Word-classes in performance

RICHARD HUDSON

1 The question

How many of the words in an English text would you expect to belong to the various word-classes? If you are a typical linguist, your immediate reaction to this question is probably one of impatience. The answer is obvious: it's a question we shouldn't even be bothering to ask. I shall present evidence in this paper that this reaction is mistaken, but let me admit from the start that until recently this would have been my reaction too. The results that I shall report below surprised me, and I only happened upon them by chance, as a by-product of a different research programme.

To set the scene, let me explore the excellent reasons why one should reject the question. First, there are a priori theoretical grounds. We all know that every word which is used is selected on the basis of many different factors. It must fit the meaning that has to be expressed, which is a matter of lexical meaning, but it also has to fit the morphological and syntactic requirements of the surrounding sentence, not to mention any stylistic and social requirements and the need to be easily comprehensible; and the meaning which has to be expressed is itself the outcome of pragmatic reasoning. As I write this paper, I delete and replace almost as many words as you read; what changes isn't just the way the message is encoded, but the message itself. How could all this coming and going possibly lead to any kind of pattern in the choice of word-classes? The choice of words is unique, and creative, in every text that we produce - every paper, essay, letter, or conversation - so we can confidently predict that this uniqueness will be reflected in the balance of word-classes.

There are even stronger reasons for rejecting the question: we already know the answer, thanks to various large-scale research projects on vast corpora such as the million words of the Brown corpus of written American English (Francis and Kucera 1982), its million- word sister corpus of written British English, the Lancaster-Oslo-Bergen (LOB) corpus(Johanson and Hofland 1989), and the half-million word London-Lund corpus of spoken British English (Svartvik and Quirk 1980). These studies have revealed considerable differences among texts of different types in their word-class constitution (as well as in many other features). So we know that the question rests on simplistic assumptions about some kind of mythical 'common English', which conflicts with the reality of variation.

In the light of these doubts about the question, the answer is all the more surprising. However I should issue an important caveat: the following discussion

is based entirely on written texts. One of the most obvious questions that it raises is whether or not it generalises to speech. I have no information at all about this question, though I have clues as to how well the answer generalises beyond English to other languages.

2 The answer

The answer is that at least some of the most general word-classes do have a remarkably constant representation across texts. I have figures for nouns, verbs and adjectives which are consistently highest for nouns and lowest for adjectives. Furthermore, the figures (as percentages of the total word-tokens in the text) are surprisingly constant. Over various corpora I have found averages of between 46% and 49% for nouns, and even very short texts of about 100 words yield similar figures, as I shall explain below. For verbs and adjectives I have somewhat less data, but they tended respectively to comprise about 18% and 7% of each text.

How can we reconcile this answer with the known fact, mentioned above, that the balance of word-classes varies considerably from text to text? The two claims may appear to be diametrically opposed, but they are not, for the simple reason that the classes concerned are different. Earlier work has tended to focus on rather small classes, such as wh-words or personal pronouns, whereas the figures that I am quoting apply to the grossest of all word-classes: nouns, verbs and adjectives[1]. My class of 'noun' includes not only common nouns but also proper nouns and all pronouns; furthermore, 'pronoun' for me includes all determiners, so THIS is a pronoun whether it is used on its own or with a following common noun (Hudson 1990:268ff), and 'common noun' includes all numerals (ibid 302ff). So far as I know, such gross categories haven't been considered directly in previous text counts, though they can often be reconstructed from the published figures for sub-classes; but the picture that emerges when they are studied is that there is less variation in the use of gross categories than there is in the use of their subdivisions. Let me give a typical example.

The LOB corpus (of a million words) is a collection of 2,000-word texts which were assigned to 15 different genres. The published account (Johanson and Hofland 1989) gives the figures for all the word-classes across these 15 genres, and after lumping their micro-categories into successively larger ones, the figures in Table 1 emerge.

---

[1] I have no figures for prepositions and adverbs because I find these classes most difficult to delimit. For example, is *inside* a preposition or an adverb in *I went inside*? Published data are unusable precisely because one does not know what decisions have been made in cases like this.

	mean	standard deviation
common nouns	21%	3.7
all nouns	47%	1.8

Table 1: common nouns and nouns in the LOB corpus, across genres.

