# Function Words in Authorship Attribution Studies

**Antonio Miranda García**
Department of English, School of Engineering, University of Málaga, Spain

**Javier Calle Martín**
Department of English, Faculty of Arts, University of Málaga, Spain

## Abstract

The search for a reliable expression to measure an author's lexical richness has constituted many statisticians' holy grail over the last decades in their attempt to solve some controversial authorship attributions. The greatest effort has been devoted to find a formula grounded on the computation of tokens, word-types, most-frequent-word(s), *hapax legomena*, *hapax dislegomena*, etc., such that it would characterize a text successfully, independent of its length. In this line, Yule's $K$ and Zipf's $Z$ seem to be generally accepted by scholars as reliable measures of lexical repetition and lexical richness by computing content and function words altogether.[1] Given the latter's higher frequency, they prove to be more reliable identifiers when isolatedly computed in p.c.a. and Delta-based attribution studies, and their rate to the former also measures the *functional density* of a text. In this paper, we aim to show that each constant serves to measure a specific feature and, as such, they are thought to complement one another since a supposedly rich text (in terms of its lemmas) does necessarily have to characterize by its low functional density, and vice versa. For this purpose, an annotated corpus of the *West Saxon Gospels* (*WSG*) and *Apollonius of Tyre* (*AoT*) has been used along with a huge raw corpus.

**Correspondence:**
Javier Calle Martín, University of Málaga, Faculty of Arts, Department of English, Campus de Teatinos, Málaga 29071, Spain.
**E-mail:**
jcalle@uma.es



## 1 Introduction

The task of attributing the plausible authorship of a literary work resembles that of an art historian anxious to find even the least salient traces in the works of a plastic artist (either an architect, a sculptor, a painter, or a designer), insofar as the more the former knows about the latter, the more connections will undoubtedly lead him/her to relate work and author. The traditional methodology in authorship attribution employs differences and likenesses to ascertain the linguistic profile of the author *behind the curtain*. However, the issue is not that straightforward in non-traditional approaches either, as these connections are often hard to validate.

*Lexical richness* may be accepted as a point of departure for identifying the stylistic idiosyncrasy of an author and, hence, for authorship attribution 'as they both (stylometrics and authorship attribution studies) share an interest in the size, coherence, and distribution of the vocabularies of texts and authors' (Hoover, 2003). Most stylometric studies are lexically based, especially because it is the level of language where repetitions may be reliably used as a basis for measurement (Holmes, 1994). In fact,



doi:10.1093/llc/fql048 Advance Access Published on 14 November 2006

the language items of a work are assumed to entrap both conscious and unconscious aspects of the author's style; it would then be an asset to find the features of the unconscious aspect, since they cannot be consciously manipulated by the author (Laan, 1995).

The relevant literature reveals that divergent opinions can be attested as regards the numerous authorship attribution studies that, over the last years, have tried to obtain the most faithful statistical procedure, be the items morphological, syntactic, or lexical. In particular, the validity of some techniques employed in authorship attribution has been, to some extent, questioned with the argument that there is 'no consensus on results, nor there is consensus as to the accepted or correct methodology or technique', notwithstanding the persistent 'wish for a single method of analysis and a verifiable unique style' (Rudman, 1998). For the state of the art, we briefly review the contributions by Tweedy and Baayen, Somers and Tweedy, Burrows, and Hoover.

From a positive standpoint, Tweedie and Baayen (1998) consider that textual similarities between and within authors may become valid for authorship attribution, as they seem to validate the empirical trajectories of texts in the plane spanned by the dimensions of *lexical richness* and *lexical repetition*, providing thus a statistical technique for constructing confidence intervals around them. In relation to their experiment, Hoover (2003) points out that the success greatly depends on the variety of texts used (thematic and chronological), because the results would have been radically different, had the selection been also different. Actually, he is rather sceptical about the possibility of relying on a one-to-one connectedness between *lexical richness* and authorship attribution and/or style description, on account of the following words: 'even though the measures of vocabulary richness may be effective in clustering texts by a same author, they cannot provide a consistent, reliable, or satisfactory means of identifying an author or describing a style' (Hoover, 2003), to conclude that 'the long-cherished goal of the measure of vocabulary richness that characterizes authors and their styles appears to be unattainable. The basic assumption that underlies it is false'. Somers and Tweedie (2003) also distinguish a pastiche from the original by means of a discriminating study based on word usage and *lexical richness*. Lastly, Burrows's Delta (2002; 2003) performs rather successfully as the rate of unlikeness is reliable for relating English Restoration poems, both of known and unknown authors. The reliability and effectiveness of Burrows's *Delta* has been assessed by Hoover (2004a) who, in turn, tests successful modifications and alternatives to upgrade the model.

For opportunity, we refer to Miranda and Calle's multivariate word and lemma-based study (2005a) of the *West Saxon Gospels* (*WSG*), which was accomplished to rank them in terms of *lexical richness* measured by lemma-counting. The technique has proven reliable in Old English texts (a highly inflected language), as the same ordering of works by lemma-counting in the main study, namely $Mt > Lk > Mk > Jn$, is replicated in the partitioning and the contrastive studies carried out, and similar results are obtained by evaluating Zipf's $Z$ or the expression $HL^3/V(N)^2$.

The present research aims to replicate the above-mentioned ranking of the *WSG* by using *function words* (*FW*) in view of the consensus around their greater reliability in stylometric studies (Burrows, 2003) and their subsequent success for authorship attribution. Therefore, our hypothesis is grounded on the assumption that the particular distribution of *FW* may constitute the salient features which can identify an author's style and differentiate it from others. Three techniques are experimented to check this hypothesis, and each of them gives rise to each part of this article, besides the Introduction and the methodological procedure themselves. The first part deals with the computation of *functional density*, the second with the scoring of the most common *function lemmas* (*FL*), and the third with the Delta-scoring of the most common *FW*. To finish, some conclusions are drawn, especially to indicate whether each model (*functional density, scoring of FL, and Delta-scoring of FW*) yields the same ranking, or whether the types/tokens approach can be complemented or optimized with studies based on *FL* and/or *FW*.

## 2 Methodology

The field study comprises two different corpora as well as two handling techniques. The computation of *functional density* and the scoring of *FL* were done with the annotated corpus containing the *WSG* and the Old English *Apollonius of Tyre* (*AoT*) used in Miranda and Calle (2005a). The benefits derived from the annotation compensate for such a time-consuming task, though it supposes 'the regrettable intrusion upon the data and avoids the interchange of information with colleagues' (Burrows, 2003), or even 'experimental corruption' (Rudman, 1998).[2] In the case of OE, unless you handle an annotated text, you cannot distinguish between (1) vowel-length minimal pairs when no length marker is allocated, as *gē* versus *ge*, *fōr* versus *for*, etc; (2) homographs as *þā* which may be nominative/accusative plural of all genders of {*se, sēo, þæt*}, accusative singular feminine of {*se, sēo, þæt*}, adverb of time, and/or conjunction of time.

The annotated corpus was then processed with the *Old English Concordancer* (henceforth *OEC*) (Miranda *et al*., 2006), which retrieves morpho-syntactic, lexical, and statistical information, both word- and lemma-based. Using this tool, all the sorting and counting operations were performed automatically, and the data were then downloaded onto Excel spread-sheets to accomplish the statistical study, and eventually plot the corresponding figures.

The study of *FW*, however, required a huger corpus for the sake of standardization, so a set of archives from the *Toronto Complete Corpus of Old English Dictionary* (Healey, 2000) was accordingly processed by means of *WordSmith Tools 3.0* (Scott, 1996). Leaving aside the matter of size, which constitutes its main advantage, to cope with an untagged corpus of Old English (henceforth OE) is not free of intrusion either, in view of the fact that you cannot automatically compute the occurrences of allomorphs which are likely to be taken as different words owing to (1) the alternative employment of ‹þ›, ‹Þ›, ‹ð›, ‹Đ› in, for instance, ‹þā›, ‹Þā›, ‹ðā›, ‹oþþe›, ‹oÞðe›, ‹oðþe›, ‹oððe›, etc; (2) the existence of inter- and/or intradialectal variants, such as ‹angēan›, ‹ongēan›, ‹agēan›, ‹hī›, ‹hīe›, ‹hig›, ‹hȳ›; and (3) the equivalence of ‹and›, ‹ond›, and the Tyronian nota, or of the forms ‹betwēonan›, ‹betwēonon›, ‹betweox›, ‹betwix›, ‹betwux›, ‹betwyx›.

To overcome these inconveniences, the results generated by *WordSmith Tools* required further replacement and grouping of the corresponding allomorphs. As in the annotated corpus, the data were also downloaded onto Excel spread-sheets for their statistical analysis.

## 3 Functional Density: the Ratio of Function Words to Content Words

The objective of this first study is to rank the *WSG* by the ratio *FW* to *content words* (*CW*), which requires to state what we understand by *FW* and *CW* in the context of OE: the set of the latter is assumed to comprise nouns, adjectives, verbs, and adverbs of manner (mostly derived with the suffix –(*līc*)*e*, such as *sōþlīce*, *witodlīce*, etc.), whereas the set of the former is assumed to comprise the rest (pronouns, prepositions, conjunctions, interjections, and other-than-manner adverbs). Although the line distributing adverbs into one set or another looks blurred, our proposal seems the most plausible insofar as only adverbs of manner ending in –(*līc*)*e* share their stems with *CW* and as such they must be considered to belong to the same open set.[3] To help understand the scope of the ratio *FW*/*CW*, let us compare it with other common related expressions, namely: (word-based) *lexical richness* = *V*/*N*, (lemma-based) *lexical richness* = *L*/*N*, *inflectional diversity* (*V*/*L*), etc., in the following examples. Examples (1) and (2) illustrate two purposely created eight-word sentences, examples (3)–(7) contain each of the sentences of a real abstract, and (8) the whole abstract (see Note 4 for the text and reference).

*That that that that student used was wrong.* (1)
*False friends lead foreign students to make mistakes.* (2)

First, the occurrences are counted and the results tabulated as in Table 1.

Next, the expressions are computed and their results illustrated in Table 2.

