Kobe University Repository : Kernel

PDF issue: 2026-10-06

英・日テキストの特徴語を自動抽出する「EJTKAN」の開発 : コーパス研究における特徴語分析の意義と可能性

石川, 慎一郎

(Citation)
統計数理研究所共同研究リポート,484:32-56

(Issue Date)
2026-03-25

(Resource Type)
departmental bulletin paper

(Version)
Version of Record

(JaLCDOI)
https://doi.org/10.24546/0100502562

(URL)
https://hdl.handle.net/20.500.14094/0100502562

英・日テキストの特徴語を自動抽出する「EJTKAN」の開発

―コーパス研究における特徴語分析の意義と可能性―

石川　慎一郎（神戸大学）
iskwshin@gmail.com

Development of EJTKAN, An Online Text Keyword Analysis Tool: Significance and Possibility of a Keyword Analysis in Corpus Studies

ISHIKAWA, Shin'ichiro (Kobe University)

Abstract

The author developed and released a new text analytic tool called English/Japanese Text Keyword Analyzer (EJTKAN). When text files are uploaded, EJTKAN automatically performs language identification, morphological analysis, word extraction, word frequency analysis, frequency comparison, calculation of the keyness statistics, and outputs a list of statistical keywords. This paper provides an overview of a keyword analysis as one of the major corpus query techniques, and it also introduces the results of a keyword analysis of story-writing essays produced by L1 Chinese students and native Japanese speakers.

キーワード
コーパス、特徴語分析、ETJKAN

1. はじめに

英語コーパス研究では、関心対象とするターゲットコーパス（target corpus）と、比較基準となる参照コーパス（reference corpus）に出現するすべての語の頻度を比較し、前者において顕著に多く（または少なく）出現している語を特徴語（keyword）として抽出する特徴語分析（keyword analysis）が広く実践されている。これは、WordSmith Tools（Scott, 1996-present）やAntConc（Anthony, 2014-present）といったコーパス処理用のコンコーダンサ（concordancer）に特徴語分析機能があらかじめ実装されているためである。特徴語分析は、コーパス間での頻度の差の有意性の判定に使う検定統計量などを根拠として、ターゲットコーパスの内容や言語様式の特性を反映した少数の語を客観的に取り出せるという点で、コーパス分析手法として多くのメリットを持つ。

ただ、特徴語分析を行うには、テキスト内で、個々の語が計量可能な単位として切り分けられていることが前提となるため、語の切れ目がない日本語のコーパス研究において、特徴語分析は英語の場合ほど普及していないように思える。また、特徴語の抽出には何らかの統計的尺度を援用することになるが、これらに対するなじみの薄さも日本語データを用いた特徴語分析の垣根を高くしていると言えるであろう。

そこで、筆者の研究室では、手元のテキストファイルをアップロードすれば、言語種別判定→形態素解析→単語抽出→単語頻度調査→頻度比較→特徴量算出→特徴語リスト出力までのプロセスを自動で行う English/Japanese Text Keyword Analyzer（EJTKAN）というツールを新たに開発・公開した。EJTKAN は Google Colaboratory（Google Colab）上のツールで、過去にリリースした English/Japanese Word Frequency Table Generator（EJWFTG）（石川, 2024；石川, 2025）の姉妹ツールとして位置付けられる。

EJTKAN の主目的は特徴語の抽出であるが、特徴度の評価指標として各種の統計量を出力するため、本ツールはまた、頻度の差の有意性や、差の実質的な大きさを示す効果量を一覧で調べるツールとしても使用可能である。

本稿は、特徴語分析の概要に触れた後、EJTKAN の概要、操作手順、採用した特徴度指標を紹介する。また、日本語学習者コーパスのデータを用いたケーススタディの結果についても報告する。

2. 特徴語分析

2.1 概要

コーパス研究における“keyword”という語は多義的で、まず、KWIC（keyword in context）検索における検索対象語と、特徴語分析で扱うテキスト特徴語を指す場合がある。また、後者に限っても、質的に選ばれた語（質的特徴語）と統計的に選ばれた語（計量的特徴語）を指す場合があり、さらには、ターゲットコーパス側で顕著に多く出現する「正の特徴語」（positive keyword）のみを指す場合と、正の特徴語と、ターゲットコーパスにおいて顕著に少なく出現する「負の特徴語（negative keyword）」を含む総体を指す場合がある。本稿では、以下、「特徴語」を「統計的に選ばれた正負の特徴語の総体」という意味で用いる。

特徴語分析の普及に大きく貢献したのは、英国リヴァプール大学（当時）の Mike Scott 氏である。同氏は、差の有意性を示す検定統計量を「特徴度」（keyness）の指標として再定義し、自身が開発するコンコーダンサの WordSmith Tools（1996-present）に特徴語の自動分析機能を実装した。

Scott（1997）は、ターゲットコーパスにおいて特異な頻度特性を示すものを特徴語と定義している。

(1) A key word may be defined as a word which occurs with unusual frequency in a given text. This does not mean high frequency but unusual frequency, by

comparison with a reference corpus of some kind. (Scott, 1997, p. 236).

ターゲットコーパス内の頻度が特異であると言うためには、参照コーパスにおける同じ語の頻度から十分に逸脱している必要がある。このことを確認するには、差の有意性を検証する仮説検定の枠組みが援用できる。

(2) In the case of a key word procedure such as that used in WordSmith, this calculation is repeated for every single type in the text we are interested in. For example, the frequency of THE in the text is compared with the frequency of THE in the reference corpus, and the *p* value is then computed of any difference. If the text has 9% of THE and the reference corpus has only 5% of THE, say, we might get a *p* value suggesting that we can believe, with little risk of being wrong, that in our text THE is prominent. This process is repeated with the frequencies of WAS, the frequencies of IS, and so on until all word-forms have been examined. (Scott, 2010, p. 48)

仮説検定では、はじめに、「差が有意ではない」という帰無仮説を立てる。次に、帰無仮説が正しい場合に予想される期待値からのずれ幅を集約した検定統計量を計算する。そして、計算で得られた統計量が事前に決めておいた有意水準（通例 5%）に対応する基準値を上回っているか、あるいは、統計量から計算される *p* 値（帰無仮説が正しい場合に手元のデータ（またはそれより極端なデータ）が偶然で得られる確率）が 5%未満になっているかが確認されると、帰無仮説を棄却し、「差は有意」と結論する。

このとき、検定統計量や *p* 値は、差があるか否かの判定に使うものであるが、これらを拡張的に使用し、特徴度、つまりは、頻度の差の大きさを示す一種の連続尺度と読み替えれば、各種の語を特徴度の大小でランキングしたり、事前に決めた閾値で特徴語をフィルタリングしたりすることが可能となる（ただし、検定統計量をこのような目的で使用することには慎重な立場もありうる）。

もっとも、比較しようとする 2 種のコーパスに含まれる語種の総数は数百、数千に及ぶこともあり、それらの各々について、そのつど、検定統計量や *p* 値を手計算で求めることは現実的ではない。そこで、WordSmith などのコンコーダンサは、こうした計算の自動化プログラムを提供している。

(3) ... (WordSmith Tool has made) it a relatively easy and rapid task for a researcher to calculate the incidences of each and every single word in the target data as well as a comparative dataset, undertake statistical comparisons between indices of the same words in order to establish significant differences, and the resulting keywords ranked according to degrees of significance of difference. (Culpeper and Demmen, 2015, p. 93)

コンコーダンサ上での操作が簡便であっても、研究上のメリットがなければ手法は広がらない。この点に関して、これまでに様々なテキストデータに特徴語分析の手法が適用されてきたが、いずれの場合も、特徴語分析で機械的に検出された特徴語には、研究者がそれまで主観的に重要だと考えていた語が網羅的に包含されていただけでなく、研究者が見落としていた潜在的重要語も含まれていた。これにより、研究者は、自身の質的なテキスト解釈に客観的な裏付けが得られただけでなく、従前は気づいていなかった新しい重要語を発見し、その解釈と検討を通して、自身の分析を拡張することもできるようになった。こうして、実行の利便性と、得られる結果の有用性があいまって、特徴語分析は（少なくとも英語や欧州諸語を対象とした）コーパス研究においては不可欠な分析手法の一つとなったのである。

2.2 論点

検定統計量などを根拠として、ターゲットコーパスの側で特異な頻度を示す語を網羅的に抽出し、特徴度順にランキングするという特徴語分析の発想はシンプルなものであるが、その実践にあたっては、いくつか、検討すべき事柄も存在する。ここでは、関連文献（Scott and Tribble, 2006; Culpeper and Demmen, 2015; Moreno-Ortiz, 2024 ほか）も参考にしつつ、(1)参照コーパスの性質、(2)参照コーパスのサイズ、(3)参照コーパスの呼称、(4)特徴語のタイプ、(5)特徴語の方向性、(6)特徴語の分布、(7)統計的尺度、という 6 つの論点を示す。

