Shakespeare's vocabulary: did it dwarf all others?

WARD E. Y. ELLIOTT AND ROBERT J. VALENZA

1. Size

Did Shakespeare's vocabulary dwarf all others?[1] Many respected scholars have said that it did and does. F. Max Müller, the great Victorian philologist and Sanskrit scholar, said it thus:

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.

By contrast, he added, English country labourers of the day supposedly 'had not 300 words in their vocabulary;' 'a well-educated person in England, who has been at a public school and at the 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 (Müller 1862: 266–7; 1891: 377–9).' Müller's magnum opus, *The Science of Language* (from which these numbers are taken), is still in print. He was widely followed in subsequent centuries; and he is still widely cited.

Notable twentieth-century supporters of the Müller thesis were Ernest Weekley, who published five editions of *The Romance of Words* and ten other widely read books on words and names; Louis Marder, longtime editor of the *Shakespeare Newsletter;* and Robert McCrum, William Cran, and Robert MacNeil, authors of the companion text to the 1986 Public Broadcasting System series, *The Story of English,* now in its third, revised edition (2002). In 1928 Weekley wrote: 'Of Shakespeare it may be said without fear of exaggeration that his contribution to our phraseology is ten times greater than that of any writer to any language in the history of the world' (Weekley 1952 [1928]: 55).

Alfred Hart, the leading mid-20th-century authority on Shakespeare's use of 'new words;' though less sweeping than Müller and the others, thought that Shakespeare's vocabulary and coinages far outstripped those of his early contemporaries, Marlowe, Greene, Peele, Kyd, and Chapman. In 1942 he wrote:

[S]hakespeare was a lifelong and insatiable word-collector . . . . Shakespeare's unmistakable sign-manual in a play is the presence of plenty of words peculiar to it alone . . . . [A]n author who is credited by the compilers of the *Oxford English Dictionary* with being the first user of about 3,200 words . . . has verbal riches compelling the employment of superlatives in describing them . . . . Shakespeare's vocabulary was fuller, more varied and expressive of finer shades of meaning than that of any of his early contemporaries . . . . [Marlowe's (and Chapman's)] inferiority in vocabulary is manifest (Hart 1942: 22, 28, 450–1).

In 1986 McCrum et al. said this to a Public Broadcasting System audience of millions: 'Shakespeare had one of the largest vocabularies of any English writer, some 30,000 words. (Estimates of an educated person's vocabulary today vary, but it is probably about half this, 15,000)' (2002 [1986], 102). The Müller thesis marches on in this century in McCrum's third edition, and others heartily concur: 'The average educated man today . . . has a working vocabulary of less than half that of Shakespeare.'(Bragg: 135);

'Take a look at Shakespeare's enormous vocabulary . . . 27,780 are different words. The average person . . . has a recognition vocabulary of about 5,000 words. Some of the greatest writers may have twice this capability.'(Ravi).

It should not be forgotten, when one is thinking of superlatives for Shakespeare's vocabulary, that English itself is said to have the richest vocabulary in the world today, with at least a million words, depending on what gets counted (Crystal 1995: 119),[2] and growing. According to McCrum et al., English is five to ten times richer in words than German or French (2002: 10). Shakespeare lived and wrote during the language's most glorious and formative years, years when its vocabulary was growing explosively at rates unmatched before or since (Nevalainen 1999: 338–9), years when, driven by patriotism, Protestantism, and the mass markets both permitted and demanded by the printing press, it grew from a collection of base, rude, rustic, spoken local dialects to replace Latin as the default language of national literary discourse (Hussey 1992 [1982]: chap. 2; McDonald 2001: chap. 1). It wasn't just that Shakespeare had the most superlative English vocabulary of his day – if, in fact, he

did. His day was the most superlative period of vocabulary growth for English, and English vocabulary today leaves all others in the dust. How much more superlative could you get?

Everyone in Shakespeare's day was adding furiously to the language, and judging from listings in the *Oxford English Dictionary* and the *Chronological English Dictionary,* Shakespeare's new-word coinages seem to have led all the rest, even after many corrections for extravagant overcounts in the early twentieth century. We argue in Part 2 that Shakespeare's coinages are still probably overcounted by a factor of at least two. For now, it is enough to note that coinages were an important component of the persistent belief that Shakespeare's vocabulary dwarfed all others in size, variety, and creativity.

Shakespeare's supposedly outsized vocabulary and coinage rates are frequently invoked in authorship controversies. Hart's studies of Shakespeare's new-word usage figured prominently in MacDonald Jackson's and Kenneth Muir's landmark ascriptions of *A Lover's Complaint* to Shakespeare, helping explain why *A Lover's Complaint*'s abundance of words new to Shakespeare strengthened, rather than contradicted the Shakespeare ascription (Jackson 1965: 8–19; Muir 1973 [1964]). The powerfully documented Jackson and Muir studies overturned the consensus of the day that *A Lover's Complaint* was not Shakespeare's; thanks to their work, the consensus today (which we and others now think is probably wrong, see Elliott and Valenza 1997 and 2004; Vickers 2007) is that Shakespeare did write *A Lover's Complaint* . Moreover, if mainstream scholars savoured the idea of Shakespeare's verbal pre-eminence with their morning coffee, anti-Stratfordian insurgents devoured it for breakfast, lunch and dinner, providing, as it seemingly did, compelling evidence that Shakespeare was far too learned to have been the lowly Stratford grain-dealer.

'Never,' said Oxfordian doyen Charlton Ogburn Jr. after a long, exclamatory discourse on Shakespeare's learning, citing Müller, Marder, and others, 'has such verbal prodigality as Shakespeare's been approached' (Ogburn 1984: 291–2).

To be sure, there was at least one notable oldtime Müller sceptic, not so much of the richness of Shakespeare's vocabulary as of the supposed poverty of others' vocabulary. The sceptic was Otto Jespersen (1860–1943). His classic *Growth and Structure of the English Language* was first published in 1905 and last revised by him in 1938. In three tightly-packed pages on the subject, Jespersen conceded to Shakespeare his customarily estimated vocabulary of 20,000+ words, but he thought that Müller's notion that a farm-labourer uses only 300 words was 'obviously wrong.' So, it seems, was his notion that educated persons used no more that 4,000–10,000 words. Two-year-olds with doting, documenting parents were known to use 489–1,121 words; Swedish peasants seemed to have more than 26,000 words; and enterprising college professors E.S. Holden and E.H. Babbitt, using randomly selected pages from a dictionary, estimated that most of their students recognized a little below 60,000 words (Jespersen: 200–1). Jespersen added:

> These statements are easily reconciled with the ascription of 20,000 words to Shakespeare. For it must be remembered that in the case of each of us there is a great difference between the words *known* . . . and the words actually *used* in conversation . . . . If Milton as a poet uses only 8,000 as against Shakespeare's 20,000 words, this is a natural consequence of the narrower range of his subjects, and it is easy to prove that his vocabulary really contained many more than the 8,000 words found in a Concordance of his poetical works. We have only to take any page from his prose writings, and we shall meet with a great many words not

in the Concordance. (202, listing 21 such words on a single page of Milton's *Areopagitica*).