**Table 1** Counting of words and lemmas in the examples

| | Tokens (*N*) | Word-types (*V*) | Lemmas (*L*) | Most frequent words (*MFW*) | Hapax Legomena (*HL*) | Hapax Dislegomena (*HD*) | Content Words (*CW*) | Function Words (*FW*) |
|---|---|---|---|---|---|---|---|---|
| 1 | 8 | 5 | 5 | 4 | 4 | 0 | 4 | 5 |
| 2 | 8 | 8 | 8 | 1 | 8 | 0 | 7 | 1 |
| 3 | 44 | 39 | 39 | 3 | 35 | 3 | 26 | 18 |
| 4 | 15 | 14 | 14 | 2 | 13 | 1 | 8 | 7 |
| 5 | 32 | 27 | 27 | 3 | 23 | 3 | 16 | 16 |
| 6 | 32 | 27 | 27 | 3 | 23 | 2 | 15 | 17 |
| 7 | 18 | 17 | 17 | 2 | 16 | 1 | 10 | 8 |
| 8 | 141 | 92 | 89 | 7 | 67 | 14 | 75 | 66 |

**Table 2** Evaluation of the expressions

| | Word-based lexical richness *V/N* | Lemma-based lexical richness *L/N* | Inflectional diversity *V/L* | Functional density *FW/CW* |
|---|---|---|---|---|
| 1 | 0.625 | 0.625 | 1 | 1 |
| 2 | 1 | 0.625 | 1.6 | 0.142857143 |
| 3 | 0.909090909 | 0.909090909 | 1 | 0.517241379 |
| 4 | 1 | 1 | 1 | 0.875 |
| 5 | 0.875 | 0.875 | 1 | 0.882352941 |
| 6 | 0.875 | 0.875 | 1 | 1.133333333 |
| 7 | 0.944444444 | 0.944444444 | 1 | 0.384615385 |
| 8 | 0.680851064 | 0.638297872 | 1.066666667 | 0.698795181 |

**Table 3** Counting of *N* in terms of class

| Class | *Jn* | *Lk* | *Mk* | *Mt* | *AoT* |
|---|---|---|---|---|---|
| Noun | 2,960 | 3,844 | 2,172 | 3,943 | 1,282 |
| Adje(ctive) | 353 | 713 | 418 | 685 | 348 |
| Verb | 3,938 | 4,833 | 2,817 | 4,477 | 1,330 |
| Adve(rb) | 1,353 | 1,584 | 1,114 | 1,664 | 610 |
| Pron(oun) | 5,367 | 5,972 | 3,459 | 5,646 | 1,766 |
| Prep(osition) | 1,295 | 1,638 | 896 | 1,513 | 533 |
| Conj(unction) | 1,805 | 2,369 | 1,456 | 2,262 | 656 |
| Excl(amation) | 5 | 36 | 18 | 40 | 0 |
| *N* | 17,076 | 20,989 | 12,350 | 20,230 | 6,525 |

In view of the results, it seems obvious that (2) is lexically richer than (1), whilst the latter happens to be denser than the former in functional terms. It goes without saying that *V/N* (Mendenhall's formula) and *L/N* (the corresponding expression in terms of lemmas) are not reliable for comparison unless the same *N* is involved, since they are text-length dependent as the relevant literature shows (Tweedy and Baayen, 1998).[5]

The maximum value for *V/N* and for *L/N* (2, 4) is the unity, although the average is less than one, *L/N* being always equal or less than *V/N*. In contrast, when *V* and *L* are rated to calculate the *inflectional diversity* (i.e. the average of word-forms per lemma), its minimum value is the unity, which is bound to increase in highly inflected languages such as Latin, OE, or Modern German. In an intermediate position, the ratio *FW/CW* ranges from below the unity to over one, likely to provide a greater discriminating power, which the other expressions cannot confer: when *N* is the same in some texts (1, 2) and (5, 6) and, therefore, no text-length dependency can be invoked, *functional density* can distinguish between them whereas the other ratios cannot. This observation made us consider the convenience of testing its reliability. In analogy to *lexical density*, which is undoubtedly bound to imply either *lexical richness* (*V/N*) or the ratio *CW/N*, we will henceforth adopt the term *functional density* to refer to the ratio *FW/CW*, avoiding thus any kind of likely terminological misunderstanding.

Once the discriminating power of *functional density* has been introduced, we need to determine the values of *FW* and *CW* in the works of the corpus to evaluate the *functional density*. For this purpose, *N* is first counted in terms of class (Table 3); then adverbs are sorted into manner and other-than-manner adverbs by means of the *OEC* (Table 4); and lastly the ratio is obtained (Table 5).

**Table 4** Counting of adverbs

| Adverbs | *Jn* | *Lk* | *Mk* | *Mt* | *AoT* |
|---|---|---|---|---|---|
| Manner | 109 | 138 | 99 | 371 | 175 |
| <>Manner | 1,244 | 1,446 | 1,015 | 1,293 | 435 |
| *Total* | 1,353 | 1,584 | 1,114 | 1,664 | 610 |

**Table 5** Functional density

| Words | *Jn* | *Lk* | *Mk* | *Mt* | *AoT* |
|---|---|---|---|---|---|
| *CW* | 7,360 | 9,528 | 5,506 | 9,476 | 3,135 |
| *FW* | 9,716 | 11,461 | 6,844 | 10,754 | 3,390 |
| *Functional density* | 1.320 | 1.202 | 1.243 | 1.134 | 1.081 |

With the values for *functional density*, the *WSG* are ranked in decreasing order as follows:

$$Jn > Mk > Lk > Mt > AoT$$

where *Jn* (*Mk* preceding) is ranked as having the greatest *functional density* and, conversely, *AoT* (*Mt* preceding) is ranked as having the lowest *functional density*, which proves our hypothesis of relating the ranking of the *WSG* by *functional density* as the inverse of the one by *lexical richness* (Miranda and Calle, 2005a), as follows:

$$AoT > Mt > Lk > Mk > Jn$$

This proof confirms the reliability of *FW* as salient features in authorship attribution tasks as we have been successful in opposing *Jn* against *Mt*, as having the highest *inflectional density* and the highest *lexical richness*, respectively; and conversely, *Mt* and *Jn* as having the lowest *inflectional density* and the highest *lexical richness*. In other words, we could somewhat oppose *inflectional density* against *lexical richness* without considering the latter as being actually the inverse of the former due to the different scope of the parameters applied: whereas all members of *V* and *L* are included in *N* (inclusion), all members of *CW* are excluded from *FW* (exclusion).

## 4 Score of Function Lemmas

Unlike *functional density*, whose computation requires to tabulate the whole set of *CW* and *FW*, the scoring of *FL* only takes into account the occurrences of the most common ones. This study is based on lemmas for the convenience of having homographs regularized, allomorphs grouped together, verb forms, and pronominal forms linked under the corresponding lemma, be they personal, demonstrative, or possessive. For this reason, the same annotated corpus is used and the seventy-eight topmost common *FL* are chosen from the union set {*Jn*, *Lk*, *Mk*, *Mt*, *AoT*}, as listed in Table 6.

As expected, the ordering of *FL* by absolute or relative frequency does differ slightly from one work to another, the differences becoming greater between *Jn*, on the one hand, and *Lk*, *Mk*, *Mt*, on the other, even in the three topmost *FL* (see *and*, *hē*, *se*, etc.). Therefore, to count on another reliable technique of ranking by means of *FL*, we resort to find out the difference between the relative frequency of each *FL* in each text and its frequency in the union set, which is taken as an ideal corpus on the assumption that their distribution and repetition rate is the standard. In other words, the greater the deviation found in a given text to the standard (which implies more occurrences of *FL* than in the ideal corpus), the less rich it will be in lexical terms and, conversely, the smaller the deviation found in a literary work, the richer it will be from a lexical point of view.

The calculation procedure is explained stepwise, as follows. First, we score each *FL* in each of the *WSG* and in *AoT* (i.e. as the percentage resulting from *frequencies/N*). Second, we score each *FL* in the ideal corpus, constituted by the union set {*Jn, Lk, Mk, Mt, AoT*} (i.e. also as the percentage of the *total of lemma frequencies/N total*, rather than as the mean of the individual percentages). Next, we subtract the score of each *FL* in each work to the score of each item in the ideal corpus. Fourth, we add up the absolute values of the differences in each text and divide the result by the number of function lemmas employed (78), as shown in Table 7 and fifth, we rank the texts in decreasing order, either by the sum or by the quotient, as follows:

$$Mt < Lk < Mk < Jn < AoT$$

**Table 6** Frequencies of 78 topmost common *FL* in the corpus

| | *FL* | *Jn* | *Lk* | *Mk* | *Mt* | *AoT* | | *FL* | *Jn* | *Lk* | *Mk* | *Mt* | *AoT* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | and | 957 | 1,502 | 1,002 | 1,414 | 364 | 40 | ofer | 20 | 59 | 27 | 59 | 1 |
| 2 | hē | 1,030 | 1,401 | 912 | 1,081 | 334 | 41 | nū | 49 | 38 | 12 | 33 | 26 |
| 3 | se | 1,102 | 1,152 | 695 | 1,317 | 481 | 42 | þus | 18 | 25 | 24 | 54 | 23 |
| 4 | bēon | 723 | 907 | 415 | 793 | 155 | 43 | hū | 23 | 37 | 24 | 36 | 1 |
| 5 | hīe | 410 | 619 | 497 | 547 | 38 | 44 | forþām | 46 | 40 | 22 | 30 | 14 |
| 6 | þā | 308 | 667 | 370 | 530 | 192 | 45 | hwilc | 16 | 41 | 24 | 34 | 16 |
| 7 | on | 359 | 583 | 320 | 507 | 111 | 46 | būtan | 28 | 23 | 24 | 35 | 7 |
| 8 | ic | 695 | 386 | 146 | 350 | 163 | 47 | twēgen | 16 | 29 | 21 | 42 | 6 |
| 9 | gē | 415 | 365 | 160 | 497 | 22 | 48 | foran | 13 | 47 | 17 | 30 | 10 |
| 10 | tō | 389 | 314 | 150 | 314 | 150 | 49 | manig | 21 | 26 | 30 | 30 | 1 |
| 11 | þe | 340 | 342 | 152 | 297 | 99 | 50 | eft | 49 | 6 | 33 | 23 | 7 |
| 12 | ne | 338 | 312 | 172 | 304 | 55 | 51 | ēower | 27 | 31 | 4 | 36 | 6 |
| 13 | þū | 267 | 305 | 119 | 306 | 103 | 52 | oþþe | 6 | 24 | 12 | 48 | 4 |
| 14 | þæt | 329 | 277 | 155 | 121 | 121 | 53 | ǣlc | 20 | 31 | 3 | 35 | 1 |
| 15 | þes | 172 | 187 | 83 | 169 | 52 | 54 | ūt | 32 | 13 | 11 | 33 | 9 |
| 16 | eall | 55 | 157 | 89 | 122 | 43 | 55 | hwā | 32 | 19 | 15 | 22 | 10 |
| 17 | of | 131 | 146 | 28 | 115 | 15 | 56 | æfter | 24 | 32 | 12 | 20 | 26 |
| 18 | sōþlīce | 29 | 111 | 83 | 191 | 121 | 57 | ǣr | 25 | 21 | 13 | 23 | 16 |
| 19 | hit | 103 | 92 | 82 | 126 | 21 | 58 | nāht | 26 | 17 | 19 | 20 | 16 |
| 20 | hēo | 74 | 117 | 76 | 71 | 111 | 59 | oþ | 6 | 24 | 11 | 36 | 5 |
| 21 | mid | 78 | 141 | 98 | 115 | 110 | 60 | þurh | 13 | 18 | 11 | 30 | 10 |
| 22 | ān | 56 | 92 | 63 | 109 | 23 | 61 | ongēan | 8 | 20 | 20 | 23 | 12 |
| 23 | wē | 92 | 79 | 49 | 91 | 17 | 62 | æt | 27 | 14 | 7 | 14 | 11 |
| 24 | swā | 69 | 112 | 97 | 128 | 62 | 63 | wiþ | 15 | 24 | 4 | 15 | 5 |
| 25 | þonne | 50 | 86 | 38 | 101 | 20 | 64 | gyt | 14 | 16 | 10 | 10 | 3 |
| 26 | mīn | 148 | 65 | 39 | 54 | 54 | 65 | hēr | 14 | 20 | 15 | 18 | 4 |
| 27 | hwæt | 53 | 85 | 52 | 57 | 26 | 66 | hwī | 14 | | 16 | 15 | 2 |
| 28 | gif | 80 | 56 | 30 | 77 | 21 | 67 | betwux | 4 | 18 | 14 | 13 | 5 |
| 29 | þīn | 43 | 76 | 44 | 69 | 33 | 68 | sōna | 4 | 14 | 35 | 7 | 10 |
| 30 | ac | 102 | 41 | 41 | 46 | 37 | 69 | git | 9 | 9 | 16 | 12 | |
| 31 | witodlīce | 77 | 27 | 16 | 103 | | 70 | hwǣr | 24 | 7 | | 7 | 4 |
| 32 | fram | 36 | 66 | 28 | 72 | 20 | 71 | ūre | 18 | 23 | 5 | 3 | 7 |
| 33 | sum | 26 | 93 | 42 | 41 | 11 | 72 | þanon | | 4 | 22 | 2 | 2 |
| 34 | nān | 82 | 49 | 28 | 38 | 9 | 73 | twelf | 6 | 12 | 15 | 12 | |
| 35 | be | 69 | 51 | 24 | 32 | 14 | 74 | ēac | 15 | 7 | 3 | 5 | 9 |
| 36 | þǣr | 65 | 41 | 36 | 62 | 31 | 75 | swīþe | 5 | 5 | | 16 | 10 |
| 37 | forþāmþe | 58 | 69 | | 36 | | 76 | intō | 6 | 11 | 3 | 11 | 9 |
| 38 | ōþer | 33 | 48 | 24 | 50 | 6 | 77 | swilce | 6 | 13 | 3 | 7 | 8 |
| 39 | for | 48 | 26 | 28 | 42 | 22 | 78 | ge | 3 | 6 | 1 | | 7 |