2.2.1 参照コーパスの性質

一般に、調査対象とするものをターゲットコーパス、ターゲットコーパスの比較相手にするものを参照コーパスと呼ぶ。このとき、参照コーパスについては、(1)大規模一般コーパスを使うか同規模同種のコーパスを使うか、また、(2)ターゲットデータを参照側に含めるか除くか、という問題がある。

(1)は、たとえば日本の英語教科書の特徴語を調べたい場合、British National Corpus（BNC）や Corpus of Contemporary American English（COCA）など、当該言語全体を代表すると目される大規模一般コーパスと比較すべきか、米国の中学校の国語（英語）教科書など、同種・同等規模のものと比較すべきか否かという議論である。英語の標準を基準として「教科書」の逸脱を検討したいのであれば BNC や COCA などと比較することになるが、「日本」の教科書の特殊性を示したいのであれば他国の英語教科書などと比較するほうが適当であろう。また、英語学習者の作文特性を調べたい場合も、標準的な英語との比較を行いたいのであれば BNC や COCA が使えるかもしれないが、多くの場合、いわゆる英語母語話者による同等の作文との比較を行うほうが有益な知見が得られるであろう。

Culpeper and Demmen（2015）は、Scott らの見解にも言及したうえで、大規模一般コーパスを参照コーパスに用いた場合、参照コーパス側にターゲットコーパスに類似したデータが十分含まれていないため、ターゲットコーパスの特性を反映した特徴語を取り出しにくいとしている。

(2)は、たとえば、各社の英語教科書の中で A 社の教科書の特徴語を調べたい場合、A 社 vs 全社で比較すべきか、A 社 vs「A 社を除く全社」で比較すべきか否かという問題である。全体との比

較という理念を重視すれば前者が、A 社の教科書の特徴性をより明確に取り出したければ後者を選ぶのが妥当だろう。

2.2.2 参照コーパスのサイズ

参照コーパスとして大規模一般コーパスを使う場合、一般的に言えば、大きければ大きいほうが言語母集団に対する代表性は高まる。しかし、均衡的に構築されたコーパスであれば、比較的小規模のものでも分析は可能である。実際、100 万語の FLOB コーパスと 1 億語の BNC をそれぞれ参照コーパスに使用した結果、取り出される特徴語がほとんど変わらなかったという報告もある（Xiao and McEnery, 2005）。

一方、参照コーパスとして同規模同種コーパスを使う場合は、数万語程度であっても分析は可能である。もっとも、「参照」という位置づけを与える以上、少なくともターゲットコーパスと同等のサイズはあったほうがよいだろう。

2.2.3 参照コーパスの呼称

ターゲットコーパスの比較相手とするコーパスの呼称については、「参照コーパス」のほか、「基準コーパス（benchmark corpus）」（Pojanapunya and Todd, 2021）、「比較コーパス（comparative corpus）」（Culpeper and Demmen, 2015）といった語も用いられる。

たとえば、中国と韓国の英語教科書を比較するといった分析目的であれば、両者のコーパスの位置付けは同等で、どちらをターゲットとするか、どちらかを参照とするかは、純粋に処理手順上の問題にすぎない。こうした場合は、「参照コーパス」や「基準コーパス」よりも、「比較コーパス」という呼び方のほうが適切かもしれない。

2.2.4 特徴語のタイプ

たとえば、英字新聞の経済記事の特徴語を調べれば、economics, finance, stock, market など、経済の内容に密接に関連したものだけでなく、各種の機能語など、内容との関連性が薄いものも含まれる。Scott and Tribble (2006)は、前者を当該テキストの内容（aboutness）に関係する特徴語、後者を言語様式（style）に関係する特徴語と呼んで区別している。

前者は Halliday（1994）の言う観念構成的メタ機能（ideational function）に関連するもので、通例、事前に予測が可能である。一方、後者は言語の表層的構成に関わるもので、多くの場合、事前の予測を超えたものとなる。内容特徴語と言語様式特徴語は、分析上、それぞれ、主観的解釈の補強と主観的解釈の拡張という異なる意義を持つ。

2.2.5 特徴語の方向性

特徴語とは、狭義では、参照コーパスと比較して、ターゲットコーパス側で顕著に多く出ている正の特徴語（positive keywords）を指す場合が多いが、広義では、ターゲットコーパス側で顕著に少ない負の特徴語（negative keywords）も含まれる。両者は相補的な関係にあり、ターゲットコー

パス側の正と負の特徴語は、それぞれ、参照コーパス側の負と正の特徴語となる。

2.2.6 特徴語の分布

たとえば、ある小説の特徴語として何らかの語が検出されたとしても、それが当該作品全体の特徴を表す語であることの保証はなく、冒頭部や終結部など、特定のシーンや場面だけに集中して出ている可能性もある。

Scott and Tribble (2006)は、ターゲット側テキストの全体にわたって満遍なく出現するものを全体的（global）な特徴語、特定場面などに限定して頻出するものを局所的（local）な特徴語と呼んで区別している。一般に、言語様式特徴語は全体的特徴語に、内容特徴語は局所的特徴語となりがちである。

特徴語の分布については、1 テキスト内の分布度だけでなく、コーパス内に含まれる複数テキストをまたいだ分布度が問題になることもある。後者に関しては、ターゲットコーパスと参照コーパスの頻度か計算される検定統計量に代えて、両者におけるレンジ数（当該語が出現しているテキスト数の比率）に基づく新しい統計量として、Text Dispersion Keyness (Egbert and Biber, 2019）などを新しい特徴度の尺度とすべきだという提言もなされている。

2.2.7 統計尺度

特徴度尺度の選択については、(1)検定統計量系を使うか効果量系を使うか、(2)検定統計量としてどの尺度を使用するか、(3)検定統計量に多重比較補正を行うか、(4)効果量としてどの尺度を使うか、といった点で方針を決める必要がある。

(1)については、仮説検定において、差があるか否かの判断の根拠となる検定統計量を使うのが一般的だが、統計量には、コーパスサイズが大きくなると値が増大し、差が出やすくなる（有意性判定の厳格性が緩和する）という問題がある。この点を踏まえると、差があるか否かを示す統計量に加え、実際にどのぐらいの差があるかを示す効果量（effect size）系の指標を使うという判断もありうる。

仮説検定では、まずもって検定統計量で有意差の有無を決定し、差がある場合に限って、補助的に効果量を使って実際の差の大小を確認するというプロセスが一般的であり、特徴語についても、同様の指針を取るのが一応の目安となるであろう。

(2)について、古くから広く使用されてきたのがカイ二乗統計量である。2×2 のクロス表の各セルについて、コーパスから得られた実測値と（2 つのコーパス間で頻度に差がないとする帰無仮説が正しい場合の）予測値の差を求める。その後、得られた差から 0.5 を引いて（Yates 補正）、二乗し、予測値で割る。最後に、得られた値を総和したものがカイ二乗統計量である。

ただ、コーパス頻度は特殊で、たとえば、頻度 1 と頻度 100 万を比較するような場合も一般にありうる。この点をふまえれば、コーパスから得た頻度をそのまま扱うのではなく、あらかじめ、自然対数（ln）を取って圧縮してから計算したほうが分布のあてはまりがよくなる。そこで、各セルについて、実測値の ln から予測値の ln を引いて実測値の ln をかけた値を出し、最後にそれらを総和したの

が対数尤度比統計量である。カイ二乗統計量には予測値が 5 未満になる場合は適用すべきでないという制約があるが、対数尤度比の場合はこの制約はないとされる。

もっとも、カイ二乗統計量と対数尤度比統計量は計算式も類似しており、有意性判断のための基準値も同一であることから、尺度として大きな違いはない。Culpeper(2009)は、両方の尺度を使って特徴語を出して両者を比較した結果、特徴度のランキングが部分的に変わる程度で、本質的な差はなかったと報告している。

さて、カイ二乗統計量と対数尤度比統計量は、ともに、頻度に立脚した頻度論統計の尺度であるが、近年、これに対し、ベイズ統計という新しい枠組みを推奨する研究者も増えている。頻度論統計では、帰無仮説と対立仮説は二択の関係にあり、統計値によっていずれかを採択するわけだが、ベイズ統計では両者を同時に考慮し、帰無仮説の尤度と対立仮説の尤度の比率値であるベイズ因子(Bayes factor)の値により、手元のデータが帰無仮説を支持しているか対立仮説を支持しているかを連続的に判断する。ベイズ因子の値はいずれかの仮説を支持する「証拠」の強さの指標とみなされる。特徴語分析においても、カイ二乗統計量や対数尤度比統計量に代え、ベイズ因子を近似するベイジアン情報量規準(BIC)を使用すべきだという提案も出されている(Wilson, 2013)。BIC は、赤池情報量基準(AIC)等と同様、モデルに対するあてはまりとモデルの複雑さの度合いを示す指標で、BIC の値が小さいと、帰無仮説の側が支持されることになり、頻度の差、つまり、特徴度は小さいと判断される。逆に、BIC の値が大きいと対立仮説側が支持されることになり頻度の差、つまり特徴度は大きいと判断される。特徴語分析にベイズ統計の考え方を組み込むことは有益だと思われるが、現時点において、BIC を特徴語分析尺度として実装したコンコーダンサはほとんどなく、その使用は制約的である。

