Can We Quantify Domainhood? Exploring Measures to Assess Domain-Specificity in Web Corpora

Marina Santini[1(✉)], Wiktor Strandqvist[1,2], Mikael Nyström[1,2], Marjan Alirezai[3], and Arne Jönsson[1,2]

[1] RISE SICS, Linköping, Sweden
{marina.santini,mikael.nystrom,arne.jonsson}@ri.se,
wiktor.strandqvist@gmail.com
[2] Linköping University, Linköping, Sweden
{mikael.nystrom,arne.jonsson}@liu.se
[3] Örebro University, Örebro, Sweden
marjan.alirezaie@oru.se

Abstract. Web corpora are a cornerstone of modern Language Technology. Corpora built from the web are convenient because their creation is fast and inexpensive. Several studies have been carried out to assess the representativeness of general-purpose web corpora by comparing them to traditional corpora. Less attention has been paid to assess the representativeness of specialized or domain-specific web corpora. In this paper, we focus on the assessment of domain representativeness of web corpora and we claim that it is possible to assess the degree of domain-specificity, or *domainhood*, of web corpora. We present a case study where we explore the effectiveness of different measures - namely the Mann-Withney-Wilcoxon Test, Kendall correlation coefficient, Kullback–Leibler divergence, log-likelihood and burstiness - to gauge domainhood. Our findings indicate that burstiness is the most suitable measure to single out domain-specific words from a specialized corpus and to allow for the quantification of domainhood.

1 Introduction

Web corpora are text collections made of documents that have been retrieved and downloaded from the web. Building web corpora is convenient because the whole process of corpus creation is automated, fast and inexpensive, while the construction of traditional corpora (like the British National Corpus a.k.a. BNC) is normally expensive, time-consuming and require considerable amount of human expertise to decide the ideal combination of documents to store in the corpus. Needless to say that these investments in time, financial resources and human knowledge are well paid-off because traditional corpora are high-quality and long-lasting collections (e.g. the Brown corpus created in the 60's is still used today).

© Springer Nature Switzerland AG 2018
M. Elloumi et al. (Eds.): DEXA 2018 Workshops, CCIS 903, pp. 207–217, 2018.
https://doi.org/10.1007/978-3-319-99133-7_17

Web corpora are also invaluable, but often time restrictions and limited funding make the manual evaluation of web corpora painstakingly hard and impractical. For this reason, several studies have focussed on quantitative methods and statistical measures to automatically assess the quality of web corpora (e.g. see [8]). Pioneerly, before the start of the web corpora era, Kilgarriff had already spotted this need when he stated: "Measures are needed not only for theoretical and research work, but also to address practical questions that arise wherever corpora are used: is a new corpus sufficiently different from available ones, to make it worth acquiring? When will a grammar based on one corpus be valid for another? How much will it cost to port a Natural Language Processing (NLP) application from one domain, with one corpus, to another, with another?" [14]. More recently, substantial research has been carried out to show that general-purpose web corpora show the same qualities as traditional general-purpose corpora (e.g. see [2,9]). When compared with general-purpose corpus evaluation, the evaluation of domain-specific web corpora is less advanced (see Sect. 2). We would like to start filling this gap because specialized web corpora are widely used in several linguistic disciplines (e.g. in translation studies and lexicography) and they are an important building block of language technology applications (e.g. machine translation, terminology extraction and lexicon induction). Both in linguistics and in language technology, the quality of the results depends on the domain representativeness of the web corpus itself. The research question we would like to start addressing in this paper is then the following: "is it possible to quantify automatically the *domainhood* of web corpora?". The answer to this question is not straightforward. We make the argument that *a domain-specific web corpus (whatever its domain granularity) should be gauged against a list of core terms that well represent the domain of interest.* We stress that domainhood is only one dimension of the overall corpus quality assessment, and that "corpus quality" is a relative (and not absolute) concept, since the corpus quality should be defined in relation to the purpose for which a corpus has been built and the kind of linguistic phenomena it is meant to represent. In this paper, we present a case study where we compare a medical domain-specific web corpus to a general-purpose traditional reference corpus. We apply well-established measures based on word frequency lists. The measures that we explore are: the Mann-Withney-Wilcoxon Test, Kendall correlation coefficient, Kullback–Leibler divergence, log-likelihood and burstiness.