The standard deviation is a standard statistical measure of the range of variation in a set of figures, so the higher the standard deviation, the more variation there is. If the amount of variation had been proportionately constant between these categories, it should have been roughly twice as great for nouns as for common nouns; but it is in fact just half as great[2].

This pattern is repeated in every corpus I have studied, and not only for common nouns but also for other sub-classes of nouns. It seems therefore to be robust, and raises some very interesting theoretical issues. Why should there be any constancy at all in the representation of word-classes? And why should it be greater for super-classes than for their sub-classes? Unfortunately I have no answers to these questions, but I can at least present the evidence in more detail.

3 The evidence

I have three sources of data. I have culled information about the major word-classes from the published reports on the Brown and LOB corpora, including the 15 genre subdivisions of the latter, which amounts to a total of 2 million words of written English (from USA and UK respectively); I have also analysed the figures for sub-types of nouns in the LOB corpus, though not in the Brown one. In addition to these two gigantic sources, I have some which are absolutely tiny, but equally interesting precisely because of their small size: 29 texts of about 100 words each

[2]In some cases one would expect more variation in the figures for sub-classes than in those for super-classes for the simple reason that the former are so much smaller and therefore subject to more random variation. However this can hardly be the reason for the differences in the LOB corpus, where the figures concerned are simply vast even for the sub-classes - e.g. the smallest number of common nouns for any of the genre-samples is 3820, and the largest is 32008. (The discrepancy between these two figures reflects gross differences in the sizes of the various genre-samples.)

covering a wide range of genres, which were independently selected and analysed[3] by students in a second-year undergraduate class on syntax. These are interesting because they show how constant the balance of word-classes is even over very short texts.

Let's start with the means for all nouns across the corpora, shown in Table 2.

Brown	49%
LOB	47%
minitexts	46%

Table 2: means for all nouns in all three corpora.

The similarities among these figures are quite striking, especially considering how small the third corpus is. However, even a difference as small as 2% over a million words requires an explanation, which at present I can't give.

The Brown and LOB corpora are both divided into two main sub-corpora, called 'informative' and 'imaginative' - roughly, fact and fiction - for which the figures are as shown in Table 3.

	informative	imaginative
Brown	50%	48%
LOB	49%	46%

Table 3: Means for all nouns in the informative and imaginative divisions of the Brown and LOB corpora.

The figures for the sub-corpora are still very close, but reveal an interesting trend: in both cases, informative texts contain slightly more nouns than imaginative ones do. I have no explanation for this trend, though it is clear enough (given the size of the corpora) to call for an explanation; but it suggests that it may be possible to make the percentage figures even more precise by relating them to specific genres. This possibility is still compatible with the general claim that the percentage of

[3]The texts which I have included in my corpus were the ones with the fewest errors, but I have corrected even these analyses wherever necessary.

nouns is remarkably constant across texts, while allowing even greater constancy within individual genres.

Another way of presenting the similarities among the genres is to compare the figures for the 15 genres recognised in the LOB corpus - genres such as 'press: reportage', 'skills, trades and hobbies' and 'science fiction'. I mentioned these earlier in Table 1, which showed that they all clustered rather closely round the mean of 47%, with a standard deviation of a mere 1.8%. The genre with the lowest percentage for nouns scored 44.6%, only 2.4% below the mean, but the highest was much more deviant, with 52%. This was the genre of press reportage, which is 2% higher than the next highest (press reviews), so quite out of step with the other figures. If we exclude this untypical genre, and measure the relation between the number of nouns (of all types) and the number of words in each of the genre-categories, we find a stunning correlation of .999, which shows up as a virtually straight line on a graph. Very few generalisations about human behaviour come anywhere near to this figure. If the figure can be made even more precise by specifying the genre concerned, this is indeed impressive!

