Problematising characteristicness

A biomedical association case study

Sheryl Prentice[1], Jo Knight[1], Paul Rayson[1], Mahmoud El Haj[1], and Nathan Rutherford[2]

[1]Lancaster University | [2]Royal Holloway University of London

Keyness is a commonly used method in corpus linguistics and is assumed to identify key items that are characteristic of 1 corpus when compared to another. This paper puts this assumption to the test by comparing case study corpora in the fields of genetic, immunological and psychiatric biomedical association studies, using what we refer to as a 'K-FLUX' analysis to produce a set of key items. Experts from within these fields are asked to evaluate the extent to which identified key items are characteristic of their discipline. The paper concludes that less than 50% of the items identified by the method are rated as highly characteristic by experts and that this ranges between types of association study. Further, there is difficulty in reaching a consensus over what is deemed to be 'characteristic', thus posing a challenge to the ultimate aim of the keyness method. The paper demonstrates the value of supporting corpus linguistic studies with expert assessments to evaluate whether (and which) items can be said to be indicative of a particular field.

Keywords: key items, keyness, characteristic, evaluation, biomedical

1. Introduction

Keyness is a method used within corpus linguistics to identify items that are statistically significantly overused or underused in 1 corpus when compared to another (Rayson, 2008). These items (commonly referred to as 'keywords', though we shall be using Wilson's (2013) preferred term of 'key item' throughout), are often used in applied keyness studies "to characterize the genre or text under consideration" and are said to be indicative of its "aboutness" (Pojanapunya & Todd, 2018). However, while much has been written on the statistics typically used within keyness analysis (see Gabrielatos, 2018), what of the output of this method, i.e. the key items themselves, and the extent to which they can truly be said to characterise a particular genre or text? It is this question that the current paper seeks to answer.

To begin, what does it mean to define an item as 'characteristic'? Within corpus linguistic studies applying the keyness method for the purpose of discovering genric or textual aboutness, characteristicness appears to be understood in terms of relative frequency and/or consistency. However, does this accord with how individuals in other disciplines understand characteristicness? The answer to this question has implications for the applicability of the approach to texts from a variety of disciplines. To illustrate this point, this paper will first use an adapted version of keyness to identify a set of candidate characteristic items based on frequency and consistency by comparing academic paper abstracts from the fields of genetic, immunological and psychiatric biomedical association studies. The research team's biomedical expert will initially be consulted before a wider evaluation exercise is conducted with a group of biomedical experts to determine whether the candidate items are 'characteristic' according to their respective understandings.

To this end, the paper begins with a review of selected corpus linguistic studies of aboutness, which have sought to characterise the language of specific genres or texts. The paper then moves to a description of the corpora constructed to pursue comparisons across and between genetic, immunological and psychiatric forms of biomedical association enquiry. An overview of a linguistic resource created to assist the analysis of texts within the biomedical domain is then given, before moving to the employment of an adapted version of keyness to identify characteristic items in the biomedical association literature. An evaluation study then follows, in which experts in genetic, immunological and psychiatric biomedical association are consulted for their opinions on what constitutes a characteristic item in their respective fields, and whether the keyness method is successful in extracting such items. The paper concludes with a discussion of the evaluative findings and their implications for corpus studies of aboutness.

2. Using keyness to determine characteristicness

The keyness method has come under some scrutiny in recent years. Criticisms include the need for researchers to acknowledge methodological decisions, such as subjectivity introduced into the keyness analysis process via (i) the setting of frequency, effect size and statistical significance thresholds, (ii) the selection of linguistic units for analysis and (iii) comparison corpus attributes, as well as the need to consider the appropriacy and limitations of particular metrics (Gabrielatos, 2018, p. 26), how the use of differing metrics impacts on output (Gabrielatos & Marchi, 2012) and suitability for particular purposes (Pojanapunya & Todd, 2018), and that the approach demonstrates partiality in its focus on difference rather than also considering similarity (Taylor, 2013; 2018). Scott (2010: 52) discusses the limitations of claims that can be made on the basis of using the approach and argues that its output (i.e. key items) is bound by context, influenced by the size of the context considered (e.g. part or whole texts) and affected by the reference corpus chosen, and therefore that one cannot state that a given key item list is definitive. Further, Scott (2010: 51–52) points out that machine processing of a text is not the same as a human's ability to understand and distinguish particular textual nuances, and therefore the keyness method cannot as readily identify related forms (such as anaphoric and cataphoric reference, synonyms and antonyms) unless taught to do so. Even when taught, machine processing will still be prone to a degree of error.

However, while varying limitations of the keyness method have been considered, that it can determine characteristic items within a text or genre, is very much taken as given. As Conway (2009: 23) observes, using key items to characterise texts is a central method within corpus linguistic studies of literature and genre difference in particular. Scott & Tribble (2006: 60), for example, use the keyness method to compare Shakespeare's *Romeo and Juliet* with a corpus of all of Shakespeare's plays and find that "there are some KWs reflecting important themes which really characterize what the play is about", which include the key items *love*, *lips*, *light*, *night*, *banished*, *death*, *poison*, while KWs such as pronouns (*thou*) and exclamations (*O*), are taken as markers of style.

The output from such studies demonstrates that key items fall into 2 categories: those which characterise aboutness and those which characterise style (Bondi, 2010). Scott (2001) has also noted that proper names tend to feature heavily in key item lists. Style relates to the "communicative purpose" of the text, while aboutness refers to a text's "conceptual structures" (Bondi, 2010: 7). It is these conceptual structures that we are particularly interested in in the present paper. Phillips (1989: 1–2) explains that aboutness concerns "the reader's ability to state what the text is about independently of his or her ability to recall the actual wording of the text", that "large scale patterns of textual organization contribute to this ability", and that textual organisation is "the kind of patterning which is characteristic of text and lexis" (Phillips, 1989: 3–4). Phillips views automated text analysis as a means to objectively identify these characteristic patterns.

Corpus linguists employing the keyness method take a similar view, as Gabrielatos (2018: 225) explains: "the notion of keyness is closely related to the notion of aboutness, that is, the understanding of the main concepts, topics or attitudes discussed in a text or corpus". This raises the notion of the reader referred to in Phillips' (1989) work. Presumably, if the reader is a machine and that machine is capable of detecting patterns used by humans in their understanding of what a text is about, then both human and machine readers should pick out the same major underlying topics, themes and concepts. In other words, both types of reader should essentially state that a given set of texts are about similar things.

However, we know that there are differences between machine and human judgements. For example, Alderson (2007) has demonstrated differences between human and machine judgements of word frequencies. Of course, as Scott (2010: 46) points out, it would be impossible to have precise agreement between human and machine readers, because humans themselves "do not consistently agree on the key words of a given text". Nevertheless, there should at least be a broad consensus in order for us to support the notion that keyness "certainly does point to fundamental elements in describing specialised discourse or in placing a text in a specific domain" (Bondi, 2010: 3). This raises the question of what these fundamental elements are and, linked to this, who decides what they are? The corpus linguist, or the domain specialist?

Other researchers have already questioned the suitability of keyness in deriving characteristic patterns of aboutness when compared to alternative methods. Cheng (2007; 2009) has highlighted that the use of keywords as a unit of analysis misses a great deal of meaningful content and prefers the use of phraseological units. Meanwhile, Conway (2009) has observed that a frequency analysis performs better than a keyness analysis in characterising a set of biographical texts. There are, of course, other approaches to establishing textual aboutness, including computational linguistic approaches to automated content summarisation (see Nenkova & McKeown, 2012 for a review of methods) and social tagging (Kehoe & Gee, 2011). However, keyness is the focus of the current investigation.

Scott & Tribble (2006) state that keyness performs better in determining aboutness as texts become more domain specific. They look at 5 academic genres, including humanities, natural science, medicine, politics, law and education, and technical and engineering, and find that the keyness method works better for medical texts because the vocabulary (such as, *clinical*, *patients*, *disease* and *diagnosis*) is more specialised (Scott & Tribble, 2006: 82–83). This links to Scott's (2001) observation that proper names tend to be among one's key items. However, these assessments of performance are based on the corpus linguists' judgements and not the judgements of experts working within the disciplines being studied.