2 Related Work

How and to what extent can we assess the quality of web corpora? It seems that a notion like "corpus quality" has become too vague when referring to text collections produced by recent technology. An obvious rebuttal would be: what's your definition of "corpus quality"?

At the beginning of corpus linguistics, when the collections were manually built and each text was discussed by experts, the quality of a corpus was assessed as the overall representativeness of a corpus for a certain language. This was the main idea behind the construction of the Brown corpus in the 60's, the British National Corpus in the 90's, and of all the other national general-purpose corpora built in the last decades. Much effort was put into the definition of the parameters that can guarantee the "representativeness" of language use (e.g. see [3]).

Nowadays, when we talk about web corpora, it seems more appropriate to talk about "qualities" rather than a monolithic notion of "quality". "Qualities" can be defined as dimensions of variation. Domain, genre, style, register, medium, etc. are well-known dimensions of language variation. Although many researchers have worked on the design and assessment of web corpora, no standard metrics has been agreed upon for the automatic quantitative assessment of the different "qualities" of web corpora. In this study, we focus on the dimension of "domain", that we define as the "subject field" or "area" in which a web document is used. Our aim is somewhat similar to SPARTAN, a technique for constructing specialized corpora from the web by systematically analysing website contents [22]. However, our purpose is not to analyze the domain-specificity of websites as a whole, rather we focus on web documents about chronic diseases. In our experiments, we rely on measures that are well-established and allow for experimental reproducibility, and in this section we provide a short overview of studies where these measures were used.

In his seminal article, Kilgarriff motivates his review of approaches to corpus comparison by asking two crucial questions: "how similar are two corpora?" and "in what ways do two corpora differ?" [14]. Kilgarriff focuses on comparison techniques based on word frequencies, although he acknowledges (as we do) that "a full comparison between any two corpora would of course cover many other matters" [14]. He reviews various statistical measures reaching the conclusion that the Mann-Withney (also known as Wilcoxon) ranks test is a "suitable technique". He also reviews log-likelihood (a measure based on entropy [7]) and acknowledges that it is "mathematically a well-grounded and accurate measure of surpriseness" but it is more difficult to interpret than the Mann-Withney-Wilcoxon ranks test.

Less skeptical about log-likelihood's interpretability are Rayson & Garside, who show that this measure can be safely used as a "quick way to find the differences between corpora" and it is more robust than other measures because it is insensitive to corpus size [18].

The Kendall correlation coefficient helps determine whether the observed patterns of two corpora show significant correlations [10].

Another possibility is to use the Kullback–Leibler (KL) distance to assess the "randomness" or "unbiased-ness" of general-purpose corpora and the "bias-ness" of domain-specific corpora, as in [5], where the authors compare domain-specific suparts of the BNC against the whole BNC corpus, showing that KL divergence reliably indicates the difference between them.

Recently, Strandqvist et al. [21] have profitably applied three corpus profiling measures - namely, rank correlation (Kendall and Spearman), Kullback–Leibler divergence, and log-likelihood - to compare the domain-specificity of two medical web corpora, one bootstrapped with hand-picked term seeds, and the other one bootstrapped with automatically extracted term seeds.

Burstiness has been used in information retrieval and in terminology extraction [13], and more recently for corpus evaluation [20]. Burstiness is a measure that can be utilized for inducing specialized lexicon that is not evenly distributed in a corpus, but appears "in bursts" [23]. Burstiness indicates "how peaked a word's usage is over a particular corpus of documents" [17], and "bursty words are topical words that tend to appear frequently in documents when some topic is discussed, but do not appear frequently across all documents in a collection" [12]. While bursty words are feared and filtered out when assessing general-purpose corpora [20], we think that they could give a good indication of domain-specificity, and for this reason we include burstiness in our experiments.

