From Word Keyness to Phraseological Keyness: Version 2 of the Web-Based Corpus Analyzer (WBCA)

Tatsuya Ishii
Kochi University
Kochi, Japan
t-ishii@g.kochi-u.ac.jp

Abstract

This paper presents Version 2 of the Web-Based Corpus Analyzer (WBCA), a free, browser-based tool designed to support transparent and reproducible analysis of do-it-yourself corpora. Version 1 offered a KWIC-centered workflow covering corpus upload, target/reference selection, concordance, collocation, high-frequency feature extraction, and word-level keyness analysis (Ishii, 2025). Version 2 substantially extends the phraseological and statistical coverage of this workflow by applying a common analytical framework to words, lemmas, n-grams, p-frames, and POS-grams, and by adding multi-corpus and multi-measure keyness analyses. The extension responds to recent developments in corpus statistics and phraseology: keyness research has moved beyond raw frequency toward measures sensitive to statistical distinctiveness, text-level dispersion, and effect size (Brezina, 2018; Egbert & Biber, 2019; Larsson et al., 2025), while phraseological research has extended from fixed n-grams and lexical bundles (Biber et al., 2004; Hyland, 2008) to variable p-frames (Gray & Biber, 2013; Şahin Kızıl et al., 2024) and grammatically abstracted POS-grams (Lin & Liu, 2021). This paper describes the system's eight core functions and its keyness measures, together with worked-example checks of the implemented statistics. The system is intended to support more consistent and inspectable phraseological keyness analysis within a unified workflow.

Keywords: corpus analysis software, keyness analysis, phraseology, p-frames, POS-grams

1. Introduction

Keyword and keyness analyses are widely used in corpus linguistics, including register research, discourse analysis, and studies of second-language writing. In its simplest form, a keyness analysis compares the frequency of a word in a target corpus against a reference corpus and ranks words by a statistic such as log-likelihood (Dunning, 1993). Over the last decade, however, two independent strands of research have complicated this simple picture. First, corpus statisticians have argued that raw frequency comparison can be insufficient when used without text-level distributional information, and have proposed measures that additionally account for text-level dispersion and effect size (Brezina, 2018; Egbert & Biber, 2019; Larsson et al., 2025). Second, phraseologists have argued that keyness should not stop at the single word: recurrent multi-word sequences (n-grams and lexical bundles), variable-slot sequences (p-frames), and grammatically abstracted sequences (POS-grams) are equally, and sometimes more, informative about register and discipline (Biber et al., 2004; Gray & Biber, 2013; Hyland, 2008; Lin & Liu, 2021; Şahin Kızıl et al., 2024).

In practice, these two strands of research are difficult to combine. A researcher who wants to compute dispersion-sensitive keyness over p-frames, for example, typically needs to combine a concordancer, a bespoke n-gram/p-frame extraction script, and a separate statistics package, with no guarantee that frequency counts, frequency normalization (e.g., conversion to per-million-word rates), and tokenization are handled consistently across tools. This paper reports on Version 2 of the Web-Based Corpus Analyzer (WBCA), a free, browser-based tool designed to address this practical gap. Version 1 is distributed as an open-source release at https://github.com/ishitatu/Web-based-corpus-analyzer-; Version 2, newly developed and reported in this paper, is publicly available at https://github.com/ishitatu/Web-based-corpus-analyzer-VER2, with source code and documentation at https://github.com/ishitatu/Web-based-corpus-analyzer-VER2. Section 2 reviews the statistical and phraseological background that motivates the design. Section 3 describes the eight core functions of WBCA Version 2, together with the fine-grained settings each function provides and the analyses those settings enable. Section 4 discusses the contribution and limitations of the tool, and Section 5 concludes.

2. Background

2.1 Keyness Statistics Beyond Raw Frequency

Traditional keyness analysis ranks words by a frequency-based log-likelihood statistic computed from a 2×2 contingency table of raw token counts (Dunning, 1993). This approach is sensitive to "burstiness": a word that occurs many times in a small number of files can receive an inflated keyness score even though it is not characteristic of the target register as a whole. To address this, two families of text-sensitive approaches have been proposed. Text dispersion keyness compares the number of texts in which a feature occurs, thereby reducing the influence of repeated occurrences in a small number of texts (Egbert & Biber, 2019; cf. the document-count statistics of Paquot & Bestgen, 2009). Mean text frequency keyness instead compares the typical text-level frequency of a feature across the target and reference corpora (Larsson et al., 2025); the latter is implemented in WBCA as MTK. Alongside significance-based statistics, effect-size measures such as Log Ratio (Hardie, 2014) and the percentage difference between normalized frequencies (Gabrielatos & Marchi, 2012) are increasingly reported alongside, or instead of, log-likelihood, because statistical significance alone conflates corpus size with the size of a genuine difference (Brezina, 2018). Bayesian alternatives, such as the BIC-approximated Bayes factor (Wilson, 2013), have also been proposed as an evidence-based alternative to fixed p-value thresholds.