This review has highlighted inherent assumptions made in keyness-based corpus linguistic studies of aboutness: the first is that frequency-based approaches can determine characteristic lexis within a particular text or genre, and the second is that characteristicness is therefore defined by relative frequency of occurrence. To address the validity of these inherent assumptions, this paper seeks to answer the following research questions (RQs), using the biomedical genre of association studies as a test case:

1. Do items extracted with the keyness method accord with expert judgements of what characterises language within their discipline?
2. How is characteristicness conceived by biomedical experts and how does this compare with corpus linguists' frequency-based assumptions?
3. Finally, what implications do the answers to these questions have for the suitability of using keyness to determine characteristic lexis within the variety of genres in which keyness is deployed?

As the research questions demonstrate, the position taken throughout the paper is that analyst assessments based on computational measures should align with expert assessments based on in-depth knowledge of a given field, in order to strengthen or support claims that a set of items represent characteristic lexis in said field.

3. Data

To answer the research questions, a test set of data from the field of biomedical association studies was sourced. This field was selected due to its prevalence of discipline specific vocabulary, which, given Scott & Tribble's (2006) observation that keyness performs better at characterising content when the subject-matter is more domain-specific, should provide optimum conditions for the method. The test set consists of 4 comparable biomedical association study corpora with 500 academic abstracts each from genetic association studies in immunology, genetic association studies in psychiatry, non-genetic association studies in immunology, and non-genetic association studies in psychiatry.

Data were selected in this manner to allow for comparisons between (i) genetic and non-genetic association studies, (ii) immunological and psychiatric association studies, and (iii) psychiatric and immunological association studies. These comparisons were conducted in order to evaluate the performance of the keyness method in identifying items that can be said to characterise 3 specific genres: (i) genetic biomedical association studies, (ii) immunological biomedical association studies and (iii) psychiatric biomedical association studies. Further details on why corpora were compared to one another rather than a general reference corpus can be found in Section 5 of this paper.

Genetic association literature in psychiatry is far less prolific than the other forms of association study considered here. Therefore, data collection began with the genetic psychiatry association literature, using the following search terms in the PubMed (NCBI, 2018) interface:

> ((((((Humans[MeSH Terms]) AND (assoc*[Title/Abstract]) AND (psychi*[Title/Abstract]) AND (geneti*[Title/Abstract] OR gene[Title/Abstract] OR genot*[Title/Abstract]) NOT (review[Publication Type] OR immunol*[Title/Abstract] OR immunog*[Title/Abstract] OR immune[Title/Abstract]))))))

The search was designed to draw out papers on human subjects (as opposed to animals) that contained within their title or abstract the term 'associate' or its variants, as well as 'genetics', 'gene', or 'genotype' or their variants. The search was further designed to exclude review papers (given that we were concerned with original research). In addition, the terms 'immunology', 'immune' and their variants were excluded to avoid the inclusion of papers collected for the immune corpora. Queries were formulated with the assistance of the team's biomedical association study expert (who has a background in genetics). As shown in Table 1, the search resulted in a total of 4,082 papers (as of 28th November 2018). The results were sorted by publication date and the first 500 most recent results were selected and their abstracts downloaded.

As illustrated by the search hit values presented in Table 1, the number of non-genetic psychiatric association studies, non-genetic immunological association studies and genetic immunological association studies greatly out-number those published within the field of genetic psychiatry, meaning that their most recent paper abstracts will cover a shorter time scale. Therefore, the psychiatric genetic association study corpus described above was used as the basis for the design of the remaining corpora. This was to ensure that all data were drawn from the same time frame, thus reducing the potential for time-based differences in language use.

The year and month of each abstract in the psychiatric genetic association corpus was ascertained using the downloaded metadata. The corpus spanned around 2.5 years, with the most recent abstract published on 30th December 2018, and the oldest abstract published on 1st July 2016. A full breakdown of the number of texts per month/year in this corpus is provided in Appendix A. The following queries were used to search for the remaining corpora, which were collected in the same manner as the genetic psychiatry corpus:

> (((((Humans[MeSH Terms]) AND (assoc*[Title/Abstract]) AND (psychi*[Title/Abstract]) NOT (review[Publication Type] OR geneti*[Title/Abstract] OR gene[Title/Abstract] OR

genot*[Title/Abstract] OR immunol*[Title/Abstract] OR immunog*[Title/Abstract] OR immune[Title/Abstract]))))) [PsychGeneral search terms]

(((((Humans[MeSH Terms]) AND (assoc*[Title/Abstract]) AND (immunol*[Title/Abstract] OR immunog*[Title/Abstract] OR immune[Title/Abstract]) NOT (review[Publication Type] OR geneti*[Title/Abstract] OR gene[Title/Abstract] OR genot*[Title/Abstract] OR psychi*[Title/Abstract]))))) [ImmGeneral search terms]

(((((Humans[MeSH Terms]) AND (assoc*[Title/Abstract]) AND (immunol*[Title/Abstract] OR immunog*[Title/Abstract] OR immune[Title/Abstract]) AND (geneti*[Title/Abstract] OR gene[Title/Abstract] OR genot*[Title/Abstract]) NOT (review[Publication Type] OR psychi*[Title/Abstract]))))) [ImmGenetic search terms]

The results for each search were filtered to include only those papers published between 1st July 2016 and 30th December 2018, in order to match the time frame of the psychiatric genetic association corpus. The number of search hits produced before and after this filtering process are presented in Table 1.

Once again, results were ordered by publication date (most to least recent). Using the papers' metadata, texts were selected at random to match the date design of the genetic psychiatry corpus provided in Appendix A. The 500 randomly selected abstracts for each data type were downloaded from PubMed with their meta data in .xml format. Abstract texts were stripped from this format using resources described in El-Haj et al. (2018). Henceforth, the 4 collected corpora will be referred to as the PsychGenetic corpus, the ImmGenetic corpus, the PsychNonGen corpus, and the ImmNonGen corpus (see Table 1 for word counts).

Table 1. Metadata on the biomedical association study corpora

	PsychGenetic	**PsychNonGen**	**ImmGenetic**	**ImmNonGen**
No. search hits	4,082	34,859	25,143	85,788
No. search hits after date restriction	n/a	3,854	3,154	8,835
No. sample texts	500	500	500	500
No. words	85,934	59,934	93,932	77,731

4. Words, lemmas and word families

Considering the data at a word level would result in items that one might logically combine (such as *polymorphism* and *polymorphisms*) being considered as separate units. Therefore, this paper works with a predefined lemma list. In a field such as genetics, the likelihood of such a lemma list missing a large number of terms contained in one's corpus is high, due to the number of specialised terms that will not be found in general lemma lexicons. Automated lemmatisation was not used as a solution in this case, in order to avoid erroneously combined entries.

In addition, lemmas are part-of-speech specific, and therefore the singular and plural versions of the noun *association*, for example, are considered separately from the verb *associate* and its variant forms. However, one might conceivably refer to there being *an association between X gene and Y trait*, or that *gene X is associated with trait Y*. Both statements express the same meaning. Therefore, a way of grouping terms with a similar meaning, regardless of part-of-speech, was required. In other words, the project required something akin to what Bauer & Nation (1993) describe as 'word families'.

To this end, the corpus linguistic software tool *WordSmith Tools* (Scott, 2019) was used to generate and join frequency lists for each of the 4 corpora. The total frequency of each term across all of the corpora was established using this method, and the list was then ordered from high to low overall term frequency. This list was then trimmed by ordering the list first by the number of texts a term appeared in, and then by frequency of occurrence. Terms had to appear in at least 3 texts from any of the corpora to qualify for inclusion in the word family list.

