The validity of lemma-based lexical richness in authorship attribution: A proposal for the Old English Gospels[1]

Antonio Miranda-García and Javier Calle-Martín
University of Málaga

Abstract

The measure of vocabulary richness to determine the authorship of literary texts has resulted in the formulation of an innumerable set of constants with a varied level of success. Our main objective is to prove that the translatorship of the West Saxon Gospels (WSG henceforth) accomplished by traditional methods cannot be held as valid in view of the statistical data generated[2]. To fulfil our aim, we have analysed the four gospels for lexical richness through a set of statistical values and constants. Besides the analysis of each of the gospels, we have also carried out a partitioning and randomized study, a cumulative study and a contrastive study. The main innovation rises, however, from the lexical items used, as we rely not only on word-types but also on lemmas. The results have allowed us to rank the four gospels in terms of lexical richness, to distinguish between them and even to cluster any of their partitions. Finally, we conclude with a new proposal for the translatorship of the WSG.

1 Introduction

The question of authorship of the West Saxon Gospels, despite the due attention paid by scholars, is still contradictory insofar as there is not a convincing explanation of the similarities between one another and, more importantly, the number of translators involved. As a matter of fact, the traditional methodology has worked out two hypotheses, pointing either to a single or a multiple composition of the pieces at hand.

In this vein, the most recent contribution is that of Liuzza (2000: 102–19), who provides a comprehensive account of these antagonistic positions, that is to say, Drake's multiple authorship conception on the one hand and Bright's unity of authorship on the other. While Bright grounds his position on the existence of

the same errors when translating similar Latin constructions, Drake explores vocabulary use to conclude that their authorship is at least dual, and probably triple, *Mt* being exclusively by one translator, *Mk* and *Lk* by another and *Jn* by a third one (Liuzza 2000: 102). To check the validity of these hypotheses, Liuzza grouped the translation differences in the gospels around three broad categories: a) semantic (errors and misconstructions in one gospel not found in parallel passages in the others), lexical (recurring words translated differently in each version) and syntactic (common Latin constructions treated differently in each gospel). Liuzza (2000: 119) agrees with Drake's conclusions in the sense that "several hands worked on the translation". As a result of his analysis, the clearest distinction may be drawn between *Mk* and *Lk*, on the one hand, and *Mt* and *Jn*, on the other, while, at the same time, further distinctions may be made between *Mt* and *Jn* in several areas.

In our view, however, we consider that a study exploring lexical richness and repetition, both of words and lemmas, could actually afford new and revealing data concerning the authorship of the *WSG*. In the light of this, the present paper has been organized as follows. The first part briefly summarizes the methodology used. The second describes the similarities and differences within and between the *WSG* in an attempt to rank them accurately in terms of lexical richness and so distinguish between them, and/or to be able to cluster their sections or parts. For this purpose, parameters such as the rate of lexical richness and repetition, Yule's and Zipf's characteristics (Yule 1944: 57; Orlov 1983: 154–233), along with some others, will be used. The conclusions reached after the investigation, both those referring to the validity of the methodology itself and, in particular, our proposal for the authorship of the *WSG*, close the paper.

2 Methodology

An annotated corpus (consisting of both lemma and tagging) was used, from which the data were retrieved automatically by the *OEC – Old English Concordancer* – (Miranda et al. 2004), a software application able to solve any kind of query by means of Boolean filters, both word and lemma-based.

The process was as follows. First, the corpus, manually macronized, was morphologically tagged by *MAOET – Morphological Analyser of Old English Texts* – (Miranda, Triviño and Calle 2000: 127–145), which generated all the possible tags for each word, regardless of context. Second, the corpus was manually disambiguated according to context. Next, the resultant corpus was the input for the *OEC*, hence providing all kind of lexical information including lists, indexes, concordances, lexical profile and statistics.

The statistical study of the data retrieved from the *OEC*, which was carried out using an Excel spread-sheet, stands out as an added asset for stylometric (Burrows 2003:10) and authorship attribution studies inasmuch as the accurate figures of both lemmas and function words may constitute a revealing line of investigation. The etiology of this statement lies perhaps in the features of function words: a) they constitute a closed set or inventory; b) most of them are invariable (as their form is not affected because of flexion or accidence); and c) they have a higher frequency than lexical words. Lemmas, in turn, resemble some of the properties of function words, at least from a quantitative perspective because of the following facts: a) their inventory will always be smaller than that of word-types; b) they will be as invariable as function words; and c) their frequency will always be at least equal or greater than their related word-types.

