Specificities and other applications of the Fisher's exact test to textual data: What's the matter with lexical frequencies?

Bénédicte Pincemin

Specificities and other applications of the Fisher's exact test to textual data: What's the matter with lexical frequencies?

Bénédicte Pincemin

CNRS, IHRIM UMR5317, ENS de Lyon – benedicte *dot* pincemin *at* ens-lyon *dot* fr

Abstract

This article aims to make four contributions: *(i)* to disseminate internationally that the textometric measure of Specificity is a Fisher's exact test; *(ii)* to explain why this measure is still state-of-the-art for textual data analysis; *(iii)* to review and respond to the main criticisms levelled at it; and *(iv)* to provide guidelines for its proper use.

Keywords: association measures, keyword research, collocations, dispersion, corpus linguistics, textometry.

1. Introduction to the Specificity measure and brief history of its reception

The Specificity model in textometric analysis was introduced by Lafon (1980, 1984) ("characteristic elements" in Lebart et al. (1998)). Considering a corpus divided into parts, the Specificity measure evaluates if the frequency of a word in a part is it unexpectedly high or low? The statistics precisely model the probability of any frequency for a word in a part as the proportion, in all possible allocations of words into parts (retaining part sizes and word total frequencies), of instances with the observed frequency (that is, a hypergeometrical distribution, in mathematical terms).[1] In case the observed frequency is more (resp. less) than the frequency for equal distribution, a positive (resp. negative) Specificity is computed from the proportion of cases with the observed frequency or more (resp. less). For better readability and hermeneutic efficiency (since proportions here are often very low and linguistic material is more qualitative than quantitative), proportions are converted to their order of magnitude (Log_{10}). Thus, a Specificity score of +4 just means that, if words were distributed randomly in the corpus, there would be a 1 in 10,000 chance to get this frequency or more.

The Specificity measure is a core feature of the textometric methodology since its very beginning (Lebart et al. 1998). Its complex and intensive computation used to be a challenge 40 years ago: indeed, a sophisticated implementation is required because a straight application of formulae meets computational boundaries. The advancing of current hardware power and available open-source efficient implementations[2] have now made this calculation accessible widely. Nowadays, most textometric software offers a Specificity analysis according to Lafon's model (Lebart et al. 2019). Specificities are used to find characteristic words in corpus parts (keyword-type application). Later on, the same measure has been applied to collocation analysis[3]: the collocates for the *w* word are the characteristic words for the part composed of all the contexts of the *w* word, in contrast to the rest of the corpus (words that are not in the neighborhood of the *w* word). It should be noted that *mathematically*, the calculation is exactly

1 See Lafon (1980:139) for an elementary example, and our conference slides for a picture-based presentation.
2 For instance, the dhyper function in R, or the R packages textometry or corpora. For end users, note the availability of the r/PlotSpecif utility in TXM software, that outputs the Specificity score and a chart for any given set of size and frequency values (corpus size *T*, part size *t*, word frequency *F* in corpus, word frequency *f* in part).
3 Yet Lafon had developed another statistical model for collocations, see last part in (Lafon 1984).

the same for a keyword-type application (characteristic words for parts in a corpus) and for a collocation-type application (words that are attracted by a given word). But *linguistically*, a part made up of a set of *word contexts* is not of the same nature as a part made up of a set of *texts*: the closer or morphosyntactically well-defined the context, the more constrained the words, the further from random. Thus "keyword" Specificity results do not have quite the same properties as "collocation" Specificity results.