**Table 7** Sum and quotients of the differences of 78 topmost *FL*

| | *Jn* | *Lk* | *Mk* | *Mt* | *AoT* |
|---|---|---|---|---|---|
| Sum | 13.79447508 | 7.213282449 | 12.89773301 | 6.495542931 | 22.72113168 |
| Quotient | 0.174613609 | 0.091307373 | 0.163262443 | 0.082222062 | 0.287609262 |

In view of the ranking, we can state that *Mt* is the least unlike the ideal corpus, which is also applicable to *Lk* (second after *Mt*), and so on till *AoT*, which ranks as the most unlike. Recalling that the *WSG* are ranked as $Mt > Lk > Mk > Jn$ in terms of *lexical richness*, we must admit that our hypothesis holds for the ranking of the *WSG*, whose order is as expected, and the discriminating value of the scoring of *FL* might be accepted in a near future if other authorship attribution studies validate its general reliability. In this particular case, some arguments may question its validity. One has to do with the limited size of the ideal corpus, which might be criticized as not being wholly representative, and the other can be established in the sense that the level of (dis)similitude with respect to the ideal corpus derives from the nature of its origin: the union of the constituent subsets. Therefore, to rule out any randomness, the technique must be experimented with larger corpora in English (Middle, Modern, or Contemporary) with the inherent inconvenience of their annotation.

## 5 Delta-score of Function Words

In this third study, we attempt to check the hypothesis again by testing the Delta model, and, to gain accuracy, the annotated corpus (77,170 tokens) used in Sections 3 and 4 is replaced with a raw corpus (934,719 tokens, approximately one-fourth of the whole *OE* extant production) where lemmas are replaced by word-types, arresting thus the comparison and exchange of frequencies.[6] This corpus (henceforth *entire corpus*) is formed by representative prose works (838,795 tokens) and a selection of poetry (95,924 tokens). The latter is included to compensate for any likely overweighting influence of the prose works selected, which are either authored or anonymous, some of which are translations from Latin.[7]

First, a frequency hierarchy for *FW* is established (Table 8) to contain the 108 topmost common *FW*, ending with those words that occur 3 times in every 2,000. Their frequencies, which result as the sum of susceptible individual function words, that is, ‹and›, ‹þa›, ‹þæt/þat›, ‹on›, ‹he›, ‹þe/þec›, etc., regardless of context, amount to 440,112 tokens (47.084%) of the *entire corpus*, after having skipped over 57,107 tokens, which are related to *CW*.[8] The homographs are not tagged.[9]

Out of the works of the *entire corpus*, we select some for the *main* and others for the *test corpus*, taking into account that the main objective is to find out the level of (un)likeness between pairs of related works from each corpus. Thus, *Jn* and *Mt*, being the most unlike, are included in the *main corpus*, whereas *Lk* and *Mk* will be used in the test corpus. Also with a similar criterion, works by the same author (*Ælfric*) or similar versions of the same work (the Benedictine Rule for monks and for nuns) are assigned to each corpus to check their level of (un)likeness, thus following Hoover's suggestion of using works of known authorship to test the efficiency of Delta (Hoover, 2004b). The other works are included to serve as dummies. Accordingly, the distribution of works in each corpus is shown in Table 9, and the occurrences of each *FW* in the hierarchy are counted and added up ($\sum$). For space, only the total values are provided along with the ratio *FW*/*N*, the *mean* ($\mu$), and the *standard deviation* ($\sigma$).

The procedure for calculation of Burrows's Delta (2003) is described stepwise, as follows. First, we score each *FW* in each work, as well as in the *main corpus*, where the scores are given as percentages of the frequencies of each *FW* to the total word-count (*N*), so that the larger do not exert an undue influence. In addition, we calculate $\mu$ and $\sigma$ in the *main corpus*, as indicated in Table 10, which contains, in descending order, the values for the thirty-six top *FW* in *Jn*, *Mt*, *VRS*, *BenR*, *Æhom1*, *Whom*, and *ÆCHom2*, though the list expands downwards for 108 *FW*. Thus, for *and*, $\mu$ is 5.016501463, and $\sigma$ is 1.602026783, respectively, and the score for *Jn* is 5.573751316, for *Mt* is 6.963542941, and so on. The scores result from rating the occurrences of such *FW* to *N*: in the case of *Jn*, $and = 953$ and $N = 17.098$. For space, the scores of only six works are included in the table.

Second, we calculate the *z-score* of each *FW* by subtracting its score in each work from $\mu$ in the

**Table 8** Frequencies of 108 topmost *FW* in the *entire corpus*

| | *FW* | Tokens | | *FW* | Tokens | | *FW* | Tokens |
|---|---|---|---|---|---|---|---|---|
| 1 | and/ond[10] | 44,571 | 37 | heo/hio | 2,837 | 73 | na | 1,100 |
| 2 | þa | 24,879 | 38 | me/mec | 2,832 | 74 | þas | 1,097 |
| 3 | þæt/þat | 23,451 | 39 | ealle | 2,716 | 75 | soþlice | 1,086 |
| 4 | on | 21,128 | 40 | gif/gyf | 2,648 | 76 | oþre | 1,070 |
| 5 | he | 19,986 | 41 | eac | 2,598 | 77 | þis | 1,038 |
| 6 | þe/þec | 19,371 | 42 | þær/þar | 2,574 | 78 | þy/þi | 1,026 |
| 7 | to | 15,980 | 43 | ær | 2,465 | 79 | þus | 1,020 |
| 8 | se | 12,542 | 44 | us/usic | 2,455 | 80 | hire/hyre/hiere | 983 |
| 9 | his/hys | 11,938 | 45 | seo/sio | 2,320 | 81 | þæra/þara | 978 |
| 10 | ne | 11,191 | 46 | æfter | 2,277 | 82 | oþ | 972 |
| 11 | mid | 11,033 | 47 | eft | 2,230 | 83 | mihte/meahte | 961 |
| 12 | swa/sua | 10,928 | 48 | wæron | 2,023 | 84 | sona | 943 |
| 13 | þam/þæm | 9,870 | 49 | man/mon | 1,995 | 85 | sume | 873 |
| 14 | him/hym | 9,149 | 50 | þeah/þeh | 1,895 | 86 | sum | 857 |
| 15 | hi/hy/hie | 8,001 | 51 | hwæt | 1,880 | 87 | siþþan/syþþan | 854 |
| 16 | is/ys | 7,009 | 52 | fram/from | 1,806 | 88 | þætte | 845 |
| 17 | wæs | 6,979 | 53 | eow | 1,804 | 89 | min/myn | 842 |
| 18 | ic | 5,983 | 54 | hu | 1,790 | 90 | æfre | 820 |
| 19 | þonne/þænne | 5,677 | 55 | oþþe/odþe | 1,593 | 91 | oþer | 781 |
| 20 | for | 5,217 | 56 | þan | 1,578 | 92 | þissum/þisum/þysum | 766 |
| 21 | þæs/þes | 5,136 | 57 | æt | 1,550 | 93 | oþrum | 724 |
| 22 | þære/þare | 4,879 | 58 | wære | 1,497 | 94 | oft | 685 |
| 23 | ac | 4,765 | 59 | wiþ | 1,432 | 95 | sind/synd/syndon/sint | 670 |
| 24 | þu | 4,634 | 60 | nan | 1,376 | 96 | beo/bio | 647 |
| 25 | we | 4,509 | 61 | an | 1,376 | 97 | magon | 642 |
| 26 | of | 4,035 | 62 | beoþ/bioþ | 1,353 | 98 | fela | 642 |
| 27 | þone/þæne | 4,022 | 63 | ofer | 1,332 | 99 | her | 622 |
| 28 | hine/hyne/hiene | 3,817 | 64 | swiþe/swyþe/suiþe | 1,319 | 100 | anum | 616 |
| 29 | ge | 3,786 | 65 | buton/butan | 1,256 | 101 | forþæm/forþam | 608 |
| 30 | þurh | 3,732 | 66 | wolde | 1,240 | 102 | nis/nys | 607 |
| 31 | hit/hyt | 3,578 | 67 | beon/bion | 1,223 | 103 | ælc | 606 |
| 32 | be/bi/by/bie | 3,169 | 68 | sy/si/sie | 1,205 | 104 | witodlice | 570 |
| 33 | heora/hiora | 3,090 | 69 | ure | 1,184 | 105 | þon | 549 |
| 34 | biþ/byþ | 3,018 | 70 | eallum | 1,136 | 106 | ongean/angean | 504 |
| 35 | nu | 2,862 | 71 | sceal/sceall | 1,125 | 107 | hira/hyra | 374 |
| 36 | in | 2,840 | 72 | mæg | 1,111 | 108 | ymb/ymbe/embe | 348 |
| | | **346,755** | | | **65,031** | | | **28,326** |
| | | | | | | | | **440,112** |

*main corpus*, and dividing the result by $\sigma$, as shown in Table 11. Thus, the *z-score* for *and* in *Jn* is 0.347840533, which derives from the referred expression. For space, the *z-scores* of only six works are included in the table.