For all these reasons, we will work with word-types and lemmas, whenever possible, but our intuition – though strongly pivoted on experience – advises us to work with lemmas for the sake of reliability, especially when dealing with texts of a highly inflected language like Old English. Although we acknowledge that tagging involves "the regrettable intrusion upon the data and avoids the interchange of information with colleagues" (Burrows 2003: 10), or even "experimental corruption" (Rudman 1998: 357), we consider that the reliability of the results will compensate for the risks of accomplishing the time-consuming task of lemmatization.

3 Lexical richness in the WSG

This part explores vocabulary richness in the *WSG* using different mechanisms, first as a whole, then using similar-sized text-series, a set of randomly-selected samples, a cumulative study, and finally a lemma-based analysis.

3.1 A unitary study

The unitary study first starts with the calculation of the running words or tokens *N*, the different words or word-types *V(N)*, the lemmas *L(N)*, the *hapax legomena HL*, the *hapax dislegomena HD*, the most frequent word *MFW* and the most frequent lemma *MFL*. In addition, the lemmas considered as *hapax legomena HLL* and *dislegomena HDL* were also calculated. Then, we continued with the computation of a set of constants for vocabulary richness, namely, a) the ratio *V(N)/N* or Mendenhall's characteristic (1827: 237–49); b) the ratio *L(N)/N*; c) *V(N)/L(N)*; d) Yules's characteristic *K*; e) Zipf's constant *Z* as well as some others.

Table 1: Absolute values in the *WSG*

	Jn	*Lk*	*Mk*	*Mt*
N	17,082	20,989	12,350	20,230
V(N)	2,223	3,802	2,477	3,792
L(N)	977	1,532	1,174	1,527
MFW	947	1502	1007	1414
HL	1163	2298	1484	2256
HD	309	537	374	564
MFL	1103	1502	1007	1413
HLL	383	613	589	660
HDL	143	224	175	564

The ratios *V(N)/N* and *L(N)/N* for *Lk* and *Mt* practically coincide (their divergence < 10^{-2}), an expected consequence given the similitude of the values for *N*, *V(N)*, and *L(N)*. These values are slightly exceeded in *Mk*, due to its shorter text-length, and, on the contrary, are not reached by *Jn*, which is undoubtedly the least rich of the four. Accordingly, their ranking in terms of *V(N)/N* would be

$$Mk > Mt > Lk > Jn \qquad (r_1).$$

The results obtained with the other formulae – such as Dugast (1979: 23), Maas, Guiraud or Brunet (cited in Tweedie and Baayen 1998: 326–28) – are also similar, as all of them have proven to be text-dependent. If, for example, we follow Guiraud, the results modify the text-dependency of the first expression and thus rank the texts as

$$Lk > Mt > Mk > Jn \qquad (r_2),$$

the same as would result if *L(N)/N* were the ratio used[3].

The proportion *V(N)/L(N)* indicates the index of *flexionability*, or of allomorphism, of the text, the minimum value being 1, when the number of lemmas equals that of word-types, and the maximum approaching 3. Although this constant cannot alter the above rankings (as its values depend both on *V(N)* and *L(N)*), it allows us to guess the behaviour of each depending on whether the text presents a high, medium or low number of inflections.

We have also calculated the lexical repetition through Yule's characteristic *K* (Yule 1944: 57) with the following formula (Tweedie and Baayen 1998: 330; Hoover 2003: 174)[4],

$$K = 10^4 \left[-\tfrac{1}{N} + \sum_i V(i,N)(\tfrac{1}{N})^2 \right]$$

and the values for *K* also originate the ranking r_2 above.