Lafon's Specificities implement a Fisher's exact test, and this is the name under which this approach is best known internationally. Pedersen (1996) popularized its application to textual data, arguing the Fisher's exact test is the most appropriate statistical test to dependent word pairs identification. He was followed by Stefanowitsch and Gries (2003) and Evert (2009) notably, in corpus linguistics studies. The Fisher's exact test was pointed out as theoretically superior, so that Evert took it as a "reference point" (2009) to which other measures could be evaluated. Nevertheless, the Fisher's exact test was much more computationally intensive, and other popular tests, such as Log-Likelihood, appeared as adequate approximations as soon as frequencies are large enough (Evert 2009). Therefore, in practice, most studies used alternative association measures, whether statistical *significance* tests like the Fisher's exact test (log-likelihood, chi-square, t-test, z-score and so on) or intuitive *effect-size* measures (such as odds ratio) or *heuristic* formulae (such as TF-IDF) (Evert 2022). Later on, when the computational aspect was no longer an issue, fundamental criticisms have been levelled at properties of the calculation. Even researchers who had supported the Fisher's exact test for years are now questioning its effectiveness and developing alternatives (Gries 2022, Evert 2022).

The aim of this paper is to consider the main current difficulties or objections to the Specificity measure. We would like to take stock of these critical aspects and explain why and how we believe that, taking these properties into account, Specificities still are fully relevant to the textometric approach today.

2. Disputed features

2.1. Sensitivity to high lexical frequencies

The Specificity measure is more responsive to high frequency words: the same relative variation, or the same proportion of word occurrences in a part (f/F), gets a higher score when frequency is higher (Labbé and Labbé 2001). Or the equivalent, a same score may be assigned to deviations that are in proportion stronger for low frequencies than for higher ones (Lafon 1980:159). It is as if the higher the frequency, the easier it is to increase the score. This could be seen as a bias towards high frequency words: it would be a weakness of the model, it would reveal a drawback – as if designers of the measure had not known, described and commented this feature as a plain fact that deserves attention and interest (Lafon 1980:158).

In fact, this behavior is a very natural consequence of a statistical approach based on significance testing (all other statistical measures relying on p-values share this property as well), and it can be understood: the more occurrences you observe, the more confident you are in your judgment, the lower the probability can be when a deviation is observed. This is the flip side of the reliability provided by these statistics: common fluctuations are identified as such and receive bad scores, so that there is no need for bootstrap confidence intervals for instance. With that in mind, the promotion of more frequent words is not to be *corrected* actually (since the measure does what it is designed to do), nevertheless this definitely has to be taken into account in uses (for instance, having a look to lower scores for lower frequency words) and may be complemented with other descriptive tools (see §3.2).

Thus, when applied to all words in a corpus, it is no surprise that many grammatical words occupy top spots, as well as some high frequency nouns, proper names and uninflected words, as soon as they all meet the statistical criteria about their distribution. But identical scores may be interpreted differently, taking into account different categories of words, and processing them separately (Labbé and Labbé 2001). The measure detects grammatical words variations, the user is free to consider them or not, according to their focus of interest. Another consequence of the selection of high frequency words is that the characterictic words for a part may span quite a large amount of tokens in the part (Habert 1985).

Conversely, the Specificity measure somehow penalizes low frequency words. These may not be able to reach the significance threshold even in the case of a notable effect size (a relative frequency much higher in the part than in the whole corpus). There is a (good) reason for this: the measure embeds useful statistical considerations, that is, are there enough occurrences to make a quantitative judgment? For few occurrences, one cannot exclude that a relatively high (or low) frequency would be due to common fluctuations. Thus, low frequencies that could generate unreliable results get low scores inherently. As well as for high frequencies, reading Specificity results with a view on global and local frequencies is useful to better understand and qualify each case: can the low score be due to low frequency, could the frequency distribution of the word draw my attention even if not statistically salient?

Consequently, the analysis of the basic vocabulary, that is, the words with low specificity scores in all parts (Lebart et al. 1998:133), should focus on high and medium frequency words. Low frequencies are not that relevant because they cannot get a high specificity score anyway, their low score being primarily a consequence of their low frequency.