Third, we calculate the difference between the *z-score* of each *FW* in the *main corpus* and that of each work in the *test corpus*. Then, we obtain the *Delta-score* by adding the absolute values of the differences obtained, and dividing the result by the number of *FW*.

Fourth, we rank each work in the *test corpus* according to Delta values, in such a way that the smaller the value, the least unlike it will become to the work of the *main corpus* under scrutiny. For the sake of clarity, we summarize all the values of Delta, along with those of $\mu$ and $\sigma$ in Table 12. To the right of the values for Delta-*score*, the ones for Delta-*z-score* are also included.

From the analysis of the values for Delta in the table, two main facts must be highlighted inasmuch as they seem to validate both points of our

**Table 9** Frequency weight of the 108 *FW* in each work

| **Main corpus** | | | | ***Test corpus*** | | | |
|---|---|---|---|---|---|---|---|
| Works | *FW* | *N* | *FW*/*N*% | Works | *FW* | *N* | *FW*/*N*% |
| *Jn* | 9,736 | 17,098 | 56.94233244 | *Ad* | 3,095 | 6,345 | 48.7785658 |
| *Æhom1* | 47,573 | 89,676 | 53.04986842 | *Alc* | 624 | 1,092 | 57.14285714 |
| *ÆCHom2* | 49,986 | 100,954 | 49.51363988 | *Alex* | 3,443 | 6,690 | 51.46487294 |
| *Mt* | 10,746 | 20,435 | 52.58624908 | *AoT* | 3,446 | 6,546 | 52.64283532 |
| *CP* | 38,687 | 69,922 | 55.32879494 | *ÆCHom1* | 54,330 | 107,220 | 50.67151651 |
| *BenR* | 11,270 | 22,763 | 49.51017001 | *Lk* | 11,295 | 21,126 | 53.46492474 |
| *VRS* | 32,473 | 95,924 | 33.85284183 | *Ælet* | 14,678 | 30,299 | 48.44384303 |
| *Whom* | 14,026 | 28,799 | 48.70307997 | *Mk* | 6,762 | 12,423 | 54.43129679 |
| | | | | *Or* | 25,771 | 51,097 | 50.43544631 |
| | | | | *BenRW* | 9,505 | 20,500 | 46.36585366 |
| $\sum$ | 214,497 | 445,571 | 48.13980264 | $\sum$ | 132,949 | 263,338 | 0.504860673 |
| $\mu$ | | | 49.9358625 | $\mu$ | | | 51.38420122 |
| $\sigma$ | | | 7.125803781 | $\sigma$ | | | 3.169190176 |

hypothesis: (1) to confirm the level of (un)likeness among the *WSG*, as conjectured; (2) to relate works by the same author, as expected. With respect to (1), Luke's study evinces that *Mt* is the least unlike to *Lk*, whereas *Jn* is the second. Analogously, Mark's study shows that *Mt* is also the most like to *Mk*, whereas *Jn* is also the second. In addition, *Mt* and *Lk* look like each other the most. From these results the *WSG* can be ranked in terms of Delta the same way as by *lexical richness*, that is, *Mt* > *Lk* > *Mk* > *Jn*.

With respect to (2), our analysis is restricted to *ÆCHom1*, *Ælet*, and *BenR*, whose association with works in the main corpus was evident, and the results confirm the proof clearly: *ÆCHom2* and *Æhom1* are the least unlike to *ÆCHom1*; *Æhom1* and *ÆCHom2* are the least unlike to *ÆLet*; and *BenRW* is the least unlike to *BenR*. However, it is worth pointing the likeness observed by *Whom* with respect to *Ælfric*'s works as well as by *Or* with respect to the latter and *BenR*. Following the scholarly suggestion of an anonymous reviewer, the Delta study was repeated by replacing *Mt* in the main corpus alternatively by *Lk* and *Mk*, and *Jn* by *Lk* and *Mk*, and *Lk* and *Mk* by *Mt* and *Jn* in the test corpus to gain reliability with the results of the six possible combinations of *Jn*, *Lk*, *Mk*, and *Mt*. For clarity, only the results of the Gospels, *Ælfric*'s works, and *BenR* are included in Table 13, and for space some decimals have been suppressed.

As the results do not require a further interpretation, their values haven been plotted to ease their comparison and to help draw some conclusions (Fig. 1).

It is true that we have used the Delta technique for our purposes, though with scientific rigour and, in light of this, we must join Hoover in recognizing the excellence of the methodology proposed by Burrows, as it proves to operate efficiently and reliably. In this line, Hoover's detailed and (prospective) publications on Delta have encouraged us to check one of his proposals, in particular, the modification of the range of *MCW* to be considered and/or the possibility of eliminating pronouns from the counting, etc. Thus, when the values for Delta proved their efficiency for our study, we repeated the experiment in four ways, that is, leaving out (1) the topmost common *FW* (in number of 10, 20, 30, 40, 50, 60, 70, 80); (2) the bottommost common *FW* (in number of 10, 20, 30, 40, 50, 60, 70, 80); (3) the top and bottommost common *FW* (in number of 10, 20); and (4) all personal pronouns and possessives (*hē, his, hī, him, ic, wē, þū, hine, hit, heora, hēo, ūs, mē, ēow, ūre, hire, mīn, hira*). Although no homogeneous result can be expected, there is one constant common to all of them: the replication of the firstly ranked like work. In addition to this, more likenesses are observed in some of the variants experimented on test works, as shown in Table 14.

**Table 10** $\mu$, $\sigma$, and *score* of the 36 topmost *FW*

| | | ***μ*** | ***σ*** | ***Jn*** | ***Mt*** | ***BenR*** | ***Æhom1*** | ***Whom*** | ***ÆCHom2*** |
|---|---|---|---|---|---|---|---|---|---|
| 1 | and | 5.016501463 | 1.602026783 | 5.573751316 | 6.963542941 | 4.9158722 | 4.615504706 | 7.29886454 | 4.538700794 |
| 2 | þa | 2.324586792 | 0.950464853 | 2.88922681 | 3.944213359 | 1.1773492 | 2.446585486 | 1.16670718 | 2.645759455 |
| 3 | þæt | 2.568470047 | 0.641575273 | 2.63188677 | 2.069977979 | 2.8906559 | 2.274856149 | 2.91676794 | 2.164352081 |
| 4 | on | 2.352682728 | 0.294637863 | 2.111358054 | 2.485931001 | 2.6226772 | 2.873678576 | 2.15979722 | 2.400103017 |
| 5 | he | 2.151551934 | 0.606903717 | 2.433033103 | 2.123807194 | 2.2536573 | 2.286007404 | 1.84034168 | 1.925629495 |
| 6 | þe | 2.186766744 | 0.533904066 | 2.836589075 | 2.319549792 | 1.8275271 | 2.894865962 | 2.31952498 | 2.176238683 |
| 7 | to | 1.765414408 | 0.348423955 | 2.275119897 | 1.546366528 | 1.5551553 | 1.885677327 | 1.92715025 | 1.907799592 |
| 8 | se | 1.390915825 | 0.437815206 | 2.20493625 | 1.600195743 | 1.027984 | 1.528837147 | 1.06948158 | 1.660162054 |
| 9 | his | 1.263381163 | 0.403684728 | 1.07030062 | 1.389772449 | 1.1553837 | 1.768589143 | 1.11809438 | 1.754264318 |
| 10 | ne | 1.503849607 | 0.372830844 | 1.965142122 | 1.507218008 | 1.9900716 | 1.150809581 | 1.55908191 | 1.077718565 |
| 11 | mid | 0.96644616 | 0.416161666 | 0.438647795 | 0.567653536 | 1.6518034 | 1.215486864 | 0.82641758 | 1.234225489 |
| 12 | swa | 1.116984416 | 0.433495869 | 0.438647795 | 0.861267433 | 1.0455564 | 1.765243766 | 1.36115837 | 1.479881926 |
| 13 | þam | 1.260876655 | 0.452256644 | 0.918236051 | 1.086371422 | 1.5595484 | 1.866720193 | 0.88891975 | 1.488796878 |
| 14 | him | 1.110030325 | 0.509022863 | 2.152298514 | 1.463175924 | 0.7775776 | 1.126276819 | 0.48960033 | 0.812251124 |
| 15 | hi | 1.423475917 | 0.656160365 | 1.590829337 | 1.90359677 | 1.3838246 | 1.169766716 | 0.94794958 | 1.206490085 |
| 16 | is | 0.917744428 | 0.24718481 | 1.07030062 | 1.120626376 | 1.076308 | 0.855301307 | 1.15629015 | 0.84890148 |
| 17 | wæs | 0.532826556 | 0.279607944 | 0.924084688 | 0.557866406 | 0.0571102 | 0.578750167 | 0.37501302 | 0.52499158 |
| 18 | þonne | 0.798004037 | 0.785265169 | 0.298280501 | 0.494250061 | 0.7160743 | 0.453856104 | 0.86114101 | 0.318957149 |
| 19 | ic | 0.716369251 | 0.57077095 | 2.532459937 | 1.008074382 | 0.2416202 | 0.438244346 | 0.2187576 | 0.530934881 |
| 20 | for | 0.494658482 | 0.194475209 | 0.397707334 | 0.254465378 | 0.4832403 | 0.747134127 | 0.45834925 | 0.735978763 |
| 21 | þæs | 0.549270008 | 0.138116886 | 0.479588256 | 0.391485197 | 0.7556122 | 0.574289665 | 0.39584708 | 0.575509638 |
| 22 | þære | 0.426354986 | 0.235841897 | 0.350918236 | 0.259358943 | 0.562316 | 0.507382131 | 0.15278308 | 0.626027696 |
| 23 | ac | 0.480321444 | 0.171621544 | 0.579015089 | 0.200636163 | 0.4568818 | 0.604398055 | 0.49307268 | 0.541830933 |
| 24 | þu | 0.47318544 | 0.349262575 | 1.128786993 | 0.748715439 | 0.188903 | 0.394754449 | 0.07639154 | 0.465558571 |
| 25 | we | 0.518090319 | 0.256746131 | 0.380161422 | 0.239784683 | 0.3778061 | 0.735982872 | 0.97572832 | 0.65277255 |
| 26 | of | 0.454937293 | 0.17584225 | 0.772020119 | 0.567653536 | 0.3206959 | 0.585440921 | 0.36112365 | 0.437823167 |
| 27 | þone | 0.420296398 | 0.14268359 | 0.426950521 | 0.538292146 | 0.3294821 | 0.606628306 | 0.27431508 | 0.423955465 |
| 28 | hine | 0.534181818 | 0.262183061 | 1.029360159 | 0.6704184 | 0.4393094 | 0.579865293 | 0.34028959 | 0.494284526 |
| 29 | ge | 0.587500137 | 0.467765031 | 1.427067493 | 1.238071935 | 0.3953785 | 0.398099826 | 0.29862148 | 0.361550805 |
| 30 | þurh | 0.42952148 | 0.269801446 | 0.076032284 | 0.146806949 | 0.421737 | 0.449395602 | 0.9410049 | 0.593339541 |
| 31 | hit | 0.493619721 | 0.245520068 | 0.602409639 | 0.572547101 | 0.5008127 | 0.302199028 | 0.71183027 | 0.266457991 |
| 32 | be | 0.404669171 | 0.27614718 | 0.397707334 | 0.156594079 | 1.0104116 | 0.47504349 | 0.24653634 | 0.385324009 |
| 33 | heora | 0.251324046 | 0.228129489 | 0 | 0.078297039 | 0.562316 | 0.483964494 | 0.30903851 | 0.45664362 |
| 34 | biþ | 0.353603409 | 0.234686223 | 0.146215932 | 0 | 0.4964196 | 0.404790579 | 0.42709816 | 0.302117796 |
| 35 | nu | 0.264713049 | 0.132946214 | 0.32167505 | 0.166381209 | 0.0439309 | 0.428208216 | 0.36459599 | 0.267448541 |
| 36 | in | 0.159270848 | 0.314100872 | 0.023394549 | 0.02936139 | 0.0439309 | 0.063562157 | 0.11111497 | 0.055470808 |