(3)について、特徴語分析ではテキストに含まれる数百、数千の語を取り上げて網羅的に頻度を比較し、差の有意性を調べるわけだが、このとき、見方を変えれば、同じ検定を数百回ないし数千回繰り返していることになる。有意水準を 5%に決めている場合、1 回だけ検定を行うなら問題はないが、それを何千回も繰り返すなら、分析全体で見れば、非常に大きな誤りの確率を認めているようにも見える。この点を問題視する研究者は、たとえば、2 種のテキストに含まれる語種の総数が 100 語であるなら(つまり検定を 100 回繰り返すのであれば)、あらかじめ有意水準を 5/100=0.05%に引き下げておく(Bonferroni 法)など、何らかの多重比較補正の適用を推奨している。

最後に、(4)について、効果量として最もシンプルなものは、ターゲットコーパスと参照コーパスにおける語の出現率(相対頻度)の比を取ったリスク比(risk ratio)である。リスク比は、ターゲットコーパスにおけるある語の出方が、参照コーパスの場合と比べて何倍になっているかを示すもので、たとえば、100 語のコーパスが 2 つあり、ある語の頻度がターゲット側で 20 回、参照側で 10 回だったとすると、リスク比は 0.2÷0.1=2(2 倍の意)となる。このほか、リスク比を対数(底は 2)に変換することで値を圧縮して扱いやすくした LogDice(Hardie, n.d.)などもある。また、ある語の頻度を全体頻度と比較して相対頻度化するのではなく、全体頻度から当該語頻度を引いた値と比較して相対化するオッズ比などもある。

2.3 最近の文献に見る特徴語分析

Culpeper and Demmen（2015）は、特徴語分析が 1990 年代の終わりにコンコーダンサの機能として普及し始めた後、「2000 年代半ばごろまでにはほぼ成熟期に入った」としている。2010 年代以降、主として統計処理の面で、特徴度の尺度としていくつか新しい提案がなされているものの、コーパス研究における特徴語分析の基本的な方法論や方向性は現時点においてもほとんど変わっていない。

以下は最近刊行された研究文献に見る特徴語分析に関する記述の一例である。

(4) We can ... compare the frequencies of words in the *target* corpus (i.e. the corpus which is being examined) with those in another corpus that represents some kind of language standard (i.e. a reference corpus). By determining which words appear unusually frequently in the target corpus compared to the reference corpus using a statistical measure such as log likelihood, we can identify so-called *keywords*... (Anthony, 2022, p. 109).

(5) Keyness statistics are generated by comparing a specialised corpus with a general corpus. Words that appear significantly more frequently in the specialised corpus are considered key to that corpus. (Rees, 2022, p. 399)

(6) [Keyword analysis] allows us to look at the frequency of items in one corpus in comparison to another, normally a larger, reference corpus, and check which words occur significantly more (positive keywords) or significantly less (negative keywords) in our corpus compared to a reference corpus. The corpora must be comparable to make keyness searches worthwhile, and it is important to be clear about what we are comparing and why. For example, if we compare the USTC [i.e. UCLan Speaking Test Corpus] data with another corpus of spoken learner data, we need to ensure it contains data similar to interactive spoken exams and not monologues. (Jones, 2022, pp. 130-131)

(7) The comparison between keywords in the target corpus and the keywords in the reference corpus will determine the final keyword candidate list. For Baker (2004), keywords direct 'researchers to elements in texts that are unusually frequent (or infrequent), helping to remove researcher bias and paving the way for more complex analyses of linguistic phenomena' (p. 348). Keywords can thus be used to discover the lexical items that are characteristic of a target corpus when compared

to a reference corpus (Pérez-Paredes, 2020), identifying language use that is salient 'even though humans may not have noted them as especially frequent, so they can provide a relatively objective way of helping to direct attention to parts of the corpus that we wouldn't have thought to consider' (Baker et al. 2021, p. 56)…. (Pérez-Paredes, 2024, p. 48)

どの説明においても、特徴語分析の大枠の理解は変わらないが、いくつかの点で違いが見られる。まず、参照コーパスの在り方に関しては、(4)が言語における基準性(language standard)に、(5)が一般性(general)に、(6)が大規模性と比較可能性(comparable)に言及している。参照コーパスとして、大規模一般コーパス以外に同規模同種コーパスを使うこともある点を考えると、特徴語分析の価値が両コーパスの内容的な類似性(similar)に左右されるという(6)の指摘は重要である。

次に、特徴語の頻度については、(4)が異常に高頻度(unusually frequently)、(5) が有意に高頻度(significantly more frequently)、(7)が特に高頻度(especially frequent)としている。これらはいずれも正の特徴語に限った記述だが、(6)は有意に高頻度または低頻度(significantly more or significantly less)とすることで、負の特徴語にも言及している。

2.4 適用例

書かれたものであれ、話されたものであれ、あらゆる種類のテキストを対象とする特徴語分析は、狭義の言語学はもちろん、言語教育、文学、政治学、社会学、医学など、広範な分野に応用可能である。

一例を示せば、言語研究であれば、地域・時代・産出モード・ジャンルを比較し、各々の特徴語を取り出すことで、実際の言語使用の特性を多面的に明らかにすることができる。

言語教育であれば、教材の言語と実際の言語、初級・中級・上級学習者による第 2 言語産出と母語話者による第 1 言語産出、あるいは多様な学問領域の論文の特徴語を取り出すことで、教材の課題と改善のヒントを明らかにしたり、第 2 言語習得の過程をモデル化したり、専門目的英語(English for Specific Purposes)の指導に活用したりすることもできる。

文学研究であれば、たとえば、シェイクスピア作品を同時代の他の劇作品や前後の時代の劇作品と比較して特徴語を取り出すことで、シェイクスピアの特徴を多元的に分析することもできる。

政治研究であれば、各政党の選挙用マニフェストや候補者の演説を比較して特徴語を調べることで、各党の政治的立場(より厳密に言えば、各党が有権者に信じ込ませようとしている政治的イメージ)を分析することができる。

社会学であれば、マイノリティに関する報道をメディア別・時代別に比較して特徴語を取り出すことで、共同体におけるマイノリティの表象やフレーミング、また、マイノリティに対する人々の意識の変化を調査することもできる(一例として、Baker et al. 2013)。

また、医学であれば、たとえば自閉症スペクトラム(ASD)児と非 ASD 児の発話を比較することで、ASD 児の言語特性を抽出し、診断に役立てることもできる。日本語の場合、ASD 児の発話には共

感を示す終助詞が過少出現することが知られており、特徴語分析はこうした傾向の抽出にも適用可能であろう(ASD 者の発話コーパス開発に関しては、鈴木他 2024 を参照)。

以下、近刊の論文集からいくつか特徴語分析の実践例を見ておこう。まず、Lutzky(2021)は X(旧 Twitter)における飛行機利用者のツイートと、それに対する航空会社の返信ツイートを集め、相互を比較した結果、顧客側には no(no response/info/explanation など)、why、just などが、航空会社側には hi, sorry, hear, hope などが多いことを示した。

また、Williams (2021)は、米国における English Language Unity Act(ELUA)をめぐる賛成派と反対派の議員の発話をコーパス化して両者を比較し、賛成派の弁論には I/my/me, common, it, together などが、反対派の弁論には students, only, programs, adult, literacy, education などが多いことを示した。

さらに、Wilkinson(2021)は、性的志向によって母国で迫害を受けている亡命希望者(LGBT refugees and asylum seeker:LGBT RAS)を報じた英国の 2009 年～2013 年の新聞記事と、2014 年～2018 年の新聞記事コーパスを比較し、2009 年以降は gay や homosexual などの語が、2014 年以降はは LGBT, gender などの語がそれぞれの時代の特徴語になっていることを報告している。

Williams や Wilkinson の研究はいわゆる批判的談話分析(critical discourse analysis:CDA)の範疇に入るものであるが、統計的特徴語を議論の根拠にすることで、CDA にありがちな主張の恣意性や証拠の希薄さという制約をうまく解決している。

