DE GRUYTER

Corpus Linguistics and Linguistic Theory

CLLT ‘vs’ Corpora and IJCL: A (half serious) keyness analysis

Journal:	*Corpus Linguistics and Linguistic Theory*
Manuscript ID	Draft
Manuscript Type:	Article
Classifications:	
Keywords:	CLLT, association-based keyness, dispersion-based keyness, Kullback-Leibler divergence, log-likelihood ratio

CLLT 'vs' *Corpora* and *IJCL*: A (half serious) keyness analysis

Abstract

In this introduction to the special issue celebrating CLLT's 20th anniversary, we look back and forward in time. To look back, we present the results of a (tongue-in-cheek) corpus-linguistic analysis of about 10 years worth of data of research published in *CLLT*, *IJCL*, and *Corpora* in order to distill the "essence" of *CLLT* for the reader. As an added bonus, we use the opportunity to discuss ways to improve established ways of performing keyness analyses. To look forward, we asked six (teams of) researchers who all have shaped corpus linguistics and thus the journal to give us their take on what the most significant developments in the field have been, and where they see the most impactful opportunities and challenges arise. This introduction briefly summarizes their contributions.

Keywords

CLLT, association-based keyness, dispersion-based keyness, Kullback-Leibler divergence, log-likelihood ratio

1. Introduction

The inaugural issue of *Corpus Linguistics and Linguistic Theory* (*CLLT*) as published in 2005. Published initially with Stefan Th. Gries and Anatol Stefanowitsch as editors-in-chief, the journal was under Stefan's stewardship between 2010 and 2015; in 2016, Stefanie Wulff took over as editor-in-chief while Stefan continued to serve as general editor. Initially, the journal was published twice a year, with typically six papers per issue; since 2020, we publish three issues yearly with typically six to seven papers per issue. There are two main reasons for this expansion. The first is that *CLLT*'s impact factor has steadily risen since its inception. Its 2023 impact factor is 1.6, which makes it the highest-ranking corpus-linguistic journal. The second reason is that, maybe partially as a result of the journal being perceived as impactful, submission rates to the journal have risen steadily, especially in the past decade. In 2023, the last complete year at the time of writing this, there were 132 first submissions to *CLLT*. More importantly, the caliber of the vast majority of these submissions is, as assessed by the editors, editorial board members, and reviewers, quite high. Correspondingly, we can estimate our average acceptance rate to be around 15%, which we consider rather competitive for a comparatively specialized journal like *CLLT*.

The approaching 20th anniversary of *CLLT* presents an opportunity for reflection, both looking back and looking ahead in time. How has corpus linguistics as a discipline evolved over time? What is the role and place of the work published in *CLLT* in corpus linguistics? And what are emerging developments in the field that stand to impact the research published in the journal? Are there any lessons to be learned from the past or the future on how to best steward *CLLT* into its next 20 years? It is these questions that the present special issue aims to address. As a broad lens, Stefan and I open this special issue by offering the results of a (only half-serious) corpus-linguistic analysis below. We analyzed about 10 years worth of data of research published in *CLLT*, *IJCL*, and *Corpora* in order to distill the "essence" of *CLLT* for the reader. As an added bonus, we use the opportunity to discuss ways to improve established ways of performing keyness analyses (a "signature move" of one of us two authors – the reader familiar with the field will be able to tell which one). To zoom in more closely, and crucially, to give varied perspectives on what lies ahead, we asked several colleagues who all have shaped corpus linguistics and thus the journal to give us their take on what the most significant developments in the field have been, and where they see the most impactful opportunities and challenges arise. Section 5 briefly introduces these authors and their contributions.

2. A broad lens: A (half-serious) keyness analysis

To determine how *CLLT* is different from its two main companion journals in corpus linguistics – *Corpora* and the *International Journal of Corpus Linguistics* (IJCL) – we proceeded with a keywords analysis, specifically, a newly-developed version of keywords analysis that was first proposed in Gries (2021) and that aims to overcome some of the shortcomings of existing versions.

For more than 20 years now, keywords analysis has essentially been an extension of collocation research. In collocation research, an association measure quantifies how strongly a node and each of its collocates are attracted to each other; in keywords analysis, association measures correspondingly quantify how strongly a corpus and each of the words occurring in it are attracted to each other. In collocation research, that often meant computing G^2, or the log-likelihood ratio, (Dunning 1993) on tables such as Table 1, and the analogous application in keywords analyses then involved computing G^2 on tables such as Table 2.

Table 1: Schematic co-occurrence table for measuring the association of collocates

	Node word *n*	All other words	Sum
Collocate *x*	*a*	*b*	*a+b*
All other collocates	*c*	*d*	*c+d*
Sum	*a+c*	*b+d*	*a+b+c+d*

Table 2: Schematic co-occurrence table for measuring keyness

	Target corpus *T*	Reference corpus *R*	Sum
Word *w*	*a*	*b*	*a+b*
All other words	*c*	*d*	*c+d*
Sum	*a+c*	*b+d*	*a+b+c+d*

That is, for Table 1, G^2 would be used to indicate how much the co-occurrence frequency of *n* with *x* (cell *a*) differed from its null hypothesis expectation, and collocates would be considered interesting if observed *a* was higher than expected *a* with a high G^2-score. Correspondingly, for Table 2, G^2 would be used to indicate how much the frequency of occurrence of *w* in the target corpus (cell *a*) differed from its null hypothesis expectation, and words would be considered key of the target corpus if observed *a* was higher than expected *a* with a high G^2-score. Thus, regardless of the keyness measure used – see Baron et al. (2009), Paquot & Bestgen (2009), Gabrielatos (2018), Rayson & Potts (2020), or Cvrček & Fidler (2022) for overviews and comparisons – keyness has traditionally been a one-dimensional construct allowing us to sort words by their position on a continuum from 'most/highly key for the target corpus' via 'not key for either corpus' to 'highly/most key for the reference corpus'.