The corpus linguistic software tool *AntConc* (Anthony, 2018) was used to produce a list of lemmas for each of the terms. Specifically, the Someya Lemma List (no hyphens)[1] was used for this purpose. This provided an initial pass over the data.[2] Items in the lemma list were grouped manually into non-genetic and genetic word families with the assistance of the team's biomedical association expert. For example, the term *antigen* formed the head word of a group containing the related terms *antigens*, *antigenic*, and *antigenicity*. Terms had to have the same senses in the association literature in order to be combined into a group. Combining terms in a manual fashion avoided the aforementioned problems with the automated combination of terms and allowed for human driven assessments of which terms should be considered collectively within the association study context.

The resulting word family list was then employed in the keyness comparison described in the section that follows to arrive at sets of characteristic items. The word family list is available as a project resource for fellow researchers conducting health-related research.[3] The list is suitable for use in both *AntConc* and *WordSmith Tools*.

5. Generating key items for evaluation

Key item (family) fluctuation analysis (henceforth K-FLUX analysis) is an adaption to the standard keyness approach, which allows a user to see which corpora a word or word family is salient in, and those in which it is not. This resembles Scott's (1997) key keyword approach in that it is designed to search for areas of consistency across corpora, but also simultaneously allows one to identify areas of inconsistency between corpora. It should be noted that 1 disadvantage of this approach when compared to a key keyword approach is that it works with overall corpus frequencies, unlike Scott's approach, which takes account of individual text differences. The precise nature of the approach's similarities to and differences from the key keyword method will be outlined in Section 5.1. K-FLUX analysis is suitable for use with 3 or more corpora.

The K-FLUX approach was utilised here in order to compare all 4 corpora of sample abstracts (i.e. ImmGenetic, ImmNonGen, PsychGenetic and PsychNonGen) and establish a set of words that were common to the genetics association study corpus pair (i.e. PsychGenetic and ImmGenetic), the psychiatry association study corpus pair (PsychGenetic and PsychNonGen), and the immunology association study corpus pair (ImmGenetic and ImmNonGen), in comparison to the remaining corpora. As with the applied keyness studies discussed in Section 2 of this paper, we are assuming that these sets of words will be characteristic of their respective genres, an assumption that will be put to the test in the evaluation featured in Section 6.

The corpora were compared to one another rather than to a general reference corpus. Had a general reference corpus been used, it is likely that most of the lexis contained within the genetic, immune and psychiatric abstracts would be labelled as characteristic. This is because much of the lexis is highly specialist and therefore unlikely to be found in general language use. By comparing corpora in a similar field, one gains a clearer idea of characteristic and uncharacteristic items *within* the association research domain. In addition, the psychiatric and immunological non-genetic corpora act as reference corpora for the psychiatric and immune genetic corpora, in that the latter are a particularly specialist subset of the former (e.g. genetic immunological association studies are a form of immunological association study).

5.1 Procedure

Word family lists were produced for each corpus before joining (i.e. placing alongside one another) the 4 frequency lists using *WordSmith Tools'* (Scott, 2015) consistency analysis option, which itemised each word family and its frequency in each of the corpora. The resulting list was subsequently imported into UCREL's multi-corpus comparison spreadsheet (Rayson, 2016) in order to calculate log-likelihood and effect size metrics for word family items.

Word family items were ordered according to their log-likelihood value (high to low). Log-likelihood is a measure of how likely, in this case, a word family is to occur in 1 corpus or multiple corpora relative to their comparison corpora and is calculated here as: $LL = 2\times((a\times\ln(a/E1)) + (b\times\ln(b/E2)) + (c\times\ln(c/E3)) + (d\times\ln(d/E4)))$, where a, b, c and d equal the observed frequency of a word family in each corpus, and E1, E2, E3 and E4 equal the expected frequency of a word family in each corpus (see Rayson, 2008 for a more detailed description). However, while this method indicates which word families differ across the 4 corpora, it does not indicate which corpus or corpora is responsible for the observed difference, hence the development of the K-FLUX approach. Following this approach, if a corpus' observed frequency of a word family was higher than its expected frequency, this was recorded as an instance of overuse. If a corpus' observed frequency was lower than its expected frequency, this was recorded as an instance of underuse.

In this way, the K-FLUX approach differs from the key keyword approach, in which each text within a study corpus is compared with a reference corpus in order to generate a series of keyword lists. Items occurring across keyword lists are said to mark consistent differences between the study and reference corpus, whilst indicating consistencies across study corpus texts (see Scott, 1997, for more information on the key keyword method and its implementation). Therefore, whilst both the K-FLUX and key keyword method look to identify areas of consistency, the key keyword method does so by comparing multiple study texts to a single reference corpus, while the K-FLUX approach does so by comparing multiple study corpora to multiple reference corpora.

Word families marked as being overused in the PsychGenetic and ImmGenetic corpora should indicate items that are linguistically characteristic of genetic association literature. Similarly, word families marked as being overused in the ImmGenetic and ImmNonGen corpora should suggest items that are characteristic of immunological association literature, while those overused in the PsychGenetic and PsychNonGen corpora should highlight items that are characteristic of psychiatric association literature. All such items are anticipated to be those that will also be deemed to be characteristic to biomedical experts working within the fields of genetic, immunological and/or psychiatric association.

With this in mind, the log-likelihood list was first sorted (in descending order) according to each word family's approximated Bayes Factor. Approximated Bayes is a measure that indicates the degree of evidence against the null hypothesis (Wilson, 2013). Following Wilson (2013: 5–6), who utilised earlier work by Kass & Raftery (1995), Bayes Factors were approximated using Bayesian Information Criterion (BIC) scores with the formula BIC ≈ LL − (df x log(N)), where LL is the log-likelihood of a word family, df is the degrees of freedom (in this case, 5), and N is the sum of comparison corpus word totals. According to Wilson (2013), an approximated Bayes of 10 or above indicates very strong evidence against the null hypothesis, a score of 6–10 indicates strong evidence, a score of 2–6 indicates positive evidence, and a score of 0–2 is negligible evidence against the null hypothesis. The opposite is true of negative approximated Bayes, i.e. the higher the negative approximated Bayes, the stronger the evidence for the null hypothesis (at the same magnitude as positive approximated Bayes).

Word families with positive approximated Bayes were first separated from word families with negative approximated Bayes. The positive approximated Bayes items were colour-coded according to 3 categories: (i) items overused in both psychiatric association corpora (PsychGenetic and PsychNonGen), (ii) items overused in both immunological association corpora (ImmGenetic and ImmNonGen), and (iii) items overused in both genetic association corpora (ImmGenetic and PsychGenetic). The aim of this process was to generate 3 lists of characteristic items (genetic association, psychiatric association and immunological association).

5.2 Results

The items in the left column of Table 2 are the top 20 word families that are linguistically characteristic of the non-genetic and genetic psychiatric association corpora (namely, PsychNonGen and PsychGenetic), but differ from the immune association corpora (ImmNonGen and ImmGenetic). The items in the centre column are the top 20 word families that are overused in the immune association corpora and underused in the psychiatric association corpora. Finally, the items to the right of the table are the top 20 word families overused in the genetic association corpora (PsychGenetic and ImmGenetic) and underused in the non-genetic association corpora (PsychNonGen and ImmNonGen). Word families are listed with their approximated Bayes values. Theoretically, items listed in Table 2 will be items that are characteristic of the corpus pairs. The method results in the identification of 344 characteristic items. The results presented in Table 2 tell us, from a corpus linguistic perspective, which items are most characteristic of the language used in recent psychiatric, immunological and genetic biomedical association studies, respectively.