3. EJTKAN の概要と操作手順

前節で概観したように、特徴語分析は非常に有益なテキスト研究手法であるが、日本語については、単語頻度の測定の前提となる形態素解析の手順が必須となるため、英語をはじめとする諸言語に比べると、特徴語分析の使用は少ない。

EJTKAN の開発にあたっては、こうした日本語特有の制約を解消するため、言語種別判定→形態素解析→単語抽出→単語頻度調査→頻度比較→特徴量算出→特徴語リスト出力に至る処理過程を自動化したうえで、できるだけシンプルな操作系となることを目指した。

図 1

EJTKAN の画面

以下 4 節のケーススタディの準備として、「多言語母語の日本語学習者横断コーパス」(I-JAS)(迫田, 2020)のダウンロード版に含まれる CCH(中国語母語学習者)50 名による SW1 タスク(ピ

クニックをテーマにした 5 コマの絵を見てその内容を作文する)と、JJJ(日本語母語話者)50 名による同一タスク産出を比較し、双方の特徴語を抽出したい。

EJTKAN を用いた特徴語分析において、ユーザー側で必要となるのは、後述する Step 1 と Step 2 の 2 段階の処理のみである(別途解析済みテキストデータを入手する場合のみ Step 3 が加わる)。

3.1 Step 1 ツールのドライブへのコピー

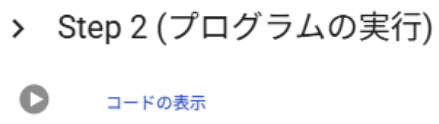
EJTKAN は Google Colab 上のツールであり、実際の処理はユーザー個々の Google ドライブ空間で行われる。そのため、ユーザーは自身の Google アカウントにログインしてツールにアクセスした後、Step 1 として、図 1 中央上部にある「ドライブにコピー」を押して、本ツールを各自のドライブ内に仮想的にコピーする必要がある。

3.2 Step 2 ファイルのアップロードと処理

次に、Step 2「プログラムの実行」の箇所で、行頭の▸ボタンを押す。

図 2

プログラムの実行



これにより、言語判別ライブラリ(テキストが英語か日本語かを自動識別する)の読み込みが始まり、完了すると、下記のようにターゲットコーパス用テキストをアップロードするための「ファイル選択」ボタンが表示される。

図 3

ターゲットコーパス用テキストのアップロード画面

上記の「ファイル選択」を押し、各自の PC 内にあるターゲットコーパス用テキストファイル(1 個でも複数個でもよい。数百個程度のファイルであっても比較的短時間で処理可能)を指定する。ここでは、I-JAS の CCH による SW1 産出 50 ファイルを指定する。

アップロードが終わると、言語種別の判別が自動で行われ、続いて参照コーパス用テキストをアップロードするための「ファイル選択」ボタンが表示される。

図 4

言語判別結果表示と参照データアップロード画面

上記の「ファイル選択」を押して、参照コーパス用テキストファイル(上記同様、1 個でも複数個でもよい)を指定してアップロードする。ここでは、I-JAS の JJJ による SW1 産出 50 ファイルを指定する。

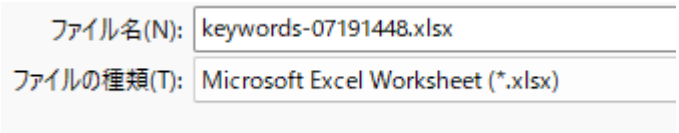
なお、ターゲットコーパス側、参照コーパス側ともに、アップロードするすべてのテキストファイルは UTF-8 形式で保存されている必要がある。UTF-8 以外のものが含まれている場合はエラーが出るので、UTF-8 に修正したうえで再度実行する必要がある。

正しくアップロードできると、システム上で、日本語処理に必要な MeCab (mecab-python3)と UniDic (1.1.0)が自動で導入され、テキスト解析、ついで、特徴量の計算がなされる。なお、特徴「語」を抽出するという目的に沿い、「補助記号」「空白」「数詞」と一部の記号は分析対象から除外している点に注意されたい。

一連の処理が終わると、以下の画面がポップアップで表示されるので、各自のデスクトップに保存する。ファイル名はデフォルトでは処理した日・時間となっているが自由に変更可能である。

図 5

特徴語分析結果シートのダウンロード

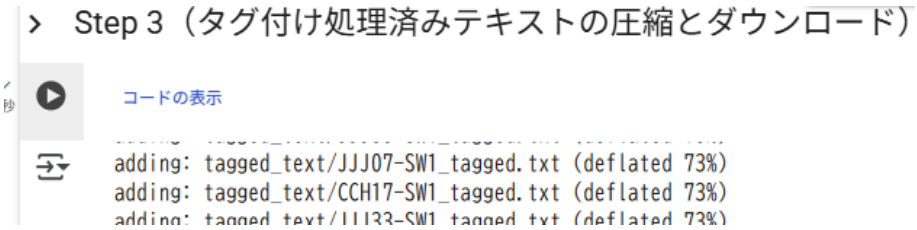


3.3 Step 3 解析済みテキストファイルのダウンロード

以上で特徴語分析の手順は終わりだが、この後の Step 3(任意)において、形態素解析済のテキストファイルをダウンロードすることもできる。たとえば、抽出された特徴語が実際のテキストで具体的にどのように使われていたかを確認したい場合は、解析済みテキストがあったほうが便利であろう。

図 6

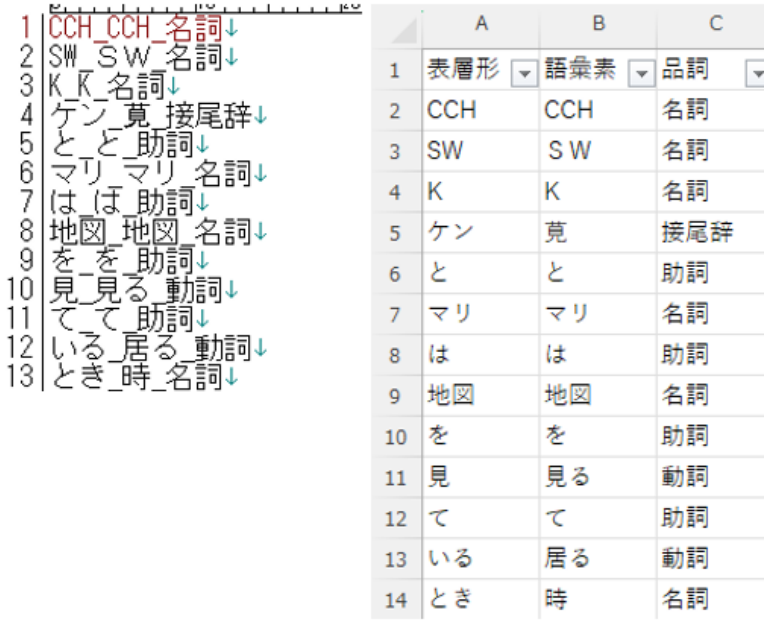
形態素解析済みデータのダウンロード



ダウンロードした形態素解析済みファイルでは、「見_見る_動詞」のように、表層形・形態素・品詞がアンダーバーでつながった状態で記録されている。このままでは使いにくいが、アンダーバーをタブに置換してエクセルに貼り付け、フィルタなどを設定すれば、特徴語がどのような前後文脈で使用されているか確認することができる。

図 7-8

形態素解析済みデータ(左:ダウンロードファイル、右:Excel に転記した状態)



1 CCH_CCH_名詞
2 SW_ＳＷ_名詞
3 K_K_名詞
4 ケン_莧_接尾辞
5 と_と_助詞
6 マリ_マリ_名詞
7 は_は_助詞
8 地図_地図_名詞
9 を_を_助詞
10 見_見る_動詞
11 て_て_助詞
12 いる_居る_動詞
13 とき_時_名詞

	A	B	C
1	表層形	語彙素	品詞
2	CCH	CCH	名詞
3	SW	ＳＷ	名詞
4	K	K	名詞
5	ケン	莧	接尾辞
6	と	と	助詞
7	マリ	マリ	名詞
8	は	は	助詞
9	地図	地図	名詞
10	を	を	助詞
11	見	見る	動詞
12	て	て	助詞
13	いる	居る	動詞
14	とき	時	名詞

図 8 は、1 行目に見出し行を追加し、フィルタをつけている。これにより、たとえば、「X で始まる語」や「Y で終わる語」を抽出したり、品詞を指定して「助詞」だけを取り出したりすることもできる。