Over time, three suggestions were made to improve such keywords analyses. They all involve including more information about the distribution of *w* in the target corpus. One suggestion involves the idea to add dispersion information, either by requiring a minimal amount of dispersion in the target corpus (e.g. Millar & Budgell 2010) or by using what has come to be called key keywords, i.e. "words that are key in a large proportion of texts in a corpus" (e.g., Scott 1997, Scott & Tribble 2006). That is, a separate keyness analysis is done for each part of a corpus, and words that are key for the target corpus in many texts are considered key keywords (see Baker 2004 for critical discussion).

Another suggestion made by, among others, Paquot & Bestgen (2009) is to use statistical tests that do not just aggregate file-/text-specific results within a corpus but instead work on relative word frequencies per file, which can then be submitted to, say, a *t*-test or a *U*-test. In their comparative evaluation, they conclude the *t*-test outperforms not only the standard measure of G^2, but also the non-parametric *U*-test.

The last suggestion is Egbert & Biber's (2019) notion of dispersion keyness. It involves the computation of G^2 on tables such as Table 2, but the 4 cell frequencies *a*, *b*, *c*, and *d* are not frequencies of occurrence of *w* and not-*w* in the target and the reference corpus, they are the numbers of texts/parts of the two corpora that *w* and not-*w* are attested in (i.e. what has been called the ranges of *w* and not-*w*). In other words, Egbert and Biber are not so much *complementing* keyness with dispersion like the approaches just mentioned, they are *replacing* frequency by dispersion as measured by range. While this is an interesting proposal, given how dispersion is indeed an important and underutilized construct, their approach is, however, not optimal: First, the dispersion measure they use is very coarse-grained because it considers only the number of texts in a corpus that a word is attested in, but does not consider the sizes of the corpus parts. Second, their dispersion measure also does not consider how often the word is attested in the corpus parts it is attested in, which means that the numeric value of 'frequency of occurrence' (absolute or relative) is reduced to a binary value: 'yes' (the word is in a corpus part) vs. 'no' (the word is not in a corpus part). Third, the above notwithstanding, the fact also remains that range as a dispersion measure is still so very highly correlated with frequency that, on the whole, their results are still more informed by frequency than by actual dispersion. Correspondingly, Gries (2022, to appear) shows that, across six widely used corpora, range is correlated with frequency (as measured by R^2s from GAMs) between 0.9104 and 0.9619.

These considerations underscore the general relevance of the notion of corpus homogeneity for keywords as mentioned, for instance, by Baron et al. (2009:264): "Homogeneity (Stubbs 1996:152) within each of the corpora is important since we may find that the results reflect sections within one of the corpora that are unlike other sections in either of the corpora under consideration (Kilgarriff 1997)"; or by Brezina & Meyerhoff (2014), who essentially echo this kind of previous research and point out how, from a more sociolinguistic perspective, the aggregation of frequency counts per corpus emphasizes between-group/corpus differences and ignores within-group variation.

For our present purposes, Paquot & Bestgen (2009:264) put it best in their conclusion: "Keywords of a specific corpus are lexical items that are evenly distributed across its component parts (corpus sections or texts) and display a higher frequency and a wider range than in a reference corpus of some kind." To consider such issues properly, in this paper, we are extending a proposal developed in Gries (2021), who makes two suggestions. First, rather than *replacing* frequency with dispersion, he proposes to *augment* frequency with dispersion, which changes keyness from a one-dimensional construct to a two-dimensional one with essentially an association-based and a dispersion-based component of keyness. Second, he proposes to use the Kullback-Leibler divergence (KL-divergence, D_{KL}) for the dispersion- and association-based components of keyness. This is because (i) D_{KL} is an extremely easy-to-compute directional measure of how one vector of proportions (the so-called posterior distribution) differs from another one (the so-called prior distribution), and because (ii) Gries (2022) shows that, of all dispersion measures compared there, D_{KL} is the one least correlated with frequency and, thus, most likely to make an original contribution above and beyond frequency. From these two characteristics, a third one follows, namely that this approach is the first to be able to distinguish

- words that are key for a target corpus only because they are relatively more frequent in the target corpus than in the reference corpus;
- words that are key for a target corpus only because they are relatively more evenly dispersed in the target corpus than in the reference corpus;
- words that are key for a target corpus because they are both relatively more frequent and more evenly dispersed in the target corpus than the reference corpus.

The next section outlines how we computed our keyness scores.

