Was Shakespeare's Vocabulary the Richest?

Cyril Labbé, Dominique Labbé

Was Shakespeare's Vocabulary the Richest?

Cyril Labbé[1], Dominique Labbé[2]

[1] Laboratoire d'Informatique de Grenoble - Université Joseph Fourier (cyril.labbe@imag.fr)

[2] Laboratoire PACTE (CNRS - Institut d'Etudes Politiques de Grenoble)
(dominique.labbe@umrpacte.fr)

In Née Emilie, Daube Jean-Michel, Valette Mathieu, Fleury Serge (dir.). *Proceedings of the 12th International Conference on Textual Data Statistical Analysis*. **Paris: June 3-6 2014, p 323-336**.

Abstract

It is generally assumed that the vocabulary of W. Shakespeare is exceptionally rich and his work contains a very large number of different words. We present a method to compare the extent of the vocabularies of several authors' works of unequal length. Applied to the theater of Shakespeare's time, it shows that the vocabulary of Shakespeare is not exceptional and that some or his contemporaries – like B. Jonson or T. Dekker – used a larger vocabulary.

Résumé

Il est généralement admis que le vocabulaire de W. Shakespeare est remarquablement riche. Son œuvre contiendrait un très grand nombre de mots différents. On présente une méthode qui permet de comparer *I*montre que le vocabulaire de cet auteur n'a rien d'exceptionnel et que certains contemporains – comme B. Jonson ou T. Dekker – utilisaient un vocabulaire plus étendu.

Keywords: lexicometry ; type-token ratio ; vocabulary richness ; vocabulary growth ; vocabulary specialization ; English theater ; Shakespeare

"Shakespeare, who displayed a greater variety of expression than probably any writer in any language, produced all his plays with about 15,000 words. Milton's works are built up with 8,000 ; and the Old Testament says all that it has to say with 5,642 words". English country laborers of the day had not 300 words in their vocabulary" "a well-educated person in England, who has been at a public school and at university, who reads his bible, his Shakespeare, and the Times… seldom uses more than about 3,000 or 4,000 words in actual conversation… and eloquent speakers may rise to a command of 10,000"

(Muller F. M., *Lectures on the Science of Language*. New York: Scribner, 1862, p. 377-379, quoted by Elliott & Valenza 2004)

"By comparison with other writers of the time, Shakespeare has a large recorded vocabulary".

(Maguire L. & Smith E. *30 Great Myths about Shakespeare*. Oxford: Wiley & Sons, 2013, p. 138).

1. Introduction

Among many others, the two above quotations expresses a wide-spread opinion: the vocabulary of W. Shakespeare (1564-1616) is the richest. During the 19th century, this richness was considered as absolute; nowadays this large vocabulary is relative to the Elizabethan era, but the idea remains.

The figures displayed to support this assumption, show that the richness is understood as the size of the vocabulary used in a work. It is proposed to test this opinion by studying the plays by Shakespeare: are their vocabularies more extensive than that of other plays from the same age but by different authors? This opinion has already been challenged by several studies (Eliott & Valenza 2004; Craig 2011). It should also be noted that we are only interested in the vocabulary actually observed and not in an estimation of the total vocabulary known by these authors (Efron & Thisted 1976; Thisted & Efron 1987).

Following the common intuition, one can define the vocabulary richness as the number of different words that can be found in a text or in the authors' oeuvre. The more they are, the greater the vocabulary richness or, inversely, the lower it is, the poorer the vocabulary.

This definition raises two important considerations.

First, one must "standardize" the spelling, so that a word is always written the same way: "One word, one spelling". This is quite important: at the beginning of the 17th century, spelling convention in books wasn't as strict as it is nowadays (see examples given by Elliott & Valenza 2004). A careful spelling standardization is a time consuming process. W. Elliot and T. Merriam have mostly done it and kindly provided us the 89 plays used in this experiment.

Secondly, the richness should not be considered as an absolute value but as a relative value helping comparison between authors, plays or corpora. In other words, it is not necessary to know if Shakespeare vocabulary is the "richest" – as stated by F.M. Muller – it is sufficient to test if it is richer (or poorer) than the vocabulary used by others authors of its time (of whom

there are at least two plays by them): F. Beaumont (1585-1616), T. Dekker (1572-1632), G. Chapman (1559 – 1634), J. Fletcher (1579-1625), R. Greene (1558-1592), B. Jonson (1572-1637), T. Kyd (1558-1594), C. Marlowe (1564-1593), T. Middleton (1580 ? -1527), G. Peele (1556-1596). Other authors like S. Daniel, J. Ford, T. Heywood, J. Lyly, T. Nashe and H. Porter are omitted from this experiment because we presently have only one play by each of them.

To achieve this goal we have used a corpus composed of 89 plays written during the Shakespeare's lifetime - so called "Elizabethan-Jacobean" or "Early Modern" period, EM in the following (annex; for more information on this period, see: Chambers 1923). Texts have been processed following the norm OCP ("Oxford Concordance Program": Hockey & Martin 1998).

2. Vocabulary Richness and text lengths

The relation between the text length (N) and the vocabulary size (V) is known as "Type-Token-Ratio ". It is a well-studied question in "linguistic computing" (Wimmer & Altmann 1999). In the corpus EM, a visible relation exists between these two variables. For example, *Hamlet* is the longest play by Shakespeare (29 549 tokens) and it is the one which contains the largest number of different word types (4 663); *The Comedy of Errors* - the shortest play by Shakespeare (14 358 tokens) - contains the fewest different words types (2 504).

In Figure 1 each play of the corpus is shown as a point having as coordinates its length (number of tokens) and its vocabulary (number of word types). The vocabulary clearly grows with the length (the thin black line is the trend).

However the scatter plot shows an important dispersion around the trend, which can be referred to the differences (variations) of "vocabulary richness" (R).