**Table 11** *z-score* of the 36 topmost *FW*

| | *FW* | *Jn* | *Mt* | *BenR* | *Æhom1* | *Whom* | *ÆCHom2* |
|---|---|---|---|---|---|---|---|
| 1 | and | 0.347840535 | 1.215361378 | −0.06281369 | −0.2503059 | 1.424672237 | −0.29824762 |
| 2 | þa | 0.594067226 | 1.704036254 | −1.20702789 | 0.12835687 | −1.21822455 | 0.337911141 |
| 3 | þæt | 0.098845335 | −0.77698142 | 0.502179331 | −0.45764528 | 0.542879239 | −0.62988395 |
| 4 | on | −0.8190552 | 0.452244227 | 0.916360243 | 1.768258307 | −0.65465284 | 0.160944315 |
| 5 | he | 0.463798723 | −0.04571523 | 0.168239728 | 0.221543329 | −0.51278357 | −0.37225417 |
| 6 | þe | 1.217114405 | 0.248702072 | −0.67285424 | 1.326266762 | 0.248655606 | −0.01971901 |
| 7 | to | 1.462888764 | −0.62868203 | −0.60345768 | 0.345162602 | 0.464192653 | 0.408654978 |
| 8 | se | 1.859278557 | 0.478009706 | −0.82896119 | 0.315021773 | −0.73417789 | 0.614976879 |
| 9 | his | −0.47829539 | 0.31309405 | −0.26752913 | 1.251491436 | −0.35990161 | 1.216006258 |
| 10 | ne | 1.237270259 | 0.009034663 | 1.304135665 | −0.94691744 | 0.148143068 | −1.14296081 |
| 11 | mid | −1.2682532 | −0.95826372 | 1.646853284 | 0.59842298 | −0.33647639 | 0.643450254 |
| 12 | swa | −1.56480527 | −0.58989486 | −0.16477211 | 1.495422211 | 0.563267092 | 0.837141797 |
| 13 | þam | −0.75762426 | −0.38585444 | 0.660403201 | 1.339601189 | −0.82244652 | 0.503962133 |
| 14 | him | 2.047586198 | 0.693771584 | −0.65311934 | 0.031917021 | −1.21886469 | −0.58500162 |
| 15 | hi | 0.255049571 | 0.731712671 | −0.06042927 | −0.38665731 | −0.72471054 | −0.33069025 |
| 16 | is | 0.617174624 | 0.820770291 | 0.641477994 | −0.25261714 | 0.965050082 | −0.278508 |
| 17 | wæs | 1.399309784 | 0.089553426 | −1.70136916 | 0.164242869 | −0.56441005 | −0.02802129 |
| 18 | þonne | 2.208751858 | 0.267515171 | −0.70852992 | −0.45813784 | −0.73764438 | −0.3401006 |
| 19 | ic | −0.7324983 | −0.38915644 | −0.00051671 | −0.45992731 | 0.253642482 | −0.69627248 |
| 20 | for | −0.49852703 | −1.23508341 | −0.05871255 | 1.298240774 | −0.18670366 | 1.240879402 |
| 21 | þæs | −0.50451291 | −1.14240059 | 1.493967721 | 0.181148427 | −1.1108195 | 0.18998133 |
| 22 | þære | −0.31986153 | −0.70808472 | 0.576492367 | 0.343565526 | −1.15998008 | 0.846637992 |
| 23 | ac | 0.575065596 | −1.629663 | −0.13657762 | 0.722966408 | 0.074298553 | 0.358401909 |
| 24 | þu | 1.877102211 | 0.788890704 | −0.81395035 | −0.22456168 | −1.13609051 | −0.02183706 |
| 25 | we | −0.53721899 | −1.08397208 | −0.54639277 | 0.848669275 | 1.782453366 | 0.524573556 |
| 26 | of | 1.803223204 | 0.641007733 | −0.76341964 | 0.742163087 | −0.53351025 | −0.09732659 |
| 27 | þone | 0.046635514 | 0.826974899 | −0.63647364 | 1.305909865 | −1.02311217 | 0.025644623 |
| 28 | hine | 1.88867404 | 0.519623889 | −0.36185561 | 0.174242664 | −0.73952994 | −0.15217342 |
| 29 | ge | 1.794848486 | 1.390808967 | −0.4107226 | −0.40490481 | −0.61757215 | −0.48304024 |
| 30 | þurh | −1.31018273 | −1.04786144 | −0.02885252 | 0.073662029 | 1.895777148 | 0.607180069 |
| 31 | hit | 0.443099901 | 0.321470177 | 0.029297 | −0.77965396 | 0.888768698 | −0.92522673 |
| 32 | be | −0.0252106 | −0.89834375 | 2.193549327 | 0.254843517 | −0.57263968 | −0.07005381 |
| 33 | heora | −1.10167277 | −0.75845962 | 1.363225751 | 1.019773682 | 0.252989926 | 0.900013301 |
| 34 | biþ | −0.88367981 | −1.50670714 | 0.608541131 | 0.218108971 | 0.313161776 | −0.21938063 |
| 35 | nu | 0.428458993 | −0.73963625 | −1.66068744 | 1.229784304 | 0.751303411 | 0.020575933 |
| 36 | in | −0.4325881 | −0.41359152 | −0.36720658 | −0.30470686 | −0.15331342 | −0.33046721 |

These results serve to confirm the validity of Delta, independent of the range of frequency adopted without enlarging the initial scope of 108 function words, at least for the particular cases investigated.

## 6 Conclusions

We have proven, out of any reasonable doubt, both by lemma-counting and by Zipf's *Z*, that the *lexical richness* of the *Gospel according to Saint Matthew* is greater than that of the other three, which leads to the following ranking: *Mt* > *Lk* > *Mk* > *Jn* (Miranda and Calle, 2005a). The inverse ranking *Jn* > *Lk* > *Mk* > *Mt* (resembling a mirror image of the previous one) has been obtained by means of *functional density*. The same ranking has also been generated from the study of *FL*, and the values for Delta-*score* and Delta-*z-score* also point to this direction insofar as *Mt* and *Lk* are estimated as the least unlike whereas *Mt* and *Jn* were taken as the most unlike. From these results, it follows that the four works can be equally ranked in terms of *functional density* or by means of lemma-scoring and successfully clustered, thanks to Delta-*score*.