2.2 From Word Lists to Phraseological Units

A parallel development has broadened the unit of keyness analysis beyond the single word. Lexical bundles — the most frequent contiguous n-grams in a register — have been used to characterize the phraseology of academic writing, teaching, and disciplinary discourse (Biber et al., 2004; Hyland, 2008). Because fixed n-grams cannot capture systematic lexical variation within an otherwise stable frame (e.g. the * of the), p-frames generalize n-grams by allowing one or more variable slots within a recurrent sequence, and have been proposed as a more linguistically motivated unit for phrase-list development (Gray & Biber, 2013; Şahin Kızıl et al., 2024). POS-grams represent a further level of abstraction, replacing lexical items with part-of-speech categories to capture recurrent grammatical patterning independently of the specific words that instantiate it, and have been used to identify discipline-specific lexicogrammatical tendencies in research article introductions (Lin & Liu, 2021). Taken together, this body of work implies that a keyness-analysis tool should ideally support the same statistical machinery — frequency, dispersion, effect size, and significance — across all of these units, from single words to POS-grams, rather than treating word-level keyness and phraseological keyness as separate problems requiring separate tools.

3. WBCA Version 2: System Overview

WBCA is a client-side, browser-based application: a corpus never leaves the user's computer, which may make the tool suitable for analyzing unpublished or otherwise sensitive DIY corpora. Version 1 implemented a KWIC-centered workflow of corpus upload, target/reference selection, concordance, collocation, high-frequency feature extraction, and word-level keyness analysis (Ishii, 2025). Version 2 keeps this workflow intact and extends the principal frequency, concordance, and keyness modules to a broader range of lexical and phraseological representations — although the exact feature options vary by module. This paper focuses on the five principal feature types used in the keyness modules: word (surface form), lemma, n-gram, p-frame, and POS-gram. The version described in this paper corresponds to the WBCA v2.0 release archived in the project repository (August 2026); the public release may additionally include experimental input formats and feature types not discussed here. The following subsections describe the eight core functions, the settings each function exposes, and what those settings enable (Figures 1–5; Tables 1–2).

3.1 Function 1: Data Type and Upload

Function 1 defines what kind of data enters the tool and how it is tokenized. Two input modes are supported — plain text, and tagged text (word_POSd_POSs_lemma, produced with an accompanying Google Colab tagging notebook) — and an auto-detect option samples the first file to select the mode automatically. In tagged mode, every token carries two levels of part-of-speech annotation alongside its lemma: POSd is a detailed, Penn Treebank-style tag that distinguishes inflectional forms (e.g. NN vs NNS for singular and plural nouns, VBD vs VBG vs VBN for verb forms, JJR for comparative adjectives), while POSs is a simple, Universal Dependencies-style category (NOUN, VERB, ADJ, ADV, PRON, DET, ADP, etc.); a full correspondence table is provided in the tool's documentation. All POS-based analyses in later functions can be run at either level of granularity. Two upload routes are provided. Folder upload is used when the corpus already has a folder structure such as Move1/Move2: the analyst selects one parent folder, and each sub-folder placed directly under it — with the text files inside — becomes an independent corpus whose name is used throughout Functions 3–8, while the parent folder itself never becomes a corpus name (Figure 1). The hierarchy must therefore be prepared before upload, and the resulting corpus names should be checked against the folder list shown from Function 3 onwards. Multiple files upload, by contrast, is intended for trying out a small number of individual text files: since these carry no folder information, all uploaded files are grouped into one virtual corpus named “Ungrouped”. Because the choice of mode determines which feature types are available later (lemma-, POS-, and POS-gram-based analyses require tagged data), Function 1 effectively fixes the analytical vocabulary of the whole session.

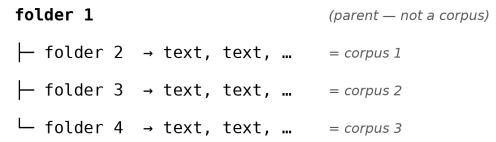


Figure 1: Folder structure for folder upload. Each sub-folder directly under the selected parent folder becomes an independent corpus; the parent folder itself does not.