The greatness of Shakespeare's mind is therefore not shown by the fact that he was acquainted with 20,000 words, but by the fact that he wrote about so great a variety of subjects . . . that he needed this number of words in his writings (ibid.).

Jespersen's cautionary book went through ten editions and must have been read and admired by thousands of people – and you can still find his cautions reflected and updated in the writings of at least two modern English-language encyclopaedists, Tom McArthur (1992: 1092), and David Crystal (1995: 123; 2004: 317). But Jespersen (unlike Müller) was out of print for decades; not enough people have paid attention to McArthur and Crystal; and it would seem from the continued prevalence of Müller's view that the old-line bardolatry still holds sway.

But is it right? Our short answer, based on our analysis of the Claremont Archive of Renaissance Texts and on type-token figures for other writers graciously supplied to us by Professor David Hoover, of New York University, is that Müller and company were right that Shakespeare had a big vocabulary, but wrong in supposing that it was bigger or better than other writers' vocabularies, either in his own day or since. Much of Shakespeare's apparent pre-eminence over others is due to the greater accessibility of his writing. He wrote more than most others and was better recorded, catalogued, and anthologised. The people who wrote the *Oxford English Dictionary* could get to him in ways that they could not get to other writers. Müller, for example, could choose between three Shakespeare concordances, notably Mary Cowden Clarke's *Complete Concordance to Shakespeare* (1845), which had been 16 years in the making. His estimates for works with lexicons – Shakespeare, the Bible, and parts

of Milton – were in the right ballpark; but, where he lacked a lexicon, as with farm labourers, he was way off; and, where, as with Milton, he had a lexicon based on a much shorter, less comprehensive corpus than Shakespeare's, he had no way to correct for it.

What is true of Müller was largely true of his uncomputerized successors, even the best of them, Hart, who was well aware of the limitations of handmade lexicons but could only partially compensate for them. It takes a lexicon, an inventory of word types, to count someone's total manifest vocabulary; handmade lexicons took years to produce and were unavailable for most other writers; and, even when available, they offered no easy way to correct for differing corpus size, which is crucial for vocabulary comparisons since, other things equal, larger corpora yield larger vocabularies of distinct words.

The most authoritative modern Shakespeare lexicons are Marvin Spevack's (1968, 1973), a long one and a short one, both meticulously compiled on a computer from the meticulously edited *Riverside Shakespeare* (Evans et al., 1997 [1974]) – though it would be surprising if even the *Riverside* preserved every countable distinction between one archaic spelling and another – burthen/burden; murther/murder; fadom/fathom; vild/vile; Dolphin/Dauphin; crocadile/crocodile; wreck/wrack, and so on. Every Shakespeare edition, in a sense, is a translation from Early Modern to contemporary English, and variations among different editors' commonizing zeal produce defendable variations among the countable different words to be found in each one's Canon (Taylor 1989: 254, 316). Spevack machine-counted 29,066 different words – 'types' – from a *Riverside* Shakespeare Canon of 884,647 total words – 'tokens' (Spevack 1973: v). His is considered the best available count of Shakespeare's manifest vocabulary, taken from words Shakespeare actually recorded,

but even his count has limitations typical of machine counts, our own included. It is utterly reliant on the editor's orthographical judgement; it includes some words supplied by editors, and it includes words from parts of the Canon – maybe as much as a tenth of it – which were not written by Shakespeare solo but by some combination of Shakespeare and a co-author (see Vickers 2002b; Wells 1987; Elliott and Valenza 1996).

Most important, it counts every different word inflection found, not just each word's root (also called its *lemma*, or *headword*). 'Horse' and 'horses' are two words by this convention, not two variants of one dictionary headword.[3] Both of these conventions take a large view of Shakespeare's vocabulary, especially using an inflected lexicon, which is much larger – maybe two-thirds larger – than the lemmatized vocabularies used by hand-counters like Hart.[4] If the two-thirds correction factor is good, Shakespeare's lemmatized, manifest vocabulary is closer to 17–18,000 than to 29,000 words (Hart 1943b = 17,677; Scheler 1982: 89 = 17,750; WordHoard = 18,090[5]) – close enough, given the many arguable ways of counting words, to make Müller's estimate of 15,000 look remarkably accurate.

On the other hand, Spevack's count, like ours, understates Shakespeare's vocabulary in three ways, two small, one large. The first small one is multiple meanings of homographs, different words with the same spelling. Our *Webster's New Collegiate Dictionary* (which was new in 1950) lists eight meanings for *spring* as an intransitive verb, eight more for *spring* as a transitive verb, nine for *spring* as a noun, and three for *spring* as an adjective, 28 different possible meanings in all, and up to 28 possible different words to the context-sensitive hand counter – but only one word to most computers, including Spevack's and ours. Correcting for homographs, if you could do it, would modestly increase the total, probably by at least the 700+ words

listed in Spevack (1968, vol. IV, App. D). The second small understatement is linked word phrases like 'come on' or 'grown up,' lexical units which are functional equivalents of a single word and mean something distinguishable from 'come' and 'grown,' yet normally don't get caught and counted by computers.

The largest understatement of any kind of vocabulary count, machine or hand, is the one touched on by Jespersen, that they don't measure how many words an author *knows* – his latent vocabulary – but only those he has *used and recorded* – his manifest vocabulary. Some writers use the term 'recognition vocabulary' or 'passive vocabulary' as near-synonyms for 'latent,' and 'use vocabulary' or 'active vocabulary' as near-synonyms for 'manifest.' Latent /passive/ recognition vocabularies are always larger than manifest/active/use vocabularies. Works which cover many different subjects, as Jespersen noted, should show larger vocabularies than works with fewer subjects. And large corpora (other things equal) should normally show larger vocabularies than small ones because each new token (token = countable word, including every repetition) added to a given corpus offers a chance of adding a new type (type = distinct word type, doesn't count repetitions) as well, by transferring the type from the author's latent vocabulary to his manifest vocabulary. Hart well understood this last point and, with the help of concordances, calculated Shakespeare's rates of introducing once-used, or 'new' words for Shakespeare's poems and a few plays. Such new-to-Shakespeare words amounted to ten percent of the total word types in *King Lear* and appeared about once every ten lines of text in *Lear* and *Hamlet*, once every twenty lines in earlier plays (Hart 1943a: 128–31, summarized in Jackson 1965: 11–12).

Hart did his best to control for date and text length, but he was badly hampered by the limited concordancing available to him, which was skimpy for everybody but

Shakespeare, and which could not easily control for corpus or sample size. Like Müller and everyone else between the nineteenth century and the computer era, Hart had the nineteenth-century deconstructing genius for breaking down Shakespeare's and others' writing into their constituent components – words, grammar, metrics, and so on (see Taylor 1989: 193) – but he couldn't apply it as freely and systematically as you can with computers today. All these stalwart old stylometricians were roadbound and oxcart-borne, amazingly resourceful and proficient in using the slow, judgemental tools they had, and hugely better than their predecessors who lacked even roads (lexicons) and oxcarts, but gravely limited, nonetheless, by the technology of their time.