In recent studies, Gries (2022) puts forward that significance testing-based association measures (log-likelihood, and Fisher's exact test as well) are strongly correlated to cooccurrence frequency. This leads him to reject measures like Specificities, and to design another measure aiming at focusing on association information only, without any relationship to frequency, so as to get clear measures and avoid information conflation. We propose three reasons for seeing things differently in our textometric context and maintaining our interest in the Specificity measure. First, the assessment of measures in Gries' analysis depends on the meaning of "association". If we define the strength of the association between a word and a corpus part as how much more densely present the word is in the part (in comparison with in the corpus), then odds ratio and related measures are better than log-likelihood or Fisher's exact test, because they exactly implement this meaning. The Fisher's exact test could still be an interesting measure but targets a different meaning: it rather implements "with a frequency that is improbable at random", which in fact is not completely equivalent to "with a high relative frequency compared to the corpus". Second, notwithstanding their relationship to frequency, Specificities fulfill the requirement for clarity, since scores directly reflect the underlying model.[4] In both cases, either Specificities or Gries' pure association measure, frequency must be considered too, in the latter case because it is independent and useful information, and in the former case because it qualifies Specificity scores according to different profiles and refines interpretation. So, there is no big difference, it is just a different way to combine association and frequency, somehow a kind of addition or subtraction. Gries' solution sounds easier (addition of two elementary features) but Specificities help more directly to distinguish common fluctuations. Third and last reason to think Specificities are still relevant, the decried

[4] In our view, Gries aims for *analytical* clarity, that describes a complex reality through the combination of separate and basic dimensions. We are arguing here for *hermeneutical* clarity, that is grounded on a relevant and overall understanding. We believe both these forms of clarity are valuable in a scientific approach.

correlation to frequency may be mostly effective in the case of collostruction association (and collocation in short contexts) rather than in the case of keyword association. In collocation-type applications, scored words are more constrained (syntactic dependencies, semantic prosody, distributional relations), which makes a kind of preselection, so that frequency ends up being one of the main factors left in the rating.[5]

In keyword-type applications, a high frequency makes a high responsiveness to frequency variation, but this affects positive as well as negative specificities. In other words, when you consider a high frequency word, especially a common grammatical word, you can suspect it may be characteristic, but you have to compare its observed frequency to its expected frequency to figure out whether it is over- or underused. A word can have a high frequency in a part without being characteristic (it is very present but not discriminatory), and conversely a negative specificity does not mean a low frequency or absence (Habert 1985): this is actually a non-intuitive effect of semiotics of Specificity bar charts. All of which goes to show the difference and complementarity between Specificity and frequency.

2.2. Sensitivity to part sizes

This section deals with the case exactly symmetrical to that of the previous section, reflecting the computational symmetry in data (contingency table) and processing. Instead of taking the point of view of words, we are now taking the point of view of parts. So, in short, the bigger the part, the higher the scores of its characteristic words can raise – up to half the corpus. This half-corpus limit can be deduced from the fact that when a corpus is divided into two parts, the specificity scores are exactly opposite. In fact, the probability of a word appearing in a part with a given frequency is necessarily equal to the probability of the word appearing in the complementary part with the complementary frequency: these are two views on the same allocation, so the proportion of allocations are strictly identical. Thus, the highest potential for specificity scores is the case of a half-corpus part.

A first practical consequence is that Specificity scores are not comparable from one part to another when the parts are of quite different sizes (Salem 1987). This does not mean that Specificities are inapplicable to corpora divided into unbalanced parts. The mathematical model has no requirement about part sizes, it is valid in any case. The warning stands not for the computation step but for the interpretation one: numerous high scores may not especially denote originality if occurring in a part much larger than others, and conversely for a poorly characterized part you have to consider whether this might just be due to it being too small. Maybe you can revise your corpus and boost parts that appear too narrow; maybe you cannot really, because data is lacking or because it would distort reality you want to describe. The Specificity measure does take part sizes into account and can fit any corpus configuration. This is an excellent thing, because from a descriptive point of view, it is first and foremost up to the model to adapt to the data, and not up to the data to conform to the model. A fair balance therefore needs to be struck between corpus relevance and corpus effectiveness, after which an essential role must be given to interpretation when reading the results.

