の『八』なし」・「『賛』

は上の部分の『夫』が『先』(中国語の字体)」のようにコメント（但し、韓国の大学生／台湾の大学生の文法の誤用に関するコメントはない）が追加されている。

手書きされた書きことばの場合、本稿で考察した「YNU 書き言葉コーパス」だけでなく、「日本・韓国・台湾の大学生による日本語意見文データベース」も、漢字の誤りやひらがな表記の語などの問題を含んでおり、タグや特記事項の情報があることで、形態素解析後の人手による処理が容易になると思われる。今後は、語彙抽出のためのバージョンも構築していく中で、どのような形式がより望ましいのか、試行錯誤していきたい。さらに、本稿の結果が反映されたバージョンが公開できるような手立てを考える予定である。

「コーパスと統計は相性が悪い」(伊藤 2003) と言われているが、コーパスを用いた日本語教育学研究を進めて行くには、少しでも統計との親和性を向上させていくことが大切である。そのためにも、コーパス側からできることとして、被調査者の意図を反映した正確な品詞数が算出できるコーパスの作成はその一歩になると考えられる。

参考文献

S. Ishikawa(2018)"The ICNALE Edited Essays: A Dataset for Analysis of L2 English Learner Essays Based on a New Integrative Viewpoint". English Corpus Studies, vol. 25, pp. 117-130

石田基広(2017)『Rによるテキストマイニング入門（第2版)』、森北出版

伊藤雅光(2003)「コーパスと統計」〔特集 コーパス言語学―〈1. コーパス言語学の発展〉〕『日本語学』22(5)、明治書院、pp. 26-35

小木曽智信(2014)「第5章 形態素解析」前川喜久雄監修・山崎誠編集『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

金澤裕之編(2014)『日本語教育のためのタスク別書き言葉コーパス』、ひつじ書房

窪薗晴夫(1999)『現代言語学入門2 日本語の音声』、岩波書店

末吉美喜(2019)『テキストマイニング入門 Excel と KH Coder でわかるデータ分析』、オーム社

長谷川守寿・西尾広美（2017）「語彙・文型調査を目的とした『幼稚園の配布文書コーパス』の作成について―特定の目的コーパスの作成例として―」、『人文学報』513-7、首都大学東京人文科学研究科、pp. 55-71

細井陽子・八木豊(2020)「第4章 音声データのテキスト化」迫田久美子・石川慎一郎・李在鎬編著『日本語学習者コーパス I-JAS 入門―研究・教育にどう使うか―』、くろしお出版、pp. 48-61

参照サイト

「日本・韓国・台湾の大学生による日本語意見文データベース」
http://www.tufs.ac.jp/ts/personal/ijuin/koukai_data1.html

（はせがわ もりひさ・東京都立大学）