Computers and sizeable machine-readable text archives changed this situation profoundly and, in 1976, made possible an astonishingly sophisticated, landmark estimate of Shakespeare's total recognition vocabulary. However, the landmark appeared in *Biometrika*, a distinguished statistical journal unfamiliar to most number-shy Shakespeare and language scholars. It was a landmark all but invisible to regular Shakespeare scholars but of surpassing visibility and importance to us. Even without their own text archive, and using only the then-new Spevack concordance, plus a clever methodology first propounded by Sir Ronald Fisher in the 1940s, and massive crunching with an IBM 360, master statisticians Brad Efron and Ronald Thisted, were able to make a wholesale extrapolation of Shakespeare's new-word productivity. They concluded that Shakespeare, had he doubled his historic output of 884,647 tokens, would have added 11,460 new types, plus or minus 150, to his known 31,564. Even more astonishingly, they showed that Shakespeare's latent vocabulary had to be at least 35,000, besides the 'known' 31,564, for a total vocabulary of at least 66,000 inflected words (Efron and Thisted 1976: 435). If David Crystal is right in supposing

that the English language contained about 150,000 lemmas in Shakespeare's day (Crystal 2004: 317), and if our rough estimate of Shakespeare's inflected-to-lemma ratio at 1.67 to one is close to correct, it means that Shakespeare knew about 40,000 lemmas, more than a quarter of words then in the English language.[6]

Like Fisher's original study of how to forecast the discovery of new butterfly species from known past discovery rates (1943), the Efron-Thisted study is still seen as a triumph of statistical ingenuity, a landmark demonstration of how accurately you can guess the unknown from the known, with the right techniques. No one has tried more elaborately than we have to retest Efron's and Thisted's main techniques, assumptions, and conclusions from the ground up. We found some limits for microlevel applications (below) but concluded that their findings for Shakespeare in the large are still sound and well-deserving of their landmark status among statisticians (Valenza: 1990). This chapter is, in part, an effort to make it a landmark for language and literature scholars as well.

A decade after their unheralded breakthrough, responding to Gary Taylor's ascription of *Shall I Die?* to Shakespeare, Thisted and Efron attempted to apply the same methods they had applied wholesale in 1976 at retail to 500-word Shakespeare and non-Shakespeare samples comparable in length to *Shall I Die?* They concluded that *Shall I Die?* had enough Shakespeare-new words to be Shakespeare's, but other sample poems by Marlowe, Jonson, and Donne did not (Thisted and Efron: 1987).[7]

Though it was the second of the two Thisted-Efron articles (1987) that started the Shakespeare Clinic's work, it may well be the first (1976) which finishes it. Working from the second article, we found that 'new words' were one of the Clinic's best tests – but not in the way that the Müller thesis would have led us to expect. Shakespeare did not invent and use new words so furiously as to dwarf his peers; quite the

contrary. After we calculated Shakespeare's normal, expected new-word ranges, using our Intellex software to isolate 1,500-word blocks from the rest of the Shakespeare corpus and to make some necessary corrections for type-token ratios, we found (unsurprisingly, since we drew our Shakespeare profiles to keep false negatives at or below 5%) that none of our 3,000-word blocks from Shakespeare's poems, and only 5% of those from his plays, exceeded Shakespeare's expected new-word range. The surprise was that 64% of such blocks by other poets, and 56% of such blocks by other playwrights *exceeded* Shakespeare's normal expected range of words new to him. In other words, most of other people's poems and plays have *more* Shakespeare-new words than 95% of Shakespeare's. Also surprisingly, not a single other-authored block fell *below* Shakespeare's expected new-word range.

How could this be, when everyone knew from Müller and company that Shakespeare's vocabulary and new-word coinage rates dwarfed everyone else's? Our first thought was that the other writers must have different inventories of favorite words, so different as to overbalance Shakespeare's supposedly overwhelming advantage in vocabulary richness. Our second thought was to wonder whether Shakespeare's supposed advantage was as overwhelming as everyone believed. Both of these thoughts are probably true. The second one is what prompted us to write this chapter. We didn't and don't doubt that he knew 29,000 words; they were all there in the 884,000-word Canon. Or that he coined hundreds, if not thousands of words. But did it actually set him above his peers and posterity? Only if the traditional low estimates for other writers were accurate. We believe they were not.

To get a better estimate of how Shakespeare's vocabulary compared with pertinent others, we first controlled for size, taking large blocks of about 40,000 words from Shakespeare, eight of his contemporaries and Milton. We gave them three tests, one

traditional, two novel and doable only with our own program, Intellex, to test raw variety, distinctiveness, and total inferred vocabulary. Table 1 and Appendix One, from which it is drawn, show the results: no matter which test you use, Shakespeare comes out toward the middle of the pack. If anyone towered over the others, it was Milton and perhaps Spenser, not Shakespeare.

Table 1. Three Vocabulary Tests: Shakespeare and Eight Others

Author: number of blocks	Types	New Words	Inferred Vocab.
Milton: 2	6,500	1,691	131,953
Shakespeare: 5	5,470	905	62,569
6 Others: 9	5,223	839	66,132
Fletcher: 2	4,444	430	49,983

Three Vocabulary tests, Shakespeare (shaded), seven contemporaries, and Milton, 18 40,000-word blocks averaged for each author or author-group in each category. Milton leads the group; Fletcher trails; Shakespeare and six others – Jonson, Dekker, Marlowe, Middleton, Chapman, and Greene – fall in between. All these tests measure inflected, not root vocabularies. 'Types' and 'New Words' are per block, averaged.

Test One, *Types to Tokens*, is a well-known but imperfect one. It measures *raw variety*. Milton used an average of 6,500 different words (or 'types') in each of the two 40,000-word (= 40,000-token) halves of *Paradise Lost*. Shakespeare used an average of 5,470 words per block in four 40,000-word blocks of plays and one of poetry. Types-to-tokens was too variable, overlapping, and sensitive to sample size for us to have used it as a reliable author identifier for the Shakespeare Clinic, but, with sample size very large and well controlled, it says clearly enough that, in terms of raw variety of words used, Milton was at the top, Fletcher at the bottom, and Shakespeare and the others in between.

Test Two is *New-to-the-Group Words*, employed for the first time in this chapter. If types-tokens measures *variety*, New-to-the-Group Words measure *distinctiveness*. Using Intellex, we compiled a baseline lexicon from a corpus of 600,000 words, 160,000 by Shakespeare, the rest by seven other authors. The 600,000 tokens yielded 28,747 different types, 12,265 of them 'new words' found in only one of the fifteen 40,000-word blocks. Shakespeare, with 27% of the group's tokens, contributed 31% of the group's new words, a bit more than his share, but far less per block than Milton, who was, again, at the top, with Fletcher at the bottom, and Shakespeare and others in between.[8]

Test Three, *Total Virtual Inferred Vocabulary* (TVIV), is our newest and most ambitious. From earlier work, (Valenza: 1990), we had two pieces of solid-looking ground, to which we now add a new way of inferring from them a writer's total vocabulary, from a much smaller baseline and with much less elaborate computation than it took Thisted and Efron to do similar calculations for all of Shakespeare in 1976. The first piece of solid ground was Thisted and Efron's own 'bootstrap' calculations, which, using two completely different methodologies, came up with the same low-range estimate for Shakespeare's latent vocabulary of 35,000 words. This, with his manifest vocabulary of 31,000 words, added up to a well-supportable estimated total vocabulary of at least 66,000 words – or perhaps something closer to 64,000, if added to Shakespeare's corrected manifest vocabulary of 29,000 words per Spevack's 1974 count. Either would do for our purposes, but, for consistency with Thisted-Efron's earlier article, we use the original higher one, of 66,000 words. The second piece of solid ground was Valenza's own massive recalculations of Thisted-Efron's data in 1990 for all of Shakespeare and all of the King James Bible. Both of these large corpora, though differing greatly in provenance, genre, subject matter, and

vocabulary richness, turned out to have normalized word frequency curves which tracked each other almost exactly across the entire frequency range.