Table 2. Top 20 key item families in the psychiatric corpora compared to the immune corpora (left), the immune corpora compared to the psychiatric corpora (centre), and the genetic corpora compared to the non-genetic corpora (right)*

Word	**Approx. Bayes**	**Word**	**Approx. Bayes**	**Word**	**Approx. Bayes**
PSYCHIATRIC	1874.23	CELL	1778.01	GENE	3705.45
DISORDER	1836.87	IMMUNE	1537.01	POLYMORPHISM	550.54
SCHIZOPHRENIA	1010.56	INFECT	1050.33	VARIANT	540.46
DEPRESSION	930.59	T	609.01	SNP	365.69
SYMPTOM	538.90	VIRUS	571.68	ALLELE	195.11
SUICIDE	452.94	INFLAMMATION	556.68	LOCUS	169.99
ANXIETY	413.05	IL	465.08	MUTATION	101.82
BEHAVIOR	357.15	TUMOR	460.71	NUCLEOTIDE	79.75
COGNITIVE	316.34	RESPONSE	432.89	SEQUENCE	65.89
PSYCHOSIS	310.66	CYTOKINE	275.12	ANALYSIS	65.46
BRAIN	302.59	B	260.50	IDENTIFY	65.37
SOCIAL	298.90	CD4	234.93	WE	64.91
BIPOLAR	297.25	ACTIVATE	223.49	DNA	62.39
COMORBID	278.18	PROTEIN	220.89	MIR	50.38
MDD	244.32	ANTIGEN	213.35	HAPLOTYPE	39.83
RISK	185.46	AUTOIMMUNE	210.09	REGION	29.36
ADOLESCENT	184.15	IMMUNOGLOBULIN	205.98	EPIGENETIC	25.48
TRAUMA	178.50	HIV	197.91	ENRICH	22.64
ALCOHOL	177.49	CANCER	195.16	SINGLE	18.14
EMOTION	162.58	HUMAN	193.94	NETWORK	14.40

* Terms are capitalised to represent word families.

6. Evaluation studies

This section details 2 studies conducted to evaluate the ability of the keyness method to identify items that are characteristic of genetic, immunological and psychiatric association studies, as judged by experts in the fields. The first is a pilot study, in which the team’s biomedical association expert is asked to rate the characteristicness of items generated by the keyness approach. The second is a wider evaluative study, in which 15 different biomedical experts working in genetic association, immunological association and/or psychiatric association are consulted.

6.1 Study 1: Pilot study

The aim of the pilot study is to provide initial insights into the effectiveness of the keyness method in identifying characteristic items within a specific domain and to generate some preliminary answers to the paper's research questions that can be explored as part of a wider evaluative exercise.

6.1.1 *Procedure*

Following the generation of lists of key item families detailed in Section 5, our team's biomedical association expert was asked to look at the 3 top 20 word family lists given in Table 2 (genetics items, psychiatry items and immunology items) and to state whether or not the items were characteristic of genetic, psychiatric or immunological association studies, respectively. This was done with the aim of addressing RQ1. They were further asked to provide their reasoning for rating particular word families as characteristic or not characteristic of the disciplines in order to gain an understanding of their definition of characteristicness, in answer to RQ2.

Items rated as uncharacteristic were explored via collocation and concordance analyses to investigate potential reasons for their occurrence. Measures of collocation are given as mutual information (MI) values. MI is a measure of the strength of co-occurrence between 1 word (or word family in this case) and another word. Items with a MI value of 3 or more were considered. Note that some methodological procedures will be covered in further detail in the discussion section that follows to allow for exemplification of the analytical process.

6.1.2 *Results and discussion*

In terms of the psychiatry word families listed to the left of Table 2, the team's biomedical association expert provides some positive confirmation as to the method pulling out items that they would rate as characteristic. These include the names of specific conditions, such as *anxiety*, *bipolar*, *depression*, *MDD* (Major Depressive Disorder) and *schizophrenia*, subject specific items such as *psychiatric* and *psychosis*, traits such as *behaviour* and *emotion*, causal factors such as *alcohol* and *trauma*, brain-based items such as *brain* and *cognitive*, and outcomes including *suicide*, and *adolescent* (the typical age of onset for psychiatric disorders).

However, the list also contains items that our biomedical association expert does not rate as characteristic, for example, the word families *comorbid*, *symptom* and *risk*, which they state are not specific to psychiatry. Nevertheless, similar findings in relation to *risk* have been made by Saber (2012: 53), who observes that the term is salient in the introduction sections of psychiatry papers and typically features in units such as *the risk of*, *increased risk of*, and *risk factor for*. This paper bears out Saber's observations on a word family level, with popular immediate collocates of the word family including *increased* (MI = 5.92) and *high* (MI = 5.48). Similarly, the *symptom* word family collocates with the terms *severity* (MI = 7.83) and *severe* (MI = 5.69). Concordance examples of the *risk* and *symptom* word families are given in Examples (1), (2) and (3) below.

(1) Following adjustment for comorbid psychiatric disorders, women with PCOS were still at a *significantly* <u>increased</u> **risk** for bulimia, schizophrenia, bipolar disorder, depressive and anxiety disorders, personality disorders [PsychNonGen, 406.txt]

(2) Different factors might be related to the *very* <u>severe</u> trajectories of emotional **symptoms** and peer relationship problems [PsychNonGen, 53.txt]

(3) In 22q11DS, chronically poor PAS trajectories and poor childhood and early adolescence academic domain and total PAS scores *significantly* <u>increased</u> the **risk** of prodromal **symptoms** or overt psychosis [PsychGenetic, 315.txt]

The collocates *severe* and *increased* enhance the risks and symptoms described in Examples (1)–(3). These are further amplified by the use of adverbs such as *significantly* and *very*. The tendency to present findings in this manner may have implications for how the general public perceive individuals with psychiatric disorders when compared to immunological disorders, when such research is translated by the media, given that *risk* can imply danger (see Hamilton et al., 2007 for a more detailed discussion of the term *risk* in genetic discourse).

The majority of items extracted from the immune association corpora (ImmNonGen and ImmGenetic) when compared to the psychiatric association corpora (centre of Table 2, Section 5.2) are rated as characteristic by the team's biomedical association expert. In their judgement, these contain well-known immunological word families such as *cell*, *immune*, *autoimmune*, *antigen*, *response* (as in immune or host response) and *activate* (as in the activation of a particular cell or response), the names of particular diseases or disorders (e.g. *cancer*, *tumor*, *hiv*), terms relating to immune responses to viral infections (i.e. *infect* and *virus*), and more specialist terms, such as *immunoglobulin*, *IL* (interleukin), *CD4* (a protein found on immune cells), and *cytokine*, *B* and *T* (referring to types of cells).

However, the team's biomedical association expert judges the term *human* to be uncharacteristic, on the grounds that it is not an immune specific term. Looking at concordance examples of the *human* word family – see Examples (4) and (5) below – it would appear that this occurs due to a need to specify that the findings relate to human rather than animal immunology.

(4) In doing so, these viruses have developed profound mechanisms that mesh closely with our **human** biology [ImmGeneral, 358.txt]

(5) Correlative studies from checkpoint inhibitor trials have indicated that better understanding of **human** leukocytic trafficking into the **human** tumor microenvironment can expedite the translation of future immune-oncologic agents [ImmGenetic, 333.txt]

The lack of such specifications in psychiatry literature compared to immune literature may be due to us having a capacity to do laboratory research on human cells involved in the immune system, but not so much of an ability to work with human brain cells.

The method further pulls out items that the team's biomedical association expert would describe as being characteristic of genetic association literature (left of Table 2, Section 5.2), and which they would expect to differ from non-genetic association literature, including the word families *gene*, *polymorphism*, *allele* and *haplotype*. The *variant* word family refers to gene variants, the *mutation* word family to gene mutations, and the *sequence* word family to *DNA* sequences. The list also contains more specialist items such as *MIR* (micro-RNAs) and *SNP* (single nucleotide polymorphism – indeed, the majority of the *single* word family refer to this unit). However, the items *analysis*, *identify*, and *we* are rated as uncharacteristic by the team's biomedical expert, as these are deemed to be general rather than domain-specific items. Nevertheless, previous studies have looked at the use of *we*, for example, in English biomedical journal articles (Williams, 2012). Examples (6) and (7) present instances of *we* in the PsychGenetic and ImmGenetic corpora.