3 What's in a Specialized Corpus? A Case Study

Since "words are not selected at random" [14], we assume that the words included in a corpus represent its content and language use. In our experiments, we use two corpora of Swedish texts, namely a reference corpus and a specialized corpus.

The reference corpus is the Stockholm-Umeå corpus [11] (henceforth SUC, a.k.a. the National Swedish Corpus), while the specialized corpus is a medical web corpus (henceforth *eCare_Sv_01* [19]) which was recently bootstrapped from the web using 155 SNOMED CT[1] terms (unigrams and bigrams) in Swedish indicating chronic diseases. The size of the SUC amounts to 1 million words, whereas the size of *eCare_Sv_01* is approximately 700 000 words. Both corpora are relatively small.

The 155 SNOMED CT terms that we used as seeds were chosen by a domain expert and they well-represent the domain of interest, since they indicate chronic diseases that are classified as such in the SNOMED CT ontology. To build the corpus, we used these terms as queries in Google.se search engine. The web corpus was downloaded with BootCat [1] (Customized URLs option). Using regular search engines (like Google, Yahoo or Bing) and term seeds (as queries) to build a corpus is handy, but it also has some caveats that depend on the design or distortion of the underlying search engine [22]. These caveats affect the content of web corpora since it might happen that irrelevant documents are included in the collection, especially when searching for very specialized terms. Since manual and qualitative inspections are often prohibitive, the automatic assessment of the domain-specificity of a corpus crawled from the web is potentially very useful.

For the evaluation, we used a gold standard term list containing the tokenized SNOMED CT term seeds, without duplicates. This means that SNOMED CT terms like "kronisk anemi" (en: chronic anemia) and "kronisk artrit" (en: chronic arthritis), in the gold standard will be represented by three entries, namely "kronisk", "anemi" and "artrit". All in all, the gold standard term list includes 165 terms[2].

[1] SNOMED CT browser is available at http://browser.ihtsdotools.org/.

[2] The lists of the selected 155 SNOMED CT terms and the tokenized gold standard (165 entries) are available here: http://santini.se/eCareCorpus/home.htm.

It makes sense to use domain-specific terms for both bootstrapping the web corpus and for evaluating its domainhood because the terms used as seeds (source terms) should be found in non-trivial proportions in the final corpus to be sure that the corpus is representative of the domain.

4 Metrics

In the experiments presented in this paper we take the bag-of-words approach and we used metrics based on word frequency lists. We report results and discussion for the following measures: Mann-Withney-Wilcoxon Test, Kendall correlation coefficient (τ), Kullback–Leibler (KL) divergence, log-likelihood and burstiness.

Word Frequency Lists. A word frequency list (a.k.a. unigram lists) can be seen as a "compact representation of a corpus, lacking much of the information in the corpus but small and easily tractable" [16]. (We used the R packages "tm" and "quanteda" to create frequency lists).

Mann-Withney-Wilcoxon Test. The Mann-Whitney-Wilcoxon Test (non-parametric test) helps decide whether the distributions of the two corpora are identical without assuming them to follow the normal distribution. (We used the R function "wilcox.test()" to calculate the test).

Kendall Correlation Coefficient (τ). Kendall correlation coefficient (τ) is a non-parametric measure of correlation between two rankings. τ is a probability value which indicates the difference between the probability that the observed data are in the same order and the probability that the observed data are not in the same order. There are several variations of Kendall's τ (they differ only in the way that they handle ties). (We used the R function "cor.test()" with method = "kendall" to calculate the test).

Kullback–Leibler (KL) Divergence. The convenience of KL divergence (a.k.a. relative entropy) lies in its ability to quantify how "distant" an estimation of a distribution may be from the true distribution. The KL divergence is non-negative and equal to zero if the two distributions are identical. In our context, the closer the value is to 0, the more similar two corpora are. (We used the R function "KL.empirical()", (log_2), package "entropy" to compute KL divergence).