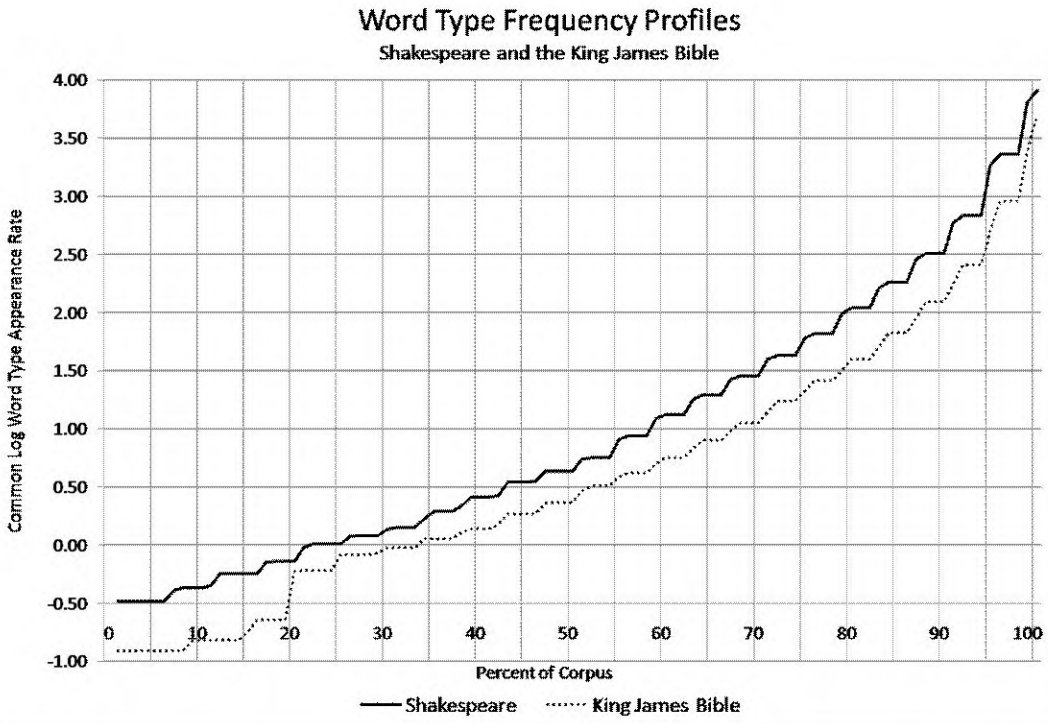

FIGURE 1. A comparison of normalized word type frequency profiles for Shakespeare and the King James Bible. *Shakespeare introduces new types at roughly twice the rate at which they occur in the Bible across the entire profile.* Source: Valenza, 1990, 32.

If two corpora as different from each other as Shakespeare and the Bible track each other that closely, you should be able to infer a plausible comparable frequency curve for other poets and playwrights. From that, you should be able to infer their total vocabularies simply from the number of word types they have to go through to get through half the tokens in their corpus or sample. Another way of putting this is that

(1) if an author has a big vocabulary of different word types, an analyst should go through more types before exhausting all the author's tokens; and (2), if different authors' word-frequency curves do track one another, the analyst should also go through more types before exhausting half of each author's tokens, and this adjustment makes the counting task much easier. If you accept the Thisted-Efron figure of 66,000 for Shakespeare's total inferred vocabulary as reasonably solid, as most do, and suppose that most word frequency curves track one another, as is certainly so of Shakespeare and the King James Bible and could well be so for others, you can then make a plausible guess about a given author's total inferred vocabulary simply from counting the types it takes to get through half the tokens and applying a correction factor to place the block's frequency curve relative to Shakespeare's total vocabulary of 66,000 words. See Table 1, last column; Shakespeare is shaded; in Appendix One the 18 blocks by 9 authors are ranked by TVIV.

The correction factors are such that taking 77 types to go through 20,000 tokens would put a given block's TVIV just about at the level of Shakespeare's, 66,000. Higher or lower type counts at the halfway mark would put a block's TVIV proportionally higher or lower, as can be seen by comparing Milton1, with 127 types at the 20,000-token halfway mark, implying a TVIV of 140,000 types, and Fletcher1, with only 63 types at the halfway mark, implying a TVIV of 49,000 types. The extra 'V,' for 'virtual,' is our way of underscoring the speculative nature of our assumptions that word frequency curves track each other.

Do these tests make any sense? Our first move was to scan Table 1 and Appendix One for consistencies and discrepancies. The consistencies are many: both Milton samples cluster at the top by all three tests. Both Fletcher samples cluster at the bottom by all three tests. Everyone else falls in between by all three tests. Shakespeare

himself, with five blocks measured, shows the greatest range of inter-block discrepancy, with three blocks of tragedies and romances very tightly clustered around a TVIV midpoint of 60,000 or so, but with two outliers, poems at 84,000 and early comedies at 49,000. Neither of the two Shakespeare outliers is shockingly distant from Shakespeare's mean, the Thisted-Efron-estimated TVIV of 66,000: the poems are +36%, the comedies -26%. Neither the long outlier nor the short one clashes with common-sense expectations, that poems would be richer than plays, and tragedies richer than comedies – an expectation which is also confirmed by the other two tests.

The mean of all five Shakespeare blocks tested is 62,569 words, just 5% lower than the 66,000 calculated for the entire Canon by Efron and Thisted (1976), and only 2% lower if Shakespeare's actual vocabulary is 64,000 per Spevack's later, lower estimate. Most two-block samples by the remaining authors, three out of the four remaining pairs, are separated by no more than ± 5,000 words (± 7%) from the two-sample midpoints.

In short, though the three tests measure richness in very different ways – one (types-tokens) with no reference at all to frequency curves, one (new words) based solely on the bottom 1%, of the frequency curve (least frequent), and one (TVIV) based solely on the top, most frequent, half – they are remarkably consistent with one another and with prior work by Thisted and Efron – and in remarkable agreement that Müller and company were wrong about other writers' vocabularies relative to Shakespeare's. Once you remove the gross biases of corpus size from the calculations, it becomes clear that, if anyone's vocabulary dwarfed others in size, it was Milton's, and maybe Spenser's, not Shakespeare's. If so, we would conclude that Shakespeare's

pre-eminence was much less a matter of how many words he knew, than of how he used them.