5 Here we are relying mainly on a kindly communication from Gries, in anticipation of the forthcoming publication of his new book *Towards a revision of most corpus-linguistics statistics: Rethinking frequency, dispersion, association, keyness, and more*, that extends (Gries 2022). In a keyness analysis applied to spoken *vs* written language, log-likelihood exhibits a rank correlation (Spearman) of 0.63 with the logged frequency of the word in the part; whereas for collostructional studies, correlation with the logged cooccurrence frequency is over 0.9 – observations range from 0.934 (in Gries 2022) to 0.991 (in the book to appear).

The second practical consequence is that the Specificity measure is hardly effective for short texts or small subcorpora. Indeed, for short pieces of text, one cannot exclude that a relatively high (or low) frequency would be due to common fluctuations. In addition, the part size determines the minimum frequency at which an absence becomes significant (Salem 1987:214): the smaller the part, the higher the frequency required for a word to have negative specificity. This may create a deficit in negative specificities in small parts. After considering the impact of single small parts, we can also draw the implications for the level of corpus division. The more finely divided the corpus, the narrower the range of Specificity scores, the fewer characteristic words are found and even fewer negative Specificities (Labbé and Labbé 2001), the smaller the basic vocabulary (Habert 1985). This does not mean that big parts and coarse-grained slicing are better (the risk there is being overwhelmed with characteristic words), but that you have to know that a fine splitting of the corpus may hinder the Specificity scores.

Habert (1985:140-141) also pointed out a consequence of the lexical frequency sensitivity that shows up when considering part characterization: if a part features a higher lexical diversity, which leads to lower frequencies (less repetitions), then the harvest of characteristic words will tend to be less abundant in the part or to get mitigated scores. In other words, richness in characteristic words is no evidence of lexical diversity.

2.3. A bag-of-word model: the dispersion of a word inside a part is neglected

A pleasant way of putting it could be: the Specificity statistics have no clear idea of what is going on inside. They are focused on measuring contrast – how different the part is from the whole – yet within the part they make no difference whether the characteristic word's occurrences are all squeezed in a single text or are nicely distributed all over the part. Specificities are an *inter* measurement and have to be complemented with an *intra* measurement. Corpus linguistics responded to this need with the development of dispersion measures, either to add relevant information and refine analysis (Gries 2019), or to supersede association measures in keyword research (Lijffijt et al 2016, Egbert and Biber 2019).

The textometric tradition also had developed solutions. Features are dedicated to displaying word occurrences as the corpus progresses: *carte des sections* (map of sections) in Lexico and Trameur, *topologie* in Hyperbase, *progression* in TXM. Another solution consists in recursively applying the Specificity computing at different scales, typically on the part level then on the text level (Mayaffre et al. 2018): characteristic words for a part are checked to see if they are also characteristic in most texts in the part. Note that this solution is not a change in the model, rather in the way of using it. From that perspective, the limitation may not come from the measure itself but from the way it is used. In brief, the textometric user is expected to combine views and analyses so as to gradually make sense and build up their interpretation.

Considering text frequency (i.e. in how many different texts the word occurs) instead of plain lexical frequencies (i.e. how many times the word occurs) could be a way to adapt the measure and integrate dispersion into association measures (Egbert and Biber 2019, Evert 2022). However, in the case of the Fisher's exact test, this transformation does not fit the underlying model nicely: words would have a fixed number of texts in which they can occur, texts would have a fixed vocabulary size, a word could not be randomly assigned twice in a text, and so on – the model is not intuitively interpretable any longer. However, text frequency itself could be an additional piece of information in Specificity tables, alongside word frequency and part size.