Log-Likelihood (LL). Log-likelihood (a.k.a. G^2) [7] can be used for corpus profiling [18]. It is a measure based on a contingency table and compares the expected values in two corpora under observation. The larger the LL score of a word, the more different its distribution in the two corpora.

Burstiness. Burstiness helps identify words that are important in certain documents, but that are unevenly distributed in the corpus as a whole. Several burstiness formulas exist. In the experiments reported in this paper, we use Church & Gale's formula [4], including the modification proposed by [12] (i.e. the use of relative frequencies rather than absolute frequencies).

5 Experiments

In this section, we present the experiments and discuss the results.

5.1 Corpus Profile: Sorted Frequency Lists

We computed the relative frequencies of SUC and *eCare_Sv_01* after having removed stopwords (we used the standard Swedish stop list in R). Then we divided each frequency of occurrence by the total number of words in the corpus and we normalized by multiplying by 1 million to get the frequencies of words per millions (wpm) [15]. When visually compared, the sorted frequency lists immediately show the different composition of the two corpora. Common words appear at the top of SUC, while the top ranked words in *eCare_Sv_01* are domain-specific words such as kronisk (en: chronic), behandling (en: treatment), patienter (en: patients), as shown in Table 1.

Table 1. Sorted frequencies (**wpm**)

Rank	SUC	Freq	eCare	Freq
1	också (*also*)	2266.12	kronisk (*chronic*)	4224.16
2	andra (*other*)	1938.1	behandling (*treatment*)	4132.86
3	finns (*exist/be*)	1614.37	hos (*at* (locative))	3669.21
4	år (*year*)	1588.68	patienter (*patients*)	2741.92

5.2 Rank Correlation

In Fig. 1, we compare the rankings (with ties) of the top 1000 words of the two corpora. The cut-off at 1000 is arbitrary and it is simply used to skim off low frequencies. Figure 1 shows that the relation between the top 1000 words in the two corpora tends to be negative, i.e. when the rank of an *eCare_Sv_01* word is high, the rank of the same word is low in the SUC and vice versa (see the aggregation of points that are parallel to x and y axes). However, several words (see the clouds of points in the middle) show a positive relation, i.e. when the rank increases in *eCare_Sv_01*, it also increases in SUC. The solid line in Fig. 1 depicts the LOWESS smoother (a.k.a. Locally Weighted Scatterplot Smoothing). The LOWESS function (R function: "lines(lowess())") creates a smooth line through the scatter plot to help detect the relationship between

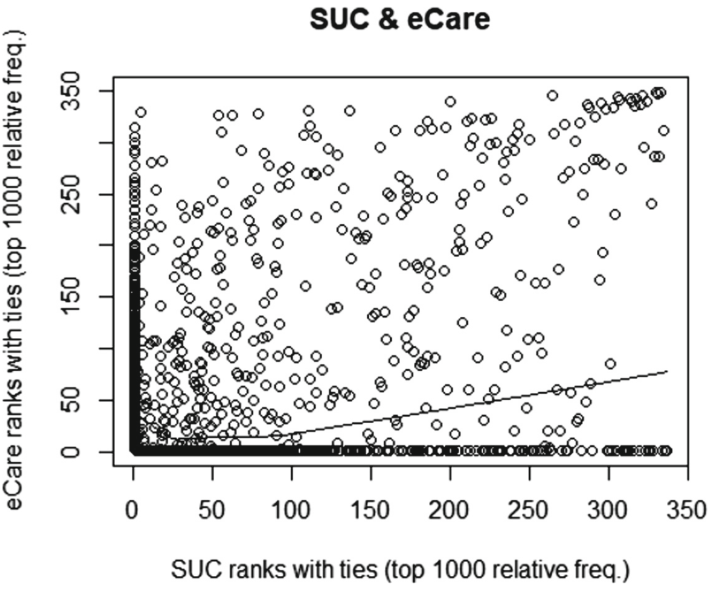

Fig. 1. Frequency scatterplot

variables and foresee trends [10]. In our scatterplot, the LOWESS smoother shows a slight positive relationship. Although informative, the scatterplot does not tell us whether a statistically-significant correlation exists between the two corpora. For this reason, we ran two non-parametric rank correlation measures, namely the Mann-Withney-Wilcoxon Test and Kendall τ and we measured their statistical significance at $p < 0.05$, two-sided.