A series of preprocessing settings then controls how tokens are defined, each of which changes what later analyses can see. P-frame settings determine the position of variable slots, with a live pattern preview. Merge short lines joins subtitle- or lyrics-style fragments into paragraphs so that KWIC lines show usable context. Word-definition options decide whether hyphenated forms (state-of-the-art), clitics (can't, we'll), or user-specified symbol strings count as one word, and a dedicated switch supports space-pre-tokenized Japanese text. A Unify apostrophes option normalizes curly and other variant apostrophes (‘ ’ ’) to the straight ' at load time, so that don't and don’t are counted as the same token in both surface and lemma counts; the setting takes effect on the next corpus load, and searches are then typed with the straight apostrophe. Number handling offers four policies — keep numerals as-is, replace pure-number tokens with #, replace every digit with #, or exclude pure-number tokens from token/type counts — allowing numerals to be either studied or neutralized. Spelling normalization can map British to American variants (colour/color) on surface forms so that the two spellings are counted as one item. Export formatting can follow APA 7th or MLA 9th conventions, and stopword lists (used by Functions 7 and 8) are fully editable, with built-in lists drawn from spaCy, scikit-learn, Gensim, and Elasticsearch plus an integrated superset, selected separately for surface and lemma matching. The tool documents which changes require re-parsing and which only require re-computation.

3.2 Function 2: Corpus Summary

Function 2 provides an immediate quality check on the loaded data. As soon as a corpus finishes loading, the tool reports the detected mode and the total number of files and tokens (punctuation and space tokens excluded), and

displays a per-file table of folder, file name, and token count, sortable by any column. This apparently simple step plays an important methodological role: encoding problems, truncated files, or accidentally included documents are visible here as anomalous token counts before any statistic has been computed, and a wrong mode detection can be corrected in Function 1 before it silently disables lemma- or POS-based analyses.

3.3 Function 3: Corpus Overview

Function 3 organizes the corpus for comparison and profiles its composition. In sub-section 3a (Corpus Information; Figure 2), each folder is assigned to a Target and/or a Reference group by checkbox — the same folder may serve on both sides, and an "All" toggle switches a whole column at once. For every folder the tool reports files, sentences, and paragraphs; minimum, maximum, median, and mean tokens per file; and tokens, types, and TTR, with drag-and-drop reordering that is shared with the keyness sections. Setting the STTR and MATTR window sizes and pressing Re-compute metrics produces a full battery of lexical richness and diversity indices — TTR, Guiraud's R, CTTR, Herdan's C, Maas a^2, Dugast's U, Brunet's W, STTR, MATTR, MTLD, HD-D, vocd-D, and Yule's K and I — together with mean word length, mean sentence length, and (for tagged data) lexical density, and the results can be visualized as bar charts or boxplots. Folder selections are synchronized in two independent pairs (3a with 8a; 8b with 8c), while Reference selections are shared across all four keyness locations; merge-target groups defined here let several folders be treated as one target corpus. Sub-section 3b (Tag Distribution) reports the per-folder distribution of part-of-speech tags, switchable between raw frequency, per million words, and percentage of folder tokens, in either table orientation.

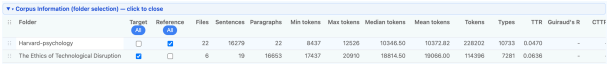


Figure 2: Corpus Information (Function 3a): Target/Reference selection with per-folder descriptive statistics.

3.4 Function 4: KWIC Concordance

Function 4 is the tool's central concordancer, and its settings define both what is matched and how the evidence is displayed. Three search modes are available — Exact (with | for alternatives, e.g. cell|cells), Wildcard (* and ?, e.g. *ing, b?t), and full regular expressions (stud(y|ies|ied)) — each with optional case insensitivity. In tagged mode an Advanced pattern format, surface_POSdetailed_POSsimple_lemma, matches any combination of surface form, detailed POS, simple POS, and lemma (e.g. *_NN|NNS_NOUN_* for common nouns), turning the concordancer into a lexicogrammatical query tool. Display settings control the maximum number of lines, the number of words shown left and right of the node, the search scope (Target only, Reference only, all folders, or a specific folder), a surface versus tagged view, and a color scheme that assigns colors to positions L1–L5, node, and R1–R5 and is inherited by Functions 5 and 6. Positional filters on the left context, node, and right context accept exact, partial, wildcard, POS (simple or detailed) or POS-gram conditions at specified positions, position ranges, or as exclusions — so that, for example, only hits with a noun immediately to the right are kept. Concordance lines and node lists can be copied or exported to Excel and CSV, and every row of the keyness tables (Function 8) opens its own KWIC view with one click.