**Table 12** Works ranked in descending order by Delta-*score* and Delta-*z-score*

| ***Ad*** | **Δ-*score*** | **Δ-*z-score*** | ***Alc*** | **Δ-*score*** | **Δ-*z-score*** | ***Alex*** | **Δ-*score*** | **Δ-*z-score*** |
|---|---|---|---|---|---|---|---|---|
| *Whom* | 1.607630889 | 1.459423071 | *Whom* | 1.607630889 | 1.459423071 | *VRS* | 1.941414074 | 1.621772993 |
| *Æhom1* | 1.643213468 | 0.946323726 | *Æhom1* | 1.643213468 | 0.946323726 | *ÆCHom2* | 1.985796452 | −17.0395037 |
| *ÆCHom2* | 1.656853496 | −23.89176 | *ÆCHom2* | 1.656853496 | −23.89176 | *Æhom1* | 2.097954566 | 0.278547568 |
| *BenR* | 1.709373544 | −0.00770188 | *BenR* | 1.709373544 | −0.00770188 | *CP* | 2.137733003 | −0.06277887 |
| *CP* | 1.714798392 | −0.08592796 | *CP* | 1.714798392 | −0.08592796 | *Whom* | 2.159908333 | −0.2530585 |
| *VRS* | 1.755397694 | −0.67136825 | *VRS* | 1.755397694 | −0.67136825 | *Jn* | 2.225937655 | −0.81963564 |
| *Mt* | 1.776415641 | −0.97444619 | *Mt* | 1.776415641 | −0.97444619 | *Mt* | 2.244754491 | −0.98109707 |
| *Jn* | 1.807032331 | −1.41593761 | *Jn* | 1.807032331 | −1.41593761 | *BenR* | 2.249835081 | −1.02469204 |
| $\mu$ | 1.708839432 | | $\mu$ | 2.317452888 | | $\mu$ | 2.130416707 | |
| $\sigma$ | 0.069348323 | | $\sigma$ | 0.090485188 | | $\sigma$ | 0.116540745 | |
| ***AoT*** | **Δ-*score*** | **Δ-*z-score*** | ***ÆCHom1*** | **Δ-*score*** | **Δ-*z-score*** | ***Lk*** | **Δ-*score*** | **Δ-*z-score*** |
| *ÆCHom2* | 1.145756764 | −8.18061939 | *ÆCHom2* | 0.328103971 | *ÆCHom2* | *Mt* | 0.475261193 | 1.95323673 |
| *Æhom1* | 1.256157716 | 0.699006029 | *Æhom1* | 0.519617857 | *Æhom1* | *Jn* | 0.819544058 | 0.782988425 |
| *VRS* | 1.284450835 | 0.496995229 | *Whom* | 1.007994064 | *Whom* | *ÆCHom2* | 0.970644479 | −3.29930755 |
| *Jn* | 1.30599764 | 0.343152618 | *Mt* | 1.008261252 | *Mt* | *Æhom1* | 1.050966589 | −0.00363749 |
| *Mt* | 1.335551102 | 0.132143062 | *BenR* | 1.013388759 | *BenR* | *VRS* | 1.17505739 | −0.42543324 |
| *Whom* | 1.418122663 | −0.45741183 | *CP* | 1.094359994 | *CP* | *Whom* | 1.228777702 | −0.6080334 |
| *BenR* | 1.518976829 | −1.1775032 | *Jn* | 1.094859707 | *Jn* | *BenR* | 1.317633024 | −0.9100606 |
| *CP* | 1.567456242 | −1.52364265 | *VRS* | 1.161962061 | *VRS* | *CP* | 1.361287179 | −1.05844498 |
| $\mu$ | 1.354058724 | | $\mu$ | 0.903568458 | | $\mu$ | 1.049896452 | |
| $\sigma$ | 0.140057459 | | $\sigma$ | 0.305254983 | | $\sigma$ | 0.294196423 | |
| ***ÆLet*** | **Δ-*score*** | **Δ-*z-score*** | ***Mk*** | **Δ-*score*** | **Δ-*z-score*** | ***Or*** | **Δ-*score*** | **Δ-*z-score*** |
| *Æhom1* | 0.5567814 | 1.516348812 | *Mt* | 0.670328977 | 1.74755156 | *BenR* | 0.517866923 | 2.373167693 |
| *ÆCHom2* | 0.567785101 | −2.13759695 | *Jn* | 0.889686888 | 0.89372523 | *ÆCHom2* | 1.511840204 | −3.40960444 |
| *Whom* | 0.929675474 | 0.112477466 | *ÆCHom2* | 1.038015514 | −4.04036021 | *Æhom1* | 1.625092551 | −0.12392254 |
| *BenR* | 1.010329086 | −0.1911672 | *Æhom1* | 1.039869082 | 0.309157704 | *VRS* | 1.652768592 | −0.18633942 |
| *Mt* | 1.036780856 | −0.2907528 | *VRS* | 1.245856592 | −0.49262582 | *Whom* | 1.743589603 | −0.39116512 |
| *VRS* | 1.135030297 | −0.66064223 | *Whom* | 1.280458099 | −0.62730835 | *CP* | 1.782758626 | −0.47950176 |
| *CP* | 1.186425085 | −0.85413329 | *BenR* | 1.37608883 | −0.99954036 | *Mt* | 1.799638316 | −0.51756998 |
| *Jn* | 1.253605177 | −1.10705285 | *CP* | 1.414058345 | −1.14733248 | *Jn* | 1.927601266 | −0.80616069 |
| $\mu$ | 0.95955156 | | $\mu$ | 1.119295291 | | $\mu$ | 1.57014451 | |
| $\sigma$ | 0.265618409 | | $\sigma$ | 0.256911627 | | $\sigma$ | 0.443406334 | |
| | | | ***BenRW*** | **Δ-*score*** | **Δ-z-score** | | | |
| | | | *BenR* | 1.042748372 | 2.116493607 | | | |
| | | | *ÆCHom2* | 1.482854715 | −5.53965446 | | | |
| | | | *Æhom1* | 1.558279397 | 0.190570783 | | | |
| | | | *Whom* | 1.609871292 | −0.00216642 | | | |
| | | | *Mt* | 1.701461287 | −0.34432869 | | | |
| | | | *VRS* | 1.767317142 | −0.59035324 | | | |
| | | | *CP* | 1.797120968 | −0.7016945 | | | |
| | | | *Jn* | 1.9146779 | −1.14086414 | | | |
| | | | $\mu$ | 1.609291384 | | | | |
| | | | $\sigma$ | 0.267680002 | | | | |

Therefore, it is worth enhancing their complementary feature for tasks of ranking and clustering in authorship attribution studies.

The efficiency of *functional density* has been proven as an indirect measure to rank texts in terms of *lexical richness* from the least rich to the most, the ranking obtained being similar to the one resulting by means of *Z*; in other words, *functional density* may be a helpful marker to relate works by the same author, once the technique has been refined and tested with larger corpora. An in-progress research based on an annotated corpus of the *Federalist Papers* seems to reveal traces confirming the validity of *functional density* in the attribution of authorship

**Table 13** Works ranked in descending order by Delta-*score*

| **2nd study *Jn* and *Lk*** | | **3rd study *Jn* and *Mk*** | | **4th study *Lk* and *Mt*** | | **5th study *Mk* and *Mt*** | | **6th study *Lk* and *Mk*** | |
|---|---|---|---|---|---|---|---|---|---|
| ***ÆCHom1*** | ***Δ-score*** | ***ÆCHom1*** | ***Δ-score*** | ***ÆCHom1*** | ***Δ-score*** | ***ÆCHom1*** | ***Δ-score*** | ***ÆCHom1*** | ***Δ-score*** |
| *ÆCHom2* | 0.325776 | *ÆCHom2* | 0.320598 | *ÆCHom2* | 0.341022 | *ÆCHom2* | 0.335463 | *ÆCHom2* | 0.332518 |
| *Æhom1* | 0.517612 | *Æhom1* | 0.496124 | *Æhom1* | 0.544182 | *Æhom1* | 0.517041 | *Æhom1* | 0.516929 |
| *Lk* | 0.991538 | *Whom* | 0.987619 | *Whom* | 1.036373 | *Whom* | 1.027869 | *BenR* | 1.047549 |
| *Whom* | 0.994752 | *BenR* | 1.008894 | *Lk* | 1.041115 | *Mt* | 1.040511 | *CP* | 1.120692 |
| *BenR* | 1.002550 | *Mk* | 1.054923 | *Mt* | 1.045140 | *BenR* | 1.053916 | *Lk* | 1.037608 |
| *CP* | 1.079067 | *CP* | 1.079768 | *BenR* | 1.047844 | *Mk* | 1.083725 | *Mk* | 1.090885 |
| *Jn* | 1.093989 | *Jn* | 1.089985 | *CP* | 1.128949 | *CP* | 1.127017 | *VRS* | 1.199552 |
| *VRS* | 1.149178 | *VRS* | 1.152916 | *VRS* | 1.202198 | *VRS* | 1.198391 | *Whom* | 1.017836 |
| $\mu$ | 0.894308 | $\mu$ | 0.898853 | $\mu$ | 0.923353 | $\mu$ | 0.922992 | $\mu$ | 0.920446 |
| ***ÆLet*** | ***Δ-score*** | ***ÆLet*** | ***Δ-score*** | ***ÆLet*** | ***Δ-score*** | ***ÆLet*** | ***Δ-score*** | ***ÆLet*** | ***Δ-score*** |
| *Æhom1* | 0.556263 | *Æhom1* | 0.539412 | *Æhom1* | 0.575250 | *Æhom1* | 0.554755 | *Æhom1* | 0.557326 |
| *ÆCHom2* | 0.562786 | *ÆCHom2* | 0.556398 | *ÆCHom2* | 0.594242 | *ÆCHom2* | 0.587587 | *ÆCHom2* | 0.585490 |
| *Whom* | 0.924538 | *Whom* | 0.925670 | *Whom* | 0.949785 | *Whom* | 0.947763 | *Whom* | 0.946296 |
| *BenR* | 1.009186 | *BenR* | 1.009798 | *BenR* | 1.036715 | *BenR* | 1.039371 | *BenR* | 1.048854 |
| *Lk* | 1.060437 | *Mk* | 1.103473 | *Mt* | 1.082660 | *Mt* | 1.083937 | *Lk* | 1.116699 |
| *VRS* | 1.126672 | *VRS* | 1.141735 | *Lk* | 1.117185 | *Mk* | 1.135830 | *Mk* | 1.144354 |
| *CP* | 1.178421 | *CP* | 1.176385 | *VRS* | 1.179883 | *VRS* | 1.192051 | *VRS* | 1.194545 |
| *Jn* | 1.254020 | *Jn* | 1.242305 | *CP* | 1.208654 | *CP* | 1.203810 | *CP* | 1.210030 |
| $\mu$ | 0.959040 | $\mu$ | 0.961897 | $\mu$ | 0.968047 | $\mu$ | 0.968138 | $\mu$ | 0.975449 |
| ***Mt*** | ***Δ-score*** | ***Mt*** | ***Δ-score*** | ***Jn*** | ***Δ-score*** | ***Jn*** | ***Δ-score*** | ***Jn*** | ***Δ-score*** |
| *Lk* | 0.605078 | *Mk* | 0.624428 | *Mt* | 1.365191 | *Mt* | 1.686223 | *Lk* | 1.446777 |
| *Jn* | 0.906291 | *Jn* | 0.761435 | *Lk* | 1.405802 | *Mk* | 1.799534 | *Mk* | 1.524938 |
| *Æhom1* | 1.042117 | *ÆCHom2* | 0.982350 | *ÆCHom2* | 1.718466 | *ÆCHom2* | 2.032705 | *ÆCHom2* | 1.754757 |
| *ÆCHom2* | 1.045090 | *Æhom1* | 1.084829 | *Æhom1* | 1.849307 | *Æhom1* | 2.160880 | *Æhom1* | 1.879825 |
| *VRS* | 1.249480 | *Whom* | 1.194722 | *Whom* | 1.972625 | *VRS* | 2.275701 | *Whom* | 1.999863 |
| *Whom* | 1.276425 | *VRS* | 1.200277 | *VRS* | 1.976360 | *Whom* | 2.286760 | *VRS* | 2.012682 |
| *BenR* | 1.378135 | *BenR* | 1.234368 | *CP* | 2.071564 | *CP* | 2.402522 | *CP* | 2.123087 |
| *CP* | 1.412151 | *CP* | 1.358265 | *BenR* | 2.120602 | *BenR* | 2.450025 | *BenR* | 2.170008 |
| $\mu$ | 1.114346 | $\mu$ | 1.055084 | $\mu$ | 1.809990 | $\mu$ | 2.136793 | $\mu$ | 1.863992 |
| ***Mk*** | ***Δ-score*** | ***Lk*** | ***Δ-score*** | ***Mk*** | ***Δ-score*** | ***Lk*** | ***Δ-score*** | ***Mt*** | ***Δ-score*** |
| *Lk* | 0.477990 | *Mk* | 0.575626 | *Lk* | 0.658944 | *Mt* | 0.515407 | *Lk* | 0.551254 |
| *Jn* | 0.773175 | *Jn* | 0.822087 | *Mt* | 0.727092 | *Mk* | 0.621508 | *Mk* | 0.710431 |
| *ÆCHom2* | 0.994175 | *ÆCHom2* | 0.968821 | *ÆCHom2* | 1.079849 | *ÆCHom2* | 1.015977 | *VRS* | 1.301363 |
| *Æhom1* | 1.097400 | *Æhom1* | 1.044130 | *Æhom1* | 1.084331 | *Æhom1* | 1.102217 | *BenR* | 1.364239 |
| *Whom* | 1.204772 | *VRS* | 1.160328 | *VRS* | 1.294856 | *VRS* | 1.204210 | *Æhom1* | 1.170908 |
| *VRS* | 1.218111 | *Whom* | 1.216995 | *Whom* | 1.345670 | *Whom* | 1.302810 | *Whom* | 1.309478 |
| *BenR* | 1.235341 | *BenR* | 1.316051 | *BenR* | 1.452950 | *BenR* | 1.414650 | *ÆCHom2* | 1.066385 |
| *CP* | 1.344573 | *CP* | 1.346741 | *CP* | 1.478661 | *CP* | 1.442436 | *CP* | 1.462047 |
| $\mu$ | 1.043192 | $\mu$ | 1.056347 | $\mu$ | 1.140294 | $\mu$ | 1.077402 | $\mu$ | 1.117013 |
| ***BenRW*** | ***Δ-score*** | ***BenRW*** | ***Δ-score*** | ***BenRW*** | ***Δ-score*** | ***BenRW*** | ***Δ-score*** | ***BenRW*** | ***Δ-score*** |
| *BenR* | 0.975306 | *BenR* | 1.005156 | *BenR* | 1.003975 | *BenR* | 1.025948 | *BenR* | 0.970164 |
| *ÆCHom2* | 1.414815 | *ÆCHom2* | 1.447568 | *ÆCHom2* | 1.477369 | *ÆCHom2* | 1.508901 | *ÆCHom2* | 1.454624 |
| *Æhom1* | 1.495624 | *Æhom1* | 1.513084 | *Æhom1* | 1.548551 | *Æhom1* | 1.558427 | *Æhom1* | 1.508144 |
| *Whom* | 1.538838 | *Whom* | 1.571288 | *Whom* | 1.576304 | *Whom* | 1.602620 | *Whom* | 1.544700 |
| *Lk* | 1.655793 | *Mk* | 1.694348 | *Mt* | 1.711224 | *Mt* | 1.742054 | *Mk* | 1.706974 |
| *VRS* | 1.696357 | *VRS* | 1.742286 | *Lk* | 1.741837 | *Mk* | 1.756083 | *Lk* | 1.717326 |
| *CP* | 1.724854 | *CP* | 1.75592 | *CP* | 1.760902 | *CP* | 1.776575 | *CP* | 1.717903 |
| *Jn* | 1.849024 | *Jn* | 1.871348 | *VRS* | 1.761349 | *VRS* | 1.811641 | *VRS* | 1.744048 |
| $\mu$ | 1.543826 | $\mu$ | 1.575129 | $\mu$ | 1.572689 | $\mu$ | 1.597781 | $\mu$ | 1.545485 |