Mann-Withney-Wilcoxon Test. For the Mann-Withney-Wilcoxon Test, the p-value of the test is 0.019, which is less than the significance level of $p = 0.05$. We can conclude that the median rank of a word in SUC is significantly different from the median rank in *eCare_Sv_01*.

Kendall $\boldsymbol{\tau}$. For Kendall τ, the p-value of the test is 0.000000003122 (p-value in R: 3.122e−09) which is less than the significance level $p = 0.05$.

Both tests indicate that the distribution of the words in *eCare_Sv_01* is significantly different from the distribution of words in SUC.

5.3 Kullback–Leibler (KL) Divergence

A smoothing value of 0.01 was applied to the normalized relative frequencies (wpm). The KL divergence between SUC and *eCare_Sv_01* is 5.80, which (unsurprisingly) indicates a large divergence between a general-purpose SUC and the domain-specific *eCare_Sv_01*.

5.4 Log-Likelihood (a.k.a. G^2)

We computed log-likelihood (LL) on smoothed frequencies. A LL score of 3.8415 or higher is significant at the level of $p < 0.05$ and a LL score of 10.8276 is significant at the level of $p < 0.001$ [6]. We used these thresholds to cut-off the list of LL scores (sorted by decreasing values). We selected the LL scores with a value of 3.8415 or higher and with a value of 10.8276 or higher. We got a list of 1542 records in the first case, and a list of 1514 records in the second case. We computed Precision@ against the gold standard for both lists. An identical value was returned for both, i.e. 0.048. The values for Precision@ are somewhat similar to the values returned by the Jaccard and Dice coefficients (i.e. 0.036 and 0.069 respectively). The intersection between the selected LL scores and the gold standard is 58, i.e. 35.15%.

This experiment provided some useful insights. First, since the words in the frequency lists are not lemmatized, some mismatches are caused by differing morphological forms. This problem can be addressed by lemmatizing the top-ranked frequent words before matching them against the gold standard. Second, it was naive to expect that the top-ranked words were only chronic illnesses. Some words often co-occur with chronic diseases and are domain-specific, like "patient" or "treatment", but they are not in the gold standard because they do not designate a chronic disease. Arguably, the current gold standard is too selective. Last but not least, we realized that the LL scores do not show the "direction" of comparison. For example, a word like "anemi" could, in principle, have a higher distribution in SUC (reference corpus) rather than in *eCare_Sv_01* (target corpus). For this reason we think that log-likelihood might not be the best approach to detect the degree of a domainhood of a target corpus.