3.5 Function 5: Text Data Views

Function 5 extends concordance reading to the level of the whole text. In sub-section 5a (Original Text), clicking any KWIC line displays the full running text of the source file, with the matched positions color-coded according to the scheme chosen in Function 4 and a surface/tagged toggle. Sub-section 5b (Concordance Plot) draws a barcode-style dispersion plot marking where each hit occurs across the span of a text; hits can be grouped by file or by folder, files or folders without hits can be hidden, and clicking a bar opens the corresponding KWIC lines. The plot accepts the same Exact, Wildcard, and Regex search modes and scopes as KWIC, and remains synchronized with the Advanced pattern. Figures can be copied or downloaded as image files.

3.6 Function 6: Collocation Analysis

Function 6 computes the collocates of a node — a word or a p-frame such as in the * — within a user-defined window of one to five words on each side, separately for the Target and Reference corpora, and displays the two collocate profiles side by side. Collocates can be counted on surface form, lemma, or POS, and more than twenty association measures are implemented, including frequency, MU, t-score, z-score, MI, MI^2, MI^3, log-likelihood (two- and four-cell), LogDice, Dice, minimum sensitivity, directional Delta P, Cohen's d, Log Ratio, log odds ratio, χ^2, Fisher's Exact, Poisson-Stirling, Log-Log, Jaccard, Cosine, Simpson, and directional Kullback–Leibler contributions; Table 1 groups these measures by what they emphasize (Brezina, 2018). A Version switch selects between uncorrected (a) and boundary-corrected (b) computation of the window, and three thresholds — minimum statistic score, minimum collocate frequency (C), and minimum co-occurrence frequency (NC) — separate low-frequency noise from statistically weak association. These settings correspond directly to the standard Collocation Parameters Notation, e.g. 4b–MI2(3), L5–R5, C5–NC1 (Brezina et al., 2015), so that results can be reported reproducibly. Collocate POS filters update the tables instantly after the first computation, and clicking any collocate opens a KWIC view restricted to the relevant window position.

Group	Measures	Typical use
Frequency-oriented	Frequency, t-score, LogDice, Log-Log	Highlight frequent, typical collocations; robust for common patterns
Exclusivity-oriented	MI, MU, Delta P (both directions), DKL contribution	Highlight rare but strongly associated pairs (technical terms, fixed phrases)
Compromise / balanced	MI^2, MI^3, Dice, minimum sensitivity (MS), Jaccard, Cosine, Simpson	Reduce the low-frequency bias of MI while retaining exclusivity
Significance tests	LL (G^2) 2-/4-cell, z-score, χ^2, Fisher's Exact, Poisson-Stirling	Test the null hypothesis of independence between node and collocate
Effect size / directional	Log Ratio, log odds ratio, Cohen's d, Delta P	Quantify the magnitude and direction of association for reporting

Table 1: Association measures available in Function 6, grouped by what they emphasize (cf. Brezina, 2018). The

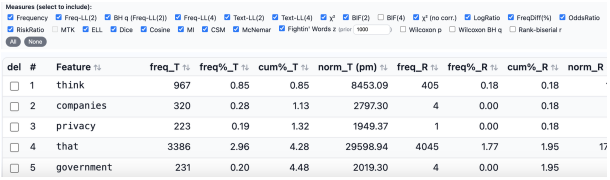
grouping is an interpretive guide rather than a mutually exclusive statistical taxonomy.

3.7 Function 7: High-Frequency Features

Function 7 extracts the most frequent items of a chosen feature type — word, lemma, n-gram, p-frame, POS-gram, or cluster (n-grams containing a given search word) — with user-adjustable n-length and a minimum frequency threshold for each corpus. For p-frames, the number of variable slots (stars) can be fixed exactly or bounded from above, so that frames of different fixedness can be compared. Tie handling is explicit: the Top-N cut-off can include ties beyond N, exclude them, or apply a strict cut, with alphabetical or generation order as the tie-breaker — settings that make frequency lists exactly reproducible. Instead of an uploaded Reference, a built-in general-corpus frequency list (BNC1994 whole, written, or spoken, BNC Consortium, 2007; BNC2014, Love et al., 2017, Brezina et al., 2021; AmE06, Potts & Baker, 2012; BE06, Baker, 2009; Brown, Kučera & Francis, 1967) can be used for word-level comparison. The stopword lists customized in Function 1 can be applied, a word-class filter restricts output to function or content words, and optional POS columns support part-of-speech filtering. A search-word filter (with *, ?, and |) extracts only n-grams containing a given word at any, initial, or final position — turning the list into a targeted phraseology of a single item. Function 7 is most useful when the analyst wants a descriptive picture of the dominant vocabulary and phraseology of each corpus before any statistical contrast — for example, the lexical bundles that structure a genre, or the most frequent frames of a discipline. Excluding stopwords brings content words to the top of the list instead of the, of, and and; the word-class filter can conversely restrict the list to function words only, for grammatical-style comparisons; POS columns allow the list to be narrowed to nouns, verbs, or any other category; and the case-insensitive option merges capitalized variants (Government/government) or keeps them distinct. Output columns report raw frequency, percentage share, cumulative percentage, normalized frequency per million words, and file range for each side, and the results can be transposed, compared through an overlap matrix of top-N lists, looked up across all corpora in a feature profile, visualized as bar charts, or exported to Excel/CSV. Clicking any feature in the list jumps to its KWIC concordance (Function 4).