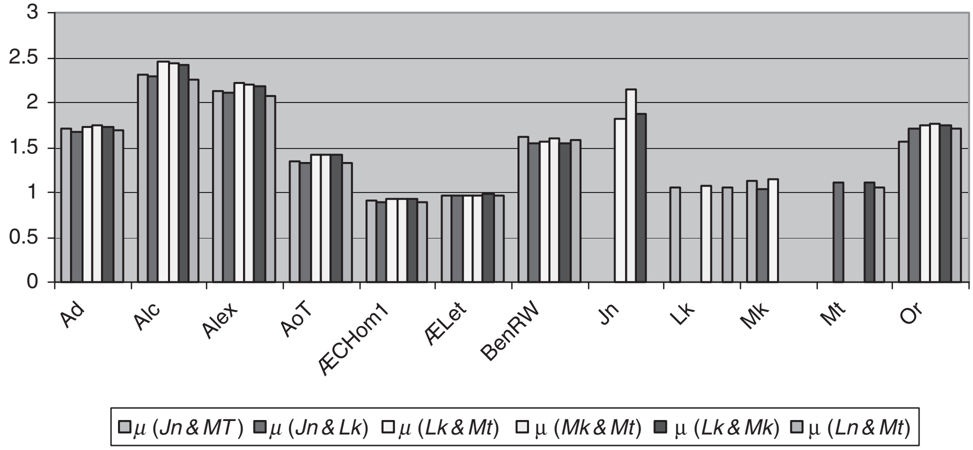


**Fig. 1** Plotting of works in terms of their Delta-*score* mean

**Table 14** Ranking by Delta-*score* after hierarchy truncation or reduction[11]

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Tops off** | | | | | | | | |
| | 10 | 20 | 30 | 40 | 50 | 60 | 70 | 80 |
| *Lk* | D | D | AG | B | B | B | B | B |
| *ÆLet* | AI | AG | AG | AG | AG | E | A | E |
| *Mk* | AI | AI | AI | AI | AI | A | A | |
| *BenRW* | A | D | D | D | D | D | D | D |
| **Bottoms off** | | | | | | | | |
| | 10 | 20 | 30 | 40 | 50 | 60 | 70 | 80 |
| *Lk* | D | D | D | D | D | D | D | D |
| *ÆLet* | F | C | C | F | C | C | C | C |
| *Mk* | E | E | D | D | A | A | B | A |
| *BenRW* | EI | E | E | B | | | | A |
| **Tops and bottoms off or pronouns off** | | | | | | | | |
| | 10 tops and 10 bottoms off | | | | 20 tops and 20 bottoms off | | | Pronouns off |
| *Lk* | C | | | | EI | | | E |
| *Ælet* | 1st and 2nd, and 7th and 8th alternate | | | | 1st and 2nd, and 7th and 8th alternate | | | F |
| *Mk* | AI | | | | A | | | A |
| *BenRW* | D | | | | D | | | E |

of these 18th-century American English texts. After having handled the *Disputed Papers* and a significant part of the undisputed contributions (all by Jay, and by Madison, and some by Hamilton), the computation of *functional density* points to the fact that Madison's and the *Disputed Papers* are the least unlike, as is generally acknowledged (Mosteller and Wallace, 1984). However, the application of *functional density* has to overcome the difficulties deriving from the requirement of an annotated corpus and an appropriate concordancer.

Likewise, the use of the top *FL* has proven to measure the divergence from an ideal corpus as an indirect technique to ascertain *lexical richness*. The technique may be helpful at least to signal the main tendency, but its accuracy depends on the availability of a large annotated corpus to avert the problems of text-length dependency.

Lastly, Burrows's methodology has been successfully tested as Delta was able to cluster the pairs of gospels under experiment as the most (un)like as well as to associate works by the same author. Considering that our aim was to corroborate what we already knew: (1) the ranking of the *WSG* by *lexical richness* (by *lemma-counting* and by *Z*) and (2) the leading candidates in the case of *Aelfric's* works (*Homilies* and *Letters*) and of the *Benedict Rule*, the hypothesis also holds true in both cases. Besides, Hoover's proposal for top and/or bottom truncation, and/or for neglecting pronominals has been successfully tested. From experience, we must highlight the simplicity of the procedure of accounting only *FW*, although it would be useful to check whether similar results can be obtained by applying Delta methodology to the hierarchies of the whole works or to their corresponding partitions (for example, of 3,000 words). In this line, we cannot help insisting on the suitability of *FW* in quantitative studies (Craig, 1999), irrespective of the aims pursued. Several facts will be argued to support our statement. One is quantitatively featured as their suitability is given by the weight within the text: only the frequencies of the top 108 most common *FW* range from 48.1398% in the *main corpus* to 50.4861% of the *test corpus* (i), the ratio *FW* to *CW* (*functional density*) ranges from 1.134867033 in *AoT* to 1.320108696 in *Jn* (ii). The other is qualitatively featured as the distribution of their use can certainly describe the style of an author.

Before closing this section, some lines are to be devoted to the nature of the works which bear this study: the translation of the *Holy Gospels* into West Saxon (henceforth WS), the most prestigious dialect of OE. It is necessary to insist on the word *translation* because it is an actual rendering from a Latin source rather than a gloss (unlike the cases of the *Lindisfarne* Gospels, written in the Northumbrian dialect, or of the Mercian *Rushworth*). Without assessing the quality of the version (not free of mistakes!), we must admit that all the evidences found in our study signal the common lexical similarities of the so-called Synoptic Gospels since *Mt* and *Lk* are supposed to have their common textual source in *Mk*, but these likenesses are not so markedly significant and regularly distributed as to counterbalance the differences that can serve to put them apart. Previous to this research, we accomplished a comparative lexical study between the Latin source and the WS renderings of the Gospels, by computing *N, V, MFW, HL, HD, K and Z* to conclude that, in general terms, the data concerning each WS gospel resembled its Latin counterpart: both are ranked as $Mk > Lk > Mt > Jn$, if Yule's *K* is applied, and as $Mt > Lk > Mk > Jn$, if Zipf's *Z* is applied. Obviously, the WS texts are numerically greater than the Latin ones, as the latter relies on a more synthetic system than OE. In fact, the ratio *N*(WS) to *N(L)* ranges from 10.63 in *Lk* to 11.16 in *Mt*. The reason for this numerical difference in *N* must be associated with the unbalanced use of certain *FW* in each language: demonstratives (acting as determiners) are more common in OE, personal pronouns acting as subjects are usually omitted in Latin, and prepositions are more than once unnecessary in Latin due to its stronger casual system. Therefore, the weight of *FW* within *N* differs from one language to another, which makes them reliable enough for authorship difference. Despite the low value of *N*, the widely known *The Lord's Prayer* can illustrate our argument better, as shown in Table 15 (de la Cruz *et al.*, 2003).

As can be observed, the difference is not too sharp due to the fact that the imperative is the most common tense used and, therefore, the pronominal subject can be omitted in both. Text-length *N* is expected to increase with decreasing level of synthesis, as happens when two other prayers in Latin are compared with their versions in Contemporary English (Table 16).

