Introducing Functional Diversity: A Novel Approach to Lexical Diversity in (Historical) Corpora

Folgert Karsdorp[1,*], Enrique Manjavacas[2] and Lauren Fonteyn[2]

[1]*KNAW Meertens Institute, Amsterdam, the Netherlands*
[2]*Leiden University, Leiden, the Netherlands*

Abstract
The question how we can reliably estimate the lexical diversity of a particular text (collection) has often been asked by linguists and literary scholars alike. This short paper introduces a way of operationalizing functional diversity measurements by means of token-based embeddings, and argues that functional diversity is not only a practically advantageous, but also a theoretically relevant addition to the Computational Humanities Research toolkit. By means of an experiment on the historical ARCHER corpus, we show that lexical diversity at the level of functional groups is less sensitive to orthographic variation, and provides insight into an important and often disregarded dimension of vocabulary diversity in textual data.

Keywords
Lexical diversity, Functional diversity, Historical text, Hill numbers

1. Introduction

With the present paper, we wish to make a case for the practical and theoretical advantages of adopting the framework of attribute diversity – which distinguishes categorical diversity from the higher-order concept of functional diversity – into Humanities research on lexical diversity. Given two sets of unique word types, set A{*cat, dog, bird, rabbit*} and set B{*cat, progesterone, remember, blue*}, approaches focusing solely on categorical lexical diversity will suggest A and B are equally diverse. However, an approach that takes the semantic distance between the items into account will also capture the higher functional-semantic or attribute diversity of set B. To help establish the latter approach in Humanities Research, we propose a way of operationalizing functional diversity estimates by means of token-based embeddings.

The question whether we can estimate lexical richness or diversity is a pertinent one in Humanities. In Linguistics, attempts have been made to estimate the vocabulary size of a particular language [12, 11], or how many words an average speaker of a particular language

CHR 2022: Computational Humanities Research Conference, December 12 – 14, 2022, Antwerp, Belgium
[*]Corresponding author.
folgert@karsdorp.io (F. Karsdorp); e.m.a.manjavacas.arevalo@hum.leidenuniv.nl (E. Manjavacas); l.fonteyn@hum.leidenuniv.nl (L. Fonteyn)
https://www.karsdorp.io/ (F. Karsdorp);
https://www.universiteitleiden.nl/medewerkers/enrique-manjavacas-arevalo/ (E. Manjavacas);
https://www.universiteitleiden.nl/en/staffmembers/lauren-fonteyn (L. Fonteyn)
0000-0002-5958-0551 (F. Karsdorp); 0000-0002-3942-7680 (E. Manjavacas); 0000-0001-5706-8418 (L. Fonteyn)

knows at different ages [5, 25]. In a similar vein, researchers have also attempted to estimate (and compare) the richness of the active vocabulary of particular authors [e.g. 16, 18] or literary works across time [e.g., 23], or the 'productivity' of linguistic structures (i.e., how many different word types are used in a particular linguistic context [2, 3]) for different individuals [e.g., 24, 1] or across time [e.g., 21]. To attain these goals, researchers often resort to corpus research, using text (excerpt) collections of varying sizes with diversity measures that rely on the number of word tokens, unique word types, and/or hapax legomena (i.e., words that occur only once), such as (variations on) Mean Word Frequency (MWF) and Type-Token Ratio (TTR) [for examples, see 27], realized/potential/expanding productivity [2], or measures that originate in Shannon entropy [26].

There is, however, a practical problem that arises with any measure of diversity that relies on hapaxes and/or unique types. In many digitized text corpora, the number of unique character strings cannot be equated to the number of unique words. A substantial amount of variation in how word types are represented in a corpus may be due to OCR errors (e.g., in historical texts, the long S character <ſ> is often mistaken for <f> or <l>, which means the word type *strength* could be represented in a corpus as at least three different character strings: <ſtrength>, <frength> and <lrength>). Furthermore, some types of corpora contain texts where authors do not (consistently) adhere to (present-day) standard spelling conventions, such as corpora of (informal) language on social media or any historical corpora that pre-date the establishment of uniform spelling conventions. This introduces a dimension of variation that makes it difficult to accurately count the number of actual hapax legomena or unique types. Of course, at least some of this unwanted variation can be tackled in corpus pre-processing through (semi-)automated spelling normalisation, but this too can prove challenging given that neither OCR errors nor non-standard spelling variation are entirely or even largely systematic.

In this paper, we argue that there are substantial advantages to relying on functional diversity measures (rather than, or as a complement to lexical diversity measures) to estimate and compare the 'lexical richness' of (collections of) text. More specifically:

- We demonstrate that functional diversity estimates are affected to a much lesser extent by spelling errors and inconsistencies than lexical diversity estimates. As such, there is a clear **practical advantage** to relying on functional diversity.
- We suggest that, even in corpora that are free from orthographic noise, there is a **theoretical advantage** to examining higher-order diversity at the level of functional groups. We propose that a theoretically relevant distinction can be made when making claims about 'vocabulary richness' or lexical diversity by taking the semantic similarity between words into account.[1] This higher-order, functional-semantic dimension of diversity is theoretically relevant, as it helps characterize diversity in terms of depth and width, and offers a perspective on diversity that is not captured by more traditional, exclusively categorical measures.

[1] The distinction between lower-order and higher-order diversity proposed here is reminiscent of the distinction between 'productivity' and 'schematicity' in [21, 13].