5.5 Burstiness

As mentioned above, burstiness singles out content words that tend to appear in some documents, but that are not spread out evenly across the whole corpus. This characterization well fits *eCare_Sv_01* where each chronic disease is discussed in some of the documents, but not in all of them. We calculated burstiness separately for *eCare_Sv_01* and for SUC and compared the results. For each corpus, we sorted the burstiness values by decreasing order and we took the top 2105 bursty words. The expectation is that *eCare_Sv_01* should contain many words listed in the gold standard, while SUC should show a limited overlap with the golden standard. This expectation is indeed met. 90 terms out of 165 (i.e. 54.5%) are top-ranked in the list of *eCare_Sv_01*'s bursty words. It is also to be noted that the range of values of bursty words differs a lot across the two corpora. In *eCare_Sv_01*, the first top-ranked 2105 words have burstiness values ranging from 0.1 to 0.005, while in SUC burstiness values range from 0.044 to 0.0022 for the same range. Table 2 reports the number of words in common with the gold standard, i.e. the intersection, (col.2), the Jaccard coefficient (col.3), the Dice coefficient (col. 4) and Precision@2105 (col. 5). Certainly, the values of the two coefficients and the value of Precision@ do not make justice of the magnitude of the overlap since their calculation takes into account the number of unmatched items, which in our case are many because the gold standard list is much shorter than the list of ranked words. These are the bursty 90 words that *eCare_Sv_01* shares with the gold standard: *andningsinsufficiens, anemi, artrit, artropati, atelektas, atrofisk, basalcellscancer, beryllios, blefarit, bronkiolit, clonorchiasis, continua, cystica, cystit, cystitis, dakryocystit, depression, dermatit, dysfagi, eksem, emfysem, exoftalmus, explosivitet, faryngit, fluoros, gastrit, giktartrit, gingivit, glomerulonefrit, glossit, hemicrania, hepatit, hyperglykemi, hyperkapni, hypernatremi, hyponatremi, infektionssjukdom, intermittent, jaccouds, juvenil, kammartakykardi, kartageners, kolecystit, kolit, konjunktivit, kontaktdermatit, kronisk, krupp, laryngotrakeit, lipoidnefros, lungembolism, lungemfysem, mastit, mastocytos, mastoidit, meningokockemi, metrit, missfall, mycetom, neutropeni, njursvikt, obliterativ, orkit, osteomyelit, ozena, pankreatit, paraplegi, parodontit, paronyki, perikardit, polyserosit, postkardiotomisyndrom, prostatit, psoriasisartrit, rhinitis, rinit, schizofreni, schnitzlers, sicca, silikos, spondyloartropati, syndrom, synovit, testistorsion, tics, trakeit, trakeobronkit, tyreoidit, urtikaria, vulvit.*

Table 2. Comparison between bursty words and the gold standard

	Intersection	Jaccard	Dice	Precision@2105
SUC	1	0.000440	0.00088	0.00001
eCare	**90**	0.04128	0.07929	0.03590

5.6 Discussion

We have seen that the comparison between sorted frequency lists does give a coarse but grounded idea of the domain-specificity of *eCare_Sv_01*. Both correlation tests confirm that the distributions of the word frequencies of the two corpora are not positively correlated and this difference is statistically significant. This finding is further supported by the value returned by the KL divergence, which indicates a large distance between SUC and *eCare_Sv_01*.

However, frequency profiling, rank correlation tests and KL divergence do not tell us how representative the *eCare_Sv_01* corpus is of the domain it is meant to represent.

Sorted LL scores single out words with different distributions in the two corpora. However, these values do not allow us to measure the degree of domain-specificity on an individual corpus.

Burstiness gives a much clearer indication of the domain topics that are discussed in *eCare_Sv_01*. The intersection between *eCare_Sv_01*'s bursty words and the gold standard is an encouraging 54%, a value that can be increased with some additional preprocessing, for example by lemmatizing the bursty words before evaluating them against the gold standard, and by including in the gold standard medical expressions such as “patient”, “treatment”, that do not indicate chronic diseases but are indeed domain-specific.

6 Conclusion and Future Work

In this paper, we explored measures to assess domainhood in web corpora. Domainhood indicates the degree of domain-specificity of a specialized corpus. We carried out comparative experiments where the domainhood of a traditional general-purpose national corpus (SUC) and a specialized medical corpus (*eCare_Sv_01*) were measured against a gold standard list of chronic diseases that represents the target domain. Although the outcome of this experiment is intuitive, we empirically showed that the intuition is supported by counts. Our findings indicate that burstiness is a suitable measure to assess the domain-specificity of a corpus.

We are currently extending these experiments to other languages and additional corpora. We are also working on a new gold standard and on the implementation of additional burstiness formulas.

Acknowledgement. This research was supported by E-care@home, a "SIDUS - Strong Distributed Research Environment" project funded by the Swedish Knowledge Foundation. Project website: http://ecareathome.se/.