The spreadsheet also has a scattering of further comparative statistics on other writers, including a few willing members of the Claremont McKenna College faculty, to the same general effect: the first five colleagues we could get to give us recognition percentages of three sample pages from the dictionary all turned out, after extrapolation, to have vocabularies five or ten times larger than the 5–10,000 words confidently and wrongly assigned to educated people by Müller and his modern followers. It also turns out, not surprisingly, that we were not the first to think of this simple, obvious, dictionary-sampling methodology, which was equally available to Müller and the others, but apparently never tried by them. Jespersen and Crystal both mention such tests, and with similar outcomes, that is, crediting run-of-the-mill college-educated moderns with Shakespeare-sized or larger inflected vocabularies of 50–80,000 words (Jespersen 1982 [1938]: 201; Crystal 1995: 123).

TVIV's project even larger inflected vocabularies for other people than Shakespeare, including two moderns well known to us who are not otherwise to be compared with Shakespeare: 132,000 words for Spenser, 96,000 for Melville, 81,000 for the *Iliad*, almost 100,000 for Elliott and Valenza's manuscript on Shakespeare, and 177,000 for Elliott's alone on population policy. Jespersen and Crystal were in no position to test people's TVIVs the way we have, but our results are entirely consonant with two of Jespersen's views, that modern vocabularies equal or exceed Shakespeare's, and that vocabulary richness should increase with the variety of subjects addressed (Jespersen, 1982 [1938]: 202); hence the much greater TVIV for Elliott's wide-ranging population manuscript than for our more tightly-focused Shakespeare book.

People who write about Shakespeare often use much fancier language than Shakespeare did himself. One need only open a late twentieth-century Shakespeare journal to find words far more abstruse than Shakespeare's -- *fetishisation*, *commodification*, *poststructuralism*, *inferred virtual vocabulary*, and *enclitic microphrases*, for example. Some of these inkhorn polysyllables are our own, and there is a place for them, but they are all far, far from Shakespeare. Shakespeare managed to do some of his best work with short, concrete words like 'seething brains,' or 'cool reason,' artfully combined:

> . . . I never may believe
> These antic fables, nor these fairy toys.
> Lovers and madmen have such seething brains,
> Such shaping fantasies, that apprehend
> More than cool reason ever comprehends.
> (*A Midsummer Night's Dream* 5.1.2–6)

The result would be the same with any set of Shakespeare's most famous lines – 'Friends, Romans, countrymen,' 'All the world's a stage,' 'Tomorrow and tomorrow and tomorrow,' and many others.

In terms of sheer numbers of words known, and by several different measures, many of us can match or exceed Shakespeare – unsurprisingly, perhaps, given the vast expansion of the language and the whole human enterprise since Shakespeare's time. So could a number of his own contemporaries who drank from the same Castalian fountain of words that Shakespeare did. Quantitatively, Müller and company were completely wrong. Qualitatively, in terms not of how many words Shakespeare knew, but of how well he used them, we would suppose that there is still plenty of room for

bardolatry. Shakespeare learned early how to strike deep, not with an outsize inventory of long, inkhorn words, but with a par-for-the-course inventory, mostly of plain words, surpassingly well chosen and put together. It's not too soon, or too late, for the rest of us to learn it too.

2. Shakespeare's word coinages

What about the other Shakespeare commonplace, that he invented more new words than anyone else? That, too, could well be a myth, but we do not yet have the evidence that would settle the question. Some of today's authorities still rely on old, extravagant overcounts of Shakespeare's new-word coinages. Seth Lerer tells us, for example, that Shakespeare 'coined nearly six thousand new words' 'at a rate unmatched by any previous or subsequent author' (2007: 129, 135). Most other modern estimates are much lower, clustering around 1,700,[9] still high enough for Shakespeare to surpass all others. But, as Giles Goodland's 'Strange Deliveries,' in Ravassat and Culpeper, 2011, Chapter One, shows, current estimates could still be twice too high, and there is little evidence that the deflationary process begun in 'Strange Deliveries' has run its full course.

Bryan Garner, using the *Chronological English Dictionary*, counted 10,302 loan-word neologisms added to the English language between 1580 and 1619, a 39-year span (1982: 151). Shakespeare's writing life, which ran from about 1590 to 1613, encompassed almost 60% of this period, good for, say, 6,000 of the whole period's new, mostly Latinate loan-words. Garner thought that Shakespeare contributed just over 600 of them, about a tenth. Latinate loan-words are only a fraction of all newly coined words, but Garner's one-tenth estimate for loan-words both matches and cites estimates by Joseph T. Shipley (1977: 28), and is based on tables in Jürgen Schäfer,

(1973: 206–20) that Shakespeare was the first user of 'well over 1,700' words, 'one new word in every ten.' 'Shakespeare,' Shipley concluded, 'was the greatest word-maker of them all' (cited, Garner 1982: 153). David Crystal, having scanned the electronic *OED* for coinages, and subtracted nonsense words like *gratillity* – but not 'parallel' words used by others at about the same time – found 2,035 lexeme coinages for Shakespeare, about a tenth of Shakespeare's 20,000 or so total headwords, and many more than Nashe (around 800 coinages), Spenser (c. 500), Sidney (c. 400), or Marston (c. 200); (Crystal, 2004, 327–8).

Following Crystal and Goodland, we would guess that the 2,000+ coined words estimate is probably still too high for Shakespeare's actual contributions of coined words relative to others', for some of the same reasons we would guess that his overall vocabulary is still overestimated relative to others': Shakespeare's words were well recorded and catalogued, and readily accessible to the compilers of the original *Oxford English Dictionary*; other people's words were not (Jespersen 1982 [1938]: 211; Schäfer: 1980; Crystal: 2004). His 'first-uses' would routinely be dated from the first recorded mention of the play whose later publication turned out to contain it. Lesser writers' new words were generally dated from first publication, not first mention. The *OED2* now dates everyone's first-uses from first publication, not first-mention, and, as Goodland has shown, fixing that accounting bias has cost Shakespeare many first uses.

Shakespeare gets credit for the first recorded mention of words like '*sblood* and *Newgate*, but it's hardly likely that he originated these terms. He also gets credit for putting together composite and compound words like *unreal*, *worthless*, *well-read*, or *worm-hole*, whose constituent elements were not his invention (Crystal 2004: 325). Others likewise get credit for such composites, but one wonders whether Shakespeare

doesn't profit disproportionately from such accounting. He gets more credit for malapropisms, nonce words, and proper names than, say, Thomas Nashe. Shakespeare had many such words, and it seems odd for us to count them as coinages if Shakespeare did not intend them as such. Fewer of his new words were overlooked in the *OED*; and more 'parallel citations' of new words from the same year are credited to him than to others (Schäfer 1980: Ch. 2; Goodland). According to Crystal, up to 644 headwords used by others within 25 years of Shakespeare's first use could be described as 'parallel citations.' If we divide these half and half between Shakespeare and others, it gives him 1,713 coinages, about the going rate today for Shakespeare coinage estimates (Crystal 2004: 326).