However, when we calculate the lexical richness through Zipf's constant (Orlov 1983: 154–233) with the formula

$$V(N) = \frac{Z}{\log(p \cdot Z)} \cdot \frac{N}{N - Z} \cdot \log(N/Z)$$

the ranking obtained is

$$Mt > Lk > Mk > Jn \qquad (r_3).$$

Although *K* and *Z* are widely accepted as the most reliable constants because they are not so text-dependent, we have also computed the values of some other constants, as shown in Table 2:

Table 2: Values of constants in the *WSG*

	Jn	Lk	Mk	Mt	Ranking
V(N)/N	0.130	0.181	0.200	0.187	r_1
L(N)/N	0.057	0.072	0.095	0.075	r_1
V(N)/L(N)	2.27	2.481	2.10	2.483	r_5
Guiraud	16.93	26.94	22.29	26.68	r_2
K	100.96	110.20	133.19	100.32	r_6
Z	12,375	47,950	33,950	86,320	r_3
HL/HD	3.7947	4.2793	3.9679	4	r_2
HL/(V)N	0.5264	0.5886	0.5991	0.5944	r_1
HL/N	0.0179	0.0255	0.0302	0.0278	r_1
$HL^3/V(N)^2$	322.86	796.21	532.65	797.24	r_3
TTR	11.97	19	15.76	18.86	r_2
REPEAT	16.6329	20.4371	12.0253	19.698	r_4

L(N)/N	0.05719	0.07299	0.0950	0.07548	r_1
HLL/HDL	2.6232	2.7366	3.0228	2.716	r_6
HLL/L(N)	0.3849	0.4003	0.4505	0.4195	r_1
HLL/N	0.02956	0.03295	0.02637	0.03518	r_5
REPEAT L(N)	0.00011	0.00012	0.00019	0.00013	r_1

The analysis of these data has led us to conclude that, in most of the cases, either r_1, or r_2 results. Only after computing Z and $HL^3/V(N)^2$, the ranking is again r_3. The rankings *Mt* > *Lk* > *Jn* > *Mk* and *Mk* > *Lk* > *Jn* > *Mt* correspond to the notation r_5 and r_6, respectively.

3.2 A partitioning study

The only accurate method to calculate the lexical richness of texts, neglecting this text-length dependency, is undoubtedly to handle texts of the same length. On the grounds of this, each gospel has been divided into blocks of 3,000 words (the Maximum Common Divisor to all the texts) so as to calculate the lexical richness of each block in terms of *V(N)* and *L(N)*, and compare them between and within the four gospels. Given that *N* varies for each gospel, it seems obvious that the number of blocks will also vary. The less-than-3,000-word partitions are shown in brackets in Table 3, but they will not be compared for the reasons above.

Along with these data, their *mean value* (μ) and the *standard deviation* (*SD*), we also include the data of a 3,000-word randomly-built passage (*RND*), composed of six 500-word passages.

Table 3: Values of *V(N)* and *L(N)* in each block of the *WSG*

	V(N) in blocks of 3,000 words									
	0–3	3–6	6–9	9–12	12–15	15–18	18–21	μ	SD	RND
Jn	770	729	674	749	670	*(602)*	------	718.4	30.65	716
Lk	971	960	922	1009	975	979	*(946)*	969.3	63.46	1005
Mk	899	890	854	965	*(236)*	------	------	902	40.07	965
Mt	960	994	952	918	918	989	*(886)*	955.1	30.14	998

	L(N) in blocks of 3,000 words									
Jn	416	392	366	420	344	*(359)*	------	387.6	29.13	383
Lk	567	523	518	583	570	538	*(541)*	549.8	24.75	575
Mk	517	521	485	554	*(167)*	------	------	519.2	30	542
Mt	572	553	511	511	509	570	*(498)*	537.7	28	564

The figures for *V(N)* in *Lk* and *Mt* (ranging from 900 to 1,000) rather approximate in the first three series (from 96.5 to 98.9%), diverge a little in the next two, and converge again in the sixth. The figures for *Mk* (ranging from 800 to 900) approach those of *Lk* and *Mt* which, in turn, significantly diverge from those for *Jn* (ranging from 600 to 800). However, it is worthwhile to point out that *Mt's SD* is the least in opposition to that of *Lk*, which is the greatest. The results of the randomly-built block are congruent with those in the series inasmuch as they could be interpolated within the value range.