2. Measuring Diversity

Functional Diversity For our measurements of functional diversity in (historical) corpora, we apply the framework of attribute diversity, which was originally developed in the context of ecological diversity [8, 6]. In ecology too, it is important to not only account for categorical diversity (the taxonomic model of species diversity), but also for attribute variation between and within species. After all, certain species (e.g., ducks vs. geese) are more similar to each other than others (e.g., ducks vs. sheep). This is not captured by taxonomic diversity, which treats all species as equally distant.

In the framework of attribute diversity [8, 6], categorical diversity is considered a special case of functional diversity, where each type (or species) is considered its own functional group and all groups are functionally equally different. In this extreme case, each functional difference results in the definition of a new functional group which is equivalent to a categorical type. More precisely, the threshold τ for defining a new functional group is set to the smallest pairwise distance between types. The framework allows researchers to specify functional groups at higher distinctiveness thresholds τ. τ then specifies the distance threshold beyond which types are considered equally distant and thus belong to different functional groups. As τ tends to infinity, types become functionally indistinct and belong to the same functional group.

Each type i contributes to the frequency of a functional group. Let n_i be the frequency of type i and a_i the frequency of a functional group, then $v_i(\tau)$ can be defined as the proportional contribution of type i to a group for a given threshold level τ. Functional diversity, then, is defined as the sum over the proportional contributions $v_i(\tau)$ of each type $i = 1, 2, \dots, k$:

$$FD = \sum_{i=1}^{k} v_i(\tau), \tag{1}$$

where $v_i(\tau) = n_i/a_i$. Note that when each type belongs to its own functional group, i.e., when the definition of functional groups and types coincide, v_i equals unity. In this case, $n_i = a_i(\tau)$ and thus the functional diversity is equal to the number of types k. When functional groups and types do not coincide, certain functional groups consist of more than one type, which in turn may belong to more than one group. To account for such many-to-many type-function relations, the abundance a_i at threshold τ is computed as the number of tokens of type i plus a fraction of the tokens of any other type j that is functionally indistinctive from type i:

$$a_i(\tau) = n_i + \sum_{j \neq i}^{k} \left(1 - \frac{d_{ij}(\tau)}{\tau}\right) n_j \tag{2}$$

here, $d_{ij}(\tau)$ refers to the distance between type i and j, which is set to τ if $d_{ij} > \tau$ and d_{ij} otherwise.

Functional Hill numbers Eq. 1 describes the functional richness of a collection, or the number of functional groups given a distinctiveness threshold τ. Richness is just one of many diversity measures which treats each functional group as equally important. However, certain functional groups may be more prominent than others which better is captured by other diversity measures, like Shannon entropy or the Gini-Simpson index. To account for other aspects of diversity, Chao and colleagues [8, 6] integrate functional diversity into a mathematically unified family of diversity indexes called Hill Numbers [14]. Hill numbers are parameterized only by q, which determines the sensitivity to the relative frequency p_i of variant type i [14, 7, 17, 9]:

$$ {}^{q}D((p_1, \dots, p_k)) = \left(\sum_{i=1}^{k} p_i^q \right)^{\frac{1}{1-q}} \tag{3} $$

The diversity values at certain orders q correspond to well-known diversity indices. The number of unique types (also called the 'richness' of a sample is equal to ${}^{0}D$. With $q = 0$ no weight is given to the relative frequency of the types, or, conversely, maximum weight is given to rare types. By setting q to 1, the weight of each type is proportional to its relative frequency. Note, however, that ${}^{1}D$ is undefined. Yet, the limit $\lim_{q \to 1}$ exists, which is equal to the exponent of Shannon entropy [cf. 7, 17, 9]. With $q > 1$, disproportionally more weight is given to more frequent types. For instance, the Hill number of order $q = 2$ is equal to the inverse of the Gini-Simpson index, which expresses the probability that two random tokens are of the same type. An interesting property of Hill numbers is that all diversity indices are expressed in terms of the effective number of types: the number of equally frequent types required to obtain a particular observed diversity value. Because the indices are on the same scale, they can easily be represented in 'diversity profiles', which chart the diversity at different order q. These profiles, then, can be used to characterized the evenness of some collection. Profiles with steep declines indicate a large disparity in the frequencies of the types, wheres flat profiles indicate a more even distribution among types.

By incorporating functional diversity into the Hill number framework, Chao, Chiu and colleagues [8, 6] show how to estimate the effective number of equally distinct functional groups at a given distinctiveness threshold τ and diversity order q. The 'effective number', sometimes called 'true diversity', represents the number of types in an idealized reference sample that all have the same frequency and distance between them of at least τ. Expanding on Eq. 1, the functional diversity of order q is defined as follows:

$$ {}^{q}FD(\Delta(\tau)) = \left(\sum_{i=1}^{k} v_i(\tau) \left(\frac{a_i(\tau)}{n} \right)^{q} \right)^{\frac{1}{1-q}}, \tag{4} $$

where n refers to the total number of tokens in the collection.

Example To obtain a better intuition of what functional diversity measures entail, and specifically how the measure responds to the parameter τ, we present the following example.[2] Consider these four words and their corresponding frequencies: *apricot* ($n_1 = 20$), *pineapple* ($n_2 = 15$), *digital* ($n_3 = 10$), *information* ($n_4 = 5$). For each word, Table 1 lists whether it co-occurs with any of ten context words. Each word can thus be represented as a binary

[2] Our example is a translation of [6] into a linguistic context.