2.1 Our approach step 1: association-based keyness

The association-based keyness component is computed from the same kind of table as traditional keyness analyses have used, i.e. 2×2-tables like Table 2. Since D_{KL} is a directional measure, one can actually compute two D_{KL}-values on such tables: one that quantifies how much each word changes the proportional distribution of the two corpora, and one that quantifies how much the target corpus changes the proportional distribution of each word. We are using the former here, which is conceptually related to the question of how much better you can predict which journal you're looking at when you sampled a certain word; also, the D_{KL}-values for that direction of keyness are less correlated with frequency than the latter. Thus,

- the posterior distribution for each word is the proportional distribution of *w* across the corpora, i.e. it is the vector of $a/(a+b)$ and $b/(a+b)$;
- the overall prior distribution is the sizes of the target and the reference corpus as proportions, i.e. it is the vector of $(a+c)/(a+b+c+d)$ and $(b+d)/(a+b+c+d)$.

Thus, the KL-divergence for *w* there is computed like this:

$$D_{KL} = \left(\frac{a}{a+b} \times log_2\left(\frac{a}{a+b} \div \frac{a+c}{a+b+c+d}\right)\right) + \left(\frac{b}{a+b} \times log_2\left(\frac{b}{a+b} \div \frac{b+d}{a+b+c+d}\right)\right)$$

As readers might recognize, this is actually very similar to the formula of G^2, but while G^2 uses the actually observed frequencies, D_{KL} uses percentages, which is precisely why it is much less correlated with the actual frequencies of Table 2 (both the frequency of *w* in the target and the marginal totals) and, thus, a bit more informative in its own right. Here's an example showing how straightforward this is to compute in R (input is in regular font, output is in italics):

```
log2.0 <- function(values) { ifelse(values==0, 0, log2(values)) }
(input <- matrix(c(106, 117183, 26, 445704), ncol=2, dimnames=list(
   WORDS=c("college", "OTHER"), CORPUS=c("Clinton", "Trump"))))
         CORPUS
WORDS     Clinton  Trump
  college     106     26
  OTHER    117183 445704

(corpus.sizes.prop <- colSums(input)/sum(input))
  Clinton     Trump
0.2083216 0.7916784

(input.prop <- prop.table(input, 1)[1,])
  Clinton     Trump
0.8030303 0.1969697

sum(input.prop * log2.0(input.prop / corpus.sizes.prop)) # DKL
[1] 1.167906
```

These D_{KL}-values can then be normalized to the interval [0, 1] using some heuristic transformation like this one:

$$D_{KL\text{norm}} = 1\text{-exp}^{-DKL}$$

Of course, these computations are made for either each word occurring in the target corpus or for each word occurring in at least one of the two corpora. As a measure of deviation of one distribution from another, this value, like G^2, is by definition 0 or positive, but, as with G^2, we can turn it into a signed version of itself by multiplying it with -1 if the word is more key for the reference corpus than the target corpus.

2.2 Our approach step 2: dispersion-based keyness

The dispersion-based component of keyness in the present approach is similar but a bit more complex and proceeds in three steps. Step 1 is to compute for each word in each corpus its dispersion in each corpus; this addresses the degree to which words are homogeneously distributed through the corpora mentioned above. For each corpus,

- the posterior distribution is the proportional distribution of a word across all the parts of the corpus;
- the prior distribution is the corpus part sizes as proportions.

(If one has a term-corpus matrix – all words in the rows, the two corpora in the columns, and frequencies of occurrence in the cells – then the posterior distribution corresponds to the column percentages and the prior corresponds to the row sum percentages.) This will return for each word in each corpus a value between 0 (if the word is very evenly distributed across the corpus parts) and, theoretically at least, +∞ (if the word is extremely clumpily distributed in the corpus parts).

Step 2 is to again transform these values into a value range that is more useful for the current analysis, the range from 0 to 1, with 0 meaning ‘clumpy distribution’ and 1 meaning ‘even distribution’; we do this the same way as above.

Step 3 is to compute for each word the difference $D_{KL\text{norm}}$ in the target corpus minus $D_{KL\text{norm}}$ in the reference corpus such that

- if a word is attested in both corpora, we compute the above difference;
- if a word is attested in only one of the two corpora, we set its dispersion value in the other corpus to 0 and then compute the difference.

This will lead to

- positive values (in the interval (0, 1]) indicating dispersion-based keyness for the target corpus;
- negative values (in the interval [-1, 0)) indicating dispersion-based keyness for the reference corpus;
- values of 0 indicating no dispersion-based keyness for either corpus; and
- the more the value differs from 0, the stronger the keyness preference.

This, too, is done for either each word occurring in the target corpus or for each word occurring in at least one of the two corpora.

2.3 A single continuum (just for simplicity’s sake)

As a result of all this, each word comes with two pairs of two values:

- a frequency with which it occurs in the target corpus and a frequency with which it occurs in both corpora;
- a frequency-based keyness value and a dispersion-based keyness value (for both of which positive and negative values indicate keyness for the target corpus and the reference corpus respectively).