Likewise, the figures for *L(N)* in *Lk* and *Mt* (ranging from 509 to 572) nearly overlap in the first and third block ($\delta < 1\%$), and diverge a little in the rest ($\delta < 7\%$). The values for *Mk* (ranging from 485 to 554) approximate to those of *Lk* more than to *Mt's*. The figures for *Jn* (ranging from 344 to 420) diverge more significantly from the others. As the μ of *V(N)* for *Lk* is slightly greater than that for *Mt*, and the former's *SD* doubles the latter's, there is not a clear distinction between *Lk* and *Mt*. Therefore, it seems that the most accurate ranking in terms of *V(N)* would be

$$Lk/Mt > Mk > Jn \qquad (r_4),$$

both variants of r_2 or r_3. These results seem to confirm, to a certain extent, the tendency of the values for lexical richness in section 3.1.

Accordingly, by means of *V(N)* we can successfully distinguish between any 3,000-word passage of *Lk* and *Mt*, either from *Mk*, or from *Jn*. By applying $HL^3/V(N)^2$ in each block, the same distinction is achieved. In terms of *L(N)*, however, we can successfully distinguish between any 3,000-word passage of *Lk*, *Mk*, and *Mt* from *Jn*, as a similar ranking is replicated. Therefore, the use of *V(N)* seems to be more accurate than that of *L(N)* for the purpose of distinguishing or clustering 3,000-word passages.

In order to check these rankings, the blocks were decreasingly ranked in terms of the values for *MFW, HL, HL/HD* and K_2 (Yule's characteristic calculated in terms of lemmas), to find out their coincidence with the rankings presented so far, but no clear systematic evidence was found. Lastly, the values for

K (*x*-axis) against *Z* (*y*-axis) of each block were plotted and the figure obtained made us realize that the rankings by *Z* coincide with the one by *L(N)*. Although this could be an expected result, whenever same-sized texts are treated, this fact comes to strengthen our initial intuition that lemmas can be used as an indicator of lexical richness.

In addition, the increasing percentage of the values of one block with respect to the previous were also calculated, as shown in Table 4. For example, the cell (*Jn*, 3–6) contains the increasing percentage found between the second block (3,000–6,000 words) and the first one (0–3,000 words), and so on in any direction.

Table 4: Rate of *V(N)* and *L(N)* along the blocks of the *WSG*

	V(N) % increase						
	3–6	6–9	9–12	12–15	15–18	18–21	Total
Jn	51.16	24.82	20.09	13.81	11.37	------	287
Lk	65.08	30.03	25.65	16.95	12.54	10.21	392
Mk	63.07	31.44	24.75	3.03	------	------	275
Mt	71.04	34.28	19.22	15.44	14.49	9.20	395
	L(N) % increase						
Jn	41.1	19.59	17.23	9.47	9.76	------	237
Lk	44.8	20.95	18.52	11.55	8.37	7.58	270
Mk	50.9	22.93	20.85	1.82	------	------	227
Mt	49.3	21.19	15.07	11.92	9.22	8.03	275

As in the case of *V(N)*, *L(N)* also shows a regular increasing of the values from one block to the next, where the percentages for *Lk* and *Mt* are similar. On the contrary, the number of times that the initial lemmas in block 0–3,000 have been increased is obviously less than in *V(N)*. In other words, *Mt* is the one that increases the most (275%), whereas *Mk* is the one with the least increase (only 227%).

It goes without saying that the size of the increase gets smaller from one to the other block (both for *V(N)* and *L(N)*); so, it may become illustrative to observe the plotting of the relative values of *V(N)* and *L(N)* as the richness rate decreases with increasing vocabulary, and vice versa.

3.3 A cumulative study

A cumulative study has also been carried out to find out a) whether the same tendency is maintained or not; b) whether the lexical growing from one to the next block is positive or negative (as repetition is not allowed); and c) whether the increasing or decreasing percentage is regular or systematic. The same size for the passages (3,000 words) has also been adopted. In this fashion, the first block contains the first 3,000 words, the second contains the first 6,000 words, and so on.