Table 1
Co-occurrence table supporting the example in Figure 1 which illustrates how functional diversity can be calculated at different levels of τ.

	boil	data	sugar	pizza	water	hat	tourist	kiosk	camera	photo
apricot	1	0	1	0	1	0	0	0	0	0
pineapple	1	0	1	1	1	1	0	0	0	0
digital	0	1	0	0	0	0	0	0	1	1
information	0	1	0	0	0	0	1	1	0	0

context vector, which can be used to compute the distance between two words. For example, computing the pairwise distances between all four words using the Jaccard distance yields the following distance matrix Δ:

$$\begin{array}{l} \text{apricot} \\ \text{pineapple} \\ \text{information} \\ \text{digital} \end{array} \begin{bmatrix} 0 & 0.4 & 1 & 1 \\ 0.4 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0.6 \\ 1 & 1 & 0.6 & 0 \end{bmatrix}$$

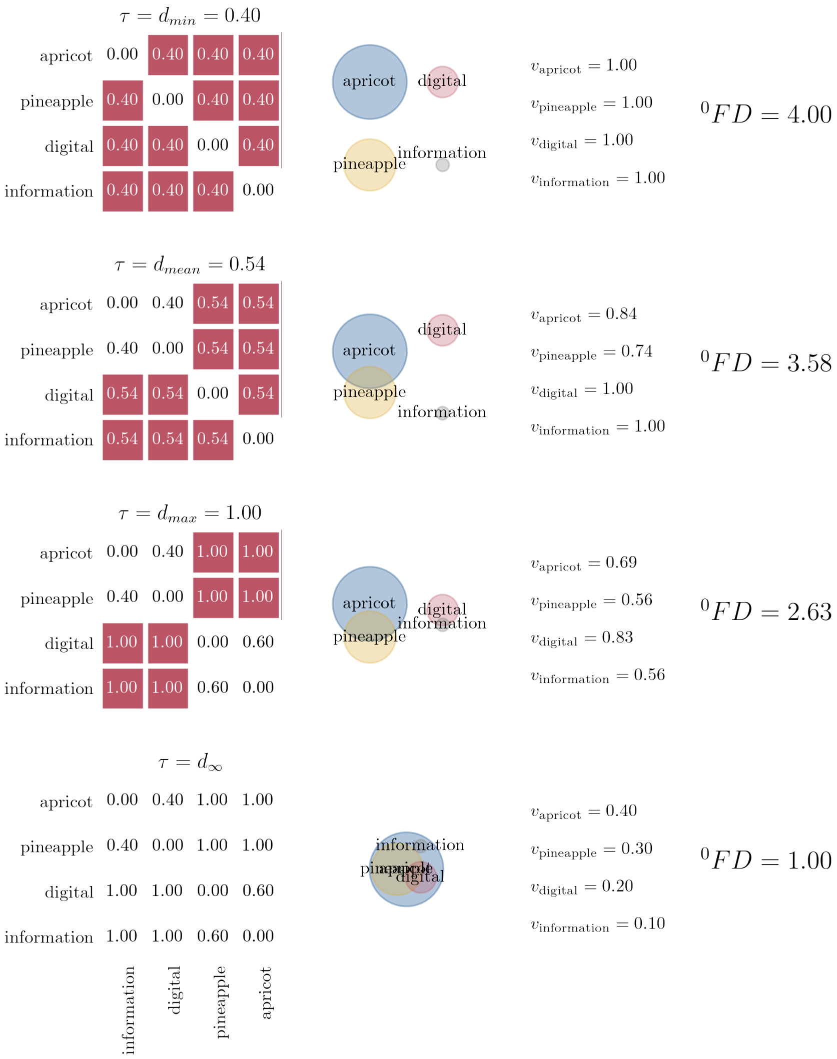
In Figure 1, we calculate functional diversity for different distinctiveness thresholds τ. We begin in the top row with $\tau = d_{\text{min}}$, which is equal to the minimum distance between different word types (i.e., intra-type distances are not considered in this example). At d_{min}, τ equals 0.4, which means that word types with at least a distance of 0.4 between them are considered functionally equally distant. This translates by truncating all distances greater than $\tau = 0.4$ to 0.4 in the distance matrix (cf. the matrix in Figure 1). In this scenario, each word is functionally equally distant and thus each type has a proportional contribution v_i of unity to its functional group, or, in other words, each type makes up for its own functional group. This is illustrated in Figure 1 with the circles whose size is proportional to their frequency. The circles do not overlap, which illustrates that they each comprise their own functional group. The functional diversity at $q = 0$, then, is $FD = 4$, which is simply the sum over the proportional contributions $v_i(\tau)$ of each type i to a functional group (cf. Eq. 1).

As the threshold value τ increases, an increasing number of types becomes functionally indistinguishable. In other words, with higher values of τ, functional groups consist of more types. Chiu and Chao [8, 6] suggest to use Rao's quadratic entropy Q for τ, which is a similarity-sensitive diversity measure representing the average distance between two randomly selected instances in a collection [22]. Q, hereafter denoted as d_{mean}, is expressed as:

$$Q = d_{\text{mean}} = \sum_{i=1}^{k} \sum_{j=1}^{k} d_{ij} p_i p_j, \tag{5}$$

where d_{ij} refers to the distance between types i and j, and p_i and p_j to their relative frequencies.

