2022/1/24

『語の文体値データ』(2022 年 2 月公開第 1 版)説明書

1. 概要

『語の文体値データ』には、『BCCWJ 図書館サブコーパスの文体情報』[1]を利用して算出した語の文体に関するデータ(専門度平均値、客観度平均値、硬度平均値、くだけ度平均値、語りかけ性度平均値)を収めている。

『BCCWJ 図書館サブコーパスの文体情報』は、文体研究を進めるために、『現代日本語書き言葉均衡コーパス』(BCCWJ)[2]の図書館サブコーパス収録のサンプルすべてに文体情報のアノテーションを行った成果を公開用にまとめたデータである。図書館サブコーパスの各サンプルに、「内容・表現の文体的特徴を表す分類指標」及び「形式・内容・表現に文体判断が単純にいかない特徴をもつものの分類指標」を付与している。

『語の文体値データ』の作成に当たっては、「内容・表現の文体的特徴を表す分類指標」である専門度 (specificity)、客観度 (objectivity)、硬度 (formality)、くだけ度 (softness)、語りかけ性度 (spokenness) を利用した。

専門度は「対象読者に想定される読解レベル(難易度)」に関わる分類指標である。客観度は「テキストの作成意図」に関わる分類指標である。硬度、くだけ度、語りかけ性度は「さまざまな文体情報」に関わる分類指標であり、硬度とくだけ度は「形式性、親疎性を問う」分類指標であり、語りかけ性度は「口語性を問う」分類指標である。これらの分類指標は、それぞれ言語データ構築経験有のおおよそ 20〜50 代の女性、延べ 9 名の作業者によって付与され、それぞれ次の 3 段階〜5 段階のいずれかの段階が付与されている。

(a)専門度　1 専門家向き/2 やや専門的な一般向き/3 一般向き/4 中高生向き/5 小学生・幼児向き

(b)客観度　1 とても客観的/2 どちらかといえば客観的/3 どちらかといえば主観的/4 とても主観的

(c)硬度　1 とても硬い/2 どちらかといえば硬い/3 どちらかといえば軟らかい/4 とても軟らかい

(d)くだけ度　1 とてもくだけている/2 どちらかといえばくだけている/3 くだけていない

(e)語りかけ性度　1 とても語りかけ性がある/2 どちらかといえば語りかけ性がある/3 特に語りかけ性はない

『語の文体値データ』では、図書館サブコーパスに使用された短単位[3]の語彙素(129,804 語)及び長

---

[1] 国立国語研究所(2015)『BCCWJ 図書館サブコーパスの文体情報』(2015 年公開第 1 版)(http://doi.org/10.15084/00003109)。『BCCWJ 図書館サブコーパスの文体情報』に関する説明は、この国立国語研究所(2015)の添付文書「BCCWJ 図書館サブコーパスの文体情報_説明書」及び 柏野和佳子(2013)「書籍サンプルの文体を分類する」『国語研プロジェクトレビュー』4(1), pp.43-53.に基づく。

[2] 国立国語研究所「概要 現代日本語書き言葉均衡コーパス (BCCWJ)」 (https://ccd.ninjal.ac.jp/bccwj/index.html)参照。

[3] 「短単位」は「用例収集に適した」単位であり「形態論的に一貫した言語単位」として認定されている単位である。「長単位」は「複合語を把握する」ことができ「サンプルの言語的特徴の解明に適した」単位である。(国立国語研究

単位の語彙素(821,510 語)のそれぞれに対して、専門度平均値、客観度平均値、硬度平均値、くだけ度平均値、語りかけ性度平均値を掲載している。ただし、専門度、客観度、硬度、くだけ度、語りかけ性度の段階が付与されていないサンプルに出現した語彙素も含まれている。この場合は、専門度平均値、客観度平均値、硬度平均値、くだけ度平均値、語りかけ性度平均値は算出できないため、「¥N」と記載している。

専門度平均値、客観度平均値、硬度平均値、くだけ度平均値、語りかけ性度平均値の算出方法の概要は次の通りである[4]。「短単位」「専門度平均値」を例として説明する。

図書館サブコーパスのサンプルすべてを対象として、ある短単位の語彙素(W_1)が出現するサンプルを特定する。出現するそれぞれのサンプルの専門度の段階の数字を『BCCWJ 図書館サブコーパスの文体情報』から得て、それを数値と見做して算術平均値を算出し、その値を W_1 の専門度平均値とする。ただし、W_1 が同一のサンプルに複数回出現する場合は、そのサンプルでの出現は 1 回と見做して算術平均値を算出する。

※専門度平均値、客観度平均値、硬度平均値、くだけ度平均値、語りかけ性度平均値及びその算出方法の詳細については、次の文書のほか後述する関連文献をご覧ください。

馬場俊臣(2018)「『BCCWJ 図書館サブコーパスの文体情報』を利用した語の文体差研究の可能性」『言語資源活用ワークショップ発表論文集』3, 国立国語研究所, pp.241-256.
(http://doi.org/10.15084/00001658)

2. 利用方法

『語の文体値データ』には、次の 8 種類のファイルが含まれている。

①WS_suw_100.csv (1.3MB)