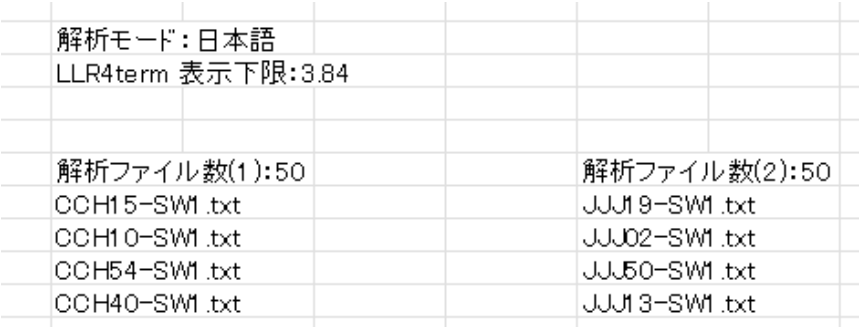
3.4 EJTKAN の出力

3.4.1 検定統計量と効果量

出力される Excel ファイルは 6 枚のシートで構成されており、1 枚目には解析条件が記録されている。デフォルトでは対数尤度比統計量(LLR)(4 セル計算)の 3.84 以上の語が特徴語と認定される。LLR の 3.84 は、多重比較検定を行わない場合の有意水準 5%に相当する値であり、つまりは、参照コーパス頻度との間に有意な差があることを意味する。

図 8

解析条件表示シート



解析モード:日本語	
LLR4term 表示下限:3.84	
解析ファイル数(1):50	解析ファイル数(2):50
CCH15-SW1.txt	JJJ19-SW1.txt
CCH10-SW1.txt	JJJ02-SW1.txt
CCH54-SW1.txt	JJJ50-SW1.txt
CCH40-SW1.txt	JJJ13-SW1.txt

EJTKAN は、姉妹ツールである EJWFTG の開発コンセプトを踏襲しているため、日本語処理の場合、特徴語についても、2 枚目以降のシートに、表層形、表層形＋品詞、語彙素、語彙素＋品詞の 4 モードでの特徴語が出力される。また、6 枚目のシートには、新しい機能として特徴品詞が出力される。

下記は、表層形の場合の特徴語リストの一部である。

図 9

特徴語出力結果画面の例

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
	LLR(4-term)の順位	Word	基本データ(T: Target, R: Reference)									検定統計量系				効果量系				
			T Freq:T内当該語頻度	T Size:T総語数	T NormFreq:T内相対頻度(%)	T ExpFreq:T予測頻度	R Freq:R内当該語頻度	R Size:R総語数	R NormFreq:R内相対頻度(%)	R ExpFreq:R予測頻度	T NormFreq > R NormFreq	χ2:カイ二乗統計量(Yates補正あり)	LLR (4-term):対数尤度比統計量(4セル)	LLR (2-term):対数尤度比統計量(2セル)	BF:ベイズ因子 Cf. Wilson (2013)	Odds Ratio:オッズ比	Risk Ratio:リスク比	Freq Diff:相対頻度差 (%) Cf. Gabrielatos and Marchi (2012)	LogRatio:対数比相対頻度比 Cf. Hardie (ind.)	Effect Size for LLR (ELL):対数尤度比効果量 Cf. Johnston et al (2006)
1	1	CCH	332	6121	5.42	170	0	5853	0.00	162	+	324.51	454.60	445.55	436.16	0.00	0.00	0.00	9.31	0.0073
2	2	ら	66	6121	1.08	34	0	5853	0.00	32	+	61.51	88.92	88.57	79.18	0.00	0.00	0.00	6.98	0.0021
3	3	彼	64	6121	1.05	33	0	5853	0.00	31	+	59.58	86.22	85.89	76.50	0.00	0.00	0.00	6.94	0.0020
4	4	さん	35	6121	0.57	18	0	5853	0.00	17	+	31.63	47.07	46.97	37.58	0.00	0.00	0.00	6.06	0.0013
5	5	行き	33	6121	0.54	18	3	5853	0.05	18	+	22.16	28.00	27.93	18.54	10.57	10.52	951.84	3.39	0.0008
6	6	時	25	6121	0.41	13	1	5853	0.02	13	+	19.38	26.55	26.50	17.11	24.00	23.91	2,290.54	4.58	0.0008
7	7	は	285	6121	4.66	233	171	5853	2.92	223	+	24.10	24.86	23.93	14.54	1.62	1.59	59.37	0.67	0.0003
8	8	食べ物	35	6121	0.57	21	6	5853	0.10	20	+	17.96	21.49	21.42	12.03	5.60	5.58	457.79	2.48	0.0006
9	9	なかっ	15	6121	0.25	8	0	5853	0.00	7	+	12.47	20.15	20.13	10.74	0.00	0.00	0.00	4.84	0.0008
10	10	とき	23	6121	0.38	13	2	5853	0.03	12	+	15.16	19.83	19.79	10.40	11.03	11.00	999.65	3.46	0.0006
11	11	つもり	11	6121	0.18	6	0	5853	0.00	5	+	8.66	14.77	14.76	5.37	0.00	0.00	0.00	4.39	0.0007
12	12	家	15	6121	0.25	8	1	5853	0.02	8	+	10.01	14.10	14.08	4.69	14.38	14.34	1,334.32	3.84	0.0005
13	13	いぬ	10	6121	0.16	5	0	5853	0.00	5	+	7.71	13.43	13.42	4.03	0.00	0.00	0.00	4.26	0.0007
14	14	ある	14	6121	0.23	8	1	5853	0.02	7	+	9.09	12.89	12.87	3.48	13.42	13.39	1,238.70	3.74	0.0005
15	15	聞い	9	6121	0.15	5	0	5853	0.00	4	+	6.77	12.08	12.08	2.69	0.00	0.00	0.00	4.11	0.0006

シート内で、情報は、「見出し」、「基本データ」、「検定統計量系の尺度値」、「効果量系の尺度値」の 4 グループで整理されている。

まず、「見出し」部については、対数尤度比統計量（4 セル）の降順に基づくランクと、実際の語が表示される。

次に、「基本データ」部では、特徴度計算に使うデータとして、ターゲットコーパス（T）と参照コーパス（R）の各々における単語頻度、コーパス総語数、相対頻度、（コーパス間に差がないと仮定した場合の）予測頻度（予測値）、また、ターゲットコーパス内相対頻度から参照コーパス内相対頻度を引いた値の正負の別が記載されている。

「検定統計量系」の尺度値の部では、左から順に、カイ二乗統計量、対数尤度比統計量（4 セル、対数尤度比統計量（2 セル）、ベイズ因子（BIC）の値が表示される。

右端の「効果量系」の尺度値の部では、左から順に、オッズ比、リスク比、相対頻度差、LogRatio、対数尤度比効果量の値が表示される。

これらの尺度の選定と計算方法については、英国ランカスター大学の Paul Rayson 氏らが開発した Log-likelihood and effect size calculator（UCREL NLP Group. n.d.）に準拠している。以下は表示される項目のリストである。 各々の尺度の計算式については本稿末尾の付表を参照されたい。

表 1

特徴語リストに表示される項目の一覧

統計量算出のための基礎データ
T Freq：T 内当該語頻度 T Size：T 総語数 T NormFreq：T 内相対頻度(%) T ExpFreq：T 予測頻度 R Freq：R 内当該語頻度 R Size：R 総語数 R NormFreq：R 内相対頻度(%) R ExpFreq：R 予測頻度 T NormFreq>R NormFreq: T/R 相対頻度差の正負
検定統計量系特徴度
χ2：カイ二乗統計量（Yates 補正あり） LLR (4-term)：対数尤度比統計量（4 セル） LLR (2-term)：対数尤度比統計量（2 セル） BF：ベイズ因子 Cf. Wilson (2013)
効果量系特徴度
Odds Ratio：オッズ比 Risk Ratio：リスク比 Freq Diff：相対頻度差(%) Cf. Gabrielatos and Marchi (2012) LogRatio：対数化相対頻度比 Cf. Hardie (n.d.) Effect Size for LLR (ELL)：対数尤度比効果量 Cf. Johnston et al. (2006)

検定統計量のうち、対数尤度比統計量については、2×2 の分割表（通例、ケースがターゲットおよび参照コーパスで、変数が X 頻度と X 以外頻度となる）において、予測値と実測値の差を 4 セルの各々から算出して総和する 4-term の計算式と、X 頻度に関わる 2 セルから計算する 2-term の計算式に基づく値を出している。近年の研究では 4-term を使うのが一般的であるが、過去の研究では 2-term を用いたものもあるため、両方を示している（AntConc でも対数尤度比統計量については両者が実装されている）。

効果量に関しては、5 種を表示しているが、まずは、言語研究に限らず幅広い分野で使用されているリスク比に注目するとよいだろう。リスク比とは、相対頻度の比のことで、コーパスサイズの差を調整した上で、ターゲットコーパス内において、当該語が、参照コーパスに比べて「何倍多く出ているか」を示す値であるため、わかりやすい。Log Ratio はリスク比を底が 2 の対数に変換した値で、指標の提唱者である Andrew Hardie 氏が自身の開発したコーパス検索プラットフォーム CQPweb に実装したこともあり、欧米のコーパス研究で普及しつつある。LogRatio の値が 0, 1, 2,