Table 5: Values of *V(N)* and *L(N)* in each cumulative block of the *WSG*

				V(N)			
	0–3	0–6	0–9	0–12	0–15	0–18	0–21
Jn	770	1,164	1,453	1,745	1,986	2,223	------
Lk	971	1,603	2,089	2,625	3,070	3,455	3,802
Mk	899	1,466	1,927	2,404	2,477	------	------
Mt	960	1,642	2,205	2,629	3,035	3,475	3,792
				L(N)			
Jn	416	587	702	823	901	977	------
Lk	567	821	993	1,177	1,313	1,423	1,532
Mk	517	776	954	1,153	1,174	------	------
Mt	572	854	1,035	1,191	1,333	1,456	1,573

The plotting of the trajectories for *V(N)* answers the three rhetoric questions above. As for the first, it is obvious that the ranking r_4 is maintained. With respect to the other two questions, an affirmative answer is put forward on account of the fact that a regular increasing rate is observed, with the exception of the last block of *Mk*.

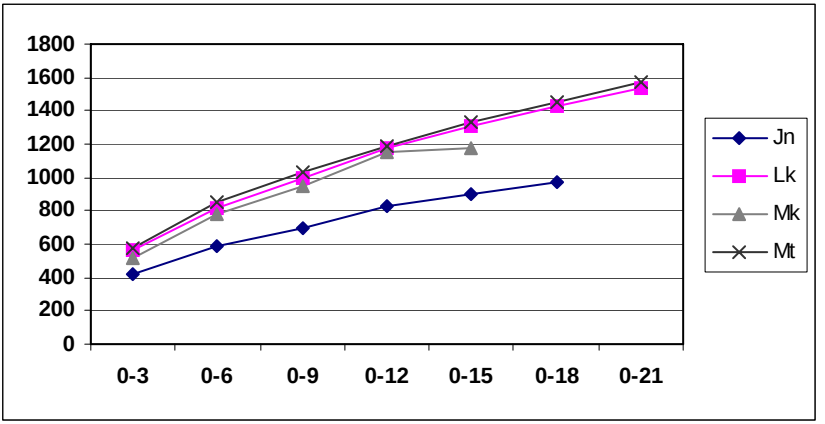

Figure 1: Plotting of L(N) along the blocks

Three clear-cut conclusions can also be drawn from these data. The first has to do with the regular increasing of the values from one block to the next. The second confirms the proximity of the figures for *Lk* and *Mt*, the maximum distance being 6. In fact, along the series of blocks, *Jn* and *Mk* nearly triplicate the number of *V(N)*, whereas *Lk* and *Mt* practically quadruplicate the values of the first block. As a consequence of the second deduction, the third replicates the ranking of r_4.

The first deduction for *V(N)* is also applied to the data for *L(N)*. However, the difference lies in the fact that the values for *Mt* are greater than those of *Lk* in each and every block and, therefore, the ranking is r_3. In this fashion, we found the solution for this particular case after having examined all the data for months without realizing that our intuition seemed to hold. Once again, we could not see the wood for the trees, as it is the number of lemmas in any of the cumulative blocks, be they 0–3,000, 0–6,000 or 0–12,000, and so on, that lets us identify and cluster any set. We have only checked the Old English *Apollonius of Tyre* for corroboration and, certainly, it does hold.

3.4 A contrastive study

We have already studied the distribution of lemmas in the *WSG* from different perspectives: as a whole, in a partitioning and in a cumulative model. Our intuition led us to consider more closely the use of lemmas in the evaluation of lexical richness for comparative purposes (think of the mismatches originated by

allomorphism). Thus, it is helpful to find out the number of *common lemmas*, that is, the lemmas occurring in the four gospels, *shared lemmas*, that is, those lemmas occurring in more than one gospel, and *hapax lemmata*, that is, lemmas occurring just in one.