The observation of this graph can be meaningful with regard to some comparisons. As an example, in the corpus EM, two plays (*Bartholomew Fair*, *A King and no King*) have lengths greater than *Hamlet* and smaller vocabularies (see Table 1). This obviously shows that the vocabulary used in these two plays is poorer than the one in Hamlet.

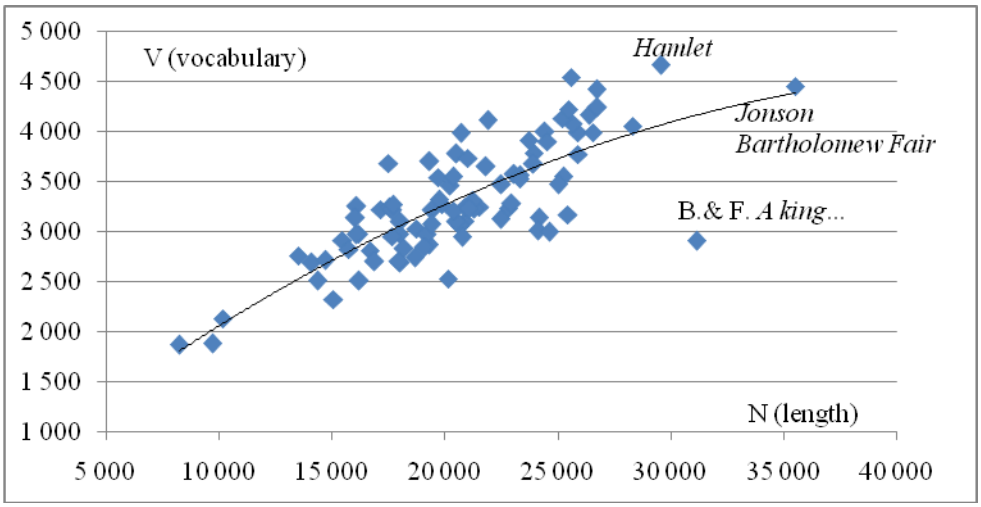

Figure 1. Relation between the number of word types (V) and the number of tokens (N) in each early modern play.

Author	Play	N	V
Jonson B.	*Barthomew Fair*	35 501	4 455
Beaumont F. & Fletcher J.	*A King and no King*	31 127	2 904
Shakespeare W.	*Hamlet*	29 549	4 663

Table 1. Lengths and vocabularies of the three longest EM plays

Given this, it is possible to write these two inequalities:
$\{V_{Hamlet} > V_{A\ King}\ ;\ N_{Hamlet} < N_{A\ King}\} \Rightarrow R_{Hamlet} > R_{A\ King}$
$\{V_{Hamlet} > V_{Barthomew\ Fair}\ ;\ N_{Hamlet} < N_{Barthomew\ Fair}\} \Rightarrow R_{Hamlet} > R_{Barthomew\ Fair}$.

However, from the above two inequalities, it is not possible to conclude that: $R_{Barthomew\ Fair} > R_{A\ King}$ because: $N_{Barthomew\ Fair} > N_{A\ King\ and\ no\ King}$

In other word the relation "is richer than" is not a complete order and two plays are often not comparable. Sometimes, the complete comparison of three plays is possible (for example Table 2).

Author	Play	N	V
Middleton T.	*The Nice Valour or The Passionate Madman*	14 095	2 687
Shakespeare W.	*Comedy of Errors*	14 358	2 504
Greene R.	*Alphonsus, King of Aragon*	15 067	2 321

Table 2 Lengths and vocabularies of three comparable EM plays

From the point of view of their "vocabulary richness", these three plays can be classified as follows:
$R_{Nice\ Valour(}$Middleton$) > R_{Comedy\ of\ Errors}$ (Shakespeare) $> R_{Alphonsus}$(Greene)

Other comparisons are possible. Given its length, *A Game of Chess* (Middleton: 17 503 tokens and 3 684 types) has a vocabulary richness greater than those of the 38 plays longer than it (Annex). This includes 8 plays by Middleton himself (out of a total of 14 plays by him in the corpus) and 14 by Shakespeare (out of his 38 plays in the corpus): *All's Well That Ends Well* (V = 3 469), *As You Like It* (3 228), *Julius Caesar* (2 840), *King John* (3 546), *Measure for Measure* (3 307), *The Merchant of Venice* (3244), *Merry Wives of Windsor* (3 226), *Much Ado About Nothing* (2 942), *Pericles* (3 218), *Richard II* (3 650), *The Taming of the Shrew* (3 208), *Timon of Athens* (3269),*Titus Andronicus* (3 319), *Twelfth Night* (3 074).

In Annex, other direct comparisons are of great interest, for example, Chapman's *Bussy d'Ambois* has a greater vocabulary richness than the Shakespeare's plays quoted above (except *Richard II* and *King John* which are longer than *Bussy* and therefore impossible for a direct comparison with it).

This suggests that, even if some Shakespeare's plays seem to have a rich vocabulary (particularly the historical ones), none of them would appear to be of an extraordinary/outstanding richness. A two by two comparison of authors can also be affected by the fact that the EM corpus contains a disproportionately large set of Shakespeare's plays (see Table 3). Nevertheless these direct comparisons can be helpful in comparing authors.

	Number of plays	N (tokens)	V (different types)
Peele G.	2	24 877	3 938
Kyd T.	2	38 231	5 064
Chapman G.	2	40 618	5 133
Dekker T.	2	43 778	5 845
Greene R.	3	51 102	5 836
Marlowe C.	7	111 858	9 164
Fletcher J. & Beaumont F.	5	116 244	7 401
Johnson B.	6	144 628	12 158
Fletcher J.	8	177 968	9 914
Middleton T.	14	263 426	13 828
Shakespeare W.	38	830 379	27 084
	89	1 843 109	

Table 3. Types and tokens in the works of the "Early Modern" authors (ranked by lengths).