3.8 Function 8: Keyness Analysis

Function 8 is the analytical centerpiece of Version 2 and consists of three sub-sections. Sub-section 8a (Standard Keyness; Figure 3) compares the Target and Reference groups defined in Function 3 for any of the five feature types under identical settings — feature type, n, p-frame stars, minimum total frequency, Top N with the same tie handling as Function 7, and optional case folding. Rather than committing to a single keyness statistic, WBCA computes a user-selected battery of measures side by side; Table 2 summarizes them by category. In the measure labels, (2) and (4) distinguish two-cell implementations, computed from the two observed frequencies and their expected values, from four-cell implementations computed over the full 2×2 contingency table; text-based variants replace token frequencies with the number of files containing the feature. Benjamini–Hochberg false-discovery-rate corrected q-values (Benjamini & Hochberg, 1995) can be attached to the log-likelihood tests; the correction is computed over the full set of candidate features meeting the minimum-frequency threshold, before any Top-N truncation. A Show p-values option adds p-value columns with the conventional UCREL critical values (3.84, 6.63, 10.83, 15.13 for $p < .05, .01, .001, .0001$ at df = 1). An Advanced Statistics option adds per-side dispersion measures — Range, Juilland's D, DP and DP_norm (Gries, 2008; Lijffijt & Gries, 2012), Carroll's D_2, Rosengren's S, and KL divergence — plus file-based contrasts (DispRatio, DispBIF, DispLogR; Egbert & Biber, 2019). Sub-section 8a is intended for analyses in which the goal is to identify what is statistically characteristic of one corpus against another — for example, which words, bundles, or frames distinguish one rhetorical move, discipline, or register from a baseline. As in Function 7, the analysis can be constrained to the vocabulary of interest: stopword exclusion removes function words, the word-class filter limits keyness to content words only (or, conversely, to function words for grammatical comparisons), POS filters restrict the table to chosen parts of speech, and case-insensitive folding determines whether capitalized variants are merged. Computation can be stopped or canceled at any point, and clicking any row of the keyness table jumps straight to its KWIC concordance (Function 4). Log Ratio is the base-2 logarithm of the ratio of relative frequencies (Hardie, 2014); FreqDiff(%) is (normT − normR)/normR × 100 (Gabrielatos & Marchi, 2012); the odds ratio uses the Haldane–Anscombe correction when a cell is zero; the Wilcoxon test is an unpaired rank-sum (Mann–Whitney U) test over per-text per-million-word frequencies, reported with a rank-biserial r effect size; and MTK implements the mean text frequency keyness of Larsson et al. (2025). Complete formulas, zero-cell treatments, logarithm bases, and implementation details are documented in the in-tool statistical reference distributed with the archived release.


Measures (select to include):
Frequency | Freq-LL(2) | BH q (Freq-LL(2)) | Freq-LL(4) | Text-LL(2) | Text-LL(4) | χ^2 | BIF(2) | BIF(4) | χ^2 (no corr.) | LogRatio | FreqDiff(%) | OddsRatio
RiskRatio | MTK | ELL | Dice | Cosine | MI | CSM | McNemar | Fightin' Words z | Wilcoxon p | Wilcoxon BH q | Rank-biserial r
All | None

del	#	Feature	freq_T	freq%_T	cum%_T	norm_T (pm)	freq_R	freq%_R	cum%_R	norm_R
☐	1	think	967	0.85	0.85	8453.09	405	0.18	0.18	
☐	2	companies	320	0.28	1.13	2797.30	4	0.00	0.18	
☐	3	privacy	223	0.19	1.32	1949.37	1	0.00	0.18	
☐	4	that	3386	2.96	4.28	29598.94	4045	1.77	1.95	17
☐	5	government	231	0.20	4.48	2019.30	4	0.00	1.95	

Figure 3: Standard keyness (Function 8a), excerpt: measure selection (top) and the left part of the resulting keyness table (bottom; word-level example).