References

1. Baroni, M., Bernardini, S.: BootCat: bootstrapping corpora and terms from the web. In: LREC (2004)
2. Baroni, M., Bernardini, S., Ferraresi, A., Zanchetta, E.: The wacky wide web: a collection of very large linguistically processed web-crawled corpora. Lang. Resour. Eval. **43**(3), 209–226 (2009)
3. Biber, D.: Representativeness in corpus design. Literary Linguist. Comput. **8**(4), 243–257 (1993)
4. Church, K.W., Gale, W.A.: Poisson mixtures. Nat. Lang. Eng. **1**(2), 163–190 (1995)
5. Ciaramita, M., Baroni, M.: A figure of merit for the evaluation of web-corpus randomness. In: Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (2006)
6. Desagulier, G.: Corpus Linguistics and Statistics with R. Springer, Heidelberg (2017). https://doi.org/10.1007/978-3-319-64572-8
7. Dunning, T.: Accurate methods for the statistics of surprise and coincidence. Comput. Linguist. **19**(1), 61–74 (1993)
8. Ferraresi, A., Zanchetta, E., Baroni, M., Bernardini, S.: Introducing and evaluating ukWaC, a very large web-derived corpus of English. In: Proceedings of the 4th Web as Corpus Workshop (WAC-4) Can We Beat Google, pp. 47–54 (2008)
9. Fletcher, W.H.: Implementing a BNC-compare-able web corpus. Building and Exploring Web Corpora, pp. 43–56 (2007)
10. Gries, S.T.: Elementary statistical testing with R. In: Krug, M., Schlüter, J. (eds.) Research Methods in Language Variation and change (2013)
11. Gustafson-Capková, S., Hartmann, B.: Manual of the Stockholm Umeå corpus version 2.0. Stockholm University (2006)
12. Irvine, A., Callison-Burch, C.: A comprehensive analysis of bilingual lexicon induction. Comput. Linguist. **43**(2), 273–310 (2017)

13. Katz, S.M.: Distribution of content words and phrases in text and language modelling. Nat. Lang. Eng. **2**(1), 15–59 (1996)
14. Kilgarriff, A.: Comparing corpora. Int. J. Corpus Linguist. **6**(1), 97–133 (2001)
15. Kilgarriff, A.: Simple maths for keywords. In: Proceedings of the Corpus Linguistics Conference, Liverpool, UK (2009)
16. Kilgarriff, A.: Comparable corpora within and across languages, word frequency lists and the KELLY project. In: Proceedings of the 3rd Workshop on Building and Using Comparable Corpora, pp. 1–5 (2010)
17. Pierrehumbert, J.B.: Burstiness of verbs and derived nouns. In: Santos, D., Lindén, K., Ng'ang'a, W. (eds.) Shall We Play the Festschrift Game?, pp. 99–115. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-30773-7_8
18. Rayson, P., Garside, R.: Comparing corpora using frequency profiling. In: Proceedings of the workshop on Comparing Corpora, pp. 1–6. Association for Computational Linguistics (2000)
19. Santini, M., Jönsson, A., Nyström, M., Alireza, M.: A web corpus for eCare: collection, lay annotation and learning-First results. In: Proceedings of the 2nd International Workshop on Language Technologies and Applications (LTA17). FedCSIS (2017)
20. Sharoff, S.: Know thy corpus! Exploring frequency distributions in large corpora. In: Diab, M., Villavicencio, A. (eds.) Essays in Honor of Adam Kilgarriff. Text Speech and Language Technology Series. Springer, Heidelberg (2017)
21. Strandqvist, W., Santini, M., Lind, L., Jönsson, A.: Towards a quality assessment of web corpora for language technology applications. In: Proceedings of TISLID18 - Languages For Digital Lives and Cultures. Ghent University, Belgium (2018)
22. Wong, W., Liu, W., Bennamoun, M.: Constructing specialised corpora through analysing domain representativeness of websites. Lang. Resour. Eval. **45**(2), 209–241 (2011)
23. Zhao, Z., Mei, Q.: Questions about questions: an empirical analysis of information needs on twitter. In: Proceedings of the 22nd International Conference on World Wide Web, pp. 1545–1556. ACM (2013)