Data in Table 3 (bold lines) lead to the following interpretation:
$R_{Dekker} > R_{Greene}$
$R_{Marlowe} > R_{Fletcher\&Beaumont}$
$R_{Johnson} > R_{Fletcher}$

Again, when directly using *N* and *V*, it is impossible to set up a complete comparison of the works of each author. If we consider that the observed size of the vocabulary (*V*) is a function of the text's length (*N*) and of its vocabulary richness (*R*), then to compare *R* in two texts - of unequal lengths - one must be able to neutralize the impact of *N* on *V*. This can be done by modelling the way the vocabulary grows with the number of tokens used.

3. Modeling the vocabulary growth

Given, a text or a corpus, let:

N : total number of tokens in this text or corpus ;

The *V* types, in the whole work, are graded in order of frequency into *n* frequency bins.

V_i: the number of types which occur *i* times.

Example: Shakespeare's *King John*: *V* = 3 546 (types), *N* = 20 375 tokens.

The problem is to predict how new words will appear while the text is growing. To study this phenomenon, the text *King Joh*n is divided in 204 slices of 100 tokens. At each interval of 100 words, the different types are counted from the beginning of the corpus. For the K milestones - 100, 200, …, 204 - let:

N_k be the number of tokens counted from the beginning of the texts until the k_{th} milestone. N_k varies from 0 to 204 (N_{204} = 20 375);

$u_k = \frac{N_k}{N}$; u_k varies from 0 (beginning of the text) to 1 (u_{204});

V_{*k} be the number of different types counted since the beginning of the texts until the k_{th} milestone; V_{*k} varies from 0 to 3 546.

Figure 2 presents the vocabulary growth in *King John* divided in slices of 100 tokens.

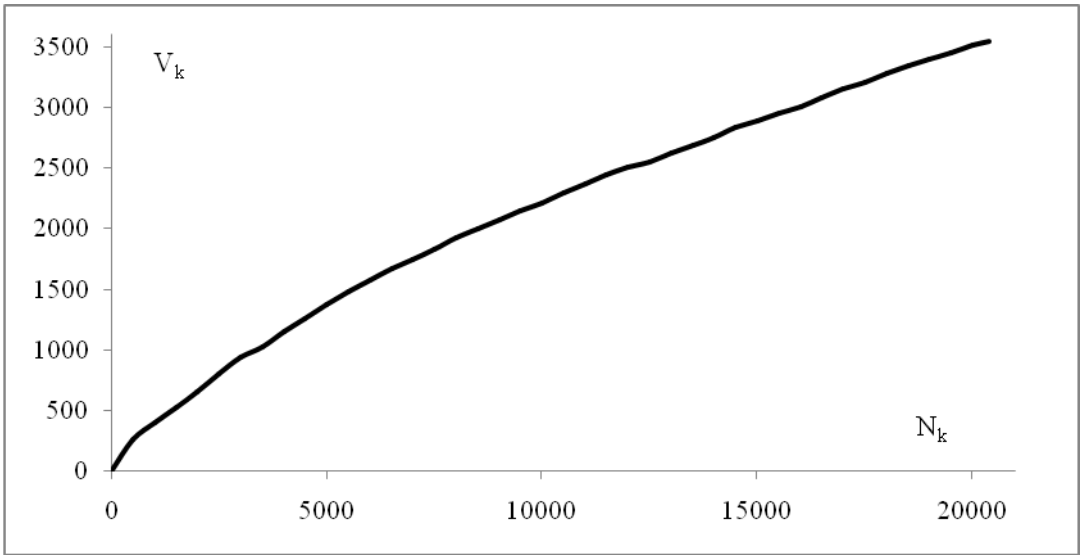

Fig. 2 Vocabulary growth in Shakespeare's *King John*

The slope of the curve slowly decreases as *N* grows and it is very similar to the one of the Figure 1: *V* is a decreasing non-linear function of *N*. To compare the vocabulary of *King John* with the one of another play of size *N'* (with $N' < N_{King\ John}$), Muller proposes to estimate the number of types (*V'*) as a random sample of size *N'* drawn out of *King John* (Muller 1977; Ule 1985):

$$(1)\quad V'(u) = V - \sum_{1}^{n} V_i Q_i(u) \quad \text{with} \quad u = \frac{N'}{N} \quad \text{and} \quad Q_i(u) = (1 - u)^i$$

The equation (1) is based on the assumption of a sampling without replacement (hypergeometric law: Hubert & Labbé 1988a). Of course, natural languages do not strictly follow this assumption and this leads to a systematic bias that is illustrated by Figure 3.

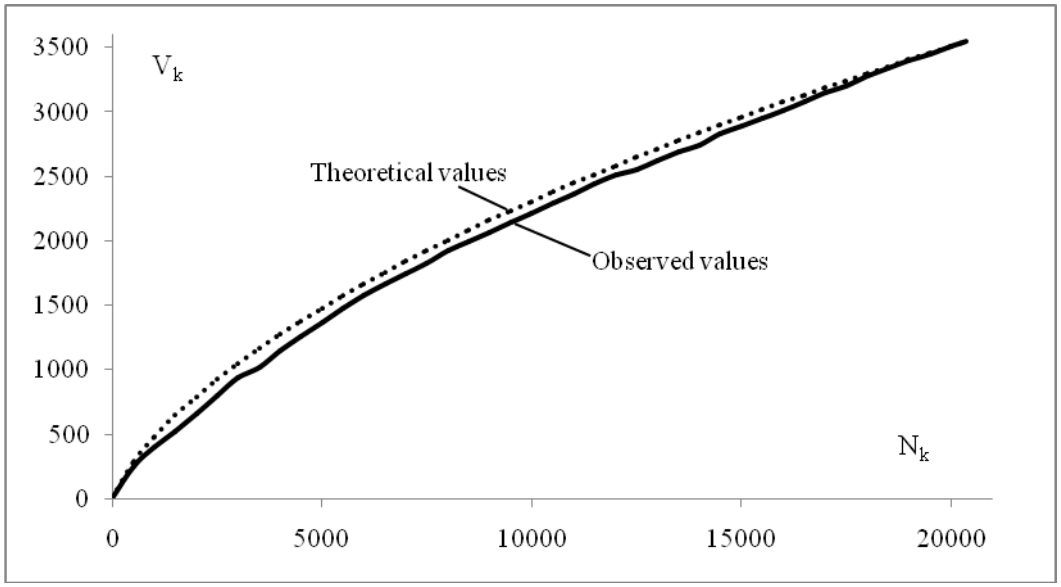

Fig. 3 Vocabulary growth in Shakespeare's *King John* (observed values (bold line) and theoretical values calculated with the help of hypergeometric model (dotted line).