As shown in Figure 1, setting τ at $d_{\text{mean}} = 0.54$ decreases the functional diversity to $FD = 3.58$. At the threshold of 0.54, *apricot* and *pineapple* become functionally less distinct, contributing to a shared functional group (illustrated by the overlapping circles). By contrast,



$\tau = d_{min} = 0.40$

	information	digital	pineapple	apricot
apricot	0.00	0.40	0.40	0.40
pineapple	0.40	0.00	0.40	0.40
digital	0.40	0.40	0.00	0.40
information	0.40	0.40	0.40	0.00

$v_{\text{apricot}} = 1.00$
$v_{\text{pineapple}} = 1.00$
$v_{\text{digital}} = 1.00$
$v_{\text{information}} = 1.00$

$^0FD = 4.00$

$\tau = d_{mean} = 0.54$

	information	digital	pineapple	apricot
apricot	0.00	0.40	0.54	0.54
pineapple	0.40	0.00	0.54	0.54
digital	0.54	0.54	0.00	0.54
information	0.54	0.54	0.54	0.00

$v_{\text{apricot}} = 0.84$
$v_{\text{pineapple}} = 0.74$
$v_{\text{digital}} = 1.00$
$v_{\text{information}} = 1.00$

$^0FD = 3.58$

$\tau = d_{max} = 1.00$

	information	digital	pineapple	apricot
apricot	0.00	0.40	1.00	1.00
pineapple	0.40	0.00	1.00	1.00
digital	1.00	1.00	0.00	0.60
information	1.00	1.00	0.60	0.00

$v_{\text{apricot}} = 0.69$
$v_{\text{pineapple}} = 0.56$
$v_{\text{digital}} = 0.83$
$v_{\text{information}} = 0.56$

$^0FD = 2.63$

$\tau = d_{\infty}$

	information	digital	pineapple	apricot
apricot	0.00	0.40	1.00	1.00
pineapple	0.40	0.00	1.00	1.00
digital	1.00	1.00	0.00	0.60
information	1.00	1.00	0.60	0.00

$v_{\text{apricot}} = 0.40$
$v_{\text{pineapple}} = 0.30$
$v_{\text{digital}} = 0.20$
$v_{\text{information}} = 0.10$

$^0FD = 1.00$

Figure 1: Example of how functional diversity is operationalized. The figure, inspired by [6], shows for increasing values of τ (i.e., d_{min}, d_{mean}, d_{max}, d_{∞}), the corresponding truncated distance matrix $\Delta(\tau)$, an illustration of the overlap between functional groups, the proportional contribution of each word type to a functional group, and the total functional diversity.

at $\tau = d_{\text{mean}}$, *digital* and *information* remain functionally equally distant and as such belong to their own functional group. Note that with $FD = 3.58$, the functional diversity at $\tau = d_{\text{mean}}$ is larger than 3. This is because the co-occurrence profile of *apricot* and *pineapple* is considered partially overlapping but not identical. When τ is set to the maximum distance in the distance matrix ($\tau = d_{\text{max}} = 1$, however, *digital* and *information* contribute to a shared functional group. Note that when there are many different word types, $\tau = d_{\text{max}}$ is less informative than $\tau = d_{\text{mean}}$, because functional diversity is then often close to unity [cf. 6]. Finally, as τ tends to infinity, all words become part of the same functional group, which is expressed by a functional diversity of $FD = 1$.

3. Data and pre-processing

Archer Corpus For our experiments, we use ARCHER 3.2 [28], a corpus of historical English registers (3.3M words). The corpus covers a period of almost 400 years (1600-1999), and contains texts from 12 different genres or registers: advertisements, drama, fiction, sermons, journals, legal text, medicine, news, early prose, science, letters, and diaries. In terms of spelling, ARCHER 3.2 contains the original spelling of published editions normalized with VARD2 [4]. In contrast to many other historical corpora, ARCHER 3.2 is a well-balanced, cleaned (and relatively small) corpus, and hence it constitutes the ideal starting point for our experiment.

Simulating Errors To mimic different degrees of text noise, we 'pollute' each text in the clean ARCHER corpus by simulating errors. In this simulation procedure, each token of each text is modified with a probability p. The modification involves replacing each letter by a random ASCII letter with probability s. With $s = 0.2$, a word like `diversity` is replaced with `diversizy`. We experiment with $p \in 0, 0.1, 0.2, 0.35, 0.5, 0.75$, and chart the import of having a more distorted text on the stability of the diversity measures. [3] In all experiments s is set to 0.2. With this procedure, each text is manipulated five times per p value. The reported diversity measurements are computed by taking the mean diversity over these five different texts.

Embeddings For the present study, we use token-based embeddings to obtain semantic similarity estimates between the words in a given text. These embeddings are computed on the basis of MacBERTh [20, 19], a Large Language Model that follows the architecture of BERT-base uncased [10], which is pre-trained on historical English (1450-1950) using a custom vocabulary. Token-based embeddings are expected to be more robust than type-based embeddings in the presence of noise, since they take the sentential context in which the target word appears into account. This means that they can associate (even lower frequency) variants of the same word with each other, where the sentential context is expected to match. Moreover, thanks to the built-in adaptive tokenization approach, MacBERTh is also able to compute embeddings for words that were not seen during training, which is an invaluable feature for texts with large amounts of orthographic variation.[4]

In order to compute the type-level distance matrix between all word types in a corpus, we

[3] Note that texts resulting from $p = 0.75$ are perhaps less realistic than lower values. To illustrate, the OCR error rate in Eighteenth Century Collections Online (ECCO) has, for instance, been estimated to at approximately 25% [15].

