Universal Dependenciesコーパスを利用したレジスター研究の試み

伊藤 薫
九州大学大学院 言語文化研究院
ito@flc.kyushu-u.ac.jp

1 はじめに

人間はコミュニケーションの状況に応じて語彙や文法を使い分ける。このことは2節で紹介するように、ジャンル研究や文体研究として半世紀以上に渡り研究が積み重ねられてきた。語彙項目や構文の機能を探る上でコミュニケーションを取り巻く状況の考慮は欠かせないが、個別状況における使用傾向を探るにはまとまった言語データが必要となる。特に、統計的機械学習の隆盛により言語資源の需要は高まっており、以前にも増して様々なデータが蓄積・共有されるようになった。言語学においても言語使用を基盤とした実証的な研究が求められるようになり、広く公開されたデータの活用によって発展が見込まれる分野は相応にあると思われる。過去の文体研究者は手作業で言語的特徴の計量を行っていたが、コーパスの普及により以前と比較して簡便に計量的研究が行えるようになり、近年はオープンデータの増加によって更にこの流れは加速しつつある。

しかし、工学向けに構築されたデータを言語学に応用するにはいくつかの障壁が存在すると思われる。本論では**Universal Dependencies (UD)** [1] プロジェクトの一環として構築されたコーパスとそれに関連したコーパスを用いたレジスター研究を行う。UDは通言語的に統一された方法で依存構造や品詞などについてアノテーションが付与されたコーパス群を開発する国際プロジェクトである。依存構造が内容語主辞で付与されていること、**Universal POS (UPOS)** と呼ばれる品詞体系や係り受け関係の種類が定義されコーパスに付与されていることなどを特徴とする。

UDはそもそも工学向けのデータを提供するプロジェクトであるため、人手アノテーションのしやすさやパージング精度の高さなど、言語学的妥当性以外も考慮して設計されている [4]。このため言語学の一部の分野では応用が難しいと考えられるが、ジャンル・レジスター研究のような巨視的視点が取られる分野であれば、UD特有の事情に注意しながら活用が可能ではないかと思われる。そこで、本論では一番粒度の粗い品詞体系に焦点を絞り、UPOSを用いてレジスターの特徴を明らかにする。具体的には、UDの枠組みによるパラレルコーパスを用いた翻訳の影響の調査 (3節)、日本語書き言葉均衡コーパスを用いた既存の品詞体系とUPOS体系による文書分布の差についての調査 (4節) を行う。

2 先行研究

ある自然言語の内部に見られる変異は**言語変種**と呼ばれ、方言や文体など、様々な変異が包摂される。言語変種の下位分類は研究者の興味の対象により様々に分類され、Biber and Conrad[1] はジャンル・レジスター共に状況に依存した言語変種を対象とするが、**ジャンル**がある1つのテクスト全体を対象とするのに対し、**レジスター**はある言語変種に浸透し、変種全体に見られる特徴を興味の対象とすると定義している。

品詞とレジスターの関係についての研究として、樺島 [3] は早くも1954年にレジスター別に無作為抽出した文 (1レジスターあたり300文) について品詞割合の計量を行い、レジスターごとにその割合が違うことを指摘している。樺島は特に名詞の比率に着目し、文脈依存度の違いによって比率に差が出ると論じた。また、本論でも使用する日本語書き言葉均衡コーパスについては、山崎 [5] がBCCWJの全データを用いて樺島の指摘した名詞割合を追試している。

英語については Biber and Conrad[1] が体系的な研究成果をまとめている。当該書籍では上述したジャンル、レジスターの区別などの基礎に始まり、様々なレジスターについての事例研究を提供している。研究手法としては主に言語的特徴 (品詞に限らず、派生、人称、法助動詞など多岐にわたる) を関心事に応じて頻度の単純比較を行う手法と、因子分析を用いて話し言葉—書き言葉、手続き記述—内容記述などの対立を軸にレジスターの特徴と言語的特徴の関係を調べる手法が取られている。