In this diagram, the x-axis is the length of the text (N_k) and of the excerpts (N'_k with $N'_k \leq N_k$); the y-axis is the size of the vocabularies observed (V_{*k}) and the theoretical one calculated with

formula (1) (V'_k). The dotted line represents these theoretical values whereas the bold line is the observed values in the text *King John.* The theoretical values can be interpreted as the expected numbers of different types in K simulated excerpts drawn out of *King John* from the beginning of the text until the k_{th} milestone.

It can be seen that the theoretical values (dotted line) are almost always significantly higher than the observed ones while the theoretical curve is supposed to adjust the observations... For example, it is the case for 32 out of the 38 plays of Shakespeare. This phenomenon has been reported by Muller, Ule and Cossette (1994). According to Muller, this phenomenon is due to the so-called "specialization of the vocabulary" (*p*) according to the different topics dealt in the text. The formula (1) would thus apply to a particular case: a text without vocabulary specialization *(p = 0)*. In the EM corpus, this is the case for fewer than one play out of six. Thus to compare without bias the richness of the vocabularies of the five others, it is necessary to take into account the way the specialized vocabulary impacts the vocabulary growth.

4. Specialization of the vocabulary

First, charts like Fig. 2 & 3 are adjusted by calculating *V'* — the number of different types expected in an excerpt of *N'* tokens — according to the following formula (Hubert & Labbé, 1988a) in which the second part – between brackets - is the formula (1).

$$(2)\ V'(u) = p.u.V + (1-p)\left[V - \sum_{1}^{n} V_i Q_i(u)\right] \text{ with: } p \text{ "coefficient of vocabulary partition".}$$

The coefficient of vocabulary partition (*p*) measures the relative size of the two sets of vocabulary, which are used by one author in order to compose a text. The first set contains *pV* specialized word types which are devoted to a special part of the text. It is not possible to identify precisely these words, but various experiments have shown that they are mainly nouns of figures, towns and countries, technical terms... The average growth of this first set is a linear function of *N'* (first part of the formula (2)). The second set contains *(1-p)V* types which belong to the general vocabulary. This set contains the vocabulary used whatever the topic: articles, prepositions, auxiliary and modal verbs, etc. The probability of their appearing is constant at any stage of the text and can be estimated as if they belong to a sample of size *N'* tokens randomly drawn, without replacement, from the *N* tokens of the whole corpus. The size of this second set is estimated with the help of the hypergeometric formula: second part of the formula (1).

The value of *p* is that which minimises the sum of the squared deviations between the observed values (V^*_k) and the calculated ones (V'_k):

$$(3)\ p = \frac{\sum_{1}^{K}\left[(u_k - 1)V + \sum_{1}^{n} V_i Q_i(u_k)\right]\left[V_*(u_k) - V + \sum_{1}^{n} V_i Q_i(u_k)\right]}{\sum_{1}^{K}\left[(u_k - 1)V + \sum_{1}^{n} V_i Q_i(u_k)\right]^2}$$

Formulae (2) and (3) are easy to compute. For the calculation, the K intervals are not necessarily equal or proportional. Of course, the accuracy of results depends on the number

and quality of these observations: at least ten values of $V_*(u_k)$ are necessary, evenly distributed within the texts or corpus.

Given this minimum requirement, many experiments prove that *p* is actually independent of the size and number of the excerpts. Figure 4 presents the results on *King John*: the theoretical curve – calculated with the help of this *partition model* - (dotted line) actually goes through the chart of the observed values (bold line).

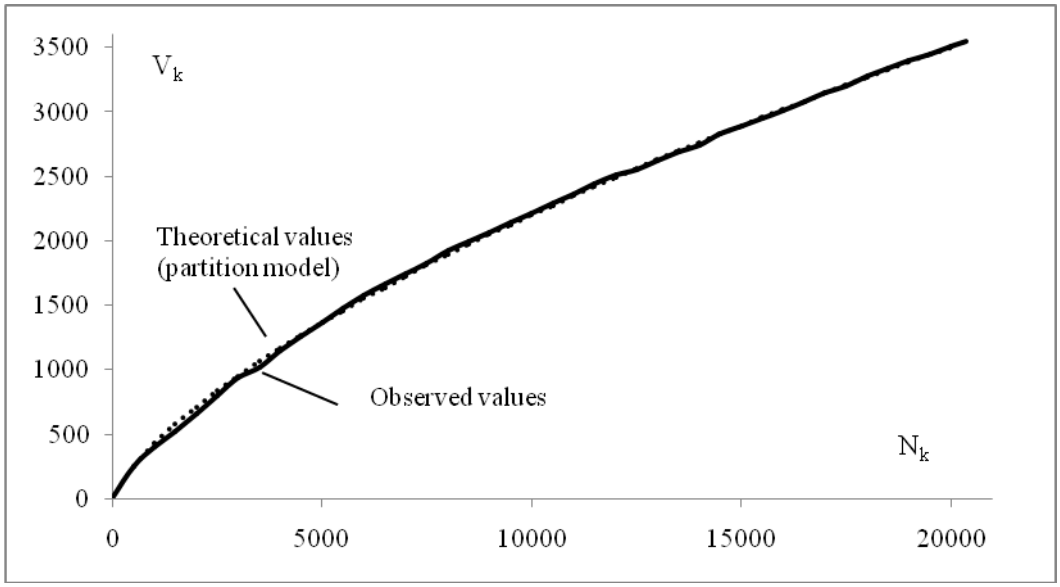

Fig. 4. Vocabulary growth in Shakespeare's *King John*. Observed values (bold line) and theoretical values calculated with the help of partition model (dotted line).