[4] The purpose of this paper is to introduce the attribute diversity framework into lexical diversity research. As a first operationalization, we resorted to token-based embeddings, which is a theoretically sound choice (as these models are sensitive to the fact that words can have multiple meanings) that comes with certain practical advantages (with respect to lower-frequency and 'unseen' items). We are, however, interested in trying out other ways of operationalizing the concept of functional groups in future work. One possibility, for instance, would be to test and compare different implementations of implement semantic similarity, comparing type and token-based approaches.

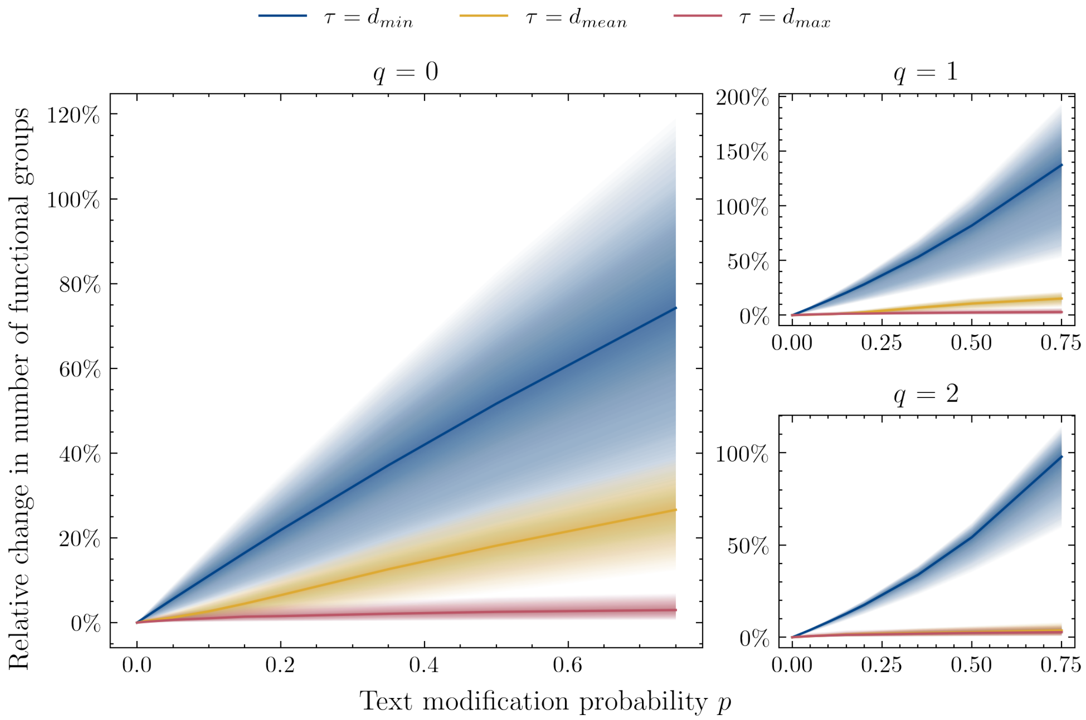


Figure 2: Relative change in the number of functional groups after modifying texts with probability p with respect to their unaltered counterparts ($p = 0$). The left panel displays the results for $q = 0$ (corresponding to functional richness), and the two smaller panels on the right present results for higher diversity orders (i.e., $q \in \{1, 2\}$, which put increasing weight on the frequency of word types.

first compute the token-embeddings of all words it contains.[5] If the same token appears multiple times in the input corpus, we compute a single embedding by averaging over the embeddings of all occurrences. Finally, we rely on the cosine distance function in order to obtain a distance value between 0 and 2.[6]

4. Results

4.1. Functional diversity is affected less by increased orthographic variation

Figure 2 shows the relative change in the number of functional groups after modifying texts with a text modification probability p with respect to their original, unmodified counterparts ($p = 0$). The left panel shows the values for $q = 0$ at three different thresholds of τ. As expected, the number of functional entities of functional entities at $\tau = d_{\text{min}}$ increases more or less

[5] Note that due to the tokenization approach of MacBERTh, input words are often split into smaller units (sub-tokens). In order to compute a single embedding in such cases, we average over the embeddings of the different sub-tokens.

[6] More specifically, the cosine distance is defined as 1 minus the cosine similarity of two given vectors—the latter being bounded between -1 and 1.