Further whittling of the total will be done as more of other people's writings gets digitized, and as co-authored parts of the Canon get more clearly distinguished from Shakespeare's parts. Both of these are happening apace – see Goodland's 'Strange Deliveries', Ravassat and Culpeper, 2011, Chapter One, and Vickers, 2002a and 2002b. If someone else wrote *A Lover's Complaint* or the first two acts of *Pericles*, Shakespeare should not get the credit for the coinages these contain. Moreover, over 900 of Shakespeare's 'coinages' were words like *adoptious*, which never caught on with others and hence, as coinages, are more like private IOU's or confederate money than like actual current coin of the realm (Crystal 2004: 326; see the discussion of nonsense and nonce words above). When George W. Bush comes up with 'Bushisms' like *subsidation*, *analyzation*, *hopefuller*, *more few*, and *explorationists*, we suppose that he is struggling to follow accepted rules of word formation (see Crystal 2004: 314–5) but has gotten in over his head. Everyone sniffs at such gaffes, and no one praises them as additions to the language (http://slate.msn.com/id/76886/). If Bush gave us words like *insultment*, *omittance*, *opulency*, *revengive*, *thoughten*, or *casted*,

these would likewise be gathered and laughed at as 'Bushisms.' But it was not Bush who gave us the second set, it was Shakespeare – and *his* gaffes are hailed as brilliant landmarks of 'linguistic daring,' fresh evidence of his peerless mastery of the language, 24-carat coinages for Shakespeare that would be dismissed as pot-metal if they came from someone else. It is probably true that the lines between obvious gaffes and permissible creative license are clearer now than they were in Shakespeare's time,[10] but a coinage still in circulation today, like *acutely*, carries more weight than one like *insultment* that never caught on and would be a gaffe or a malapropism if it came from anyone else – and, in context, was probably intended by Shakespeare to be taken as a gaffe. Again, we suspect that Shakespeare's coinage list would be considerably shorter if he were held to the same standard as others.

Counting coinages has most of the problems of counting other words – making due allowance for corpus size, latency, inflections, multiple meanings, and so on – plus special problems of its own, which are harder to do with computers and have been less explored: sorting out what counts as a word, whether it is one person's real coinage, and not another's, and deciding whether it has caught on. The bigger, better, and more searchable the corpus of available comparison texts, the more and better such sorting can be done. David Crystal has pushed this kind of analysis farther than most, but it is still in its earliest stages (Crystal 2004: 317–29). We do suspect from a Jürgen Schäfer estimate (1980: Ch. 4), that Shakespeare's rate of 'innovation units' per thousand words of corpus is lower than Nashe's, but we don't know yet how it compares with, say, Jonson, whose total corpus and inferred vocabulary seem similar in size to Shakespeare's, nor with Milton, whose inferred vocabulary looks much larger than Shakespeare's, but who wrote too late to get in on the Early-Modern orgy of word-creation. After Crystal's initial whittling down (but not Schäfer's or

Goodland's), Shakespeare still retained an impressive residue of core coinages which exceeded what would be left of Nashe and the others similarly whittled (Crystal 2004: 327) – but not nearly as other-dwarfingly as one would have thought from reading Müller, Marder, or Shipley. Schäfer's computations of authorial innovation *rates* per thousand published words (his 1980) would put Shakespeare behind Nashe, just as our three different estimates of vocabulary richness per 40,000-word block put him behind Milton and Spenser. Over the years, as analysis has grown more carefully focused, and the impacts of bardolatry better controlled, Shakespeare's coinage edge over his peers has not grown, but steadily shrunk.

Giles Goodland shows us that much more shrinkage can be expected. He started with what looks like a baseline of 122 words beginning with P, Q, and Ra-, about seven percent of Shakespeare's estimated 1,700 coinages, and reexamined them against just 40 months of then-newly-recorded Renaissance e-texts. From the baseline he subtracted 41 words antedated by the new texts (about one a month); 16 words redated by rules for the first time applied equally to Shakespeare and others; and 17 words that looked like intended nonce words, not coinages. These reductions cut the baseline by 60%, with no allowance for further antedatings or for the three surviving 'Shakespeare coinages' from *A Lover's Complaint*, which Shakespeare may not have written – 'posied,' 'pellet,' and 'phraseless.' If all of Shakespeare's estimated 1,700-word coinages were also deflated by a comparable 60%, it would leave only 680 coinages, with an overall ongoing shrinkage of about fourteen words a month. That would be less than Crystal's 2004 estimate for Nashe (2004: 327–8) and consistent with Schäfer's conclusion that Nashe had higher rates of innovation than Shakespeare (1980: Chap. 2). We would expect Nashe's coinages, like Shakespeare's, to suffer some attrition from recent and future additions to Early-Modern e-text databases but,

following Schäfer, would be surprised if his coinage counts turned out to be as grossly inflated as Shakespeare's.

In fairness to Shakespeare, again following Crystal, we should note that our simple word-counting does not give Shakespeare (or others) due credit for what could be hundreds of cases where he (or they) added a new sense or meaning to an already-known word (Crystal 2004: 318–21). It is possible that Shakespeare would do better than most by this hard-to-count, qualitative standard and might yet be found to surpass the others even after further massive deflation – but this is a different way of stating the theme of this chapter: that Shakespeare's genius is not so well measured by the number of words he used or coined, as by the way he used them.

NOTES

---

[1] Our thanks to David Crystal, David Hoover, MacDonald Jackson, and three anonymous readers for helping us try to answer this question. And our congratulations to Hugh Craig, of the University of Newcastle, Australia, for his impressive study, 'Shakespeare's Vocabulary: Myth and Reality,' forthcoming in the *Shakespeare Quarterly*, which we have had a chance to read in manuscript. He arrived independently at conclusions very similar to ours: 'Shakespeare is in fact no different from his contemporaries in the number of different words he uses.'

[2] Some encyclopaedists say that English could have as many as a billion words, counting proper names, abbreviations, and scientific terms (McArthur 1992: 1091).

[3] *Horse* can also be a *morpheme*, the smallest linguistic unit of a 'generated' or compounded headword (like *horsemanship*) that has semantic meaning.

Horsemanship has three morphemes, *horse, -man*, and *–ship*. In principle, a dictionary stripped of compounded headwords found in the language and reduced to nothing but morphemes would be even shorter than conventional dictionaries. But we know of no one who has tried to de-compound all of Shakespeare's words into morphemes, and we have not tried it ourselves.

[4] Hart's handcounting also permitted him to eliminate proper nouns from his word counts, a task which is harder for machine counts. But note that our definition of "new word" differs from Hart's. For Hart a new word is the word's actual first occurrence in a Shakespeare work. For him there can be many others afterward, but it is still new the first time. For us, following Efron and Thisted (1976), a new word is a word found only in the block tested. There can be more than one such word in a given block, but none in any other block before or after. A better description of ours might be "words unique to block," but we are in no hurry to change 25 years of our own prior usage, far less Efron's and Thisted's, to something more complicated.

[5] http://wordhoard.northwestern.edu/userman/index.html.

[6] Efron and Thisted's estimate had a problem which we consider a small one: it seems to be based on a preliminary 1968 Spevack estimate of 31,654 Shakespeare word types, and not on their final 1974 estimate of 29,066 types. No one can remember what caused the difference between the two estimates. Possibly some of the poems from *The Passionate Pilgrim* were assigned to Shakespeare in the first estimate but

statisticians made in inferring a large latent vocabulary from a smaller manifest one of either size.