2.4. Corpus parts are processed in isolation, dispersion among other parts is ignored

Following on from our previous section, we get the other side of the coin: in a way, the Specificity statistics have no clear idea of what is going on outside either… When evaluating the frequency of a word in a part, its detailed frequencies in other parts are not taken into account (for instance, whether it occurs in only one other part, or in every other – no difference). The measure is basically focused, it contrasts a specific case with the rest of the corpus taken as a whole, ignoring the structure given by the other parts. This is clearly shown by the fact that the calculation only requires the following 4 parameters: T, the total size of the corpus (number of tokens); t, the size of the part; F, the total frequency of the word (absolute number of occurrences); f, the frequency of the word in the part.

In textometric practice, it is therefore common to contextualize the Specificity results by combining two readings (Habert 1985): on the one hand, the list of characteristic words for a part, by descending score, and on the other, the series of specificity scores of a word along all the parts (typically with a bar chart).

2.5. Asymmetry between positive and negative Specificities

In case the expected frequency of a word in a part is rather low (which happens quite frequently since half of the vocabulary are hapaxes), the word cannot get a negative Specificity in this part (even if absent this would not be statistically surprising); whereas it may get a positive Specificity if its total frequency is high enough and the part size big enough too. For low expected frequencies, the probability distribution is asymmetric so that low probabilities only happen for frequencies higher than expected (Lafon 1980). As a consequence, the measure should detect more positive than negative Specificities (Labbé and Labbé 2001). The case of the division of the corpus into two parts is a borderline case, since in this case there is an exact opposition of scores between the two parts, so in total as many specificities are detected for each sign.

This imbalance between the two types of Specificity is a fact that can be explained (when absence is one of the probable cases, there is no lower frequency that could correspond to underuse). We therefore consider this to be a fact, a property of lexical distributions, not an anomaly of the measure (we do not see any reason why a balance should be normal).

2.6. Observations deviate from the model, the statistics do not represent the data

Language is not random: word occurrences are not independent events, since there are obvious contextual, syntactic and semantic interconnections. Evidence of this, the statistical test selects many more so-called unexpected events than the calculated probabilities estimate. Some critics conclude that, fundamentally, the model is improperly used, it does not suit the data (Labbé and Labbé 2001).

This is a tricky point because there is a misunderstanding about the role given to statistics. In the textometric field, Specificities are intended to serve as a gauge rather than as a predictive model. The designers of the measure totally endorse this fact (Lafon 1980:164; Lebart et al. 1998:135-136): their aim is not to fit language but to get a benchmark, a tool for measuring deviation from a situation taken as a reference. The reference has to be clear, not necessarily realistic. They know the statistics do not reflect actual word distributions, and they explicitly state this is not an issue, it is still possible and relevant to apply the model. Textometry is mainly a descriptive and exploratory use of statistics, as opposed to a confirmatory one (Lebart et al. 1998, 2019). Brunet puts it very neatly: “If I want to check whether a line is straight or not, […] I use a ruler. If […] the line has sinuosities, I am not going to break the ruler, on the pretext that

it does not fit the data, that nature rebels against ideal figures and that the 'predictability' of the ruler is always contradicted by facts. […] In lexical matters, the statistical rule also only allows us to measure. All it does is describe, not explain, and certainly not predict." (our translation of Brunet 2011 [1984]:84).

In this respect, scores cannot be understood as lexical or linguistic probabilities – probabilities in language. In textometric practice, scores do not prove the significance of a word frequency, but they are pragmatically used for a relative ranking of words in a part (sorting in descending score and focusing on top words) and for an absolute threshold typically set at 3. Indeed, words with a score less than 3 (p=1 ‰) are poor candidates, since their frequency can be due to common fluctuations. Moreover, the p=0.05 (5 %) usual significance level is inadequate since language does not work randomly (too many words would be identified as outliers), and because the test is repeated on numerous words, which raises the problem of multiple comparisons: setting a more stringent threshold is a way to deal with this problem (Lebart et al. 1998, Gries 2005, McEnery & Hardie 2012).