The observed curves for the other 88 plays of the corpus are also well fitting. This property allows one to take into account the specialization of the vocabulary in the computation of the number of types that a text would have had if it had been smaller. This will also allow the comparison of the vocabulary diversity of two texts of unequal lengths.

There are some limitations to this model. Especially, it can be assumed that, if the compared two texts are too diverse in length (one very small and one very large) the comparison would still be too "stretched" to lead to a proper comparison.

Let us consider "the", which is the most common word in all these corpora. In the whole EM Corpus, it occurs 25 239 times (F_{the} = 25 239). For this word, let consider two possibilities:

- *P(X=1)* ("all its occurrences are drawn out of the whole corpus") has no sense when considering a sample length of less than 25 239 tokens (the event X = F_{the} is impossible);

- *P(X=0)* ("none is drawn out of the whole corpus") makes sense only for a sample (N' < N - 25 239) otherwise the event (F'_{the} = 0 is impossible).

This is the reason why Daniel, Ford, Heywood, Lyly, Nashe and Porter are omitted from this experiment for the time being. Within these limits, the partition model can be used to determine the vocabulary diversity of each play of the corpus EM.

5. Diversity of the vocabulary in the corpus EM

Two solutions can be considered to compute an unbiased vocabulary diversity for each plays of the corpus. The first one would be to compute the size of the vocabulary if all the plays had been of the smallest length found in the corpus (B. Jonson: *A Tale of a Tub* 8 237 tokens). The second one is to fix a standard and interpretable length that would allow a "universal" comparison between texts. In this second solution, the vocabulary *diversity* of a text is defined as the average number of different word types found in all different excerpts of 10 000 tokens (V'_{10000}) that can be drawn from this text. This measure is computed with formula (2). This later solution seems to be more adequate in comparing vocabulary diversity of plays/authors/works.

Figure 5 shows these computed values for the corpus EM. This scatter plot should be compared to the one of figure 1. It clearly shows that the computed diversity (V'_{10000}) is not related to the lengths of the texts.

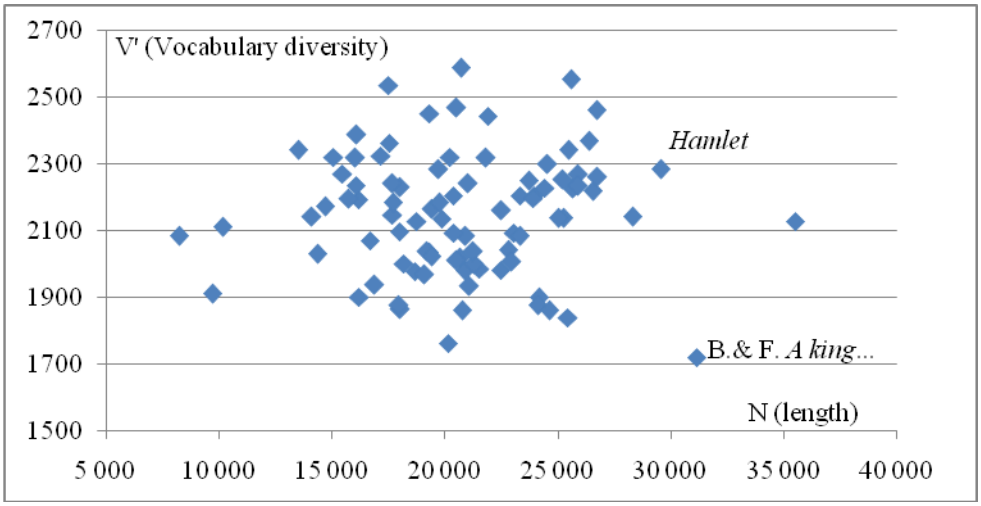

Figure 5. Relation between the vocabulary diversity (V'_{10000}) and the number of tokens (N) for each EM plays

The fact that the computed value is not determined by the length has always been found true for all the tested corpora (see for example Monière & Labbé 2008; Labbé 1998). Given this, it is now possible to compare fairly all the plays of the EM corpus. Which one has the most diverse vocabulary or the poorest one? Tables 4 and 5 give the top ten and the bottom ten. The last columns give the computed richness of vocabulary (V'_{10000}).

Authors	Plays	Length	Vocabulary	V'_{10000}
Dekker Thomas	*The Whore of Babylon*	20 711	3 989	2 587
Shakespeare William	*Henry V*	25 581	4 545	2 553
Middleton Thomas	*A Game at Chess*	17 503	3 684	2 536
Shakespeare William	*King Henry VI, Part 1*	20 518	3 782	2 469
Jonson Benjamin	*The Alchemist*	26 724	4 420	2 461
Shakespeare William	*Edward III*	19 331	3 705	2 452
Jonson Benjamin	*The New Inn*	21 890	4 116	2 443
Shakespeare William	*Macbeth*	16 085	3 256	2 388
Jonson Benjamin	*Volpone*	26 382	4 166	2 370
Marlowe Christopher	*Tamburlaine 1*	17 527	3 243	2 367