[7] After many years of crunching, we (and most others, such as Foster 1987, Pendleton 1989, and Vickers 2002a) doubt that *Shall I Die?* is Shakespeare's; but Thisted and Efron, in fact, had only claimed that it was a 'could-be' for Shakespeare, not a 'must-be.' We also doubt that Thisted-Efron tests are valid for samples as short as 500 words. But we consider these tests well-validated for samples of 1,000 words or more and copiously validated for the entire Shakespeare corpus of 884,000 words. Thisted-Efron's 1987 article was the primary inspiration for our Claremont Shakespeare Clinic, a series of student teams using computers to test the works of 37 'Shakespeare Claimants,' along with 30-odd poems and plays of the Shakespeare Apocrypha and Dubitanda, for Shakespeare authorship (Elliott and Valenza, 1996).

[8] One could argue that Milton's *Paradise Lost* had a different subject-matter vocabulary from those of the playwrights tested, and would therefore be expected to add more to the group's new-word inventory than a sample with the same subject-matter as the others – or that Shakespeare, who contributed at least a quarter of the group's tokens, might have contributed fewer words per block that were new to himself. Both could be so, but probably not enough so to bring Shakespeare to

comparison of large, standardized blocks from a variety of different authors is far more indicative of their relative vocabularies than the traditional alternative of comparing whole corpora of vastly different sizes.

[9] Schäfer 1973: 206-20; Shipley 1977: 28; Garner 1982: 151; Crystal 2004: 326; Hitchings 2008: 124; McQuain 1998: viii [1,500].

[10] It is also probably true that the Shakespeare list of "Bushisms," if read in context, could be discounted in several ways. "Revengive," *lr.*, 2.01.48, is found in two *Lear* Quartos, but the Folio has it as "reuenging." "Omittance," *AYL*, 5.05.133; "insultment," *Cym.*, 5.05.141; "opulency," *Tim.*, 5.01.387; and perhaps even "casted," *H5* 4.01.23, are spoken by characters of whom Shakespeare was making fun for trying to speak beyond their abilities: respectively, Phoebe, Cloten, Poet, and Henry V. It is doubtful that Shakespeare himself intended them as serious words, yet lexicographers count his "Bushisms," but not Bush's, at full value. It seems to us a double standard.

REFERENCES

All Shakespeare references are to Evans, G. Blakemore and J. J. M. Tobin (eds), (1997 [1974]). The Riverside Shakespeare. Boston, Houghton Mifflin Company.

Bragg, Melvyn 2003. *The Adventure of English*. New York, Arcade Publishing.

Clarke, Mary Cowden. 1881. The Complete Concordance to Shakespeare: Being a Verbal Index to All the Passages in the Dramatic Works of the Poet, Revised Edition. London: Bickers & Son.

Crystal, David 1995. The Cambridge Encyclopaedia of the English Language. Cambridge University Press.

Crystal, David 2004. *The Stories of English*. Woodstock and New York: The Overlook Press.

Efron, Brad and Ronald Thisted 1976. 'Estimating the number of unseen species: how many words did Shakespeare know?', *Biometrika* 63: 435–47.

Elliott, Ward. E. Y. and Robert J. Valenza 1996. 'And then there were none: winnowing the Shakespeare claimants', *Computers and the Humanities* 30: 191–245.

Elliott, Ward E. Y. and Robert J. Valenza 1997. 'Glass slippers and seven-league boots: C-prompted doubts about ascribing *A Funeral Elegy* and *A Lover's Complaint* to Shakespeare', *Shakespeare Quarterly* 48: 177-207.

Elliott, Ward E. Y. and Robert J. Valenza 2004. 'Did Shakespeare write *A Lover's Complaint*? The Jackson ascription revisited' in: Brian Boyd, ed., *Words That Count: Essays on Early Modern Authorship in Honor of MacDonald P. Jackson.* University of Delaware Press: 117-140.

Finkenstaedt, Thomas, Leisi, Ernst, and Wolff, Dieter. 1970. *A chronological English dictionary listing 80000 words in order of their earliest known occurrence.* Heidelberg: C. Winter, Universitätsverlag.

Fisher, Ronald A. et al. 1943. 'The relation between the number of species and the number of individuals in a random sample of an animal population', *Journal of Animal Ecology* 12: 42–58.

Foster, Donald. W. 1987. ' "Shall I die?" post mortem: defining Shakespeare', *Shakespeare Quarterly* 38 (Spring): 58–77.

Foster, Donald W. 1989. *Elegy by W.S.: a study in attribution*. Newark, Delaware: University of Delaware Press; London and Toronto: Associated University Presses.

Garner, Bryan A. 1982. 'Shakespeare's latinate neologisms', *Shakespeare Studies* 15: 149–70.

Goodland, Giles 2010. '"Strange deliveries": contextualizing Shakespeare's first citations in the *OED*', in this volume, Chapter One.

Hart, Alfred (1942). Stolne and surreptitious copies, a comparative study of Shakespeare's bad quartos. Melbourne and London, Melbourne University Press.

Hart, Alfred 1943a. 'Vocabularies of Shakespeare's plays', *Review of English Studies* 19: 128–40.

Hart, Alfred 1943b. 'Growth of Shakespeare's vocabulary', *Review of English Studies* 19: 242–54.

Hitchings, Henry 2008. *The Secret Life of Words: How English Became English*. New York: Farrar, Straus, and Genoux.

Hussey, Stanley S. 1992 (1982). *The Literary Language of Shakespeare*. Harlow, Essex: Longman Group UK Limited.

Jackson, MacDonald P. 1965. *Shakespeare's A Lover's Complaint: Its Date and Authenticity*. Auckland: University of Auckland.

Jespersen, Otto. 1982 (1938). *Growth and Structure of the English Language*. Chicago: University of Chicago Press.

Lerer, Seth 2007. *Inventing English: A Portable History of the Language*. New York: Columbia University Press.

Marder, Louis 1962. *His Exits and Entrances: The Story of Shakespeare's Reputation*. Philadelphia and New York: J.B. Lippincott Company.

McArthur, Tom. 1992. *The Oxford Companion to the English Language*. Oxford and New York, Oxford University Press.

McCrum, Robert, Robert MacNeil and William Cran 2002 (1986). *The Story of English*. New York: Penguin Books.

McDonald, Russ 2001. *Shakespeare and the Arts of Language*. Oxford: Oxford University Press.

McQuain, Jeffrey and Stanley Malless 1998. *Coined by Shakespeare: Words and Meanings First Penned by the Bard.* Springfield, MA: Merriam-Webster.

Muir, Kenneth 1973 (1964). '"A Lover's Complaint": a reconsideration' in *Shakespeare the Professional and Related Studies*. London: Heinemann, 204–19.

Müller, F. Max 1862. *Lectures on the Science of Language*. New York: Charles Scribner.

Müller, F. Max 1891. The Science of Language, founded on lectures delivered at the Royal institution in 1861 and 1863. New York: Charles Scribner's Sons.

Nevalainen, Terttu 1999. 'Early Modern English Lexis and Semantics', in: Roger Lass (ed.), *The Cambridge History of the English Language, 1476 to 1776*. Cambridge, Cambridge University Press, volume 3, 771.

Ogburn, Charlton, Jr. 1984. *The Mysterious William Shakespeare: The Myth and the Reality*. New York: Dodd, Mead, & Company.