(6) **we** investigated genetic variants affecting cytokine production in response to ex\xA0vivo stimulation in 2 independent cohorts of 500 and 200 healthy individuals [ImmGenetic, 399.txt]

(7) **We** believe that the continued development of mouse mapping populations, genetic tools, bioinformatics resources, and statistical methodologies should remain a parallel strategy by which

> to investigate the genetic and environmental underpinnings of psychiatric disorders and other diseases in humans [PsychGenetic, 250.txt]

Among the most frequent collocates of *we* are *study* (116 occurrences, MI = 4.52), *found* (102 occurrences, MI = 5.86), *investigated* (83 occurrences, MI = 6.51), *identified* (80 occurrences, MI = 5.32), and *performed* (59 occurrences, MI = 6.13). Interestingly, *believe* is 1 of the strongest collocates of *we* in the genetic association corpora (MI = 7.46). This is what Plappert (2017) would describe as encoding a claim in genetics discourse. These active rather than passive constructions suggest that, in this form of biomedical discourse at least, or in its most recent studies, there is an expression of ownership of actions and ideas.

In sum, the K-FLUX method brings a number of items to the fore that the team's biomedical association expert would rate as characteristic of a given comparison. Through this evaluation process, however, it is revealed that the team's expert is determining characteristicness as a function of how familiar an item is to them and via an item's subject specificity. This immediately raises a challenge for the keyness approach. Familiarity is a subjective phenomenon that could be influenced by a number of external factors, such as one's level of knowledge and experience in a particular field. This is something that the keyness method, with its objective measurements, cannot control for.

What, then, of specificity? As revealed in the analysis of Examples (1)–(7), there is a discord between linguistic and biomedical interpretations of specificity, which results in a misalignment of terms judged to be characteristic within each academic discipline. A keyness-based approach would typically view an item as being specific to a particular language variety if its usage is marked in comparison to another language variety. Hence, items discussed in Examples (1)–(7) would be characteristic under this interpretation. However, the team's biomedical expert has a more subject-specific view of specificity, which results in some of the K-FLUX items being rated as not characteristic.

In terms of answering RQs 1 and 2, these preliminary findings would suggest that the items the method pulls out accord, to some extent, with expert judgements (particularly with regard to immunology items), but not exclusively so, that alternative definitions of what constitutes characteristicness exist, and that linguistic and biomedical perspectives on characteristicness do not necessarily accord with one another. Do such observations hold if we subject the lists to a wider range of judgements? Study 2 will explore this in some detail.

6.2 Study 2: Wider evaluative study

The pilot study observations in Section 6.1 are based on subjective judgements formed on the basis of consultation with 1 biomedical expert using clipped lists of items. This section reports on a wider evaluative study, which was conducted in order to: (i) more objectively establish the extent to which the keyness method can identify characteristic items in the domain of biomedical association studies; (ii) explore the potential range of definitions of characteristicness that exist and whether a consensus can be reached; and (iii) assess whether corpus linguistic and biomedical perspectives on characteristicness are complementary.

The pilot study informed the design of the wider evaluative study in 3 ways: (i) in the pilot study, the expert's judgements of characteristic items within biomedical association studies were restricted to those they were presented with. This raised the question of what an expert might judge to be characteristic if they were asked to generate items on their own volition. Therefore, the wider study was designed to look at both expert generated items (and whether the K-FLUX approach captured these) *and* expert opinions of computationally generated items. To prevent experts' generation of items being influenced by K-FLUX generated items, experts were asked to provide their items first. (ii) The pilot only asked the expert whether or not they deemed an item to be characteristic and not the degree to which they viewed the item as characteristic. On occasion, the expert found items difficult to categorise in absolute terms. Therefore, the wider study introduced a grading system for items. (iii) While the pilot expert provided their reasons for labelling an item as characteristic or not, the study highlighted the need to gain a more concrete understanding of the rationale behind this decision-making process and to formalise the criteria being used. Therefore, the wider evaluative study was designed to elicit written responses on experts' reasoning, which could be manually coded. These 3 design criteria are reflected in the descriptions of Evaluation Tasks 1 to 3 in the section that follows.

6.2.1 *Procedure*

For the evaluation exercise, 15 participants were sourced from the professional network of the team's biomedical expert, who were selected on the basis of their subject-specific knowledge: 5 with a working knowledge of genetic association; 5 with a working knowledge of immunological association; and 5 with a working knowledge of psychiatric association. Participants ranged in age, gender, ethnic background, country of origin and career stage. Permission was sought from participants, who received a full description of the task (shown in Appendix B). The evaluation task took the form of 3 stages:

i. Evaluation Task 1: Those who agreed to take part were first sent an email pertaining to stage 1 (see Appendix B), in which each expert was asked for 3 items that they would describe as characteristic of their assigned literature type (genetic association, immunological association or psychiatric association).

ii. Evaluation Task 2: Once a response to the first stage was received, participants were then sent an email pertaining to stages 2 and 3 (see Appendix B). In stage 2, experts were shown a combined, randomised list of key item families generated by the K-FLUX method for their assigned literature type (e.g. those with a working knowledge of immunological association were shown the immunological key item family list). Participants were asked to rate whether or not each of the terms present on their list was, in their opinion, a) “highly characteristic”, b) “somewhat characteristic”, or c) “uncharacteristic” of their assigned literature type.

iii. Evaluation Task 3: This task was linked to participants’ task 2 responses. The experts were asked for the subjective criteria they were using to ascertain whether or not an item was characteristic of their assigned literature type.

Term	**Response**	**Response**	**Response**	**Response**	**Response**	**HC**	**SC**	**NC**
CD4	Highly Characteristic	Highly Characteristic	Highly Characteristic	Highly Characteristic	Highly Characteristic	5	0	0
INFILTRATE	Somewhat Characteristic	Somewhat Characteristic	Somewhat Characteristic	Somewhat Characteristic	Somewhat Characteristic	0	5	0
HUMAN	Not Characteristic	Not Characteristic	Not Characteristic	Not Characteristic	Not Characteristic	0	0	5
INFLAMMATION	Highly Characteristic	Highly Characteristic	Highly Characteristic	Highly Characteristic	Highly Characteristic	5	0	0
TUMOR	Not Characteristic	Highly Characteristic	Somewhat Characteristic	Somewhat Characteristic	Not Characteristic	1	2	2

Figure 1. Example Evaluation Task 1 output

The results of Evaluation Task 1 were cross-referenced with the word families output from the K-FLUX approach and a count was conducted to establish how many of the approaches’ key item families matched experts’ a-priori judgements. Inter-rater reliability was calculated for suggested genetic association, immunological association, and psychiatric association terms, respectively, using Fleiss’ Kappa (Fleiss, 1971), which is described below. This was to establish whether there was agreement on suggested characteristic items. For Evaluation Task 2, participants' lists of responses were placed alongside each other, as illustrated in Figure 1.

Using a COUNTIF function, "Highly characteristic" (HC), "Somewhat characteristic" (SC) and "Not characteristic" (NC) responses for each item were summed. Summed responses were then used to calculate levels of inter-rater agreement for each item as follows, where HCT, SCT and NCT are the total of highly, somewhat and not characteristic observations for the item, respectively, and N delineates the number of raters (in this case 5): =((HCT$_1$^2+SCT$_1$^2+NCT$_1$^2)-N)/(N*N-1).

The overall proportion of inter-rater agreement for each of the HC, SC and NC categories was then calculated as CT/(CN*N), where CT is the category total, CN is the number of categories, and N is the number of raters. This process was repeated for each of the 3 sets of items (genetic association, immunological association, and psychiatric association) offered by the K-FLUX approach. Levels of inter-rater agreement were recorded, with a score of 0 indicating no agreement and a score of 1 indicating complete agreement. Counts and percentages were derived for items with scores above and below a threshold of 0.5. Items scoring above the threshold in the "Highly Characteristic" category were marked as characteristic.