Table 4. The ten plays with the most diverse vocabulary

Authors	Plays	Length	Vocabulary	V'_{10000}
Beaumont & Fletcher	*A King and no King*	31 127	2 904	1 719
Beaumont & Fletcher	*The Second Maiden's Tragedy*	20 139	2 525	1 762
Fletcher John	*The Loyal Subject*	25 433	3 171	1 838
Fletcher John	*Valentinian*	24 623	2 997	1 861
Shakespeare William	*Much Ado About Nothing*	20 758	2 942	1 861
Beaumont & Fletcher	*Philaster*	18 012	2 685	1 863
Middleton Thomas	*A Trick to Catch the Old One*	17967	2 706	1 876
Fletcher John	*Demetrius and Enanthe*	24 112	3 007	1 876
Beaumont & Fletcher	*The Humourous Lieutenant*	24 162	3 138	1 899
Marlowe Christopher	*Massacre at Paris*	9 718	1 880	1 909

Table 5. The ten plays with the least diverse vocabulary

Some authors like W. Shakespeare, T. Middleton or C. Marlowe can be found in both tables. In addition, it is interesting to note that some plays of which vocabularies are the richest are of debated origin: *Henry V*, *Henry VI*, *Edward III* or *King John*, *MacBeth* or *Timon of Athens* would not be entirely by Shakespeare (Merriam 2000, 2002a & b, 2003, 2004; Craig & Kinney, 2009).

This means that the "author" may not be the most important factor in order to explain the diversity of a text vocabulary. Among the factors that influence this diversity, the "genre" of the play seems to be of some importance. For W. Shakespeare and his contemporaries, comedies would mobilize less vocabulary than more serious plays such as tragedy as shown in Table 6.

Genre	V'_{10000}	Indice
Historical plays	2 288	100
Tragedies	2 235	97
Comedies	2 083	90
Mean	2 191	95

Table 6. Diversity of the vocabulary for Shakespeare's plays according to their "genre"

But this may not be taken as a general rule: for example two Jonson's plays can be found within the top ten plays with the most diverse vocabulary (Table 4) and they actually are comedies. Within each genre, diversity of vocabulary seems to be the result of stylistic and thematic choices that cannot be addressed within the limited scope of this paper. Nevertheless the proposed tool is of a real utility in comparing corpora of different lengths. The Table 7 shows the diversity of the vocabulary for the EM works by author.

To have a better appreciation of the importance of the observed differences, it is useful to consider the standard deviation of the different observed sizes of vocabulary for excerpts of 10 000 tokens lengths (in the last column of Table 7). This gives an idea of differences that can be imputed to a "normal" or non-exceptional variation. A confidence interval can be associated with each value (ie with $\alpha = 0.05$; $V_{10000} \pm 1.96\ \sigma$). With less than 5 chances in 100 of being wrong, it can be considered that the vocabularies of B. Jonson and T. Dekker are significantly richer than the ones of all the others. The same conclusion can be drawn for T. Kyd (compared to the authors listed in the lines below). However, it is not the same for the pairs {Jonson - Dekker}, {Shakespeare - Marlowe}, {Marlowe - Peele}, {Peele - Chapman} and {Chapman - Middleton} whose diversities are separated by intervals which are too low.

Author	Number of plays	Number of tokens	Diversity V'_{10000}	V Standard deviation (σ)
Jonson B.	5	144 628	2 384	23,6
Dekker T.	2	43 778	2 339	25,4
Kyd T.	2	38 231	2 269	24,3
Shakespeare W.	38	830 379	2 191	23,9
Marlowe C.	7	111 858	2 148	18,9
Peele G.	2	24 877	2 139	11,6
Chapman G.	2	40 618	2 132	23,5
Middleton T.	14	263 426	2 097	21,8
Greene R.	3	51 102	2 057	19,4
Fletcher J.	7	177 968	1 913	22,3
Fletcher J. & Beaumont F.	5	116 244	1 850	22,3
Total and mean	89	1 843 109	2 139	

Table 7. Diversity of the vocabulary for each author of the corpus EM, ranked by decreasing order

An important remark is that these variations do not seem to strictly depend on the number of plays under consideration.

6. Conclusions

The vocabulary richness is now divided into two dimensions: **specialization** – proportion of word types which are devoted to a special part of the text - and **diversity** - the average number of different word types found in a large number of blocks, with a standard length, drawn randomly from this text. These two dimensions can be measured and, by adopting the measures and standards proposed in this communication, it becomes possible to compare a large number of texts in terms of their stylistic features or to identify significant stylistic changes in a work (Labbé, Labbé & Hubert 2004).

As regards the English "Early Modern" theater, the experiment presented in this paper is sufficient to reject with confidence the hypothesis that the vocabulary of the plays presented under the name of W. Shakespeare is unusually "rich." Instead, it is within the average of his contemporaries. Therefore, there is no rational basis for the idea once so prevalent that this author had an extraordinary vocabulary (if he is the author of all documents published under his name)… The champion seems to be B. Jonson, but we studied only five plays of his. It is possible that these plays are not representative of all his theatrical work... The same can be said about T. Dekker who appears to be also “richer” than Shakespeare.

The diversity of vocabulary, as its specialization, is not characteristic of the culture of an author but more probably the result of a conscious choice made for each play. Some authors chose rather restrainedly (J. Fletcher, R. Greene), others, like B. Jonson and T. Dekker, have preferred diversity. But the same author can be found at the two extremes: it is the case for W. Shakespeare, C. Marlowe and T. Middleton.

These calculations allow one to examine with a fresh eye many other issues. For example, the chronology of a work. In fact, the vocabulary of the plays published under the name of Shakespeare seems to become more restrained over time. This trend might help the discussion about the dating of some of these plays.

Finally, one can discuss the definition of "richness", considering, for example, that the vocabulary richness can also stem from higher use of idiomatic expressions and other multi-word expressions like collocations. In this case, it should be preferable to use the notion of "rarity". The feeling of "rarity" of the vocabulary of W. Shakespeare's plays could come from some unexpected words or from some "lexical creations" that are more or less extraordinary. A statistical measure of this "rarity" and of this "lexical creativity" would be possible only if we had the complete works - transcribed in modern English – of the main contemporaries of W. Shakespeare as B. Jonson, T. Middleton and J. Fletcher.