Category	Measures	What they indicate
Descriptive frequency	freq, freq%, cum%, norm (per million words), file counts and file %	Raw and normalized prominence of a feature and its text range on each side
Significance (token-based)	Freq-LL(2)/(4) (Dunning, 1993), χ^2 with/without Yates, BH q (Benjamini & Hochberg, 1995)	Whether pooled frequency differs from chance; q controls the false discovery rate over many features
Significance (text-based)	Text-LL(2)/(4) (Paquot & Bestgen, 2009), Wilcoxon rank-sum p and BH q, rank-biserial r, McNemar score as operationalized by Chujo & Utiyama (2006)	Document-count and per-text tests robust to bursty files

Bayesian	BIF(2)/(4) (Wilson, 2013)	BIC-approximated log Bayes factor (natural-log scale); BIF > 2 is read as positive evidence for a frequency difference
Effect size	Log Ratio (Hardie, 2014), FreqDiff% (Gabrielatos & Marchi, 2012), odds ratio, risk ratio, ELL	Primarily quantify the magnitude of a difference rather than statistical evidence; may be unstable at low or zero frequencies
Association / similarity	Dice, Cosine, MI, CSM, Fightin' Words z (Monroe et al., 2008)	Association between feature and corpus; regularized log-odds with a Dirichlet prior for sparse counts
Dispersion-aware	MTK (mean text frequency keyness; Larsson et al., 2025); Range, Juilland's D, DP, DP_norm, Carroll's D_2, Rosengren's S, KL; DispRatio, DispBIF, DispLogR (Egbert & Biber, 2019)	Text-level frequency contrasts and evenness of spread over texts, plus file-based keyness contrasts

Table 2: Keyness and related measures available in Function 8a, by category.

Sub-section 8b (Multi-Corpus Keyness; Figure 4) treats each Target folder (or merged target group) as a separate corpus and computes keyness for all of them independently against one combined reference, producing a summary table in which folder-by-folder differences can be scanned at once. This layout is particularly useful when three or more corpora are compared simultaneously — for example, tracking how key p-frames shift across the rhetorical moves of a genre, or which vocabulary each discipline in a multi-disciplinary corpus favors. The same stopword, word-class, and POS restrictions as 8a apply, so the comparison can be limited to content words or to a chosen part of speech. Its Target set is deliberately independent of 8a (8b and 8c share one selection, 3a and 8a another, while Reference folders are shared by all four), so two different corpus designs can be kept side by side. An Exclude Target from Reference option removes each target's own tokens from the combined reference, avoiding self-comparison; the ranking measure (any of the 8a battery, or dispersion measures) is chosen separately from the set of statistics written to the export sheets; and threshold filters on statistic value, frequency, and per-million rate prune weak rows. Clicking any cell jumps to KWIC (Function 4) with the clicked folder pre-selected, and the table can be transposed, copied with or without values, or exported.

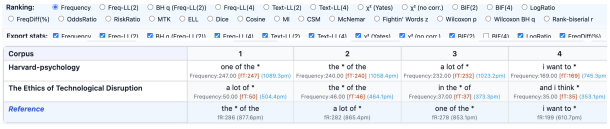


Figure 4: Multi-Corpus Keyness (Function 8b), excerpt: ranking and export-statistics selection (top) and the per-corpus summary table (bottom; 4-word p-frames against a merged reference).

Sub-section 8c (Multi-Measure Analysis; Figure 5) addresses the methodological question of how much the choice of keyness measure matters — useful both for checking the robustness of a keyword list before reporting it and for selecting an appropriate measure when building pedagogical word or phrase lists. It reproduces the comparison design of Chujo and Utiyama (2006) over user-selected measures and three list lengths (N_1–N_3). The module produces four output tables: an overlap table counting shared items among the top-N_2 lists of each measure; a Kendall's τ rank-correlation table among the measures, computed over the ranked top-N_1 features (the treatment of ties and of items absent from one list is documented in the in-tool reference); a side-by-side ranked-list table showing each measure's actual top-N_2 items, with summary rows for mean general-corpus frequency, proportion of function words, and mean word length; and a frequency-band table showing how each measure's top-N_3 selections distribute across general-corpus frequency bands. Features can be searched across all columns, and the display switched between dropdown, tabbed, and stacked modes. Like Functions 7, 8a, and 8b, 8c links every listed item back to KWIC (Function 4) — and from there to the original text (Function 5) — so that a statistically salient word, n-gram, p-frame, or POS-gram can be inspected in its own concordance without leaving the interface.