Fleiss' Kappa (*k*) values (Fleiss, 1971) were then calculated to establish levels of inter-rater reliability for K-FLUX genetic association, immune association and psychiatric association items, separately. Fleiss' Kappa values were calculated as follows, where $\overline{P}$ is the average of inter-rater agreement values and $\overline{P}_e$ is HCA^2+SCA^2+NCA^2, with HCA, SCA and NCA being the proportions of highly, somewhat and not characteristic agreement, respectively:

$$\kappa = \frac{\overline{P} - \overline{P}_e}{1 - \overline{P}_e}$$

For Evaluation Task 1, a similar process was followed for participants' suggested genetic association, psychiatric association and immunological association items, in turn. However, the categories in this case were "Yes" (the item was suggested) or "No" (the item was not suggested). Fleiss' Kappa values of < 0 are said to indicate poor to no overall agreement, while scores of > 0 indicate varying levels of overall agreement: "slight" (0.01–0.20), "fair" (0.21–0.40), "moderate" (0.41–0.60), "good" (0.61–0.80), "near perfect to perfect" (0.81–1.0) (Landis & Koch, 1977).

To process the results of Evaluation Task 3, which asked participants to explain the criteria used in their task 2 responses, each written response was qualitatively examined for its central criteria, for example:

> When trying to classify terms as highly, somewhat, and not characteristic, I tried to think of how often these terms appear in the psych literature (i) I read, and my impression of how specific these terms are to psychiatry (ii)

In this example, 2 criteria were manually identified in the participant's response: (i) frequency (how often the participant encounters a term), and (ii) specificity (how specific the term is to the participant's field of expertise). Descriptions were formed for each criterion identified at its first mention. If a subsequent participant spoke to the same description, this was recorded as an additional instance of the criterion. An overall count was conducted of how many participants referred to each of the observed criterion. Table 3 provides a full list of criteria found.

Table 3. Criteria used by participants to discern the characteristicness of candidate terms

Criteria	Description	Example
Consistency	The term features across a range of literature within the field	"how common these terms are across the spectrum of genetics research"
Familiarity	The term has been previously encountered by the participant	"Not characteristic…if I was unfamiliar with it"
Frequency	The term is repeatedly encountered by the participant in relevant literature and/or in other academic settings (e.g. conferences)	"how often I see the word in genetics articles" "how often we use these words in everyday language in a research setting"
Importance	The term denotes a core concept within a given field of specialism	"the relative importance of the term to the theme"
Interest	The level of interest associated with the term within a given field of specialism	"Put addiction and substance disorders as not much, as that is my impression of [genetic] psychiatry – and its interests"
Recency	The term features in recent advancements in a given field of specialism	"recent methods/tools that use these terms in literature"
Specificity	The term is specific to a given field of enquiry, it is not vague, and will not/rarely be found in other forms of literature	"I ranked words that were specific to immune literature as highly characteristic"

6.2.2 *Results*

Evaluation Task 1: The number of K-FLUX approach key item families matching with the 15 participants' suggested characteristic items is as follows: genetic association 5 items, immunological association 11 items, psychiatric association 4 items. As each of the 5 participants for each literature type supplied 3 items, the results listed for the genetic association, immunological association and psychiatric association categories are out of a possible 15 items, respectively. Within category inter-rater reliability scores are as follows: genetic association items (*k* = -0.21), immunological association items (*k* = -0.21) and psychiatric association items (*k* = -0.17), showing no agreement between raters.

Table 4. Evaluation Task 2 – Displaying proportion of inter-rater agreement and Fleiss' Kappa (*k*) values for key item families

	"Highly characteristic" item agreement	"Somewhat characteristic" item agreement	"Not characteristic" item agreement	*k*
K-FLUX Genetic Items	**0.54**	0.23	0.23	0.23
K-FLUX Immune Items	0.38	0.21	0.41	**0.48**
K-FLUX Psychiatry Items	0.39	0.36	0.25	0.15

Table 4 shows the proportion of inter-rater agreement and inter-rater reliability scores observed for the K-FLUX approach's key item families in each of the 3 fields of biomedical association study (genetics, immunology and psychiatry) on the highly to not characteristic scale.

Table 5. Evaluation Task 2 – Participant judgements of K-FLUX key item families

Judgement	**Genetics**	**Immunology**	**Psychiatry**	**Total**
Highly characteristic	12 (46.15%)	28 (30.77%)	16 (19.75%)	56 (28.28%)
Somewhat characteristic	1 (3.85%)	6 (6.59%)	6 (7.41%)	13 (6.57%)
Not characteristic	2 (7.69%)	33 (36.26%)	6 (7.41%)	41 (20.71%)
Terms above threshold	15 (57.69%)	67 (73.63%)	28 (34.57%)	110 (55.56%)
Terms below threshold	11 (42.31%)	24 (26.37%)	53 (65.43%)	88 (44.44%)
Total terms	**26 (100.00%)**	**91 (100.00%)**	**81 (100.00%)**	**198 (100.00%)**

Table 5 presents a summary of Evaluation Task 2 results. All items achieving a proportion of inter-rater agreement of 0.5 or above are included in the "highly", "somewhat" and "not characteristic" categories displayed. Together, these represent the total *Terms above threshold*. All items providing a proportion of inter-rater agreement of less than 0.5 are shown in the total *Terms below* *threshold*. The table is to be read vertically. Percentages shown under each literature type are out of the number of word families generated by the K-FLUX method for that literature type.

With regard to Evaluation Task 3, the number of participants referring to each of the identified criteria used to determine levels of characteristicness is as follows: consistency (2), familiarity (2), frequency (6), importance (2), interest (1), recency (1), specificity (6). Please note that only 11 of the 15 participants responded to task 3. Therefore, participant numbers are out of 11, not 15.

6.2.3 *Discussion*

The results from Evaluation Task 1 show that in 2 of the 3 domains (genetic association and psychiatric association), there is little overlap between experts' suggested characteristic items and K-FLUX key item families. A greater degree of overlap (twice that of the genetic association and psychiatric association domains) can be observed between experts' suggested immune association items and K-FLUX immune association word family items, suggesting that a keyness approach performs better in some domains than in others. However, it would appear that our method of looking for consistency across corpus pairs may have presented an obstacle to the effectiveness of the keyness method in this task. The method did identify additional characteristic items suggested by participants in 1 corpus, but not across a corpus pair. These items include *antibody*, *heritable*, *polygenic*, *PRS* and *signalling*.

A further obstacle has been presented by not considering multi-word expressions (MWEs), in that the K-FLUX method identifies items contained in many suggested MWEs, such as *health*, *expression*, *assessment*, *association*, *disorders*, *variants/variation* and *discovery*. However, a method such as key collocates would need to be employed to ascertain whether these items occur in the suggested characteristic phrases of *mental health*, *gene expression*, *risk assessment*, *genetic association*, *mood disorders*, *genetic variants/variation*, and *drug discovery*. Indeed, researchers such as Cheng (2007, 2009) have highlighted the usefulness of considering phraseological units rather than keywords, as it is via a word's associations that its meaning(s) within a discipline is/are formed. An additional methodological limitation is that the suggested item *treatment* is found across the non-genetic association corpora, which were not considered within the focus of the present analysis.

Despite these methodological obstacles, the inter-rater reliability results show no agreement between raters, suggesting that participants tend to suggest different characteristic items from one another. Therefore, even if the K-FLUX method had matched more items, one cannot reliably say that any of those suggested items would be characteristic. It is also worth considering that had more participants been included, other suggested characteristic items are likely to have emerged, particularly given the lack of agreement demonstrated in this task. Therefore, this task alone should not be the yardstick against which success or otherwise of the approach is judged.

Evaluation Task 2 asked participants to rate the K-FLUX key item families according to the degree to which they perceived the items to be characteristic of a given field of association study (genetic association, immunological association or psychiatric association). Proportions of inter-rater agreement on key item families, regardless of field, are low, as are overall kappa values. While the proportion of inter-rater agreement reaches 0.54 on genetic key item families judged to be "highly characteristic", the inter-rater reliability of genetic key item family ratings is low. Equally, while the inter-rater reliability of immunology key item family ratings reaches moderate levels (k = 0.48), the proportion of inter-rater agreement on "highly characteristic" immunology key item families is low. These results indicate a lack of consensus.

Considering Evaluation Task 2 results in terms of raw numbers (see Table 5), most of the word families generated by the K-FLUX method come from the domain of immunology. While the majority of these key item families score above the inter-rater agreement threshold (73.63%), only around a third of the 91 key item families within this domain are judged to be "highly characteristic". Just over another third of the key item families are judged to be "not characteristic".