The complete inventory of (different) lemmas in the *WSG* is 2,828, from which 509 are common to the four gospels, and 267, 453, 303, 489 are exclusive (*hapax*) in *Jn*, *Lk*, *Mk* and *Mt*, respectively. The rest of the data are included in Table 6:

Table 6: Distribution of *L(N)* in terms of common, hapax and shared lemmas

	4 gospels	*Jn*	*Lk*	*Mk*	*Mt*
Common	509	509	508	509	509
Hapax	1,512	267	453	303	489
Shared	807	201	571	362	529
Total	2,828	977	1,532	1,174	1,527

The figures for shared lemmas and hapax lemmata are quite similar in *Lk* and *Mt* as in the previous cases. Analogously, the ratio of lemmas to the total of lemmas in *Lk* and *Mt* is also similar. This similarity made us suspect that their grouping in pairs could provide useful information about their level of coincidence. In this vein, we obtained the following results: *Jn* & *Lk* 601; *Jn* & *Mk* 579; *Jn* & *Mt* 603; *Lk* & *Mk* 750; *Lk* & *Mt* 932; *Mk* & *Mt* 769.

Figure 2 below does not need interpretation as the eccentricity of the hexagon points markedly to the vertex *Lk* & *Mt*, and to a lesser extent, to the ones noted as *Mk* & *Mt* and *Lk* & *Mk*. Again, we obtain the ranking r_4 in the sense that *Lk* and *Mt* share the greatest number of lemmas and, on the contrary, when *Jn* is analysed, greater differences arise. However, taking into consideration that the pair *Mt*/*Mk* share more lemmas than the pair *Lk*/*Mk*, it is not daring to state that the ranking r_4 must be rewritten as

$$Mt > Lk > Mk > Jn \qquad (r_2).$$

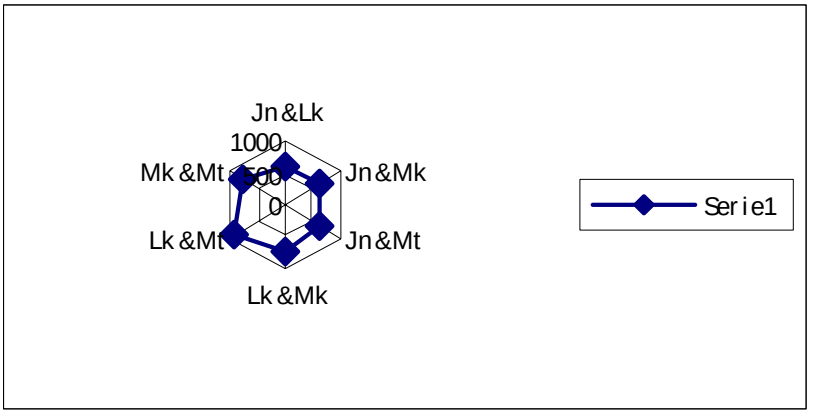

Figure 2: Distribution of shared lemmas in the WSG (in pairs)

4 Conclusions

In the preceding pages we have analysed the vocabulary richness of the *WSG* from some lexical perspectives in an attempt to provide a humble contribution to the development of authorship attribution studies. The data necessary for any further checking have also been afforded. On the basis of our analysis, we are in a position to draw the following conclusions:

First, there seems to exist grounded evidence to state that in terms of lexical richness, lexical repetition and lemma distribution, *Lk* and *Mt*, and to a lesser extent *Mk*, are lexically richer than *Jn*. This similitude should not surprise us inasmuch as, from a thematic point of view, they constitute what is known as the synoptic gospels. However, we are convinced that we have demonstrated that, using lexical richness, *Mt* is richer than the other three, so that their ranking in terms of lexical richness is *Mt* > *Lk* > *Mk* > *Jn*, hence confirming the results obtained by *Z*.

Second, our serial analysis by blocks allows us to distinguish between *Mk* and *Jn* from *Lk* and *Mt*, but the cumulative study allows us to tell them apart and, consequently, to cluster any passage on condition that the above text-lengths are maintained (3,000, 6,000, 9,000 tokens, etc.). The methodology employed is as old as the mountains and so simple that only lemma counting is required. The only difficulty arises from the need of a lemmatized corpus, which becomes a tiresome task even if you can count with tools such as *MAOET*.

Third, lemma-based studies can be as valid as those word-based, if not more accurate. Two facts can be highlighted: 1) the values for lexical richness and lexical repetition are more concentrated, and 2) the invariability of lemmas, a common feature with function words, is also an added asset.