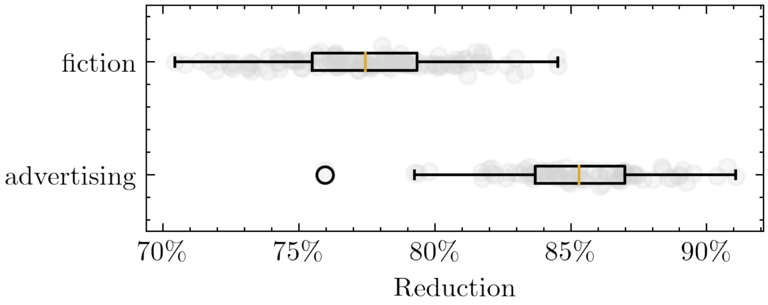


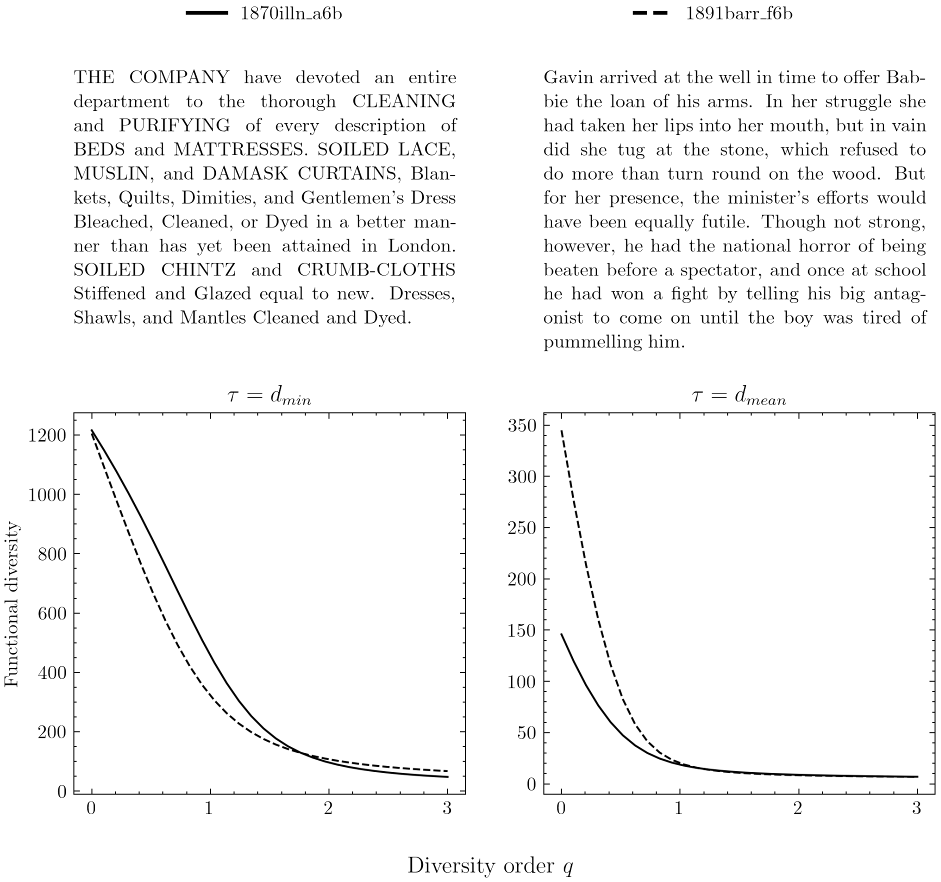
Figure 3: Box plots showing the reduction from d_{min} to d_{mean} in number of functional groups for advertisements and fiction texts.

linearly with the probability p of modifying words. Indeed, the probability of a modification yielding a orthographically unique letter combination is high, and each unique combination is taken to account for a new word type (i.e., a new functional group). The relative change in the number of functional groups is much less strong for $\tau = d_{\text{mean}}$, where the number of functional groups at extreme values of p is still relatively close to the number of groups at $p = 0$. Note that the same holds for $\tau = d_{\text{max}}$, which also remains stable with larger values of noise. However, as explained above, with $\tau = d_{\text{max}}$, estimates are often close to unity, which makes the stability of $\tau = d_{\text{max}}$ less surprising. The two right panels present the same results for higher diversity orders q. These plots show that when more weight is given to high frequency entities, functional diversity is also better able to cope with orthographic variation than lexical diversity at $\tau = d_{\text{min}}$.

4.2. Functional diversity is a theoretically relevant complement to lexical diversity

To get a firmer grip on what could be gained from integrating functional diversity estimates into discussions of lexical richness, we automatically identified text pairs with approximately the same number of unique word types (d_{min}), but a diverging number of functional groups at d_{mean}. In each of these text pairs, one text is functionally less 'condensed', using the same number of unique lexical items to cover a broader functional range. A commonly occurring type of text pairing, in that respect, is that of a fiction text with a text containing a collection of advertisements, where advertisements consistently cover a smaller number of functional groups despite being as lexically diverse as the paired fiction text. The relatively strong reduction from d_{min} to d_{mean} in advertising, illustrated in Figure 3, is intuitive, as advertisements often present a list of (functionally closely related) services and/or goods (see Figure 4), resulting in a more condensed diversity that suggests depth rather than breadth. For fiction, by contrast, there is no reason to expect a similar reduction.

Pairings of two texts from the same genre also emerged. A telling example is the pairing of *Isabel Clarendon* (1886), a fiction text by naturalist/realist author George Gissing, with *Caprice* (1917) by Ronald Firbank (see Figure 5). The excerpts in the corpus from both texts have roughly the same number of unique word types (*Caprice*: 1374 vs. *Isabel Clarendon*: 1377), but the types



THE COMPANY have devoted an entire department to the thorough CLEANING and PURIFYING of every description of BEDS and MATTRESSES. SOILED LACE, MUSLIN, and DAMASK CURTAINS, Blankets, Quilts, Dimities, and Gentlemen's Dress Bleached, Cleaned, or Dyed in a better manner than has yet been attained in London. SOILED CHINTZ and CRUMB-CLOTHS Stiffened and Glazed equal to new. Dresses, Shawls, and Mantles Cleaned and Dyed.