3, 4, 5 であるとすると、ターゲットコーパス内での相対頻度が、参照コーパスに比べて 1 倍、2 倍、4 倍、8 倍、16 倍、32 倍多いことを示す（Hardie, n.d.）。オッズ比は、頻度調整の際に、語の頻度をコーパス総語数ではなく、（総語数－当該語数）で割るものであるが、基本の発想はリスク比とほぼ同等である。これらに比べると、相対頻度差や対数尤度比効果量は現時点においてあまり広く使用されているとは言えない。

3.4.2 負の特徴語

前述のように、EJTKAN では、対数尤度比統計量（4 セル）の値の降順でランキングを行っているわけだが、対数尤度比統計量は、頻度表全体における予測値と実測値のずれ幅の合算値であり、それ自身に正負はない。つまり、純粋に統計量の降順で並べ替えを行うと、ターゲットコーパス側で多く出現している正の特徴語と、少なく出現している負の特徴語が入り混じって表示されてしまう。

そこで、EJTKAN では、「基本データ」部の最後にある「ターゲットコーパス内相対頻度から参照コーパス内相対頻度を引いた値」（TNormFreq＞RNormFreq）の正負の情報を組みあわせて、上部には正の特徴語が特徴度順に、下部には負の特徴語が同じく特徴度順に並ぶよう出力を調整している。

出力画面を下にスクロールしていけば、途中からセルに黒い背景色がつく。これは、当該後が負の特徴語になっていることを示す。

図 10

正の特徴語と負の特徴語の表示

		T NormFreq＞R NormFreq	χ2:カイ二乗統計量（Yates補正あり）	LLR (4-term): 対数尤度比統計量（4セル）	LLR (2-term): 対数尤度比統計量（2セル）
66	飛び込み	+	1.25	4.03	4.03
67	バスケ	+	1.25	4.03	4.03
68	かた	+	1.25	4.03	4.03
69	悲しい	+	1.25	4.03	4.03
1	JJJ	−	309.81	424.17	416.59
2	食べよう	−	25.24	37.28	37.22
3	出かけ	−	25.24	37.28	37.22
4	と	−	32.17	33.23	32.37
5	き	−	22.84	25.82	25.70
6	て	−	20.30	20.75	19.80

注：実際のファイルを一部加工して表示

上記の場合、正の特徴語の最終となる 69 位に「悲しい」（LLR (4-term)=4.03。基準値の 3.84 を満たす最小値）が来た後、行に黒い背景色がつき、負の特徴語（CCH が過少使用＝JJJ が過剰使用）が表示される。1 位は日本語母語話者の話者コードを示す「JJJ」で、以下、「食べよう」「出

かけ」などが続く。

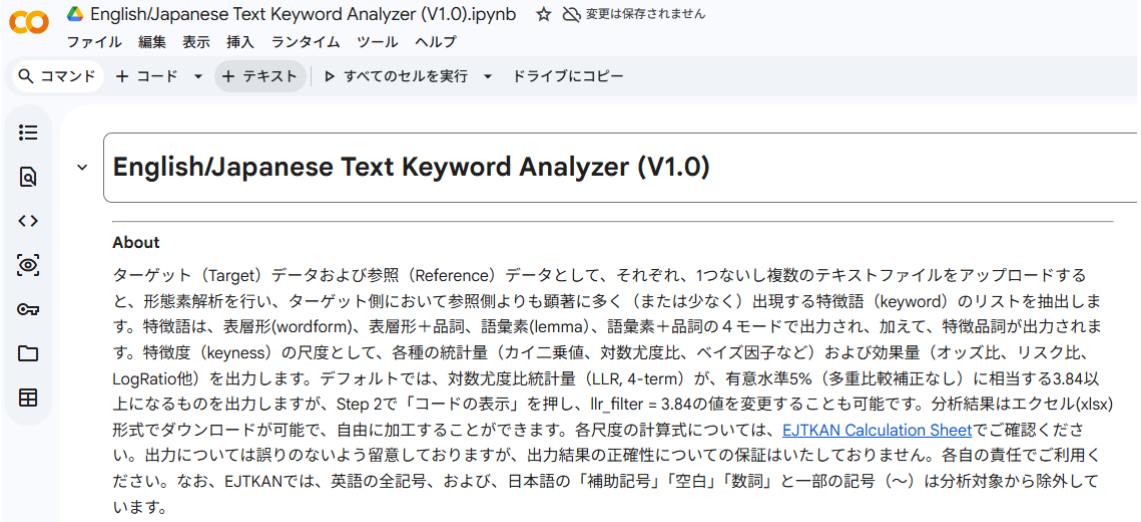
3.4.3 EJTKAN Calculation Sheet の活用

WordSmith や AntConc などのコンコーダンサや、前述の Log-likelihood and effect size calculator などを含め、特徴語を自動抽出してくれるツールは他にも多いが、いずれの場合も、ツール自体では各尺度の計算式の詳細がわからず、処理がブラックボックスになりがちである。

そこで、EJTKAN では、本ツールで採用しているすべての尺度値の計算過程をユーザーが実際に確認できるよう、計算式のすべてを再現した検証用の EJTKAN Calculation Sheet を作成し、EJTKAN の画面上部の About（概要）からリンクでダウンロードできるようにしている。

図 11

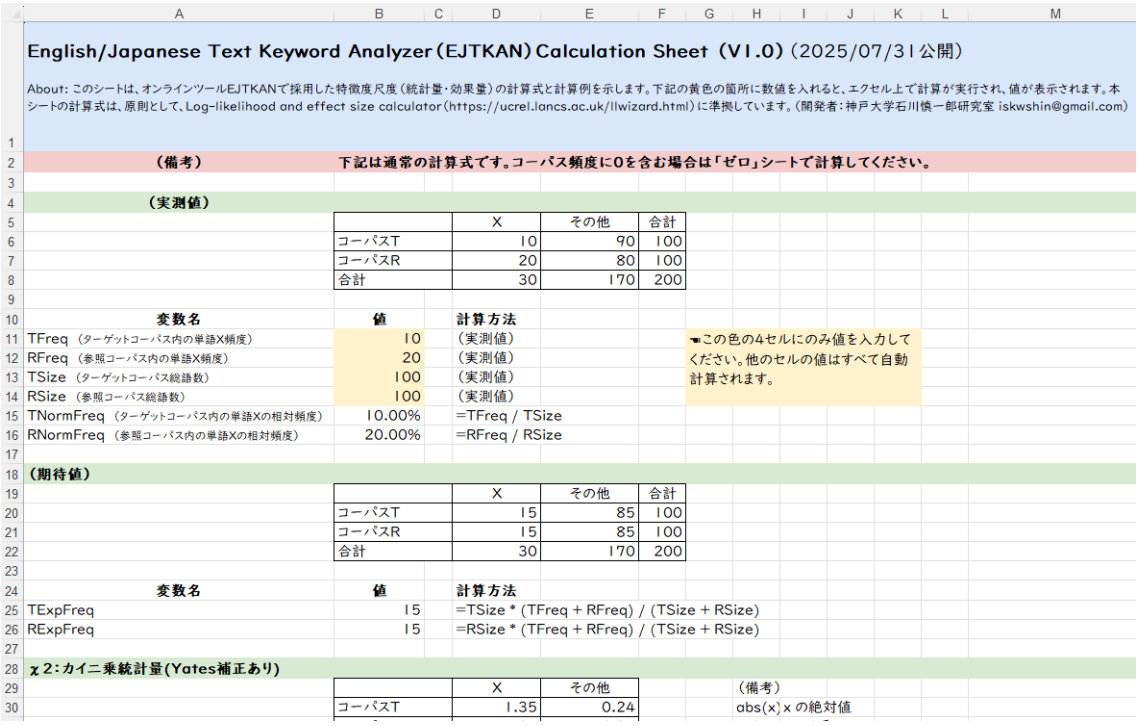
EJTKAN Calculation Sheet の入手方法（About 内のリンクをクリックする）



上記をクリックして検証用ファイルをダウンロードして開くと、以下のような画面が出てくる。

図 12

EJTKAN Calculation Sheet (V1.0)



English/Japanese Text Keyword Analyzer（EJTKAN）Calculation Sheet （V1.0）（2025/07/31公開）

About: このシートは、オンラインツールEJTKANで採用した特徴度尺度（統計量・効果量）の計算式と計算例を示します。下記の黄色の箇所に数値を入れると、エクセル上で計算が実行され、値が表示されます。本シートの計算式は、原則として、Log-likelihood and effect size calculator（https://ucrel.lancs.ac.uk/llwizard.html）に準拠しています。（開発者：神戸大学石川慎一郎研究室 iskwshin@gmail.com）