Ideally, an analyst would consider all four values, but this may not always be interesting and it is certainly not mathematically straightforward. In the interest of being able to sort all words by their keyness for the target corpus, one possibility is to try and amalgamate several of these dimensions into a single score. While any such amalgamation/conflation of course loses a lot of – too much? – information, for the purposes of this ‘playful’ paper, it will do. To get an amalgam score, we add

- the dispersion-based keyness score of each word and
- the product of the association-based keyness score and the min-max-transformed logged frequency of each word (i.e. we took all logged frequencies and normed them to the interval [0, 1] by subtracting from all logged frequencies the minimum value and dividing by the range).

This way, the association-based component of keyness is ‘weighted’ by the frequency such that association-based keyness has a higher impact on a word’s overall keyness when the word is more frequent and, for simplicity's sake, given the purposes of this editorial, we have a single amalgam score to sort all keywords by. All keywords computations were done in R.[1]

3. Data

3.1 Preparatory processing

In order to determine what is key for, or distinctively characteristic of, *CLLT* as compared to Corpora and IJCL, we downloaded all research articles – leaving out editorials, special issue introductions, reviews, and resource notes – of approximately 10 recent years from the journals' websites. For *CLLT*, we ended up with 221 pdfs, for *Corpora*, we obtained 137 pdfs, and for *IJCL*, we had 160 pdfs. From all article pdfs, we first extracted the text (using the R function `pdftools::pdf_text`), converted it to lower case, replaced all occurrences of one or more digit by "#", and tokenized them by splitting on anything that's not a letter or a hyphen ("[^-\\p{L}]+"). Then, we saved all the articles into one data table (`data.table::data.table`) with one row for each word token attested at least once in at least one journal, and three columns:

- WORD: the word token;
- PART: an ID for the article in which the word token was observed; this corresponds to the corpus parts we need for the dispersion calculations;
- JRNL: an ID for the journal in which the word token was observed: *CLLT* vs. *Corpora* vs. *IJCL*; this allows us to define the target and the reference corpus.

This data table was then split up into the target corpus (when JRNL was *CLLT*) and the reference corpus (when it was not), to which we then applied the keyness computations discussed in the previous section. The word types were then sorted by the above-described amalgamation score from most to least key for CLLT for annotation and interpretation.

3.2 Disambiguation, interpretation, and annotation

While this paper does not claim to be a full-fledged research paper, we nevertheless are using two steps that, to our knowledge, are new in keyness analyses.[2] Both of these have to do with the words returned as key because, generally speaking, a list of words sorted by keyness will

[1] Gries (to appear, esp. Section 5.4) provides an updated and more comprehensive discussion making keyness a 3-dimensional construct (comprising frequency, dispersion, and association). An R function `Keyness3D` to compute such 3-dimensional keyness scores and several amalgam scores is available from the second author upon request. `Keyness3D` requires that the *R* installation it runs on has the packages `data.table` and `Rcpp` installed; the former internally speeds up the processing of potentially large input data frames; the second is needed because the main computations in the function are 'outsourced' into C++ functions (for massive gains in speed).

[2] These two steps were proposed in the context of varieties research in Gries (under review).

likely only be the starting point of a more qualitative analysis or interpretation; as Cvrček & Fidler (2022:263) summarize,

> KW extraction is therefore normally followed by additional methods – for example, close reading of concordance lines (Egbert and Baker 2019:56), collocation analysis of KWs (e.g., Gabrielatos and Baker 2008), key multi-word expressions, *n*-grams or clusters (Partington and Morley 2004; Mahlberg 2007; and Fischer-Starcke 2009), or examining links between KWs (Scott and Tribble 2006: 73) within a larger span of words.

Our first step is mostly concerned with disambiguating the keywords in the context-free list that such analyses return. As in many keyness analyses, but maybe especially in a case like this where the target corpus consists of highly specialized academic expert writing, keywords returned by an analysis may be unclear or ambiguous. An analyst might simply not know what the keyword means; for instance many abbreviations – ones that turned out to be names of corpora (e.g., *GECO* for the German Conversations database) or linguistic terms (e.g., *centering*, which was not used in the statistical sense) or words like *nom* (nominal? nominative?) – were not immediately clear to the present authors; the same is true for many words that turned out to be names: Is *Stefanie* Stefanie Wulff or Stefanie Shattuck-Hufnagel? Is *Paul* Paul Hopper or Paul Rayson? Is *Manfred* Manfred Krug or Manfred Krifka? And *Walter* who? (DeGruyter, actually.)

To address this without having to search all articles and read potentially thousands of concordance lines, we adopted a distributional-semantics kind of approach using the *R* package wordVectors (Schmidt & Li 2015), an R implementation of the word2vec algorithm. Specifically, we trained a word2vec model on all *CLLT* articles, specifically a 300-dimensional skip-gram model with a context window size of 4, a frequency threshold of 3, and 35 training iterations. Then, we used that trained model to retrieve the top 14 most similar words (as measured by the cosine distance) in the articles for the top 3000 keywords (as per the amalgam score) and added them to the keyness results returned by the function. For any names or other words whose meaning was unclear in the context-free keywords list, we could then turn to their closest neighbors, which, following Gries (under review), we will refer to as *deep key collocates*, for context and disambiguation.