Figure 5: Multi-Measure Analysis (Function 8c), excerpt: the side-by-side ranked-list output for the first four measures (top-50 p-frames per measure).

3.9 Validation and Limitations of Scope

As an initial check on the implementation, the worked examples documented in the tool's statistical reference were verified against independently calculated values. For the reference example — a 2×2 table with a = 12, b = 3 and N_T = N_R = 1,000 — the tool's documented outputs (Freq-LL(2) = 5.782; Yates-corrected $\chi^2 = 4.299$; Log Ratio = 2.000; odds ratio = 4.036; risk ratio = 4.000) match hand calculation and the UCREL log-likelihood calculator to the displayed precision, and the collocation example for make + decision reproduces the documented t-score, z-score, MI, and LogDice values. These checks cover the core contingency-table machinery on which most of the implemented measures are built; a systematic cross-tool and cross-browser benchmark over a public test corpus, with archived expected outputs, is planned as future work and will accompany the released version.

4. Discussion

The central contribution of WBCA Version 2 is not any single statistic or feature type in isolation — most of the measures described in Section 3.8 have precedents in the corpus-linguistics literature and in existing software — but the fact that word-, lemma-, n-gram-, p-frame-, and POS-gram-level analyses are computed under one comparable statistical setting inside a single browser tab, with immediate access back to concordance and source text. Established tools such as AntConc (Anthony, 2026), CasualConc (Imao, 2026), #LancsBox (Brezina & Platt, 2026), WordSmith Tools (Scott, 2026), CQPweb (Hardie, 2012), and Sketch Engine (Kilgarriff et al., 2014) each cover parts of this workflow, often with larger corpora or richer query languages; WBCA does not claim that any individual analysis is unavailable elsewhere, but that this particular combination — phraseological units, a broad battery of keyness measures, multi-corpus and multi-measure comparison, and click-through verification — can be carried out in one client-side browser environment without scripting. This addresses a genuine methodological problem: when frequency counts, tokenization, and normalization are handled slightly differently by an n-gram script, a keyness spreadsheet, and a concordancer, results are difficult to compare and reproduce across studies. The tool's settings are also designed for reproducible reporting: collocation parameters map directly onto CPN notation (Brezina et al., 2015), tie handling in frequency and keyness lists is explicit, and the multi-measure module makes the sensitivity of results to the choice of statistic itself an object of study. Because WBCA runs entirely client-side, corpora never leave the user's computer, which also makes the tool usable with unpublished learner, spoken, or otherwise sensitive DIY corpora that researchers may be unable to upload to a server-based service.

The tool also has limitations. As a browser-based application, its computational capacity is bounded by client-side JavaScript performance, and very large corpora (tens of millions of tokens) or computationally heavy operations over very long n-grams may be slower than a dedicated desktop or command-line pipeline. The range of implemented measures, while broad, is not exhaustive, and some corpus-specific preprocessing (e.g. non-English tokenization) still depends on the external tagging notebooks referenced in Function 1. Future work will evaluate the agreement and disagreement among the keyness measures implemented in Function 8 across a range of registers, and will extend the tool's visualization components to the p-frame and POS-gram feature types introduced in this version.

5. Conclusion

This paper has described Version 2 of the Web-Based Corpus Analyzer, a free browser-based tool that extends a KWIC-centered keyness workflow from single words to lemmas, n-grams, p-frames, and POS-grams, and that implements a wide range of frequency-based, text-based, effect-size, and dispersion-sensitive keyness measures within a single, comparable statistical setting. By keeping corpus upload, concordance, collocation, frequency profiling, and keyness analysis inside one interface, and by exposing its settings in a form that maps onto standard reporting conventions, WBCA Version 2 aims to make phraseological keyness analysis as reproducible and as accessible as traditional word-level keyness analysis has become.

6. Bibliographical References

Anthony, L. (2026). *AntConc (Version 4.4.2)* [Computer software]. Tokyo: Waseda University. https://www.laurenceanthony.net/software

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57(1), 289–300.

Biber, D., Conrad, S. and Cortes, V. (2004). If you look at…: Lexical bundles in university teaching and textbooks. *Applied Linguistics*, 25(3), 371–405.

Brezina, V. (2018). *Statistics in Corpus Linguistics: A Practical Guide.* Cambridge University Press.

Brezina, V., McEnery, T. and Wattam, S. (2015). Collocations in context: A new perspective on collocation networks. *International Journal of Corpus Linguistics*, 20(2), 139–173.

Brezina, V. & Platt, W. (2026). *#LancsBox (Version 6.0.1)* [Computer software]. Lancaster University, http://lancsbox.lancs.ac.uk.