[1] https://universaldependencies.org/, 本論で使用したコーパスはのバージョンは全て UD 2.4 である。

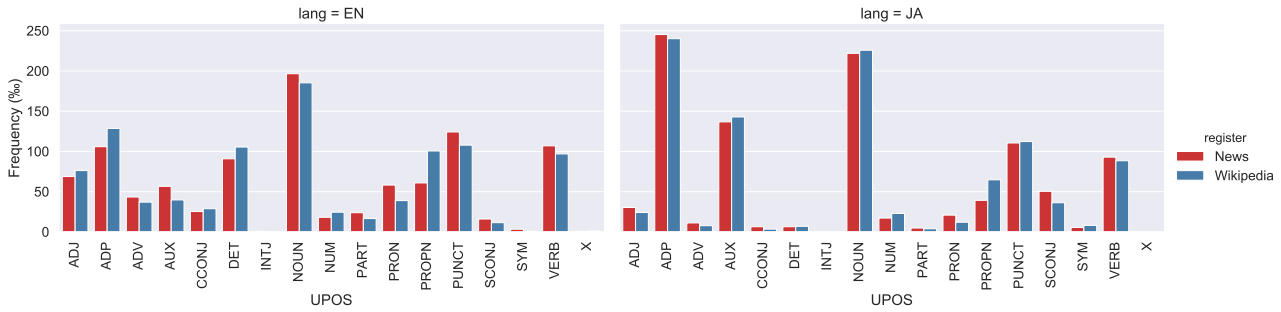

図 1: PUD ツリーバンクの UPOS 割合（左：英語、右：日本語）

3 CS1: PUD treebank

3.1 PUD treebank とは

Parallel UD (PUD) ツリーバンクは 2017 年の CoNLL Shared Task の段階で 1,000 文、18 言語からなるパラレルコーパスである [2]。1,000 文のうち 750 文は英語の文書から集められ、残りの 250 文はドイツ語、フランス語、イタリア語、スペイン語から収集された。各個別言語へは英語を介して専門家により翻訳され、手動でアノテーションが付与された。収録されている文は Wikipedia とオンラインのニュースワイヤから無作為に抽出されたものである。

3.2 手法

ニュースまたは Wikipedia の記事から無作為に抽出された 1 文が 1 文書となっているため、PUD ツリーバンクについては文書を単位として特徴を観察する研究には適切でない。一方、無作為抽出された文の集合はあるレジスターに普遍的な特徴を調査するには十分であるため、本研究では単純に各レジスターを構成する品詞の割合を比較する。比較は各言語内の構成比を概観した後、レジスター間の違いについて観察する。その後、日・英語間で品詞の構成比、レジスター間で差が見られた品詞について検討する。

3.3 結果

日・英両言語について、レジスター別の UPOS 頻度を千分率で図 1 に示す。

まず、各言語における品詞割合について述べる。英語において、単純な割合としてはどちらのジャンルでも名詞の割合が高く、概ね接置詞 (ADP)、決定詞 (DET)、句読点類 (PUNCT)、動詞の 4 つがそれに続いたが、Wikipedia レジスターでは固有名詞も上記 4 つと同程度の割合であった。レジスター間の相違としては、接置詞、決定詞、固有名詞の割合が Wikipedia レジスターで高い一方、代名詞、不変化詞 (PART)、句読点の割合がニュースレジスターで高かった。日本語では全体として接置詞、名詞の割合が高く、助動詞 (AUX)、句読点、動詞がそれらに続いた。レジスター間では割合に差がある品詞は英語ほど多く見られなかったが、固有名詞の割合が Wikipedia レジスターで高く、代名詞、従属接続詞 (SCONJ) の割合がニュースレジスターで高かった。