Acknowledgements

The authors are grateful to Ward Elliott who gave them the idea of this paper, the quotation opening it and all the plays of the Claremont Shakespeare Clinic; to Tom Merriam who also gave them a large number of plays and a constant help; to Pierre Hubert who has elaborated the partition model with them; to Tom Merriam and Jacques Savoy for their careful reading of a previous version of this paper.

References

Chambers E. K. (1923). *The Elizabethan Stage.* Oxford: Clarendon Press, 1923.

Cossette A (1994). *La Richesse lexicale et sa mesure.* Paris-Genève: Slatkine-Champion.

Craig H. (2011). Shakespeare's Vocabulary: Myth and Reality. *Shakespeare Quarterly*. 62-1, p. 53-74.

Craig, H. & Kinney A. F. (Eds) (2009). *Shakespeare, Computers, and the Mystery of Authorship*. Cambridge: Cambridge University Press.

Efron B. & Thisted R. (1976). Estimating the Number of Unseen Species: How Many Words did Shakespeare Know? *Biometrika*, 63(3), p. 435-447.

Elliott W. & Valenza R. (2004). *Shakespeare's Vocabulary : Did it Dwarf All Others_?* In Ravassat, M. and J. Culpeper (2011). *Stylistics and Shakespeare's Language - Transdisciplinary Approaches*. London, Continuum Press, p. 34-57.

Hockey S. & Martin J. (1988). *OCP Users' Manual*. Oxford: Oxford University Computing Service.

Hubert P. & Labbé D. (1988a). Note sur l'approximation de la loi hypergéométrique par la formule de Muller. In Labbé D., Serant D. & Thoiron P. *Etudes sur la richesse et la structure lexicales*. Paris-Genève: Slatkine-Champion, 77-91.

Hubert P. & Labbé D. (1988b). Un modèle de partition du vocabulaire. In Labbé D., Serant D. & Thoiron P. *Etudes sur la richesse et la structure lexicales*. Paris-Genève: Slatkine-Champion, 93-114.

Hubert P. & Labbé D. (1988c). A model of Vocabulary Partition. *Literary and Linguistic Computing*. Vol. 3, n° 4, p. 223-225.

Hubert P. & Labbé D. (1994). Vocabulary Richness. *Communication au congrès de l'ALLC-ACH*, Paris: La Sorbonne. Reproduced in *Lexicometrica*, 0, 1997 (http://www.cavi.univ-paris3.fr/lexicometrica/).

Labbé D. (1998). La richesse du vocabulaire politique : de Gaulle et Mitterrand. In Mellet S. & Vuillaume M. (eds). *Mots chiffrés et déchiffrés: mélanges offerts à Étienne Brunet*. Paris: Champion, 173-186.

Labbé C., Labbé D. & Hubert P. (2004). Automatic Segmentation of Texts and Corpora. *Journal of Quantitative Linguistics*. Vol. 11, 3, p. 193-213.

Merriam T. (2000). "Edward III". *Literary and Linguistic Computing*. 15-2. June 2000. p 157-186.

Merriam Thomas (2002a). "Intertextual Distances between Shakespeare Plays, with Special Reference to Henry V (verse)". *Journal of Quantitative Linguistics*. 9-3. December 2002. p 260-273.

Merriam T. (2002b). *Marlowe in Henry V: a crisis of Shakespearian Identity*. Oxford. Oxquarry Books.

Merriam T. (2004). "King John divided". *Literary and Linguistic Computing*. 19-2. march 2003. p. 181-195.

Merriam T. (2005). *The Identity of Shakespeare in Henry VIII*. The Renaissance Institute, Tokyo.

Monière D. & Labbé D. (2008). Des mots pour des voix : 132 discours pour devenir président de la République française. *Revue Française de Science Politique*. 58, 3 (2008), p. 433-455.

Muller C. (1977). *Principes et méthodes de statistique lexicale*. Paris: Hachette.

Thisted R. & Efron B. (1987). Did Shakespeare Write a Newly-Discovered Poem? *Biometrika*, 74(3), 445-455.

Ule L. (1985). "The Weird Ways of Vocabulary". *Literary and Linguistic Computing Journal*. Vol. 6, 1, p. 24-28.

Wimmer G. & Altmann G. (1999). Review Article: On Vocabulary Richness. *Journal of Quantitative Linguistics*. 6-1, 1-9.

Annex.

The “Early Modern” Corpus (alphabetical order) with length and vocabulary of plays