These characteristics of the Fisher's exact test application to textual data sheds light on why the Log_{10} notation provides such an efficient scale to read and rank results. Many probability values are very low and way below conventional threshold. Without such a logarithmic scale conversion, they would appear blended into a unique very low probability set. The logarithmic scale provides legible and well-distributed values with a transparent meaning.

3. Proposal: Specificities, a reasoned choice, still state of the art

3.1. What are Specificities for?

We have to precise what expectations Specificities meet, so as to derive two useful consequences: on the one hand, avoiding misuse and over-interpretation in textometric practices; and on the other hand, understanding Specificities are less relevant to other contexts, that require other kind of measures better answering other needs.

As stated in §2, Specificity scores provide a piece of information that has to be combined with other pieces so as to make sense. In corpus linguistics, Gries (2019) insists on this multi-dimensionality of analysis on textual data, the hermeneutic necessity to articulate several measures bearing elementary and complementary meanings, an approach he coins *tupleization*. In textometry, such a tupleization is implemented both spatially (e.g. multiple indicators in tables) and temporally (e.g. analytical path), that is a kind of dynamic tupleization that needs time and provides no one-shot results. The user proceeds in several steps, refining their observations through running related processes that progressively illuminate one another. The Specificity measure is a descriptive tool in a toolbox rather than an efficient and direct integrated filter. Scores do not validate results, they organize them for further exploration. In so doing, textometry does not really meet the needs of automatic analysis. The relationship to time and data is different: roughly, textometry is better suited to digital humanities users who know their corpus and wish to read them in greater depth, and is of less help for executive users who need to input big data and output main results in one go.

The point is that Specificities inform on patterns, not on meaning. Specificities have no semantic claim. Just as frequency does not determine importance, Specificity does not determine keyness or linguistic connection: it just states the word is surprisingly more present that what would happen at random. This is but one way of describing overuse; for instance, comparing relative frequencies (such as odds ratio) reflects another view and provides different

results that can be interpreted differently. Recent innovations by Gries (2022) and Evert (2022) go in this vein, with an enhanced implementation of an effect-size measure.

The Specificity measure is effective in detecting frequency fluctuations, especially concerning high frequency words; it deals with strong presence or nearly absence too, but these are not given precedence especially. Thus, when applied to the full vocabulary of the corpus, Specificities may pick a substantial portion of function words, whose interpretation may be more stylistic than thematic. Furthermore, the statistical measure may put forward phenomena that are not so obvious. Simple formulae based on relative frequencies and effect size could be too intuitive, staying somehow on the surface and grasping more or less what can already be seen with the naked eye (wide variations, or words that occur in only one part). Whereas a significance testing-based measure may reveal facts that are not that much visible: for instance, a tendency for a frequent word to be less used; or words that do not occur and for which this absence could draw our attention (nullax, see Lebart et al. 2019:129 sq.). In these cases, Specificities are a sensor for the detection of interesting absences. Hence, the choice of the measure should depend on the kind of associations sought: roughly, is it a priority not to miss out on salient features (maybe for keyword extraction?), or does the analysis aim to detect somehow deep features (maybe for stylistic studies and authorship attribution)?

Last but not least, we would like to promote an evaluation of measures in terms of *comprehension* rather than *extension*, in the sense that these terms have in logic. That is, evaluating the measure on the generic idea – what its criteria are, which interpretation it conveys – rather than on some facts – comparing its output with target results, which by the way supposes you already know what you want (this might not be obvious in an exploratory approach). Of course, facts are essential, they must be the core of any corpus approach, but interpretation is what makes sense: both have to be connected. In our case for Specificities, the formula is complex, it is not a concise ratio or percentage for instance. But the underlying model is transparent, and for users' hermeneutic concerns, the main thing is to understand the model, the principles it implements (*what* is computed), not necessarily the details of the mathematical formula (*how* it is computed).