Using a combination of our general corpus-linguistic knowledge and the deep key collocates the word2vec approach provided for each keyword, we then heuristically annotated the top 1200 keywords in terms of semantic groups of interest such as

- author names (from papers themselves and of course reference sections);[3]
- corpus names (e.g., *PDTB* (Penn Discourse Treebank), *MAONZE* (Māori and New Zealand English), or *CallHome*, etc.);

[3] Unfortunately, the three journals differ in their policies regarding the use of first names: *CLLT* provides full first names whereas *Corpora* and *IJCL* unfortunately only provide initials, so while there are some last names that are key for *CLLT*, a sizable proportion of the interpretation below is based on first names and must therefore be understood as being less contrastive (*CLLT* vs. {*Corpora* and *IJCL*} and more '*CLLT*-internal' so to speak.

- names of languages (e.g., *Māori*, *Hebrew*, or *Mandarin*, etc.);
- field/discipline (mostly linguistic, but also others, e.g., *syntax*/*syntactic*, *SDRT* (Segmented Discourse Representation Theory), *MLF* (Matrix Language Frame Model), *psycholinguistic*, *semantics*, *comprehension*, *acquisition*, etc.);
- linguistic terminology (e.g., *dative*, *compounds*, *VPCs*, *gerunds*, *scrambling*, *animacy*, *left-branching*, *persistence*, etc.);
- names of locations (e.g., *Flanders*, *Barranquilla*, *Turku*, *Newfoundland*, etc.);
- methodological/statistical terminology (e.g., *logistic*, *regression*, *TOBI*, *model*, *acceptability*, *Mandelbrot*, *confederate*, *intercept*, *fitted*, *predictor*, *multimodel*, *Kolmogorov*, *z-standardization*, *Bayesian*, etc.);
- (general) scientific terminology (e.g., *theoretical*, *explanation*, *theory*, *inhibition*, *variabilities*, *mechanisms*, *null hypothesis*, etc.).

Note, the above classification is neither obvious nor watertight. Terms like *reanalysis* could be seen as linguistic terminology or as general scientific terminology (we went with the former), same with *comments* (here, we went with the latter), and *null hypothesis* can be classified more narrowly as methodological/statistical terminology but we classified it as a more general statistical term. Similarly, *type* was used in the 'type/token' sense but also in the statistical sense in the context of (hierarchical) configural frequency analysis. Since this paper (i) is offering no more than a heuristic comparison and (ii) is not advancing bolder claims, we feel that this is less of an issue here than it would be in a general research paper.

In addition to these semantic groups of interest, there were a few groups of non-interest, so to speak, i.e. words that belonged to identifiable groups but that we will not discuss (even though some of them scored stunningly high keyness values); these include

- everyday expressions (e.g., *many*, *provide*, *potentially*, *possibility*, *trick*, *his* and *her*, *neither*, *let*, *comes*, etc.);
- expressions related to journal sections or publisher-related information (e.g., *bionote(s)*, *euppublishing*, or *ISSN*) because *CLLT* has bionotes for authors, *Corpora* is published by Edinburgh University Press, and *IJCL* article pdfs contain the ISSN of the journal;
- expressions in a foreign language, which often were examples (e.g., *sja* (as part of Russian infinitives, *sitzen* and *geben* (German for 'sit' and 'give'), *miettiä* (Finnish for 'think'), *probovat'* (Russian for 'try'), or *sentirse* (Spanish for 'feel'), etc.).

Finally, there were a variety of unclassifiable cases, many of which possibly were abbreviations and foreign language expressions or names; examples include *nne*, *tqs*, *mlf*, *pdc*, *nep*, *loridp*, *gum*, *gao*, *mcc*. For a real research paper, we would of course be able to look into the original article but for the present, more light-hearted purpose, we did not do so.

The second step, which in this case study we're only applying to the linguistic terminology, has to with using the deep key collocates to identify groups in the keywords. To see what kinds of interpretable groups our keywords form, we did an extra step of retrieving word2vec collocates (50 this time) and their cosines for the top 100 linguistic-terminology keywords; then we applied a hierarchical clustering method (distance measure: Kendall's rank correlation, amalgamation rule: complete linkage) that clustered the keywords on the basis of the similarities of their deep key collocates.

4. Results

Let's have a look at a selection of top keywords for every 'semantic domain' we used. To save space, we use regular expressions to combine multiple straightforwardly related terms into one compact representation; most importantly for readers, the question mark means the preceding unit – a single character or a parenthesized group – is optional.

4.1 (Linguistic) Field or discipline

The keywords from this domain paint a fairly clear picture: *CLLT* differs from the other journals in terms of discipline with a strong prominence of the core traditional/structural disciplines of (roughly in order of keyness) (morpho)syntax, semantics, syntax-semantics and, a bit further down, phonology and phonetics (whereas other journals have many more discourse-analytic and sociolinguistic/-cultural keywords). In terms of other disciplinary terms, psycholinguistics (especially comprehension and acquisition) and psychology, experimental work (partly involving acceptability judgments on (un)grammaticality, and cognition were central.