All in all, the data concerning the *FW* used in Latin are bound to offer such significant variations with respect to their WS counterparts that it would be daring to state that their distribution is alike in both versions. Add to this the differences observed

**Table 15** Several versions of *The Lord's Prayer* in numbers

| | Latin | Northumbrian | West Saxon | Middle English | King's James | Protestant | Catholic | $\mu$ | $\sigma$ |
|---|---|---|---|---|---|---|---|---|---|
| *FW* | 26 | 27 | 29 | 32 | 35 | 29 | 29 | 29.5714286 | 3.047247 |
| *CW* | 23 | 24 | 23 | 23 | 30 | 26 | 26 | 25 | 2.581988897 |
| *N* | 49 | 51 | 52 | 55 | 65 | 55 | 55 | 54.5714285 | 5.15936504 |

**Table 16** *Nicene Creed* and the *Glory* in numbers

| | *Nicene Creed* | | *Glory* | |
|---|---|---|---|---|
| | Latin | Catholic | Latin | Catholic |
| *FW* | 61 | 117 | 33 | 58 |
| *CW* | 102 | 115 | 50 | 71 |
| *N* | 163 | 232 | 83 | 129 |

in the translation of Latin absolute ablatives and/or the rendering of *amen, autem, vero*, etc. into OE. Therefore, we estimate that the *WS Gospels*, notwithstanding their condition of translated pieces, can be the object of an independent study of authorship, and in view of the evidences found in the use of *FW* we can even dare to affirm that the translatorship of *The Gospel according to Saint Matthew* and that of *Saint John* cannot be assigned to the same person, as Liuzza (2000) proposes. Furthermore, the features distinguishing between *Mt* and *Jn* cannot be necessarily attributed to their inclusion within or exclusion from the Synoptic Gospels, but rather to the different usage of *FW*, which amounts to 50% of *N* as an average.

## Acknowledgements

This research has been funded by the Spanish Ministry of Science and Technology (grant number HUM2004-01075/FILO). This grant is hereby gratefully acknowledged. We would also like to thank the Dictionary of Old English project at the University of Toronto for granting permission to use files of the *Toronto Complete Corpus of Old English Dictionary*. Our gratitude also to both anonymous reviewers for their suggestions as they have served to improve the presentation and clarity of the argument as well as to offer much grounded results in Delta.

## References

**Burrows, J. F.** (2002). Delta: a measure of stylistic difference and a guide to likely authorship: attribution and beyond. *Literary and Linguistic Computing*, **17**: 267–87.

**Burrows, J. F.** (2003). Questions of authorship: attribution and beyond. *Computers and the Humanities*, **37**(1): 5–32.

**Craig, H.** (1999). Authorial attribution and computational stylistics: If you can tell authors apart, have you learnt anything about them? *Literary and Linguistic Computing*, **14**: 103–13.

**Cruz, J. de la *et al.*** (2003). *Historia esencial de la lengua inglesa*. Málaga: Ágora.

**Healey, A. di Paolo** (ed.) (2004). *The Complete Corpus of Old English in Electronic Form*. Dictionary of Old English Project, Centre for Medieval Studies, University of Toronto.

**Holmes, D. I.** (1994). Authorship attribution. *Computers and the Humanities*, **28**(2): 87–106.

**Hoover, D. L.** (2003). Another perspective on vocabulary richness. *Computers and the Humanities*, **37**(2): 151–78.

**Hoover, D. L.** (2004a). Delta Prime? *Literary and Linguistic Computing*, **19**(4): 477–95.

**Hoover, D. L.** (2004b). Testing Burrow's Delta. *Literary and Linguistic Computing*, **19**(4): 453–75.

**Laan, N. M.** (1995). Stylometry and method. The case of euripides. *Literary and Linguistic Computing*, **10**: 271–78.

**Liuzza, R. M.** (ed.) (2000). *The Old English Version of the Gospels*, vols 2. Oxford: Oxford University Press.

**Miranda García, A., Triviño Rodríguez, J. L., and Calle Martín, J.** (2000). A morphological analyzer of Old English texts (*MAOET*). In Hornero, A. M. and Navarro, M. P. (eds), *Proceedings of the 10th International Conference of SELIM*. Zaragoza: Institución Fernando el Católico, pp. 127–45.

**Miranda García, A. and Calle Martín, J.** (2005a). The validity of lemma-based lexical richness in authorship attribution: a proposal for the Old English Gospels. *ICAME Journal*, **29**: 41–55.

**Miranda García, A. and Calle Martín, J.** (2005b). Yule's characteristic *K* revisited. *Linguistic Resources and Evaluation*, **39:** 287–94.

**Miranda García, A. *et al.*** (2006). The Old English *Apollonius of Tyre* in the light of the Old English concordancer. In Renouf, A. and Kehoe, A. (eds), *The Changing Face of Corpus Linguistics*. Amsterdam: Rodopi, pp. 81–98.

**Mosteller, F. and Wallace, D. L.** (1984). *Applied Bayesian and Classical Inference. The Case of The Federalist Papers*. New York: Springer-Verlag.

**Rudman, J.** (1998). The state of authorship attribution studies: some problems and solutions. *Computers and the Humanities*, **31**: 351–65.

**Scott, M.** (1996). *Wordsmith Tools 3.0*. Oxford: Oxford University Press.

**Somers, H. and Tweedie, F.** (2003). Authorship attribution and pastiche. *Computers and the Humanities*, **37**: 407–29.

**Tweedie, F. and Baayen, R. H.** (1998). How variable may a constant be? Measures of lexical richness in perspective. *Computers and the Humanities*, **32**: 323–52.

## Notes

1 See Tweedy and Baayen (1998) for a review of *K* and *Z*. See also Miranda and Calle (2005b: 287–94) for an updated interpretation.

2 The annotation was accomplished as follows. First, the corpus was typed using the TT font *Fonetik* for special characters and vowel length markers. Second, it was tagged by *MAOET* (standing for *Morphological Analyser of Old English Texts*; see Miranda, Triviño, and Calle 2000: 127–45), which provides all the possible tags for each word, regardless of context, including lemma, morphological analysis (class, type, case, gender, number, etc.), and meaning(s). Third, it was semi-automatically disambiguated according to context by means of an *ad hoc* tool.

3 Even with a clear-cut definition, the subjectivity of the disambiguator cannot be ruled out because some examples admit more than one plausible tagging.

4 John F. Burrows has proposed Delta, a simple new measure of textual difference, as a tool for authorship attribution and has shown that it has great potential, especially in attribution problems, where the possible authors are numerous and difficult to limit by traditional methods (3). In tests on prose, Delta has performed nearly as well as for Burrows's verse texts (4). A series of further tests using automated methods, however, shows that two modified methods of calculating Delta and three alternatives to or transformations of Delta produce results that are even more accurate (5). Four of these five new measures produce much better results than Delta both on a very diverse group of 104 novels and on a group of forty-four smaller contemporary literary critical texts (6). Although further testing is needed, Delta and its modifications should prove valuable and effective tools for authorship attribution (7) (Hoover, 2004a).

5 Recall that $V(N)/N$ is blamed for text-dependency. Confer, also, Guiraud's $R=V(N)/\sqrt{N}$ was established to correct the text-dependency of the results (Tweedy and Baayen, 1998).

6 Just fifty-seven of the *function lemmas* listed in Table 6 comprise the *function words* in Table 8. The twenty-one extant *function lemmas* are: *þīn*, *witodlīce*, *forþāmþe*, *hwilc*, *twēgen*, *foran*, *manig*, *ēower*, *ǣlc*, *ūt*, *hwā*, *ongēan*, *gȳt*, *hwī*, *betwux*, *git*, *hwǣr*, *þanon*, *twelf*, *intō*, *swilce*.

7 As mentioned in the methodological procedure, the *entire* corpus contains a set of files from the *Toronto Corpus of Old English Dictionary* (see Appendix).

8 *cwæþ* 4061; *godes* 3463; *god* 3208; *gode* 2092; *men(n)* 1648; *fæder* 1467; *hælend* 1437; *wearþ* 1345; *halgan* 1301; *þing* 1248; *dryhten* 1170; *c(w)om* 1097; *soþlice* 1069; *hæfde* 1040; *crist* 1032; *rice* 986; *geleafan* 966; *sunu* 959; *sona* 892; *cristes* 865; *het* 849; *cyning* 836; *mannum* 787; *eorþan* 775; *life* 761; *manna* 754; *sæde* 750; *word* 712; *lare* 691; *folc* 673; *sylf* 671; *dæge* 660; *lif* 652; *worulde* 591; *cwædon* 591; *hæfþ* 578; *witodlice* 570; *gast* 553; *mode* 543; *mod* 536; geseah 535; *micel* 527; *deaþe* 505; *swylce* 503; *lifes* 500; *criste* 499; *naman* 495; *heortan* 483; *eode* 480; *habban* 466; *don* 466; *willan* 463; *secge* 458; *mann* 453; *wurdon* 443; *stowe* 443; *wif* 441; *halga* 439; *wel* 428; *yfel* 423; *folce* 420; *tide* 413;

*ferde* 406; *gehaten* 405; *synna* 402; *lichaman* 401; *mannes* 395; *wuldres* 146; *wordum* 146; *bearn* 123; *ece* 109; *engla* 99; *foldan* 97; *wide* 94; weard 91; *waldend* 91; *halig* 84; *dryhtnes* 82; *ham* 81; *siþ* 73; *middangeard* 73; *ende* 73; *soþ* 71; *tid* 70; *hafaþ* 70; *fæste* 70; *wiht* 68; *secgan* 67; *monna* 67; *heofonum* 67; *þeoden* 63; *leoht* 61; *ful* 60; *frean* 60; *wera* 58; *helle* 58; *hæleþ* 58; *gumena* 58; *helm* 57; *ælmihtig* 56; *wuldre* 55; *ongan* 55; *frofre* 55; *hyrde* 54; *hæfdon* 53; *gewat* 53; *georne* 53; *feorh* 51; *dom* 51; *blæd* 51.

9 Burrows warns of the convenience of tagging homographs to generate more accurate results, although Hoover (2003) considers unnecessary to take this prevention.

10 Note that allomorphs are separated by slashes.

11 Capital letters denote a similar/different ranking for the works tested (represented by ordinal numbers) that is, A = 1st; B = 1st and 2nd; C = 1st, 2nd, and 3rd; D = 1st but 2nd and 3rd alternate; E = 1st, 2nd, 3rd, and 4th; F = 1st, 2nd, 3rd, 4th and 5th; G = 1st, 2nd, but 3rd and 4th alternate; H = 5th and 6th; I = 5th, 6th, 7th, and 8th.

# Appendix

**Appendix: list of abbreviations** (archives from the *Toronto Complete Corpus of Old English Dictionary*).

*Adrian and Ritheus* (Ad); Ælfric's *Catholic Homilies I* (ÆCHom1); Ælfric's *Catholic Homilies II* (ÆCHom2); *Homilies of Ælfric* (Æhom1); Ælfric's *Letters* (Ælet); Alcuin *De virtutibus et vitiis* (Alc), Alexander's *Letter to Aristotle* (Alex); *The Old English Apollonius of Tyre* (AoT); The Rule of St. Benedict (BenR); *The Rule of St. Benedict* – Winteney Version (BenRW); *Pastoral Care* (CP); *The Old English Gospel according to Saint John* (Jn); *The Old English Gospel according to Saint Luke* (Lk); *The Old English Gospel according to Saint Mark* (Mk); *The Old English Gospel according to Saint Matthew* (Mt); *The Old English Orosius* (Or); *Beowulf* and minor poems (VRS); *The Homilies of Wulfstan* (Whom).