Chujo, K. and Utiyama, M. (2006). Selecting level-specific specialized vocabulary using statistical measures. *System*, 34(2), 255–269.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1), 61–74.

Egbert, J. and Biber, D. (2019). Incorporating text dispersion into keyword analyses. *Corpora*, 14(1), 77–104.

Gabrielatos, C. and Marchi, A. (2012). Keyness: Appropriate metrics and practical issues. Paper presented at the *CADS International Conference 2012*, University of Bologna, Italy.

Gray, B. and Biber, D. (2013). Lexical frames in academic prose and conversation. *International Journal of Corpus Linguistics*, 18(1), 109–136.

Gries, S. Th. (2008). Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics*, 13(4), 403–437.

Hardie, A. (2012). CQPweb — combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics*, 17(3), 380–409.

Hardie, A. (2014). Log Ratio – an informal introduction. *CASS Briefings*, Lancaster University. Available online: https://cass.lancs.ac.uk/log-ratio-an-informal-introduction/

Hyland, K. (2008). As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes*, 27(1), 4–21.

Imao, Y. (2026). CasualConc (Version 4.1.0) [Computer software]. Retrieved from https://sites.google.com/site/casualconcj/

Ishii, T. (2025). Developing Version 1 of the Web-Based Corpus Analyzer (WBCA): A browser-based corpus analysis tool with integrated workflow of KWIC, collocation, and keyness analysis. *Journal of Corpus-based Lexicology Studies*, 8, 1–20.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P. and Suchomel, V. (2014).

The Sketch Engine: Ten years on. *Lexicography*, 1(1), 7–36.

Larsson, T., Kim, T. and Egbert, J. (2025). Introducing and comparing two techniques for key lexical bundles analysis. *Research Methods in Applied Linguistics*, 4(3), 100245.

Lijffijt, J. and Gries, S. Th. (2012). Correction to Stefan Th. Gries' "Dispersions and adjusted frequencies in corpora". *International Journal of Corpus Linguistics*, 17(1), 147–149.

Lin, L. and Liu, M. (2021). Towards a Part-of-Speech (PoS) gram approach to academic writing: A case study of research introductions in different disciplines. *Lingua*, 254, 103052.

Monroe, B. L., Colaresi, M. P. and Quinn, K. M. (2008). Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4), 372–403.

Paquot, M. and Bestgen, Y. (2009). Distinctive words in academic writing: A comparison of three statistical tests for keyword extraction. In A. Jucker, D. Schreier and M. Hundt (eds.), *Corpora: Pragmatics and Discourse*, pp. 247–269. Amsterdam: Rodopi.

Şahin Kızıl, A., Vincent, B. and McCallum, L. (2024). A methodological exploration of the p-frame approach: Implications for developing phrase lists. *Research Methods in Applied Linguistics*, 3(3), 100154.

Scott, M. (2026). *WordSmith Tools (Version 9)* [Computer software]. Stroud: Lexical Analysis Software.

Wilson, A. (2013). Embracing Bayes factors for key item analysis in corpus linguistics. In M. Bieswanger and A. Koll-Stobbe (eds.), *New Approaches to the Study of Linguistic Variability*, pp. 3–11. Berlin: Peter Lang.

7. Language Resource References

Baker, P. (2009). The BE06 Corpus of British English and recent language change. *International Journal of Corpus Linguistics*, 14(3), 312–337.

BNC Consortium. (2007). *The British National Corpus, version 3 (BNC XML Edition).* Distributed by Bodleian Libraries, University of Oxford, on behalf of the BNC Consortium.

Brezina, V., Hawtin, A. and McEnery, T. (2021). The Written British National Corpus 2014 — design and comparability. *Text & Talk*, 41(5–6), 595–615.

Ishii, T. (2025). *Web-Based Corpus Analyzer (WBCA), Version 1.* [Software]. https://github.com/ishitatu/Web-based-corpus-analyzer-

Ishii, T. (2026). *Web-Based Corpus Analyzer (WBCA), Version 2.* [Software]. https://github.com/ishitatu/Web-based-corpus-analyzer-VER2

Kučera, H. and Francis, W. N. (1967). *Computational Analysis of Present-Day American English.* Providence, RI: Brown University Press.

Love, R., Dembry, C., Hardie, A., Brezina, V. and McEnery, T. (2017). The Spoken BNC2014: Designing and building a spoken corpus of everyday conversations. *International Journal of Corpus Linguistics*, 22(3), 319–344.

Potts, A. and Baker, P. (2012). Does semantic tagging identify cultural change in British and American English? *International Journal of Corpus Linguistics*, 17(3), 295–324. [Describes the AmE06 corpus.]