4.2 Linguistic terminology

The key linguistic terminology, typically describing linguistic phenomena being studied, is compatible with the fields/disciplines discussed in the previous section. While a cluster analysis is of course only a heuristic tool, several of the clusters are nicely suggestive and fit what more qualitative inspection and browsing relevant articles reveal:

- there is one very strong alternation cluster consisting of expressions that refer to predictors in many classic alternation cases *animacy*, *(in)?animate*, *possessor*, *definiteness*, *recipient*, *givenness*, *pronominal(ity)?*, *PDC* (prepositional dative construction), and *PTC* (prepositional theme constructions);
- there is a cluster of case expressions, with also a hint of alternation research: *datives?*, *genitives?*, *nominative*, *partitive*, and *ioc* (indirect object construction) as well as *poc* (prepositional object construction);
- there is yet another cluster related to especially the dative alternation with the expressions *DOact* and *PDact* (for 'double object active' and 'prepositional dative active'), *semc* (for semantic class) and *cons* (tricky to assign, sometimes it is used as the name of a response variable meaning 'Construction', sometimes it is the word *cons* as in *pros and cons* when citing an alternation paper that has that in the title);
- there is a cluster with *(un)?grammaticality* and *(cogni|genera)tive*, which suggests a theoretical-comparison kind of group.

There is a larger number of smaller clusters that are less interesting, because they often occur in only a small number of papers. *VPCs* (for 'verb-particle constructions') is a case in point, which shows up in three papers, and other two-expression clusters are often just based on one paper. More serious exploration might pursue this further by looking at more than 100 linguistic-terminology keywords, but also by establishing a keyness dispersion threshold to make sure that clusters consist of more widely-used words. Still, fitting the results from the previous section, the explorations returns a maybe surprisingly strong emphasis of *CLLT* on cognitively, functionally/information-structurally informed syntactic alternation research.

4.3 Languages

We identified only a relatively small number of keywords in our top 1200 list that refer to languages, so we list them here in order of amalgam keyness: Māori, Hebrew, Mandarin, Shanghainese, Basque, Nepali, Malay, Guaraní, Danish, Estonian, Catalan, and Creoles as a general term. It is worth noting, however, that not much importance can be attached to those because all of these languages are key for *CLLT* virtually exclusively because of their association keyness – they are not widely dispersed at all, some are even underdispersed in *CLLT* and only score highly because of their high association (and, accordingly, frequency) in a very small number of articles (Hebrew, Danish, and Nepali).

4.4 Methodological terminology

The methodological terminology, at least as we annotated it, paints a relatively clear picture even without the use of any follow-up clustering. There is one very large group of expressions that are related to logistic regression modeling (in fact, the top two keywords in this domain are *regression* and *logistic*, which are also each other's nearest word2vec collocates):

- the frequentist regression 'group': *logistic*, *linear*, *regression*, *model(s|ed|ing)?*, *fit(ted|ting)?*, *predictors?*, *multimodel*, *predict(ions)?*, *estimate*, *intercept*, *effects*, *mle?*, *(multi)?collinearity*, *glmer/lme/mixed-effects*, d_{xy}, *Harrell*;
- expressions that might still be related to regression modeling more generally or would belong into the next, more general category: *z-standardiz(ation|ed)*, *unstandardized*, and *multifactorial*;
- other predictive modeling or generally quantitative techniques: *Akaike* (Information Criterion), *Bayesian, analogical* (*modeling*), *NDL*, *tree* (in the context of classification and regression trees), *discriminant* (analysis), *h?cfa*/*antitype*, and *behavioral profiling*;
- information-theoretical and 'distribution-related' terms: *Kullback-Leibler divergence*, *Kolmogorov*, and *Mandelbrot*;
- psycholinguistic predictors: log_{freq} and $log_{bigramfreq}$, $log_{forwardTP}$ and $log_{backwardTP}$ (*TP*: 'transition probability'), and *surprisal*;
- various terms relating to experimentation and corpus annotation: *confederates?* and *informant*, *completions*, *self-paced*, *stimuli*, *forced-choice*, and *labelers* as well as *intercoder*/*interrater* (followed by *reliability*).

4.5 Scientific terminology

This category was very heuristic and broad and many of the words in it are very general important academic words; thus, its keywords were nearly impossible to categorize or interpret in any way leading to insightful generalizations. One noteworthy finding, though, and one that jibes well with *CLLT*'s mission statement is that, after *references* (as the name of that section) and *comments* (used as parts of thanking the reviewers, see below), the next top two keywords are *explanation* and *theoretical*, and words from those word families show up a little further down the ranking as well (e.g., *explanations* and *theory*).

4.6 Authors/names

Finally, for names, here are some of the top-listed names and who the deep key collocates disambiguate them to be; we use a notation where the name in question is represented by an underscore when we identify the authors/names:

- William: _ *Labov*, but also _ *D. Raymond*, _ *Frawley*, _ *McGregor*, _ *Croft*, and _ *Pagliuca*;
- Michael: _ *A. Kirkwood*, _ *Hammond*, _ *Barlow*, _ *Tomasello*, and _ *Gradoville*;
- Robert: _ *Schreuder*, but also _ *Bayley* and *Peter _ Crosthwaite*;
- Manfred: _ *Stede*, _ *Krifka*, and _ *Krug*,
- Paul: _ *Rayson*, _ *Hopper*, and _ *Kerswill*;
- Stefan: _ *Gries*, but also _ *Engelberg*;
- Joan: _ *Bresnan* and _ *Ford*;
- Ronald: _ *Langacker* and _ *Carter*;
- Jennifer: _ *Hay* and _ *Arnold*,

Other key names/authors cited (listed here in alphabetical order) are Antti Arppe, Harald Baayen, Susan Conrad, Dagmar Divjak, Nick Ellis, Edward Finegan, Susanne Gahl, Adele Goldberg, Bernd Heine, Florian Jaeger, Geoffrey Leech, Daniel Jurafsky, Vsevolod Kapatsinski, Bernd Kortmann, Susanne Niemeier, Janet Pierrehumbert, Andrea Sand, Anatol Stefanowitsch, Sali Tagliamonte, and Graeme Trousdale.