Fourth, the data obtained from this study also allow us to question Drake's and Liuzza's proposal (Liuzza 2000: 102–103). Leaving aside the fact that *Mt* is lexically richer than the other three, there are quantitative and qualitative near-coincidences with *Lk* that strongly suggest that they were translated by the same person or that the translators shared the same lists, phrases, scriptorium, etc. All in all, we are in a position to propose a triple translatorship for the *WSG*: one for *Mt* and *Lk*, another for *Mk* and a third for *Jn*. We cannot obviate, however, those between the versions, probably because the translators shared the same materials. We hope to further this study with other works that will confirm this research, especially with texts of highly inflected languages.

Notes

1. The present research has been funded by the Spanish Ministry of Science and Technology (grant number BFF 1835/2001). This grant is hereby gratefully acknowledged.
2. West-Saxon (WS for short) is the literary dialect of Old English (OE for short) in which most extant works of the period are written.
3. Guiraud's expression is $V(N)/\sqrt{N}$.
4. An equivalent expression is offered by Smith and Kelly (2002: 414) as $K = 10000 \times \left(\sum i^2 V(i, N) / N^2 - 1 / N\right)$.

References

Biber, Douglas, Susan Conrad and Randi Reppen. 1998. *Corpus linguistics. Investigating language structure and use*. Cambridge: Cambridge University Press.

Bosworth, Joseph and Thomas Northcote Toller. 1898. *An Anglo-Saxon dictionary.* Oxford: Oxford University Press.

Burrows, John. 2003. Questions of authorship: Attribution and beyond. *Computers and the Humanities* 37: 5–32.

Dugast, Daniel. 1979. *Vocabulaire et stylistique*. Vol. I: *Théatre et dialogue*. Paris: Slaktine-Champion.

Haan, Peter de and Erik Schils. 1994. The Qsum plot exposed. In U. Fries, G. Tottie and P. Schneider (eds.). *Creating and using English language corpora*, 93–116. Amsterdam: Rodopi.

Hall, John Richard Clark. 1996. *A concise Anglo-Saxon dictionary.* Toronto-Buffalo-London: University of Toronto Press-Medieval Academy of America.

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

Kjellmer, Göran. 1994. Lexical differentiators of style: Experiments in lexical variability. In U. Fries, G. Tottie and P. Schneider (eds.). *Creating and using English language corpora*, 117–26. Amsterdam: Rodopi.

Laan, Nancy M. 1995. Stylometry and method. The case of Euripides. *Literary and Linguistic Computing* 10: 271–78.

Liuzza, Roy Michael. 1994–2000. *The Old English version of the gospels*. 2 vols. Vol. I: *Text and introduction*. Vol. II: *Notes and glossary.* Oxford: Oxford University Press.

Mendenhall, Thomas Corwin. 1887. The characteristic curves of composition. *Science* 11: 237–49.

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

Miranda García, Antonio, José Luis Triviño Rodríguez, Javier Calle Martín and David Moreno Olalla. 2001. CALLOE: A pedagogical tool for the learning of Old English. *Old English Newsletter* 34.3: 12–20.

Miranda García, Antonio, Javier Calle Martín, David Moreno Olalla and Gustavo Muñoz González. Forthcoming. The Old English *Apollonius of Tyre* in the light of the Old English Concordancer. In A. Renouf (ed.). *The changing face of corpus linguistics*. Amsterdam: Rodopi.

Orlov, Ju. 1983. Ein Modell der Häufigkeitsstruktur des Vokabulars. In H. Guiter and M.V. Arapov (eds.). *Studies on Zipf's Law,* 154–233. Bochum: Brockmeyer.

Rudman, Joseph. 1998. The state of authorship attribution studies: Some problems and solutions. *Computers and the Humanities* 31: 351–65.

Smith, Joseph A. and Colleen Kelly. 2002. Stylistic constancy and change across literary corpora: Using measures of lexical richness to date works. *Computers and the Humanities* 36: 411–30.

Somers, Harold and Fiona Tweedie. 2003. Authorship attribution and pastiche. *Computers and the Humanities* 37: 407–29.

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

Yule, George Udny. 1944. *The statistical study of literary vocabulary.* Cambridge: Cambridge University Press.