Gavin arrived at the well in time to offer Babbie the loan of his arms. In her struggle she had taken her lips into her mouth, but in vain did she tug at the stone, which refused to do more than turn round on the wood. But for her presence, the minister's efforts would have been equally futile. Though not strong, however, he had the national horror of being beaten before a spectator, and once at school he had won a fight by telling his big antagonist to come on until the boy was tired of pummelling him.

Figure 4: Excerpts from an advertisement collection (left; filename '1860illn_a6b') and fiction text (right; filename '1891barr_f6b') pairing with comparable estimates at d_{min} but diverging estimates at d_{mean}.

in *Caprice* – a minimalist novel that, unlike realist work, predominantly consists of dialogue and contains only limited descriptions of setting and character – cover a considerably smaller number of functional groups at $\tau = d_{\text{mean}}$ (213 vs. 368). Interestingly, with 5260 word tokens, the excerpt of *Isabel Clarendon* has a lower TTR than the excerpt of *Caprice*, which comprises 3753 tokens. Hence, the TTRs would suggest that *Caprice* covers more ground in fewer words. The functional diversity estimate, however, paints a different picture, which adds a theoretically relevant dimension to investigations into the lexical richness of texts.

5. Conclusion

In this short paper, we introduce a way of incorporating the notion of functional diversity into lexical diversity measurements in (historical) corpora by means of token-based embeddings.

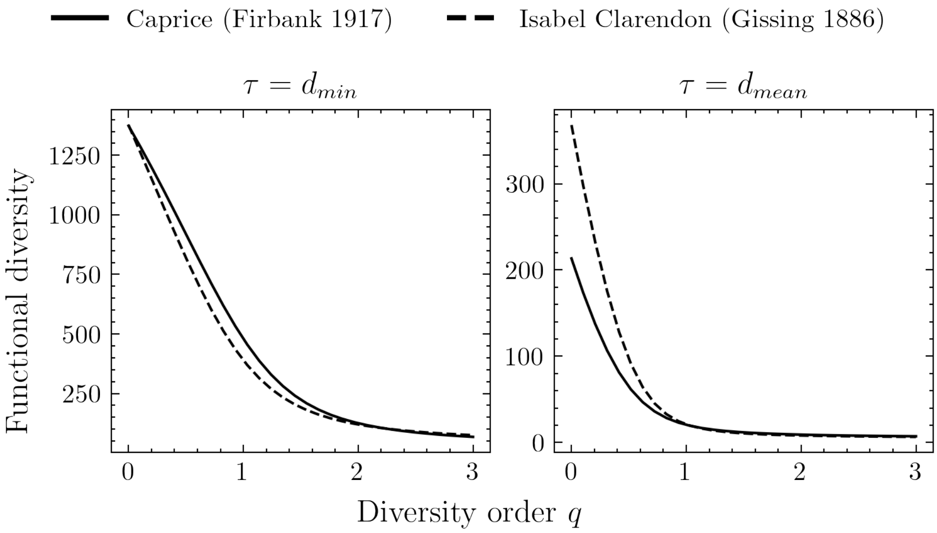


Figure 5: Comparison of the Hill number profiles at $\tau = d_{\text{min}}$ and $\tau = d_{\text{mean}}$ for two fiction texts *Isabel Clarendon* (1886) and *Caprice* (1917).

Our experiment shows that considering lexical diversity at the level of functional groups has the practical advantage of being less sensitive to orthographic noise in the data, and the theoretical advantage of adding an important and often disregarded dimension (capturing depth vs. breadth) of vocabulary diversity in textual data. As such, the framework of attribute diversity commonly used in Ecology should be considered an important addition to the Computational Humanities research toolkit.

Acknowledgments

The training of MacBERTh has been made possible by the Platform Digital Infrastructure (Social Sciences and Humanities) fund (PDI-SSH). We thank Melvin Wevers (University of Amsterdam) for his constructive feedback.

References

[1] L. Anthonissen. *Individuality in Language Change.* Berlin, Boston: De Gruyter Mouton, 2021. DOI: doi:10.1515/9783110725841.

[2] R. H. Baayen. "Corpus linguistics in morphology: Morphological productivity". In: *Corpus Linguistics: An International Handbook.* Ed. by A. Lüdeling and M. Kytö. Vol. 2. Berlin, New York: De Gruyter Mouton, 2009, pp. 899–919. DOI: doi:10.1515/9783110213881.2.899.

[3] J. Barðdal. *Productivity: Evidence from Case and Argument Structure in Icelandic.* Amsterdam, Philadelphia: John Benjamins, 2008.

[4] A. Baron and P. Rayson. “VARD2: A tool for dealing with spelling variation in historical corpora”. In: *Postgraduate conference in corpus linguistics*. 2008.

[5] M. Brysbaert, M. Stevens, P. Mandera, and E. Keuleers. “How Many Words Do We Know? Practical Estimates of Vocabulary Size Dependent on Word Definition, the Degree of Language Input and the Participant’s Age”. In: *Frontiers in Psychology* 7 (2016). DOI: 10.3389/fpsyg.2016.01116.