Equally interesting are the figures for the 29 tiny 100-word texts. One might think that the stability in the figures for the Brown and LOB corpora would emerge only over tens of thousands of words, as the many interacting influences gradually balanced one another out. The minitexts show that this is not so; roughly similar figures emerge even in a mere 100 or so words. The mean for all these texts was 46%, as already shown in Table 2, but the standard deviation reported there was only 1.8, which is tiny. Admittedly there is a greater spread of scores than for the sub-corpora, from a minimum of 40% to a maximum of 52%, but the trend towards the mean is already evident. Indeed, the sentence which I have just written scores 52% (with 16 nouns in a total of 31 words), which is already near to the mean. Here it is again, with the nouns highlighted:

> **Admittedly there is a greater spread of scores than for the sub-corpora, from a minimum of 40% to a maximum of 52%, but the trend towards the mean is already evident.**

Counting the whole paragraph from *equally* to *highlighted*, the score is 47% (74 nouns out of 156 word-tokens).

The figures quoted above suggest the following factual claim about written English texts: in any written English text, between 46% and 49% of the words are most likely to be nouns (depending on which of the corpora we take as our model). The claim could not of course be refuted by a single text, because (as we have seen) some texts do lie outside this range; but the standard-deviation figures make quite precise claims about the distribution of figures within a collection of texts:

that about 70% of them will be less than the standard deviation from the predicted figure.

Another fact which emerged from the analysis, and which I have already mentioned briefly, is that the figures for sub-classes of noun (e.g. pronoun, common, proper) vary more than do those for nouns as a whole. As Table 1 showed, in the LOB corpus there is twice as much variation, around a much lower mean, in the figures for common nouns as there is for all nouns. This pattern is repeated in the figures for the minitexts shown in Table 4.

	mean %	standard deviation
all nouns	46%	3
common nouns	22%	5
pronouns	20%	4
proper nouns	4%	4

Table 4: Percentages and standard deviations for nouns and sub-classes of noun in the 29 minitexts.

The figures in Table 5 for the two main Brown sub-corpora tell the same story.

	informative	imaginative
all nouns	50%	48%
common nouns	52%	39%
pronouns	38%	52%
proper nouns	9%	8%

Table 5: Percentage of nouns and noun sub-classes in the two main divisions of the Brown corpus.

It can be seen that the means for common nouns and pronouns vary by 13% or 14%, which is considerable compared with the almost constant figures for all nouns and for one sub-class, proper nouns. It is no wonder that earlier research has been

impressed by variation more than by constancy, since it has focussed on sub-classes.

The figures in Table 5 deserve a little more discussion, because they show that the figures for common nouns and pronouns are very closely related: as one goes up, the other goes down - in fact, they simply swap figures between the two sub-corpora. Meanwhile, proper nouns are unaffected. The inverse relation between common nouns and proper nouns is confirmed in the other two corpora. In the LOB corpus, there is an inverse correlation of -0.976 between them, an extremely significant figure, and in the minitexts there is a respectably significant figure of -0.562.

The same relation between common nouns and pronouns emerges from a factor analysis of two large corpora (the LOB corpus and the London-Lund corpus of spoken English) by Douglas Biber (Biber 1988), whose aim was to reduce the many variable features of the texts to a smaller number of more general 'factors'. The most important of these factors predicted the variation in the number of pronouns and common nouns (among other things), but pronouns were positively weighted while common nouns were negatively weighted - i.e. a text that was highly ranked on this factor would tend to have relatively many pronouns but relatively few common nouns (ibid p.102)[4].

It seems that in some sense common nouns and pronouns are in 'complementary distribution', but that the variation between them leaves the total number of nouns unaffected - a remarkable balancing act, especially bearing in mind that my category of pronouns includes determiners, each of which is typically accompanied by a common noun! Speculatively, it would seem that the mental circumstances which require a noun arise at a constant rate, but that a different set of circumstances which are much more variable then intervene to influence the choice between pronouns and common nouns.

What about word-classes other than the noun class? I have some data on adjectives and verbs (which here include auxiliary verbs), and I can report further consistency, though I have less evidence. Tables 6 and 7 summarise the data I have extracted from the Brown and LOB reports.

---

[4]Another of the variables which Biber found in his first factor was the type-token ratio, which had the opposite weighting from pronouns. This confirms a reasonably clear trend that emerged from my minitexts, which was a strong negative correlation of -.668 between the type-token ratio and the use of pronouns. According to my analysis, there is almost the same negative correlation (-.631) between the type-token ratio and the use of verbs, but Biber's data did not include a global figure for verbs.