Authors	Title	N (tokens)	V (word types)	V'_{10000}
Chapman George	*Bussy d'Ambois*	19 731	3 544	2 285
Chapman George	*The Gentleman Usher*	20 887	3 104	1 979
Dekker Thomas	*The Honest Whore, Part II*	23 067	3 575	2 093
Dekker Thomas	*The Whore of Babylon*	20 711	3 989	2 587
Beaumont F. & Fletcher J.	*A King and no King*	31 127	2 904	1 719
Beaumont F. & Fletcher J.	*The Second Maiden's Tragedy*	20 139	2 525	1 762
Beaumont F. & Fletcher J.	*Philaster*	18 012	2 685	1 863
Beaumont F. & Fletcher J.	*The Scornful Lady*	22 800	3 235	2 041
Beaumont F. & Fletcher J.	*The Humourous Lieutenant*	24 162	3 138	1 899
Fletcher John	*Sir John Van Olden Barnavelt*	21 531	3 247	1 985
Fletcher John	*Chances*	16 195	2 509	1 900
Fletcher John	*Demetrius and Enanthe*	24 112	3 007	1 876
Fletcher John	*The Island Princess*	22 456	3 126	1 982
Fletcher John	*The Loyal Subject*	25 433	3 171	1 838
Flectcher John	*Monsieur Thomas*	20 682	3 063	2 019
Fletcher John	*Valentinian*	24 623	2 997	1 861
Fletcher John	*The Woman's Prize*	22 936	3 279	2 006
Greene Robert	*Alphonsus, King of Aragon*	15 067	2 321	2 321
Greene Robert	*Friar Bacon and Friar Bungay*	16 184	2 978	2 193
Greene Robert	*James IV*	19 851	3 273	2 135
Benjamin Jonson	*The Alchemist*	26 724	4 420	2 461
Benjamin Jonson	*Barthomew Fair*	35 501	4 455	2 127
Benjamin Jonson	*The New Inn*	21 890	4 116	2 443
Benjamin Jonson	*Sejanus*	25 894	3 990	2 269
Benjamin Jonson	*A Tale of a Tub*	8 237	1 866	2 082
Benjamin Jonson	*Volpone*	26 382	4 166	2 370
Kyd Thomas	*Soliman and Perseda*	18 007	3 095	2 229
Kyd Thomas	*The Spanish Tragedy*	20 224	3 460	2 320
Marlowe Christopher	*Doctor Faustus*	15 454	2 910	2 271
Marlowe Christopher	*Dido, Queen of Carthage*	13 507	2 760	2 341
Marlowe Christopher	*Edward II*	20 508	3 098	2 010
Marlowe Christopher	*The Jew of Malta*	17 982	2 975	2 098
Marlowe Christopher	*Massacre at Paris*	9 718	1 880	1 910
Marlowe Christopher	*1 Tamburlaine the great*	17 162	3 223	2 324
Marlowe Christopher	*II Tamburlaine.*	17 527	3 243	2 363
Middleton Thomas	*A Chaste Maid in Cheapside*	16 685	2 811	2 069
Middleton Thomas	*A Game at Chess*	17 503	3 684	2 536
Middleton Thomas	*Hengist/Mayor of Queenboro*	19 427	3 218	2 165
Middleton Thomas	*The Lady's Tragedy*	18657	2 739	1 978
Middleton Thomas	*A Mad World, My Masters*	17686	2 949	2 147
Middleton Thomas	*More Dissemblers*	18 743	3 029	2 127
Middleton Thomas	*Michaelmas Term*	19 299	2 869	2 034
Middleton Thomas	*No Wit/Help Like a Woman's*	25 242	3 551	2 137
Middleton Thomas	*The Phoenix*	19 198	2 971	2 036
Middleton Thomas	*The Puritan or the Widow of Watling Street*	18171	2 827	2 001
Middleton Thomas	*A Trick to Catch the Old One*	17967	2 706	1 876
Middleton Thomas	*The Nice Valour or The Passionate Madman*	14095	2 687	2 141
Middleton Thomas	*Women Beware Women*	25 005	3 469	2 137
Middleton Thomas	*The Witch*	15 748	2 822	2 196
Peele George	*The Arraignment of Paris*	10 177	2 129	2 110
Peele George	*David and Bethsabe*	14 700	2 716	2 171
Shakespeare William	*King Henry IV, Part 1*	23 937	3 788	2 205
Shakespeare William	*King Henry VI, Part 1*	20 518	3 782	2 469
Shakespeare William	*King Henry IV, Part 2*	25 680	4 084	2 226

Shakespeare William	*King Henry VI, Part 2*	24 416	4 001	2 228
Shakespeare William	*King Henry VI, Part 3*	23 304	3 559	2 084
Shakespeare William	*Much Ado About Nothing*	20 758	2 942	1 861
Shakespeare William	*Antony & Cleopatra*	23 703	3 912	2 250
Shakespeare William	*All's Well That Ends Well*	22 481	3 469	2 160
Shakespeare William	*As You Like It*	21 292	3 228	1 999
Shakespeare William	*Coriolanus*	26 553	3 992	2 218
Shakespeare William	*Cymbeline*	26 750	4 244	2 260
Shakespeare William	*Edward III*	19 331	3 705	2 452
Shakespeare William	*Comedy of Errors*	14 358	2 504	2 030
Shakespeare William	*Henry V*	25 581	4 545	2 553
Shakespeare William	*Henry VIII*	23 325	3 529	2 204
Shakespeare William	*Hamlet*	29 549	4 663	2 283
Shakespeare William	*Julius Caesar*	19 107	2 840	1 968
Shakespeare William	*King John*	20 375	3 546	2 205
Shakespeare William	*Love's Labours Lost*	21 022	3 734	2 240
Shakespeare William	*King Lear*	25 215	4 132	2 253
Shakespeare William	*Macbeth*	16 085	3 256	2 388
Shakespeare William	*Measure for Measure*	21 260	3 307	2 037
Shakespeare William	*Midsummer Night's Dream*	16 062	2 970	2 236
Shakespeare William	*The Merchant of Venice*	20 910	3 244	2 083
Shakespeare William	*Othello*	25 891	3 774	2 234
Shakespeare William	*Pericles*	17 679	3 218	2 242
Shakespeare William	*Richard II*	21 797	3 650	2 318
Shakespeare William	*Richard III*	28 308	4 054	2 141
Shakespeare William	*Romeo and Juliet*	23 907	3 678	2 197
Shakespeare William	*The Taming of the Shrew*	20 386	3 208	2 092
Shakespeare William	*The Two Gentlemen of Verona*	16 875	2 703	1 938
Shakespeare William	*Timon of Athens*	17 713	3 269	2 183
Shakespeare William	*Titus Andronicus*	19 752	3 319	2 184
Shakespeare William	*The Tempest*	16 030	3 139	2 319
Shakespeare William	*Twelfth Night*	19 403	3 074	2 021
Shakespeare William	*Trolius and Cressida*	25 475	4 224	2 342
Shakespeare William	*Merry Wives of Windsor*	21 072	3 226	1 933
Shakespeare William	*The Winter's Tale*	24 518	3 904	2 299