日・英語間の比較では、日本語で接置詞、助動詞、従属接続詞の割合が高く、英語で形容詞、副詞、等位接続詞 (CCONJ)、不変化詞、代名詞の割合が高かった。また、レジスター間で割合に差が見られた品詞については、固有名詞と代名詞で両言語に共通した傾向が見られたものの、英語のみで接置詞、決定詞 (Wikipedia で多い)、不変化詞、句読点 (ニュースで多い) に差が見られ、日本語のみで等位接続詞に差が見られた。

3.4 考察

PUD はパラレルコーパスであるが、どの言語でも基本的に同じ品詞体系 (UPOS) でアノテーションされている。固有名詞と代名詞の使用傾向に関しては日英両言語で同じレジスター的特徴が見られたが、上述した接置詞、決定詞、不変化詞、句読点、等位接続詞に関しては一方でレジスター間に使用傾向の差が見られるものの、もう一方では差が見られなかった。これは日本語と英語の間で固有名詞と代名詞に関して機能的差異がそれほど大きくない [2] のに対し、上述したその他の品詞に関しては差異が大きいことを示していると考えられる。

例えば、接置詞に関して英語では主語と目的語が語順によって表されるため単純な構造の節では前置詞が含まれないこともあるが、日本語では主語や目的語を表す際に格助詞が必須であるため、英語で文の複雑さと接置詞の頻度の関係が強いと考えられる。また、決定詞に関しては英語では冠詞が相当するが、日本語では主に直示に関わる連体詞 (この・その・あの、等) に相

[2] ただし、代名詞に関しては英語では項を埋めるため頻度が高くなりやすく、日本語では省略される傾向が高いという違いはある。

当し、英語での使用頻度が圧倒的に高い。助動詞に関しては、英語で認識的モダリティが主に助動詞によって担われるため、世界で進行する様々な事態についての予測を述べる必要があるニュースレジスターで多く用いられる傾向があるが、日本語では相当する機能が「〜だと思われる」「〜かもしれない」等、助動詞以外の手段によって示されることも多いため、このような違いが生じないと推察される。

このように、UD では通言語的に共通の品詞ラベルを用いるが、同じ品詞ラベルであっても別の機能を担うことがあるため、レジスター的特徴を調べる際は注意を要する。また、対照研究としては言語間で共通する文法カテゴリーだと見なされる品詞についての機能差を調べることができるという点で、UD コーパスの使用は興味深い。具体的には、先に挙げた助動詞の頻度差は統語論的な定義 (主動詞に伴って用いられ、テンスやムードを表すために用いられる) によって付与されたラベルと、コミュニケーションの状況や目的に要請された機能を満たす手段 (この例では認識的モダリティを表現するための手段) に言語間で隔たりや選好の差があることが示唆される。

4 CS2: UD Japanese-BCCWJ

4.1 UD Japanese-BCCWJ の構造

UD Japanese-BCCWJ は、国立国語研究所の**日本語書き言葉均衡コーパス (BCCWJ)** のコアデータに UD の枠組みでアノテートされたコーパスである。UPOS をはじめとする UD のアノテーション情報は元となる BCCWJ の情報をコンバートして付与された。コーパスに含まれる文書は**出版書籍 (PB)**、**雑誌 (PM)**、**新聞 (PN)**、**白書 (OW)**、**Yahoo!知恵袋 (OC)**、**Yahoo!ブログ (OY)** という 6 つのレジスターに属している。これらには更に上位のカテゴリーがあり、PB, PM, PN は出版サブコーパス、OW, OC, OY は特定目的サブコーパスに分類される。UD では BCCWJ の短単位を 1 語とみなして語分割されており、短単位に付与された品詞を UPOS にコンバートすることによって品詞情報が付与されている。本論の分析では、UD Japanese-BCCWJ で用いられている UPOS および、BCCWJ の短単位品詞 (大分類) [3] を用いて分析する。各体系の品詞一覧を以下に記す。