Pendleton, Thomas A. 1989. 'The Non-Shakespearean Language of 'Shall I Die?'', *Review of English Studies* 40 (August): 323–51.

Ravi, V. S. 'No doubts about his genius', *The Hindu*, 31 August 2003.

Schäfer, Jürgen 1973. *Shakespeares Stil: Germanisches und Romanisches Vocabular*. Frankfurt am Main: Athenaeum Verlag.

Schäfer, Jürgen 1980. Documentation in the O.E.D.: Shakespeare and Nashe as Test Cases. Oxford: Clarendon Press.

Scheler, Manfred 1982. Shakespeares Englisch: Eine sprachwissenschaftliche Einführung. Berlin: Schmidt.

Shipley, Joseph T. 1977. In Praise of English: The Growth and Use of Language. New York: Times Books.

Simpson, John, and Weiner, Edmund. 1989. *The Oxford English Dictionary*. Oxford: Oxford University Press.

Spevack, Marvin 1973. *The Harvard Concordance to Shakespeare*. Cambridge, Mass.: Belknap Press of Harvard University Press.

Spevack, Marvin and William Shakespeare 1968. *A Complete and Systematic Concordance to the Works of Shakespeare*. Hildesheim: Georg Olms.

Taylor, Gary 1989. Reinventing Shakespeare: A Cultural History from the Restoration to the Present. New York: Weidenfeld & Nicholson.

Thisted, Ronald and Brad Efron 1987. 'Did Shakespeare write a newly-discovered poem?', *Biometrika* 74(3): 445–55.

Valenza, Robert J. 1990. 'Are the Thisted-Efron authorship tests valid?', *Computers and the Humanities* 25: 27–46.

Vickers, Brian 2002a. 'Counterfeiting' Shakespeare: Evidence, Authorship, and John Ford's Funerall Elegye. Cambridge, UK, and New York: Cambridge University Press.

Vickers, Brian 2002b. Shakespeare, Co-author: a Historical Study of Five Collaborative Plays. Oxford and New York: Oxford University Press.

Vickers, Brian 2007. *Shakespeare, A Lover's Complaint, and John Davies of Hereford*. Cambridge and New York: Cambridge University Press.

Weekley, Ernest 1952 (1928). *The English Language*. London: Andre Deutsch Limited.

Wells, Stanley W., Taylor, Gary, Jowett, John, and Montgomery, William. 1987. *William Shakespeare, a Textual Companion*. Oxford, UK and New York: Clarendon Press, Oxford University Press.

Winchester, Simon 2003. The Meaning of Everything: The Story of the Oxford English Dictionary. Oxford: University Press.

Appendix One: Various Measures of Vocabulary Richness

Name	Includes 40,000 words each from	Tokens	Types	NW	1/2 tokens @	Total Virtual Inferred Vocab.
MILTON1	Paradise Lost, 1-6	40,000	6671	1790	127	140415
MILTON2	Paradise Lost, 7-12	38,886	6328	1591	117	123490
DEKKER*	Whore of Babylon, Honest Whore	40,000	5742	1093	92	85567
SHPOEMS*	Venus, Lucrece, Sonnets	40,000	6054	1147	91	84180
MARLOWE1*	Tamburlaine, I & II, DF16	40,000	5444	936	83	73449
JONSON1*	Sejanus, Volpone	40,000	5452	576	81	70867
GREENE*	Alphonus, Friar Bacon, James IV	40,000	5042	667	79	68324
MIDDLE1*	Phoenix, Michaelmas, Chaste Maid	40,000	4761	883	77	65822
MIDDLE2*	Witch, Hengist, Women Beware Women	40,000	5006	833	76	64586
JONSON2*	Alchemist, Bartholomew Fair	40,000	5857	1076	74	62144
SHTEMPWT*	Tempest, Winter's Tale	40,000	5841	964	73	60938

SHCORCYM*	Coriolanus, Cymbeline	40,000	5341	879	72	59742
SHOTHLR*	Othello, Lear	40,000	5221	865	71	58556
MARLOWE2*	Jew of Malta, Edward II, Massacre	40,000	4892	720	66	52777
CHAPMAN*	Gentleman Usher, Bussy D'Ambois	40,000	4811	767	65	51652
FLETCHER2*	Barnavelt, Island Princess	40,000	4460	477	64	50536
SH3COMS	Shrew, TGV, Comedy of Errors	40,000	4892	672	63	49430
FLETCHER1*	Woman's Prize, Valentinian	40,000	4427	382	63	49430
	=Shakespeare					
Milton Average			6500	1691	122	131953
Shakespeare Average			5470	905	74	62569
6 Other Playwrights Average			5223	839	77	66132
Fletcher Average			4444	430	64	49983

* = in baseline composite lexicon for eight authors

Sh NW	3855	=31%
All NW	12265	

18 blocks by 9 authors, 15 blocks by 8 authors in baseline lexicon; 600,000 words, 4 blocks by Shakespeare, 160k words = 27%
Total composite lexicon: 28747 words, 12265 of them NW (43.1%)
All calculations done with Intellex

Spenser	Faerie Queene, Book 1, old spelling	40,000	6235	122	131827
Moliere	L'Avare, Don Juan, in French	39,253	5003	69	56215
Elliott	Politics of Population, 40k words	40,000	7067	147	177275
Elliott & Valenza	Shakespeare Book 40k words	40,000	5810	102	99984
Melville	Moby Dick Last 40k words	40,000	7056	99	95553
Iliad 3-40	Iliad Block 3 First 40k words	40,000	3985	89	81437
Odyssey 2-40	Odyssey Block 2 First 40k words	40,000	4023	58	44053

David Hoover Type-Token figures, all done with TACT

Paradise Lost 1	40,000	7005
Paradise Lost 2	40,000	6860

Sh. Poems	Sonnets, Venus, Lucrece		40,000	6098
Moby Dick 1			40,000	6554
Moby Dick 2			40,000	7339
Moby Dick 3			40,000	6855
Moby Dick 4			40,000	7171
Moby Dick 5			40,000	6937
Lewis, Main St. 1			40,000	7285
Lewis, Main St. 2			40,000	7119
Lewis, Main St. 3			40,000	6996
Lewis, Main St. 4			40,000	6381
Lewis, 367K	Babbitt, Main St., Our Mr. Wrenn		367,609	24105
Melville, 612 K	Moby Dick, Pierre, Redburn, Benito C., Omoo		612,647	31979
Other large corpora				
Shakespeare 884K	entire Canon per Spevack		884,647	29066
Homer	Iliad & Odyssey, Samuel Butler Translation		269,779	13041
King James Bible	Entire KJV		789,955	12803
Milton, all	Para Lost; Para Regained, Samson; young		117,693	14899
Milton, Para Lost	Paradise Lost		80,779	11244
Sh. Average, 6x40,000-word blocks				62569
Thisted-Efron estimate of Shakespeare's Total Vocabulary, minimum				66000
Colleagues 200 wd	Number Headwords Recognized		Lemmatized	Inflected**
N		147	47775	79784
R		138	44850	74900
W		137	44525	74357
K		90	29250	48848
S		111	36075	60245
Average		124.6	40495	67627

** inflected = lemmatized x 1.67