[6] A. Chao, C.-H. Chiu, S. Villéger, I.-F. Sun, S. Thorn, Y.-C. Lin, J.-M. Chiang, and W. B. Sherwin. “An Attribute-diversity Approach to Functional Diversity, Functional Beta Diversity, and Related (Dis)Similarity Measures”. In: *Ecological Monographs* 89.2 (2019). DOI: 10.1002/ecm.1343.

[7] A. Chao, N. J. Gotelli, T. C. Hsieh, E. L. Sander, K. H. Ma, R. K. Colwell, and A. M. Ellison. “Rarefaction and Extrapolation with Hill Numbers: A Framework for Sampling and Estimation in Species Diversity Studies”. In: *Ecological Monographs* 84.1 (2014), pp. 45–67.

[8] C.-H. Chiu and A. Chao. “Distance-Based Functional Diversity Measures and Their Decomposition: A Framework Based on Hill Numbers”. In: *PLoS ONE* 9.7 (2014). Ed. by F. de Bello, e100014. DOI: 10.1371/journal.pone.0100014.

[9] A. J. Daly, J. M. Baetens, and B. De Baets. “Ecological diversity: measuring the unmeasurable”. In: *Mathematics* 6.7 (2018), p. 119.

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423.

[11] B. Efron and R. Thisted. “Estimating the Number of Unseen Species: How Many Words Did Shakespeare Know?” In: *Biometrika* 63.3 (1976), p. 435. DOI: 10.2307/2335721.

[12] A. Ellegård. “Estimating Vocabulary Size”. In: *Word* 16.2 (1960), pp. 219–244. DOI: 10.1080/00437956.1960.11659728.

[13] L. Fonteyn and E. Manjavacas. “Adjusting scope: a computational approach to case-driven research on semantic change”. In: *Proceedings of the Workshop on Computational Humanities Research (CHR 2021)*. Vol. 2898. CEUR Workshop Proceedings. Amsterdam, 2021, pp. 280–298. URL: http://ceur-ws.org/Vol-2989/long%5C%5Fpaper26.pdf.

[14] M. O. Hill. “Diversity and Evenness: A Unifying Notation and Its Consequences”. In: *Ecology* 54.2 (1973), pp. 427–432.

[15] M. J. Hill and S. Hengchen. “Quantifying the Impact of Dirty OCR on Historical Text Analysis: Eighteenth Century Collections Online as a Case Study”. In: *Digital Scholarship in the Humanities* 34.4 (2019), pp. 825–843. DOI: 10.1093/llc/fqz024.

[16] D. L. Hoover. “Another Perspective on Vocabulary Richness”. In: *Computers and the Humanities* 37.2 (2003), pp. 151–178. DOI: 10.1023/a:1022673822140.

[17] L. Jost. “Entropy and diversity”. In: *Oikos* 113.2 (2006), pp. 363–375.

[18] M. Kubát and J. Milička. “Vocabulary Richness Measure in Genres”. In: *Journal of Quantitative Linguistics* 20.4 (2013), pp. 339–349. DOI: 10.1080/09296174.2013.830552.

[19] E. Manjavacas and L. Fonteyn. “Adapting vs. Pre-training Language Models for Historical Languages”. In: *Journal of Data Mining & Digital Humanities* Nlp4dh (2022). DOI: 10.46298/jdmdh.9152.

[20] E. Manjavacas and L. Fonteyn. “MacBERTh: Development and Evaluation of a Historically Pre-trained Language Model for English (1450-1950)”. In: *Proceedings of the Workshop on NLP4DH ICON 2021.* online: NLP Association of India (NLPAI), 2021.

[21] F. Perek. “Recent change in the productivity and schematicity of the *way* -construction: A distributional semantic analysis”. In: *Corpus Linguistics and Linguistic Theory* 14.1 (2018), pp. 65–97. DOI: 10.1515/cllt-2016-0014.

[22] C. R. Rao. “Diversity and dissimilarity coefficients: a unified approach”. In: *Theoretical population biology* 21.1 (1982), pp. 24–43.

[23] A. Riba and J. Ginebra. “Diversity of vocabulary and homogeneity of literary style”. In: *Journal of Applied Statistics* 33.7 (2006), pp. 729–741. DOI: 10.1080/02664760600708970.

[24] H.-J. Schmid and A. Mantlik. “Entrenchment in Historical Corpora? Reconstructing Dead Authors’ Minds from their Usage Profiles”. In: *Anglia* 133.4 (2015), pp. 583–623. DOI: doi:10.1515/ang-2015-0056.

[25] J. Segbers and S. Schroeder. “How many words do children know? A corpus-based estimation of children’s total vocabulary size”. In: *Language Testing* 34.3 (2017), pp. 297–320. DOI: 10.1177/0265532216641152.

[26] C. E. Shannon. “A Mathematical Theory of Communication”. In: *Mobile Computing and Communications Review* 5 (I 1948), p. 53.

[27] F. J. Tweedie and R. H. Baayen. “How Variable May a Constant be? Measures of Lexical Richness in Perspective”. In: *Computers and the Humanities* 32.5 (1998), pp. 323–352. DOI: 10.1023/a:1001749303137.

[28] N. Yáñez-Bouza. *ARCHER 3.2: A Representative Corpus of Historical English Registers.* https://www.projects.alc.manchester.ac.uk/archer/wp-content/uploads/2020/06/ARCHER_poster.pdf. 2013.