UPOS ADJ, ADP, PUNCT, ADV, AUX, SYM, INTJ, CCONJ, X, NOUN, DET, PROPN, NUM, VERB, PART, PRON, SCONJ

短単位 名詞, 代名詞, 形状詞, 連体詞, 副詞, 接続詞, 感動詞, 動詞, 形容詞, 助動詞, 助詞, 接頭辞, 接尾辞, 記号, 補助記号, 空白

[3]以下、単純に短単位 (英語標記では SUW) と称する。

品詞の細かい対応関係については浅原ら [4] の論文を参照されたい。

4.2 手法

UD Japanese-BCCWJ の文書あたり平均語数は約 733 語であり、文書単位の分析が可能であると思われる。本論では文書ごとの品詞割合を特徴量とし、レジスターと品詞割合の分布の関係を観察するため主成分分析によって次元圧縮を行い可視化した。分析には R (ver. 3.6.1) を使用し、主成分分析は prcomp, 可視化は ggbiplot パッケージにより行った。

4.3 結果

UPOS, 短単位データの主成分分析結果をそれぞれ図 2, 3 に示す。描画には各品詞体系の分析で得られた第 1 主成分、第 2 主成分を用いた。各点は 1 文書に対応し、点の色はレジスターを表す。また、原点から伸びる矢印は各主成分についての主成分負荷量を表す。図中の楕円は ggbiplot によって描画された各レジスターに対応する正規確率楕円である。なお、両図ともに紙幅の都合で描画範囲に収まらなかった文書が存在する。

寄与率に関しては、UPOS の場合第 1 主成分が 0.36、第 2 主成分が 0.25 で第 4 主成分までの累積寄与率が 0.81 であった。また、短単位の場合は第 1 主成分が 0.45、第 2 主成分が 0.31 で第 3 主成分までの累積寄与率が 0.88 であった。

4.4 考察

レジスターに関しては、両品詞体系ともに可視化結果から OW, PN, PM, PB の 4 つに関してフォーマルさについての連続した分布が読み取れる。UPOS では第 2 主成分の負方向、短単位では第 1 主成分の正方向に行くに従いフォーマルになっていると解釈できる。なお、上記 4 レジスターに関しては文書ごとの品詞割合に差異が少なかったが、OC, OY に関しては文書ごとの差が大きく図の範囲内に描画できなかった文書も多くあり、正規確率楕円も他に比べ大きくなっている。これは OC, OY がウェブユーザーが書き込み他者からの校正も入りにくく、他のレジスターに比べ自由に書くことができるのに加え、1 文書あたりの語数が少ないのも影響していると思われる。

また、両品詞体系の違いついては、名詞の細分化や助動詞に関するものが顕著である。一般的に名詞として分類される語は今回対象とした 2 つの品詞体系でさらに細かく分類されている。具体的には、短単位では接頭辞と接尾辞が名詞とは別の品詞として立てられており、UPOS では一般的に名詞に該当する語が NOUN, PROPN, NUM の 3 つに細分化されている一方で、短

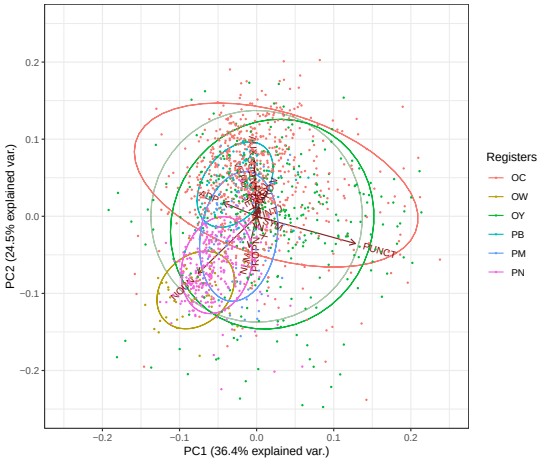

図 2: UPOS 頻度を特徴量としたレジスター別の文書分布. (原点付近)