（備考）　下記は通常の計算式です。コーパス頻度に0を含む場合は「ゼロ」シートで計算してください。

（実測値）

	X	その他	合計
コーパスT	10	90	100
コーパスR	20	80	100
合計	30	170	200

変数名	値	計算方法
TFreq （ターゲットコーパス内の単語X頻度）	10	（実測値）
RFreq （参照コーパス内の単語X頻度）	20	（実測値）
TSize （ターゲットコーパス総語数）	100	（実測値）
RSize （参照コーパス総語数）	100	（実測値）
TNormFreq （ターゲットコーパス内の単語Xの相対頻度）	10.00%	=TFreq / TSize
RNormFreq （参照コーパス内の単語Xの相対頻度）	20.00%	=RFreq / RSize

☜この色の4セルにのみ値を入力してください。他のセルの値はすべて自動計算されます。

（期待値）

	X	その他	合計
コーパスT	15	85	100
コーパスR	15	85	100
合計	30	170	200

変数名	値	計算方法
TExpFreq	15	=TSize * (TFreq + RFreq) / (TSize + RSize)
RExpFreq	15	=RSize * (TFreq + RFreq) / (TSize + RSize)

χ2：カイ二乗統計量（Yates補正あり）

	X	その他	（備考）
コーパスT	1.35	0.24	abs(x) x の絶対値

使用法はシンプルで、黄色の背景色をつけている 4 つのセル（対象コーパスにおける任意の語の頻度、対象コーパス総語数、参照コーパスにおける任意の語の頻度、参照コーパスの総語数の 4 つ）に値を入れるだけで、すべての計算が行われる。

セルには計算式を残しているので、ユーザーはどのようにして個々の尺度が計算されているか確認できるだろう。なお、いずれかの頻度値が 0 になる場合は、一部の尺度において値が計算できなくなる。この問題を解決するため、EJTKAN では、計算上、特殊な補正を加えている。これに対応し、Calculation Sheet においても、1 枚目の通常版に加え、2 枚目に頻度=0 対応版を用意している。

4. ケーススタディ：中国語母語の日本語学習者の正負の特徴語

前述のように、ターゲットコーパスを I-JAS の中国語母語話者（CCH）の SW1 作文、参照コーパスを日本語母語話者（JJJ）の SW1 作文として、対数尤度比統計量（LLR, 4-term）を基準として CCH の 4 モードの特徴語と特徴品詞を自動抽出したところ、以下のような結果を得た。

品詞はすべてを記載し、品詞以外は上位 20 種のみを特徴度降順で記載している。なお、分析したテキストでは、行ごとに話者コード（CCH_xxx、JJJ_yyy）が記載されているため、CCH と JJJ がそれぞれの特徴語の 1 位になっている。

表 2

特徴語と特徴品詞

	CCH の正の特徴語（過剰使用）	CCH の負の特徴語（過少使用）
表層形	CCH、ら、彼、さん、行き、時、は、食べ物、なかっ、とき、つもり、家、いぬ、ある、開い、入り、たら、も、全然、急	JJJ、食べよう、出かけ、と、き、て、間、確認、隙、ず、子犬、愛犬、昼、いる、が、中、リンゴ、中身、飼い犬、様子
表層形＋品詞	CCH-名詞、ら-接尾辞、彼-代名詞、さん-接尾辞、行き-動詞、時-名詞、は-助詞、食べ物-名詞、とき-名詞、家-名詞、なかっ-助動詞、つもり-名詞、いぬ-名詞、開い-動詞、入り-動詞、たら-助動詞、も-助詞、箱-名詞、あと-名詞、急-名詞	JJJ-名詞、出かけ-動詞、食べよう-動詞、と-助詞、き-動詞、て-助詞、間-名詞、確認-名詞、隙-名詞、ず-助動詞、愛犬-名詞、子犬-名詞、昼-名詞、いる-動詞、が-助詞、中-名詞、リンゴ-名詞、様子-名詞、中身-名詞、行こう-動詞
語彙素	CCH、彼、等、さん、時、は、食べ物、積り、家、全然、後、も、終わる、急、此の、箱、分かる、開く、飛ぶ、無い	JJJ、出掛ける、と、来る、て、確認、居る、間、子犬、愛犬、入り込む、昼、が、暇、仕舞う、中、中身、飼い犬、様子、忍び込む
語彙素＋品詞	CCH-名詞、彼-代名詞、等-接尾辞、さん-接尾辞、時-名詞、は-助詞、食べ物-名詞、家-名詞、積り-名詞、全然-副詞、も-助詞、後-名詞、箱-名詞、終わる-動詞、急-名詞、此の-連体詞、開く-動詞、分かる-動詞、飛ぶ-動詞、悲しい-形容詞	JJJ-名詞、出掛ける-動詞、と-助詞、来る-動詞、て-助詞、確認-名詞、居る-動詞、間-名詞、愛犬-名詞、子犬-名詞、入り込む-動詞、昼-名詞、が-助詞、暇-名詞、仕舞う-動詞、忍び込む-動詞、様子-名詞、中身-名詞、飼い犬-名詞、中-名詞
品詞	接尾辞、代名詞、形容詞、助動詞、連体詞、形状詞、接続詞、名詞、副詞	動詞、助詞、接頭辞、感動詞

CCH と JJJ の使用語彙の差を詳しく論じることは本稿の趣旨を超えるが、上表を見るだけで、(1)CCH がトキ節またはタラ節（～する時・とき、～したら）を好むのに対し、JJJ がト接続（～すると）やテ接続（～して）を好むこと、(2)CCH が話者の意図を「つもり」（～するつもり）で言語化するのに対し、JJJ は動詞意思形を使うこと（「食べよう」「行こう」）、(3)CCH の使用する名詞が具象的であるのに対し、JJJ は抽象的な空間時間を指す語を多用すること（「隙」「間」「昼」「中（身）」）、(4)犬がバスケットの中身を食べていたという事実を伝える際、CCH は単純事実として表出するのに対し、JJJ はテシマウ形で被害性を表出すること（～てしまう）、(5)CCH が人物（「ケンさん」「マリさん」「彼ら」）に繰り返し言及するのに対し、JJJ はその行動内容を中心的に描写すること、(6)CCH が形態的に単純な語を多用するのに対し、JJJ は複合的な名詞や複合動詞を多用すること（子犬、愛犬、

飼い犬、出掛ける、入り込む、忍び込む)、(7)CCH が接尾辞や代名詞など人称を中心とした記述をするのに対し、JJJ は動詞中心の記述であること、などが読み取れる。

今回のように、双方が同じイラストを描写するような比較では、内容が同等であることから、抽出される特徴語はいわゆる内容特徴語ではなく、分析者にとって事前に予測のつきにくい言語様式特徴語となる。この意味において、特徴語分析は、学習者と母語話者の日本語使用の様式の差に関して、分析者がおそらくは気づいていなかった新しい知見を提供してくれるものとなりうる。

5. まとめ

以上、本稿では、コーパス研究における特徴語分析について概観した後、筆者の研究室で開発された特徴語自動抽出ツール EJTKAN の概要を紹介した。

特徴語分析は、テキストの内容的特性や言語的特性を客観的に議論することを可能にするもので、分析者の主観的テキスト解釈を補完するだけでなく、新たな質的解釈を誘発する発見法としても価値が高い。

日本語研究における特徴語分析の使用は従来限定的なレベルにとどまっていたが、テキストデータさえあればテキストの事前処理や特徴度の計算を自動で実行する EJTKAN のようなツールがあれば、言語学、言語教育学、談話分析など、何らかの形で日本語テキストを資料として使用する幅広い研究分野において、特徴語分析をうまく取り入れ、テキスト分析の精度を高めていくことが可能になるだろう。

特徴語分析の普及の端緒となった WordSmith Tools の v1(1996)から 30 年近くを経てリリースされた EJTKAN によって、英語はもちろん、日本語の研究においても、様々なテキストを対象とした特徴語分析の実践がさらに広がっていくことを期待したい。

注

本論文は、2025 年 8 月 27-28 日に実施された「言語資源ワークショップ 2025」における筆者の口頭発表(o05)の予稿「テキストデータから特徴語を自動抽出する オンラインツール『EJTKAN』の開発―コーパス研究における特徴語分析の意義と可能性―」(https://clrd.ninjal.ac.jp/lrw/lrw2025/o05.pdf)を改題の上、加筆修正を行ったものである。ワークショップの主催機関である国立国語研究所言語資源開発センターより二重投稿に当たらないことの確認を得た上で新規の論文として公表している。