With the usual caveats (small sample, differences in editorial policies, less concordancing-based or generally qualitative follow-up than a detailed study might make desirable and others) and with an acknowledgment of the risk of pigeon-holing, these names suggest a preponderance of psycholinguistic and cognitive-linguistic researchers as well as, less strongly, sociolinguistic and typological researchers.

4.7 Conclusion: the prototype

Let's sum up the above and additional observations we haven't discussed here, and obviously the following is in jest: The prototypical *CLLT* paper (when compared to *Corpora* and *IJCL*!) is a cognitively/psycholinguistically-informed and of course corpus-based study (likely using Mark Davies's COCA) of a morphosyntactic alternation using a multifactorial predictive modeling approach (typically binary logistic regression, often in the form of a mixed-effects model) and using *R* (especially the packages `lme4`, `MuMIn`, `fpc`, `Hmisc`, and `languageR`). The predictive modeling process involves semantic, syntactic, discourse-functional, and psycholinguistic predictors (especially ones involving logged frequencies/probabilities), and the paper as a whole cites Harald Baayen, Joan Bybee, Robert Schreuder, Geoffrey Leech, and the second author of this paper with an emphasis on research they publish with Mouton/Walter de Gruyter, Stanford CSLI, Chicago University Press, and John Benjamins – sounds like we can ask ChatGPT to write a submission to *CLLT* that will sail right through!

While the focus of our not-so-serious keyness analysis has focused on describing prototypical *CLLT* papers, *CLLT* has of course thrived as well as it has for not just its 'prototype papers'. We love to get submissions outside of this narrow 'prototype' and it is gratifying to see that we do receive and publish other and outstanding submissions, as is clear once we look not at a keyness-based prototype, but at the articles that have been cited most often in the past decade. Leading that list is currently "On the 'holistic' nature of formulaic language" by Anna Siyanova-Chanturia – definitely a cognitively and/or psycholinguistically informed piece, but otherwise a marked deviation from the prototype.

Last but certainly not least, *CLLT* must have the greatest reviewers, because, based on the deep key collocates of *comments*, the prototypical paper also thanks the *anonymous reviewers/referees* for the *meticulous*, *insightful*, *constructive*, *helpful*, *valuable* (as well as *invaluable*!) *feedback* and *suggestions* on one *earlier draft* or more *earlier drafts*.

5. Contributions to this special issue

How do leaders in the field of corpus linguistics evaluate the developments of the past 20 years, and how do they envision the future of corpus linguistics? We invited six (teams of) researchers to give us their take. Here is a brief summary of what their contributions focus on.

Martin Hilpert opens our special issue with his contribution "Corpus linguistics meets historical linguistics and Construction Grammar: How far have we come, and where do we go from here?" Martin discusses several examples of where the three areas in his title meet and form the basis of a dynamic research program, including qualitative approaches, diachronic collostructional analysis, multivariate techniques, distributional semantic models, and analyses of network structure.

Jesse Egbert, Douglas Biber, Daniel Keller, and Marianna Gracheva then turn our attention to a long-standing hot topic in corpus linguistics: the impact of register. Their contribution entitled "Register and the dual nature of functional correspondence: Accounting for text-linguistic variation between registers, within register, and without registers" walks the reader through key findings from the past 20 years regarding the role that register plays in linguistic variation. Turning to the present, they note that while the majority of research to date has considered register as a variable that predicts linguistic variation, it has to be borne in mind that there is extensive situational variation in registers themselves. In four case studies, the authors show that register and situational context are related, yet make independent contributions to accounting for linguistic variation.

Monika Bednarek, Martin Schweinberger, and Kelvin Lee's contribution "Corpus-based discourse analysis: From meta-reflection to accountability" critically assess the state of corpus-based discourse analysis. They first reflect on the state of corpus-based discourse analysis with regard to core methodological issues such as triangulation and replicability of research results, and advocate for including accountability as another metric of study quality. By accountability, the authors mean being transparent about methodological choices, justifying these choices, and critically reflecting on them. One tool that aids researchers in living up to these accountability standards are Jupyter notebooks: free open-source web applications that researchers can use to document and share text, code, and analytical output.

Magali Paquot's contribution "Learner corpus research: A critical appraisal and roadmap for contributing (more) to SLA research agendas" outlines some of the core issues that learner corpus research will have to tackle in the coming years, including a diversification of the types of learner corpora available, enriching corpora with meta-data, and moving beyond contrastive interlanguage analyses towards multifactorial study designs. Magali also speaks in defense of continuing to compare learner and native speaker production, especially in the context of research that examines cross-linguistic interference, that is, the impact that a learner's first language has on the acquisition of a second language.