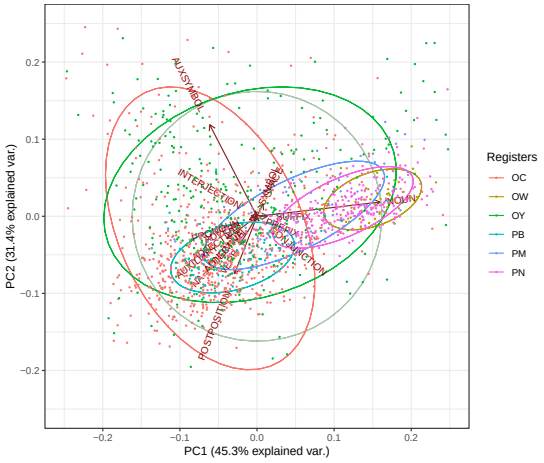

図 3: 短単位品詞頻度を特徴量としたレジスター別の文書分布. (原点付近)

単位では単に名詞としてまとめられている。短単位に関して名詞、接頭辞、接尾辞の主成分負荷量を見ると、大きさは異なるもののいずれも主に第 1 主成分の負荷が高くなっている。UPOS に関して主成分負荷量を見ると NOUN が第 1, 2 主成分ともに一定の負荷があるのに対し、PROPN, NUM に関しては主に第 2 主成分に関して負荷が高くなっている。また、助動詞に関しては UPOS で名詞と並んで第 2 主成分の負荷が大きいのに対し、短単位では第 2 主成分までの負荷は 0.2 程度とさほど大きくない。

手法が異なるため単純比較はできないものの、図 2, 3 に示したレジスターの分布に関しては、品詞体系の違いは PUD ツリーバンクで見た言語差がレジスター間の品詞割合に与える影響ほど結果に影響していないように思われる。

5 おわりに

本論では、UD コーパスとそれに関連するコーパスを対象とした事例研究を行い、UPOS 品詞体系を用いたレジスター分析の結果に与える影響について論じた。PUD ツリーバンクを用いた実験では、UPOS は同一のセットで複数言語のアノテーションに用いられるものの、同じラベルが付与されていても各個別言語の体系において、該当する品詞が実際の言語使用の場で担う機能が異なることに注意する必要があると示唆された。一方で、BCCWJ を用いた実験では、コンバージョンによる UD コーパスの作成ではある程度品詞体系が保たれているため、今回行った程度の粒度であればレジスター間の関係は UPOS を用いても捉えられることが示唆された。これらは単純に結論づけられる訳ではないが、今後も知見を積み重ねていき、言語資源の利用可能性について探りたい。

今回は品詞を特徴量とした分析を行ったが、UD では依存関係について単なる係り元・先だけでなく、係り受け関係の種類についても豊富なアノテーションが施されている。今後の研究ではこちらの応用についても可能性を探りたい。

謝辞

本研究の一部は JSPS 科研費 19K13180 の助成および国立国語研究所コーパス開発センター共同研究プロジェクトの支援を受けたものです。

参考文献

[1] Douglas Biber and Susan Conrad. *Register, Genre and Style, 2nd edition*. Cambridge University Press, Cambridge, 2018.

[2] Anders Björkelund, Agnieszka Falenska, Xiang Yu, and Jonas Kuhn. Ims at the conll 2017 ud shared task: Crfs and perceptrons meet neural networks. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pp. 40–51, 2017.

[3] 樺島忠夫. 現代文における品詞の比率とその増減の要因について. 国語学, Vol. 18, pp. 15–20, 11 1954.

[4] 浅原正幸, 金山博, 宮尾祐介, 田中貴秋, 大村舞, 村脇有吾, 松本裕治. Universal dependencies 日本語コーパス. 自然言語処理, Vol. 26, No. 1, pp. 3–36, 2019.

[5] 山崎誠. 品詞・語種の割合とテキストのジャンルとの相関. 日本語学, Vol. 33, No. 8, pp. 86–91, 2014.