EJTKAN および本論文は科学研究費補助金(23H00641)に基づく研究成果の一部である。

研究インテグリティ宣誓

本稿著者は、本論文が著者自身によるオリジナルの研究成果であって、他者の著作物の剽窃や盗用、データの捏造や改竄、二重投稿、貢献度を反映しない著者名表記といった一切の不適切行為を行っていないことを宣誓する。なお、生成 AI に関しては一切使用していない。

付表. EJTKAN における特徴度尺度

以下、EJTKAN で採用した特徴度尺度の計算式と使用されている略号を示す。

表 3

略号

タイプ	関連略号一覧
コーパス種別	T (Target):ターゲットコーパス R (Reference):参照コーパス
対象語種別	X:調査対象語 O (Others):調査対象語以外のすべての語
頻度種別	Freq (Frequency):頻度実測値 ExpFreq (Expected Frequency):(コーパス間頻度に差がないという帰無仮説に基づく)頻度予測値 NormFreq (Normalized Frequency):相対頻度(頻度／総語数)
コーパスサイズ	Size:コーパス総語数
算術処理	abs (absolute):絶対値 ln (log natural):自然対数 min (minimum):最小値選択

表 4

特徴度尺度計算式

尺度	計算過程
A: 検定統計量計尺度	
カイ二乗統計量	χ^2 XT = (abs (TFreq－TExpFreq)－0.5)^2 / TExpFreq χ^2 XR = (abs (RFreq－RExpFreq)－0.5)^2 / RExpFreq χ^2 OT= (abs (TExpFreq－TFreq)－0.5)^2 / (TSize－TExpFreq) χ^2 OR = (abs (TFreq－TExpFreq)－0.5)^2 / TExpFreq χ^2 =χ^2 XT + χ^2 XR + χ^2 OT + χ^2 OR
対数尤度比統計量 (4 セル)	LLR XT = 2*TFreq*ln (TFreq / TExpFreq) LLR XR = 2*RFreq*ln (RFreq / RExpFreq) LLR OT = 2* (TSize－TFreq) *ln ((TSize－TFreq) / (TSize－TExpFreq)) LLROR = 2* (RSize－RFreq) *ln ((RSize－RFreq)/ (RSize－RExpFreq)) LLR 4term = LLR XT + LLR XR + LLR OT + LLR OR

対数尤度比統計量（2 セル）	LLR XT = 2*TFreq* ln (TFreq / TExpFreq) LLR XR = 2*RFreq* ln (RFreq / RExpFreq) LLR 2term = LLR XT + LLR XR
ベイズ因子(BIF)	BIF =LLR 2term－ln (TSize + RSize)
B: 効果量系尺度	
オッズ比	OddsRatio = (TFreq / (TSize－TFreq)) / (RFreq / (RSize－RFreq))
リスク比	RiskRatio= TNormFreq / RNormFreq
相対頻度差（%）	FreqDiff= (TFreq / TSize－RFreq / RSize) *100 / (RFreq / RSize)
LogRatio (対数化相対頻度比)	LogRatio = log (Risk Ratio, 2) = log ((TNormFreq/RNormFreq), 2)
対数尤度比効果量	ELL=LLR 2term / ((TSize + RSize) * ln (min (TExpFreq, RExpFreq)))

引用文献

Anthony, L. (2014-present). [Computer software]. AntConc. Waseda University.

Anthony, L. (2022). What can corpus software do? In A. O'Keeffe & M. J. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 103-125). Routledge.

Baker, P. (2004). Querying keywords: Questions of difference, frequency and sense in keyword analysis. *Journal of English Linguistics*, *32*(4), 346-359.

Baker, P., Gabrielatoss, C., & McEnery, T. (2013). Sketching Muslims: A corpus driven analysis of representations around the word "Muslim" in the British press 1998-2009. *Applied Linguistics*, *34*(3), 255-278.

Baker, P., Vessey, R., & McEnery, T. (2021). *The language of violent jihad*. Cambridge University Press.

Culpeper, J. (2009). Keyness: Words, parts-of-speech and semantic categories in the character-talk of Shakespeare's *Romeo and Juliet*. *International Journal of Corpus Linguistics*, *14*(1), 29-59.

Culpeper, J., & Demmen, J. (2015). Keyword. In D. Biber & R. Reppen (Eds.), *The Cambridge handbook of corpus linguistics* (pp. 90-105). Cambridge University Press.

Egbert, J., & Biber, D. (2019). Incorporating text dispersion into keyword analyses. *Corpora*, *14*(1), 77-104. https://doi.org/10.3366/cor.2019.0162

Gabrielatos, C., & Marchi, A. (2012, September 13-14). Keyness: Appropriate metrics and practical issues. [Conference presentation]. CADS International Conference 2012,

Corpus–assisted Discourse Studies: More than the Sum of Discourse Analysis and Computing?, University of Bologna, Italy.

Halliday, M. A. K. (1994). *An introduction to functional grammar* (2nd ed.). Edward Arnold.

Hardie, A. (n.d.). Log Ratio: An informal introduction. CASS: Corpus Approach. https://cass.lancs.ac.uk/log–ratio–an–informal–introduction/

石川慎一郎(2024)「『森を見ながら木を見る』コーパス研究の意義 ： 複数テキストから統合語彙頻度表を作成する EJWFTG の開発」『統計数理研究所共同研究リポート』*469*, 95-122.

石川慎一郎(2025)「English/Japanese Word Frequency Table Generator (EJWFTG)を用いた日本語統合語彙頻度表の作成と活用」『計量国語学』*34*(6), 421-431.

Johnston, J. E., Berry, K. J., & Mielke, P. W. (2006). Measures of effect size for chi–squared and likelihood–ratio goodness–of–fit tests. *Perceptual and Motor Skills, 103*(2), 412–414. https://doi.org/10.2466/pms.103.2.412-414

Jones, C. (2022). What are the basics of analysing a corpus? In A. O'Keeffe & M. J. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 126 - 139). Routledge.

Lutzky, U. (2021). Digital media and business communication. In E. Friginal & J. A. Hardy (Eds.), *The Routledge handbook of corpus approaches to discourse analysis* (pp. 394-407). Routledge.

Moreno–Ortiz, A. (2024). *Making sense of large social media corpora*. Palgrave Macmillan.

Pérez–Paredes, P. (2020). *Corpus linguistics for education: A guide for research*. Routledge.

Pérez–Paredes, P. (2024). Frequency and keyness: What are they and how can they be used to explore representation? In F. Heritage & C. Taylor (Eds.), *Analysing representation. A corpus and discourse textbook*. Routledge.

Pojanapunya, P., & Todd, R. W. (2021). The influence of the benchmark corpus on keyword analysis. *Register Studies*, *3*(1), 88-114.

Rees, G. (2022). Using corpora to write dictionaries. In A. O'Keeffe & M. J. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 387-404). Routledge.

迫田久美子(2020)「I-JAS 誕生の経緯」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』(pp.2-13) くろしお出版.

Scott, M. (1996 - present). [Computer software]. *WordSmith tools*. Oxford University Press/Lexical Analysis Software.

Scott, M. (1997). PC analysis of key words. *System*, *25*(2), 233-245.

Scott, M. (2010). Problems in investigating keyness, or clearing the undergrowth and

marking out trails….” In M. Bondi & M. Scott (Eds,), *Keyness in texts* (pp. 43-58). John Benjamins.

Scott, M., & Tribble, C. (2006). *Key words and corpus analysis in language education*. John Benjamins.

鈴木あすみ・幕内充・和田真・中村仁洋・岩井亨視・小磯花絵（2024）「日本語を母語とする自閉スペクトラム症者の公開コーパス構築」言語資源ワークショップ 2024 予稿（o08s）. https://clrd.ninjal.ac.jp/lrw/lrw2024/o08s-paper.pdf

UCREL NLP Group. (n.d.). [Online app]. Log-likelihood and effect size calculator. https://ucrel.lancs.ac.uk/llwizard.html

Wilkinson, M. (2021). Discourse analysis of LGBT identities. In E. Friginal & J. A. Hardy (Eds.), *The Routledge handbook of corpus approaches to discourse analysis.* (pp. 554-570). Routledge.

Williams, E. A. E. (2021). Critical discourse analysis for language policy and planning. In E. Friginal & J. A. Hardy (Eds.), *The Routledge handbook of corpus approaches to discourse analysis* (pp. 481-498). Routledge.

Wilson, A. (2013). Embracing Bayes factors for key item analysis in corpus linguistics. In M. Bieswanger & A. Koll–Stobbe (Eds.), *New approaches to the study of linguistic variability: Language competence and language awareness in Europe* (pp. 3-11). Peter Lang.

Xiao, R., & McEnery, T. (2005). Two approaches to genre analysis: Three genres in modern American English. *Journal of English Linguistics*, *33*, 62-82.