3.2. How to use Specificities within a textometric approach?

First and foremost, the user must keep in mind that textometry in general, and Specificities in particular, fundamentally implement a contrastive approach (Habert 1985). Firstly, every output is related to the corpus, that determines the reference to which everything is compared. So, the user has to know their corpus that provides the background to interpret any observation. For instance, in case the corpus mixes different text genres, generic features may have a major impact in Specificities outputs. Secondly, parts have an effect on one another like communicating vessels. For instance, a word that gets a high positive Specificity score in one part may receive a series of low negative Specificity scores in other parts, just as a counterpart – what is meaningful is the singular overuse, not really underuses elsewhere.

A fair use of Specificity requires a good understanding of the underlying model (not necessarily up to mathematical formulae): that is, the idea that random allocations are simulated, and that the score expresses the order of magnitude of the proportion of cases with such a high (or low) frequency. So, we get a measure of distance from chance. It is also very useful to keep in mind the properties of the measurement, so as not to rely on an intuitive but sometimes misleading interpretation. These are all the properties we reviewed in the previous section, mainly: the influence of word frequency and part sizes, the need to consider dispersion inside the part, and the meaning given to statistics here, that provide indicators rather than validations or predictions.

We have highlighted that Specificity scores are not a comprehensive and autonomous measurement. Scores cannot be directly compared and do not provide any absolute characterization: that is why Specificity tables display not only S but also frequencies (F, f) and part sizes (t). Dispersion of words within parts can be managed through recursive specificity calculation: this is especially important in case of parts made up of a dozen or so individuals (texts, people) (say 20 or less), because in this case a strong characteristic of a single individual can show at the level of the part without representativeness. Furthermore, the textometric interpretative path combines statistical summaries and text reading. Examining words in context provides the means to refine linguistic units (phrases, patterns, topics...) (Habert 1985).

Advanced software such as TXM (Heiden 2010) opens up a wide range of possibilities for adjusting the various parameters to suit the needs of the analysis, as illustrated in (Mayaffre et al. 2018). Concerning frequency F for instance, the user can choose the lexical types (setting the type/token relationship, that is how tokens are unified and merged into types) and create complex lexical units (word sets representing topics, N-grams, morphosyntactic patterns, etc.). The T parameter, that is the entire set of units inside which random allocations are simulated, can be set to reflect potential paradigmatic subsets of variation. You can consider that any word may take the place of any other (T = corpus size), or you can consider than linguistically, choices happen within a paradigm, and focus on frequency variations within such and such word category (T= corpus restricted to the word category only, other words are temporarily just ignored). For instance, Mayaffre (2006) computes characteristic verbs only within the verb set of the corpus in order to cancel out the overall style variation towards preference for nouns or for verbs.

4. Conclusion: Summary of main ideas

The Specificity score evaluates whether the frequency of a word in a corpus part is noteworthy. To do so, it implements a Fisher's exact test. It compares the original data to a random allocation of words. Exact probabilities are computed from the proportion of all possible cases carrying out each frequency. Probabilities are cumulated so as to measure how rare it is to reach such a high (or low) frequency. Then the Specificity score converts the probability into its order of magnitude.

The Specificity measure (and likewise the Fisher's exact test) benefits from two main assets. Firstly, it is clear, that is both transparent and meaningful, as it represents a direct translation of the linguistic question into a mathematical model, so that one can fully *understand* what is measured, that is, the order of magnitude of the deviation from a random word distribution. Secondly, the measure is reliable. As an exact and non-parametric test, it does not need any external assumption about the underlying probability distribution, the full range of frequencies is managed (no validity limit for low frequencies), and the measure embeds a statistical significance evaluation (no need for confidence intervals).

A number of criticisms have been levelled at the model, they in fact address known properties: the influence of word frequency and part size, the ignorance of word dispersion or position within and outside parts, the fundamentally non-random nature of language. We have explained how we consider that these are not weaknesses for which the model should be rejected, but properties that can be meaningful and which are integrated into the conduct of the analysis. In textometry, the calculation of specificities is not applied in isolation, but is part of a fundamentally exploratory and interpretative process.