	Brown	LOB
adjectives	7%	7%
verbs	18%	18%

Table 6: Overall percentages for adjectives and verbs in the Brown and LOB corpora.

	Brown	LOB
informative	17%	16%
imaginative	21%	22%

Table 7: Percentages for verbs in the two main divisions of the Brown and LOB corpora.

These tables speak for themselves, showing remarkable consistency between Brown and LOB both in their overall percentages, and also in the differences between their main divisions. My minitexts yielded very similar figures, which are in Table 8[5].

	mean	standard deviation
adjectives	9%	4
verbs	17%	4

Table 8: mean percentages and standard deviations for adjectives and verbs in the minitext corpus.

In conclusion, then, it looks as though at least some of the other major word-classes may have a constant distribution among texts, even if they are slightly less constant than nouns.

Finally, what do we know about languages other than English? The simple answer is 'Very little', but I can report two facts which are at least suggestive, and which are based on my collection of minitexts. The most interesting of these is that

[5]Interestingly, there is some evidence from the minitexts that adjectives and verbs are alternatives to each other, namely a modest negative correlation of -0.591.

a short Welsh text[6], of 147 words, had only 30% nouns, a figure which is 10% lower than the lowest of the English texts. The text concerned is a passage from the New Testament for which a modern English translation was also analysed, and it is worth noting that nouns made up 44% of the English translation of the same passage, so the difference cannot be attributed to the genre of the Welsh text. This one figure suggests strongly that some languages may produce very different figures from English.

The second fact is that in other respects Welsh, German and Russian seem in general to show the same figures as English. In addition to the one Welsh text mentioned above I have three in German[7] and one (of 211 words) in Russian[8], whose basic data for the three main word-classes are shown in Table 9.

	Welsh	German			Russian
all nouns	30%	47%	46%	39%	49%
adjectives	5%	5%	11%	8%	10%
verbs	20%	19%	16%	19%	15%

Table 9. Frequencies of nouns, adjectives and verbs in Welsh, German and Russian minitexts.

All these figures, with two exceptions, are well within the range for English. One exception is the 30% for Welsh nouns, which I have just discussed. The other is 39% for nouns in one of the German texts, which is 1% lower than the lowest English text. Is this pure chance, or does it suggest a slightly different range for German?

[6]The Welsh text was analysed by Aled Jones.

[7]One of the three German texts reported here was analysed by Stephan Schöler.

[8]The Russian text was analysed by Dr Natalia Ignatieva.

4 The next questions

The data show beyond any reasonable doubt that there are clear statistical trends in the frequencies of word classes in texts - e.g. around 47% of words in any written English text tend to be nouns. The observed frequencies can presumably be interpreted as manifestations of probabilities for individual words: the probability of any given word in such a text being a noun, in the absence of any information about its context, is about 0.47. That, then, is the answer to my original question - an answer, I repeat, which surprised me and I suspect will surprise many readers.

But why? Why should there be any constancy at all from text to text? And why these particular constants, rather than some completely different set of figures? Why should the figure for all nouns be more consistent than that for sub-classes of noun? Why should the figures for English be so similar to those for other languages in most respects, but strikingly different from those for Welsh nouns?

I have no answer to these questions, though I can imagine research that might throw light on some of them. Indeed, I suspect it will be a very long time indeed before the combined efforts of all the human sciences will be able to explain the actual figures, e.g. why 0.47, rather than, say, 0.56? Meanwhile it is very clear (to me, at least) what our first priority should be: to gather more data of the kind reported here, to check the trends and expand the generalisations into other varieties of English (especially spoken English) and to other languages.

References

Biber, Douglas (1988) *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Francis, W. Nelson and Kucera, Henry (1982) *Frequency Analysis of English Usage: Lexicon and Grammar*. Boston: Houghton Mifflin.

Hudson, Richard A. (1990) *English Word Grammar*. Oxford: Blackwell.

Johansson, Stig and Hofland, Knut (1989) *Frequency Analysis of English Vocabulary and Grammar (based on the LOB corpus). I. Tag frequencies and word frequencies*. Oxford: Clarendon Press.

Svartvik, Jan and Quirk, Randolph (eds, 1980) *A Corpus of English Conversation*. Lund: CWK Gleerup.