The picture is slightly better for genetic key item families. While the K-FLUX method identifies fewer of these, more than half score above the inter-rater agreement threshold and those that score above this threshold are generally rated as "highly characteristic". The outcome is quite different for psychiatric key item families. Of the 81 key item families generated, only 34.57% score above the inter-rater agreement threshold. While those scoring above the threshold tend to be rated as "highly characteristic", as a proportion of the total 81 key item families generated by the K-FLUX method within psychiatry, only around 20% are rated as "highly characteristic" of this genre.

Finally, the results of Evaluation Task 3 suggest that, from the perspectives of an albeit limited number of experts, the most prevalent criteria for judging the characteristicness of words are frequency and specificity, which are both mentioned by 6 of the 11 responding participants. The participants' criterion of frequency overlaps with the interpretation of characteristicness assumed within the keyness method, while the criterion of specificity overlaps with the judgements of the team's biomedical expert in Study 1. The frequency criterion demonstrates a degree of compatibility between corpus linguistic and biomedical definitions of characteristicness.

However, as stated earlier in this paper, specificity presents a challenge to this compatibility, in that, biomedical experts have a more rigid view of what specificity entails. Within corpus linguistics, a word that can move around genres (such as *we* or *human*) can be specific to a given comparison. However, within biomedical association studies, words are only specific if they are relatively fixed to a particular genre. This observation, and the range of definitions of characteristicness being used by a limited number of participants, may explain the limited success of the K-FLUX method in identifying items that participants can agree are characteristic of biomedical association genres.

In Study 1, it was provisionally found that the key item families generated by the K-FLUX method do not always accord with an expert's judgement, that alternative definitions of characteristicness appear to exist, and that a corpus linguist's and biomedical expert's perspectives on characteristicness do not particularly align. Study 2 in some ways corroborates and in others expands upon Study 1's findings. In comparison with Study 1, in Study 2 it becomes more apparent that the word families pulled out by the K-FLUX method do not generally match with what biomedical experts would rate as characteristic of their discipline or sub-discipline. However, neither do raters judgements generally line up with one another. There are a range of definitions of characteristicness, and while a limited consensus can be reached on 1 or 2 criteria, these may be differently applied. Finally, there is some, but little cross-over between how corpus linguists employing a keyness method and biomedical experts perceive characteristicness.

7. General discussion and conclusion

This paper set out to assess the ability of the keyness method to identify items that are evaluated as characteristic within a given domain. This was tested by employing an adapted version of the keyness method (K-FLUX analysis) to produce 3 sets of key item families (genetic association, psychiatric association, and immunological association) in order to see whether the word families would be rated as characteristic of 3 different, but interrelated forms of the biomedical association study genre by experts working within the disciplines.

This was done via 2 evaluative studies: the first was a qualitative pilot evaluation study, in which an expert in biomedical association studies was asked to assist with an in-depth evaluation of how characteristic the key item families produced by the K-FLUX method were to 3 forms of biomedical association study. In the second, a wider group of experts in biomedical association studies were consulted in a more quantitative study to rate the characteristicness of the key item families to the 3 separate forms of association study considered in this paper: genetic association studies, psychiatric association studies, and immunological association studies. Those consulted were also asked to provide their reasoning for rating word families as characteristic.

RQ1 asked whether the words pulled out by the keyness method accord with expert judgements. The pilot study suggested that this was largely the case, particularly with reference to immunological association studies, with a handful of exceptions. However, the wider evaluative study suggested that less than 50% of the words extracted via the keyness method were viewed as "highly characteristic" for each form of biomedical association study. The method fared better with genetic association words (reaching nearly 50%), followed by immunological association words (around 30%), and finally psychiatric association words (where only around 20% of words were rated as "highly characteristic").

RQ2 asked how characteristicness is conceived by biomedical experts. Both the pilot and wider evaluation studies brought to light a range of definitions of characteristicness being used by those consulted, including frequency, specificity, familiarity, consistency, importance, interest and recency, with frequency and specificity being amongst the most popular. With regard to consensus, none could be reached, as evidenced by the lack of inter-rater reliability. Linked to this, RQ2 further enquired as to how corpus linguistic and biomedical expert opinions of characteristicness compare. Criteria such as frequency would appear to do so. However, some of the criteria used by biomedical experts are of a subjective nature, which the keyness method cannot measure. Others are differently applied by corpus linguists and biomedical practitioners, such as specificity, which appears to mean something different to the different academic disciplines.

RQ3 asked whether the answers to RQs 1 and 2 have implications for the suitability of the keyness method in determining characteristic items within a particular domain. This is the method's primary purpose within applied keyness studies of aboutness, where corpus statistics are used to identify reliable linguistic patterns, which are then interpreted as characterising a particular study corpus or corpora. As far as biomedical association studies are concerned, the keyness method does not appear well suited to this task. However, before concluding, it is worth pointing out that the statistical measures used in this paper may have influenced the outcome. As Pojanapunya and Todd (2018) have observed, log-likelihood is more likely to bring out general items, while use of a measure such as an odds-ratio can provide one with the means of accessing "disciplinary technical terms". Perhaps had such a measure been used in this case, participants might have rated the output items as more characteristic of their discipline.

Nevertheless, the paper has highlighted that the method's apparent lack of suitability may be largely due to the differing definitions of characteristicness that exist both within and between academic disciplines. This is an important observation that requires further investigation, given its implications for corpus research that seeks to characterise corpus content, because it suggests that observations made in such research may not hold when subjected to a range of expert judgements. While it may not be possible to entirely address this problem, corpus linguists should consider adopting a version of the evaluative process outlined in this paper in order to better establish the reliability of their linguistic findings and to strengthen claims that identified patterns are characteristic of their target material.

In sum, the paper takes the stance that "true" characteristicness lies at the confluence of analyst and expert opinion, where a computational assessment is supported or corroborated by a range of expert assessments. This is not to argue that experts' introspective judgements are of more value than those of a corpus analyst. It is rather to acknowledge that differences between analysts' and experts' opinions exist and that such differences should be made transparent in the analytical process, particularly in cases where findings are to be applied outside the discipline of corpus linguistics.

In concrete terms, the paper makes the following recommendations: (i) that corpus linguists should introduce subject experts into the analysis process, (ii) that analysts should subject their key item lists to a series of expert judgements, (iii) that the human measure of inter-rater agreement should be introduced as a further filter in the key item sorting process, with items rated as highly characteristic and scoring 0.5 or above included in the analysis (with the potential to extend this to somewhat characteristic items), (iv) that inter-rater reliability rates on key item lists should be reported, and (v) that items scoring highest across both computational and human measures should be regarded as the most characteristic. Therefore, where computational measures and human measures agree, this would indicate varying levels of characteristicness. Where these measures disagree, this would indicate insufficient evidence to support the characteristicness of an item within a given comparison.

Acknowledgements

The authors would like to thank the Wellcome Trust (project reference: 204475/Z/16/Z) for providing the funding which made this collaboration possible.

Notes

1. The Someya Lemma List was sourced from http://www.laurenceanthony.net/software/antconc/

2. Note that this process is also possible in *WordSmith Tools*. The reason *AntConc* was used in this case is due to display preferences. If a lemma list is long, *WordSmith Tools* will display a sample of the lemmas. *AntConc* displays all lemmas, which can subsequently be captured for the establishment of word family groupings.

3. The word family list can be found at: https://github.com/drelhaj/BioTextMining. This is an adapted version of Laurence Anthony's Someya Lemma List (no hyphens), originally created by Yasumasa Someya. See link in note 1. Use of the word family list should also cite the Someya Lemma List (no hyphens) on which it is based.

References

Alderson, C. (2007). Judging the frequency of English words. *Applied Linguistics*, *28*(3), 383–409. https://doi.org/10.1093/applin/amm024

Anthony, L. (2018). *AntConc* (Version 3.5.7) [Computer software]. Waseda University. Available from http://www.laurenceanthony.net/software/antconc

Bauer, L., & Nation, P. (1993). Word families. *International Journal of Lexicography*, *6*(4), 253-279.