Acknowledgment

This paper greatly benefited from stimulating discussions in Trier (February 2023, workshop organized by Christof Schöch and the Zeta project's team) and in Montpellier (September 2023, conference coordinated by Sascha Diwersy and Céline Poudat, CORLI Consortium). I am very grateful to Stefan Gries for his patient and insightful comments on a first draft.

References

Brunet É. (2011 [1984]). Le viol de l'urne. In Poudat C. (Ed.), *Ce qui compte. Étienne Brunet, Écrits choisis, tome II. Méthodes statistiques*. Paris: Honoré Champion, 79-93.

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

Evert S. (2009). Corpora and collocations. In Lüdeling A. and Kytö M. (Eds.), *Corpus Linguistics. An International Handbook*. Berlin: Mouton de Gruyter, 1212-1248.

Evert S. (2022). Measuring keyness. In Shimoda M. et al. (Eds), *Digital Humanities 2022: Conference Abstracts*, Tokyo, 202-205. https://osf.io/cy6mw/

Gries S. Th. (2005). Null-hypothesis significance testing of word frequencies: a follow-up on Kilgarriff. *Corpus Linguistics and Linguistic Theory*, 1(2), 277-294.

Gries S. Th. (2019). 15 years of collostructions: some long overdue additions/corrections (to/of actually all sorts of corpus-linguistics measures). *Int. Journal of Corpus Linguistics*, 24(3), 385-412.

Gries S. Th. (2022). What do (some of) our association measures measure (most)? Association? *Journal of Second Language Studies*, 5(1), 1-33.

Habert B. (1985). L'analyse des formes « spécifiques ». *Mots*, 11, 127-154.

Heiden S. (2010). The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme. In Otoguro R. et al. (Eds.), *Proc. of the 24th Pacific Asia Conference on Language, Information and Computation*. Tokyo: Waseda University, 389-398.

Labbé C. and Labbé D. (2001). Que mesure la spécificité du vocabulaire ?. *Lexicometrica*, 3.

Lafon P. (1980). Sur la variabilité de la fréquence des formes dans un corpus. *Mots*, 1, 127-165.

Lafon P. (1984). *Dépouillements et statistiques en lexicométrie*. Genève-Paris: Slatkine-Champion.

Lebart L., Pincemin B., Poudat C. (2019). *Analyse des données textuelles*. Presses de l'Univ. du Québec.

Lebart L., Salem A. and Berry E. (1998). *Exploring Textual Data*. Dordrecht: Kluwer Academic.

Lijffijt J., Nevalainen T., Säily T., Papapetrou P., Puolamäki K., Mannila H. (2016). Significance testing of word frequencies in corpora. *Digital Scholarship in the Humanities*, 31 (2), 374–397.

Mayaffre D. (2006). Faut-il prendre en compte la composition grammaticale des textes dans le calcul des spécificités lexicales ? Tests logométriques appliqués au discours présidentiel sous la Vème République. In Viprey J.-M. (Ed.), *JADT 2006*. Besançon: Presses Univ. Franche-Comté, 677-685.

Mayaffre D., Pincemin B., Heiden S., and Weyl Ph. (2018). L'évolution de la mémoire de la Shoah au prisme de la statistique textuelle. In Peschanski D. and Sion B. (Eds), *La vérité du témoin. Mémoire et mémorialisatio*n. Paris: Hermann & Bry-sur-Marne: Institut National de l'Audiovisuel, 93-124.

McEnery T. and Hardie A. (2012). *Corpus Linguistics. Cambridge*: Cambridge University Press.

Pedersen T. (1996). Fishing for exactness. In *Proceedings of the South-Central SAS Users group Conference*. Austin, TX, 188-200.

Salem A. (1987). *Pratique des segments répétés. Essai de statistique textuelle*. Paris: Klincksieck.

Stefanowitsch A. and Gries S. Th. (2003). Collostructions: investigating the interaction between words and constructions. *International Journal of Corpus Linguistics*, 8(2), 209-243.