短単位の語彙素のうち、出現サンプル数が 100 以上の語彙素(8,852 語)のデータ。CSV 形式。

②WS_suw_100.xlsx (1.3MB)

短単位の語彙素のうち、出現サンプル数が 100 以上の語彙素(8,852 語)のデータ。Microsoft Excel ファイル。

③WS_luw_100.csv (1.2MB)

長単位の語彙素のうち、出現サンプル数が 100 以上の語彙素(7,945 語)のデータ。CSV 形式。

④WS_luw_100.xlsx (1.2MB)

長単位の語彙素のうち、出現サンプル数が 100 以上の語彙素(7,945 語)のデータ。Microsoft Excel ファイル。

※①～④は、「Word_Stylistics-1.0.zip」(4.2MB)内にあります。

⑤WS_suw_all.csv (16.5MB)

短単位のすべての語彙素(129,804 語)のデータ。CSV 形式。

⑥WS_suw_all.xlsx (14.2MB)

短単位のすべての語彙素(129,804 語)のデータ。Microsoft Excel ファイル。

所コーパス開発センター(2015)『『現代日本語書き言葉均衡コーパス』利用の手引　第 1.1 版』(https://ccd.ninjal.ac.jp/bccwj/doc.html)に基づく。)

[4] BCCWJ の DVD 版公開データ(BCCWJ-DVD 版 Version1.1)の短単位データ及び長単位データ (「DISC2_NT」(NumTrans 版)の「TSV_SUW_NT」及び「TSV_LUW_NT」の各「LB.zip」) を用いた。

※⑤⑥は、「WS_suw_all.zip」(16.8MB)内にあります。

⑦WS_luw_all.csv (100.1MB)

長単位のすべての語彙素(821,510 語)のデータ。CSV 形式。

⑧WS_luw_all.xlsx (81.7MB)

長単位のすべての語彙素(821,510 語)のデータ。Microsoft Excel ファイル。

※⑦⑧は、「WS_luw_all.zip」(82.1MB)内にあります。

※本データは、

Creative Commons 表示 - 改変禁止 - 非営利 2.0 一般 (CC BY-NC-ND 2.0 JP)

で公開します。

(https://creativecommons.org/licenses/by-nc-nd/2.0/jp/)

3. 利用条件

『語の文体値データ』を利用した研究成果を発表される場合は、下記の情報を明記してください。

馬場俊臣(2022)『語の文体値データ』(2022 年 2 月公開第 1 版)

国立国語研究所(2015)『BCCWJ 図書館サブコーパスの文体情報』(2015 年公開第 1 版)

また、下記の文献のほか、参考にした関連文献を明記してください。

馬場俊臣(2018)「『BCCWJ 図書館サブコーパスの文体情報』を利用した語の文体差研究の可能性」『言語資源活用ワークショップ発表論文集』3, 国立国語研究所, pp. 241-256.

柏野和佳子(2013)「書籍サンプルの文体を分類する」『国語研プロジェクトレビュー』4(1), pp. 43-53.

○関連文献

馬場俊臣(2018)「接続詞の文体差の計量的分析の試み――『BCCWJ 図書館サブコーパスの文体情報』を用いて――」『北海道教育大学紀要 人文科学・社会科学編』69(1), 北海道教育大学, pp. 1-14.

馬場俊臣(2018)「接続詞の文体差の探索的分析――『BCCWJ 図書館サブコーパスの文体情報』5 指標を用いて――」『札幌国語研究』(23), 北海道教育大学国語国文学会・札幌, pp. 1-8.

馬場俊臣(2018)「複合接続表現の文体差――助動詞で始まる複合接続表現について――」『語学文学』(57), 北海道教育大学語学文学会, pp. 1-10.

馬場俊臣(2019)「BCCWJ 文体情報の各文体指標の特徴語――『BCCWJ 図書館サブコーパスの文体情報』を用いて――」『北海道教育大学紀要 人文科学・社会科学編』69(2), 北海道教育大学, pp. 1-14.

馬場俊臣(2019)「助詞・助動詞で始まる複合接続表現の文体差」『北海道教育大学紀要 人文科学・社会科学編』70(1), 北海道教育大学, pp. 1-11.

馬場俊臣(2020)「複合辞の文体差」『北海道教育大学紀要 人文科学・社会科学編』70(2), 北海道教育大学, pp. 1-12.

馬場俊臣(2020)「語の文体差を表す「文体値」について――「語彙密度平均値」と「硬度平均値」類との比較――」『札幌国語研究』(25), 北海道教育大学国語国文学会・札幌, pp.20-26.
馬場俊臣(2021)「副詞の文体の計量について――中俣論文の主成分分析結果との比較――」『語学文学』(60), 北海道教育大学語学文学会, pp.26-35.

4. 作成者、問合せ先

作成者：馬場俊臣
問合せ先：『語の文体値データ』に関するお問い合わせ、ご意見などは、馬場俊臣(baba.toshiomi@s.hokkyodai.ac.jp)まで電子メールにてお寄せください。

※本研究の一部は、令和3年度 大学共同利用機関法人 人間文化研究機構 国立国語研究所 「共同利用型共同研究（登録型)」の研究プロジェクト「「語の文体値データ」の作成及び公開」による研究成果である。