Bondi, M. (2010). Perspectives on keywords and keyness: An introduction. In M. Bondi & M. Scott (Eds.), *Keyness in Texts* (pp. 1–20). John Benjamins.

Cheng, W. (2007). Concgramming: A corpus-driven approach to learning the phraseology of discipline-specific texts. *CORELL: Computer Resources for Language Learning*, *1*, 22–35.

Cheng, W. (2009). Income/interest/net: Using internal criteria to determine the aboutness of a text. In K. Aijmer (Ed.), *Corpora and Language Teaching* (pp. 157–177). John Benjamins.

Conway, M. (2010). Mining a corpus of biographical texts using keywords. *Literary and Linguistic Computing*, *25*(1), 23–35.

El-Haj, M., Rayson, P., Piao, S., & Knight, J. (2018). Profiling medical journal articles using a gene ontology semantic tagger. In N. Calzolari et al. (Eds.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)* (pp. 4593–4597). European Language Resources Association (ELRA). https://www.aclweb.org/anthology/volumes/L18-1/

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*(5), 378–382.

Gabrielatos, C. (2018). Keyness analysis: Nature, metrics and techniques. In C. Taylor & A. Marchi (Eds.), *Corpus Approaches to Discourse: A Critical Review* (pp. 225–258). Routledge.

Gabrielatos, C., & Marchi, A. (2012, September 13–14). *Keyness: Appropriate metrics and practical issues* [Paper presentation]. Corpus-Assisted Discourse Studies International Conference, Bologna, Italy.

Hamilton, C., Adolphs, S., & Nerlich, B. (2007). The meanings of 'risk': A view from corpus linguistics. *Discourse & Society*, *18*(2), 163–181.

Kass, R.E., & Raftery, A.E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795.

Kehoe, A., & Gee, M. (2011). Social Tagging: A new perspective on textual "aboutness". *Studies in Variation, Contacts and Change in English*, *6*(5).

Landis, J. R. & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174.

NCBI. (2018). *PubMed*. National Center for Biotechnology Information, U.S. National Library of Medicine. Bethesda MD, USA. Available from: https://www.ncbi.nlm.nih.gov/pubmed/

Nenkova A., & McKeown K. (2012). A survey of text summarization techniques. In C. Aggarwal & C. Zhai (Eds.), *Mining Text Data* (pp. 43–76). Springer.

Phillips, M. (1989). *Lexical Structure of Text. Discourse Analysis Monographs: 12*. English Language Research, University of Birmingham.

Plappert, G. (2017). Candidate knowledge? Exploring epistemic claims in scientific writing: A corpus-driven approach. *Corpora*, *12*(3), 425–457.

Pojanapunya, P., & Todd, R. W. (2018). Log-likelihood and odds ratio: Keyness statistics for different purposes of keyword analysis. *Corpus Linguistics and Linguistic Theory*, *14*(1), 133–167.

Rayson, P. (2008). From keywords to key semantic domains. *International Journal of Corpus Linguistics*, *13*(4), 519–549.

Rayson, P. (2016). Log-likelihood and effect size calculator [Excel spreadsheet]. http://ucrel.lancs.ac.uk/llwizard.html

Saber, A. (2012). Phraseological patterns in a large corpus of biomedical articles. In A. Boulton, S. Carter-Thomas, & E. Rowley-Jolivet (Eds.), *Corpus-informed Research and Learning in ESP: Issues and Applications* (pp. 45–82). John Benjamins.

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

Scott, M. (2001). Comparing corpora and identifying key words, collocations, and frequency distributions through the WordSmith Tools suite of computer programs. In M. Ghadessy, A. Henry, & R. L. Roseberry (Eds.), *Small Corpus Studies and ELT* (pp. 47–67). John Benjamins.

Scott, M. (2010). Problems in investigating keywords, or clearing the undergrowth and marking out trails… In Bondi, M., & Scott, M. (Eds.), *Keyness in Texts* (pp. 43–58). John Benjamins.

Scott, M. (2015). *WordSmith Tools Manual: Consistency analysis*. Available from: https://lexically.net/downloads/version6/HTML/index.html?compare_versions.htm

Scott, M. (2019). *WordSmith Tools* (Version 7) [Computer software]. Lexical Analysis Software.

Scott, M., & Tribble, C. (2006). *Textual Patterns: Key Words and Corpus Analysis in Language Education*. John Benjamins.

Taylor, C. (2013). Searching for similarity using corpus-assisted discourse studies. *Corpora*, *8*(1), 81–113.

Taylor, C. (2018). Similarity. In C. Taylor, C., & A. Marchi (Eds.), *Corpus Approaches to Discourse: A Critical Review* (pp. 19–37). Routledge.

Williams, I. A. (2012). Self-reference in biomedical research article discussions: Further evidence for cross-cultural diversity in academic and scientific discourse. *International Journal of Corpus Linguistics*, *17*(4), 546–583.

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* (pp. 3–11). Peter Lang.

Appendix A. Structure of Corpora

Date	No. Texts	Yearly sub-totals
Dec 18	1	
Nov 18	1	
Oct 18	1	
Sep 18	1	
Aug 18	3	
Jul 18	4	
Jun 18	5	
May 18	7	
Apr 18	10	
Mar 18	9	
Feb 18	7	
Jan 18	23	72
Dec 17	27	
Nov 17	21	
Oct 17	21	
Sep 17	22	
Aug 17	21	
Jul 17	20	
Jun 17	20	
May 17	21	
Apr 17	20	
Mar 17	33	
Feb 17	28	
Jan 17	33	287
Dec 16	30	
Nov 16	18	

Appendix A. Structure of Corpora

Date	No. Texts	Yearly sub-totals
Oct 16	29	
Sep 16	24	
Aug 16	28	
Jul 16	12	141
TOTAL		**500**

Appendix B. Participant Instructions

B1. EVALUATION TASK 1

In this brief task, you will be consulted on your knowledge of terms featured in [INSERT SPECIALISM] literature.

This task has 2 stages. In this first stage, we would like to ask you:

If you had to pick 3 terms that you would say are most characteristic of [INSERT SPECIALISM] literature, what would they be? (please list)

In the second stage, you will be shown a list of terms and asked to rate (from a drop-down menu of choices) how characteristic of [INSERT SPECIALISM] literature you would say the term is. This list will consist of [X] items. Finally, you will be asked to briefly state the criteria you used to arrive at your decisions.

Thank you very much for your time.

B2. EVALUATION TASKS 2 AND 3

Please find attached a list of terms in an Excel document. In the ‘Response column, click on the cell beside each term. You will see an arrow. Click on the arrow to see a list of 3 options: ‘Highly Characteristic′, ‘Somewhat Characteristic and ‘Not Characteristic′. Select the response that most closely corresponds with your opinion of how characteristic the term is of [INSERT SPECIALISM] literature.

When you have completed your responses, please save the changes before answering the following and final question:

Briefly, what criteria did you use in deciding whether to label terms as highly, somewhat or not characteristic?

Please return your Excel sheet and your answer to the above question to [CONTACT]

Once again, we would like to thank you for your participation.

Address for correspondence

Sheryl Prentice
Department of Psychology
Lancaster University
Fylde College
Lancaster, LA1 4YF
UK

s.r.prentice1@lancaster.ac.uk

Co-author information

Joanne Knight
Lancaster Medical School
Lancaster University
Furness College
Lancaster, LA1 4YW
UK

jo.knight@lancaster.ac.uk

Paul Rayson
School of Computing and Communications
Lancaster University
InfoLab21, South Drive
Lancaster, LA1 4WA
UK

p.rayson@lancaster.ac.uk

Mahmoud El Haj
School of Computing and Communications
Lancaster University
InfoLab21, South Drive
Lancaster, LA1 4WA
UK

m.el-haj@lancaster.ac.uk

Nathan Rutherford
Department of Information Security
Royal Holloway, University of London
Egham
Surrey, TW20 0EX
UK

Nathan.Rutherford.2019@live.rhul.ac.uk