Tony McEnery and Gavin Brookes take stock of the state of connection between "Corpus Linguistics and the Social Sciences". They argue that while corpus-linguistic approaches certainly have the potential to contribute to social science research, epistemological differences have been impeding a cross-fertilization of the two disciplines. One area where the two can intersect relates to data processing and theory, such as the development of annotation software. In closing, the authors urge corpus linguists to articulate what they can offer in multi-method research designs, as this will not only allow social scientists to see the merit of including corpus-linguistic methods in their tool box, it will also help corpus linguistics remain competitive in a world of varied "big data" methods that now compete for attention.

One of these "big data" methods are large language models (LLMs) like GPT. Harald Baayen's contribution bookends the special volume with his contribution "The wompom", which references a song about an imaginary creature with infinite powers. Are large language models (LLMs) a wompom? And how do such technological advances impact the future of corpus linguistics? Harald cautions us that in order to remain competitive, we have to be more ambitious and aim for integrated humanistic cognitive computational models that allow us to make quantitative predictions not only for single phenomena, but entire arrays of data from language acquisition, processing, and change. Importantly, we will have to find ways of running such computationally costly models in ways that do not leave a massive carbon footprint – an often overlooked property of current LLMs.

In summary, these contributions all paint a picture of the future of corpus linguistics that is challenging in various ways, but also full of opportunity. As editors, we stand ready to work hard to make sure that CLLT remains an outlet for critical and innovative work that demonstrates that value of theoretically grounded, methodologically savvy corpus research in better understanding and modeling human language. Ultimately, however, editors are only navigators – they take their orders from the captain(s), and those are CLLT's authors and readers. Stefan and I want to thank you for reading and contributing to *CLLT* for two decades. We hope that you will find this special issue insightful and inspiring. We are eager to see where you steer *CLLT* in the next twenty years.

References

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

Baron Alistair, Paul Rayson, & Dawn Archer. 2009. Word frequency and keyword statistics in historical corpus linguistics. *Anglistik: International Journal of English Studies* 20(1). 41-67.

Brezina, Vaclav, & Miriam Meyerhoff. 2014. Significant or random? A critical review of sociolinguistic generalisations based on large corpora. *International Journal of Corpus Linguistics* 19(1). 1-28.

Cvrček, Václav & Masako Fidler. 2022. No keyword is an island: in search of covert associations. *Corpora* 17(2). 259-290.

Egbert, Jesse & Douglas Biber. 2019. Incorporating text dispersion into keyword analyses. *Corpora* 14(1). 77-104.

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

Gabrielatos, Costas. 2018. Keyness analysis: Nature, metrics and techniques. In Charlotte Taylor & Anna Marchi (eds.), *Corpus approaches to discourse: a critical review*, 225-258. London & New York: Routledge.

Gries, Stefan Th. 2021. A new approach to (key) keywords analysis: using frequency, and now also dispersion. *Research in Corpus Linguistics* 9(2). 1-33.

Gries, Stefan Th. 2022. What do (some of) our association measures measure (most)? Association? *Journal of Second Language Studies* 5(2). 171-205.

Gries, Stefan Th. to appear. *Frequency, dispersion, association, and keyness: revising and tupleizing corpus-linguistic measures*. Amsterdam & Philadelphia: John Benjamins.

Gries, Stefan Th. under review. Cultural keywords in varieties research. *World Englishes*.

Kilgarriff, Adam. 1997. Using word frequency lists to measure corpus homogeneity and similarity between corpora. *Proceedings 5th ACL workshop on very large corpora*, 231-245.

Millar, Neil & Brian S. Budgell. 2008. The language of public health – a corpus-based analysis. *Journal of Public Health* 16(5). 369-374.

Paquot, Magali & Yves Bestgen. 2009. Distinctive words in academic writing: A comparison of three statistical tests for keyword extraction. In Andreas Jucker, Daniel Schreier, & Marianne Hundt (eds.), *Corpora: Pragmatics and discourse*, 247-269. Amsterdam: Rodopi.

Rayson, Paul, Damon Berridge, & Brian J. Francis. (2004). Extending the Cochran rule for the comparison of word frequencies between corpora. In Gérald Purnelle, Cédrick Fairon, & Anne Dister (eds.), *Le poids des mots: Proceedings of the 7th International Conference on Statistical analysis of textual data*, Vol. II, 926–936. Louvain-la-Neuve: Presses Universitaires de Louvain.

Rayson, Paul & Amanda Potts. 2020. Analysing keyword lists. In Magali Paquot & Stefan Th. Gries (eds.), *Practical handbook of corpus linguistics*, 119-139. Berlin & New York: Springer.

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

Scott, Mike & Christopher Tribble. 2006. *Textual patterns: Key words and corpus analysis in language education*. Amsterdam & Philadelphia: John Benjamins.

Schmidt Ben & Jian Li. 2015. wordVectors: Tools for creating and analyzing vector-space models of texts. R package, <https://github.com/bmschmidt/wordVectors>.

Stubbs, Michael. 1996. *Text and corpus analysis: computer-assisted studies of language and culture*. Oxford: Blackwell.