DISCUSSION NOTE
About 37% of word-tokens are nouns

RICHARD HUDSON
*University College London**

The title of this note is a generalization that may strike readers as a joke. However, it turns out to be true (with some systematic variations) of any reasonably large body of written English, and can be matched by other generalizations about other word-classes, genres, and languages. I am as reluctant as any linguist could be to believe it; after all, the choice of word-classes in a text depends on a myriad of variable influences, from the message conveyed to the style of the author. Moreover, linguists have talked about 'nominal' and 'verbal' styles for some time (since Wells 1960), implying that the relative balance between nouns and verbs is a major source of variation among texts. More recent work on large corpora would appear to support the expectation of major differences, as shown by the counts for major word-classes in the Brown and LOB[1] corpora (e.g. Johansson & Hofland 1989:16). And yet the facts turn out to be otherwise and (in my opinion) far more interesting because they cry out for an explanation.

Table 1 shows some basic figures for the word-classes in the Brown and LOB corpora, based on the reported figures for grammatical 'word-tags' in Francis & Kučera 1982 and Johansson & Hofland 1989.[2] These overall figures are remarkably similar, though the differences are still significantly different from a statistician's point of view. Even small percentage differences have to be taken very seriously when one is dealing with tens or hundreds of thousands of cases (chi-square = 1,531, a difference which is virtually impossible simply by chance). But even if we can't ignore the differences, the similarities are sufficiently striking to suggest an underlying constancy. So far as I know, this particular constancy has not been noted before in published discussions of corpus statistics.

* I have been helped and encouraged by a number of people, including Mike Day, Robin Fawcett, Nik Gisborne, Sidney Greenbaum, Stig Johansson, Aled Jones, Geoffrey Leech, Stuart Rosen, And Rosta, Geoffrey Sampson, Itzhak Schlesinger, Clive Souter, James Tauber, and three anonymous reviewers, not to mention the students in my 'Word Grammar' class who counted word-classes so enthusiastically and efficiently.

[1] The Brown corpus of a million words of written American English was produced in Brown University and is reported in Francis & Kučera 1982. The Lancaster-Oslo-Bergen (LOB) corpus of a million words of British English is described in Johansson & Hofland 1989.

[2] These two corpora are 'tagged', i.e., each word has been assigned to a grammatical word-class. The same system of tagging was used for both corpora, and my gross classes are derived mechanically from these tags, as follows:

common noun: = CD..., NN..., AP$, APS...
proper noun: = NP..., NC, NR...
pronoun: = P..., W..., EX

The dots are variables for extra letters that may make further distinctions; for example, 'CD...' includes not only the tag CD, but also CDS, CD-CD, CD1, CD1S, CD1$ and CDS.

Word-class	Brown	LOB
Common nouns:	24%	22%
Proper nouns:	4%	5%
Pronouns:	9%	8%
Total:	37%	36%

Table 1. Three word-classes as percentage of total word-tokens in two million-word corpora of written English.

What has been discussed explicitly is the variation found among genres (e.g. Carroll 1960, Biber 1988, 1989, Dewaele 1992, Ellegård 1978, Francis & Kučera 1982, Johansson & Hofland 1989). The Brown corpus was divided into 15 genres, such as 'Press: reportage', 'Skills, trades and hobbies', 'Learned and scientific writings', 'Mystery and detective', and 'Humor'. The LOB corpus was explicitly designed as a replication of the Brown one, so the same 15 genre categories were used. The genres can be divided on a first cut into 'Informational' and 'Imaginative', and when these two supergenres are compared across the two corpora, the figures in Table 2 emerge. The similarities between the corpora are striking.

	Informational		Imaginative		Total	
	Brown	LOB	Brown	LOB	Brown	LOB
Common nouns:	26	24	19	17	24	22
Proper nouns:	5	5	4	4	4	5
Pronouns:	6	7	14	15	9	8
Total:	37	36	37	36	37	36

Table 2. Percentages for three word-classes in two subcorpora of LOB and Brown.

As can be seen, in both corpora informational texts contain 7% more common nouns than imaginative texts do, and 8% fewer pronouns. This complementarity between common nouns and pronouns works so smoothly that it leaves the total unaffected. The balance of common nouns to pronouns is already well established as one of the most important variables on which genres vary (Biber 1988:102–5, 227–8).

The superclass that includes pronouns as well as common and proper nouns is clearly what most linguists would call simply 'nouns', though the reports of the Brown and LOB figures don't recognize it as such. This is the sense in which I shall use the term 'noun' here, hence the generalization that nouns constitute about 37% of word-tokens.

Further statistical analysis confirms this picture of underlying regularity in diversity, as we shall see briefly below, but we must first stop and consider the implications of these findings for linguistic theory. If our subject had been (say) medicine and the figures had related (say) to the distribution of blood groups in two separately selected populations of a million people each, then data such as these would have been regarded as a valuable resource for testing alternative theories of blood-group distribution. As it is, of course, we have no theories which could even start to predict these figures. The nearest thing we have is in the area of quantitative dialectology (e.g. Labov 1972, Sankoff 1978, Milroy 1987); but we are still light years from a theory that could be applied to the figures for word-classes. Admittedly, we could speculate about possible explanations. For example, we might suppose that imaginative texts tend to contain a lot of dialogue, and that dialogue involves more pronouns than common nouns; or we might think that maybe fiction involves a small number of central characters, who tend to be mentioned repeatedly by pronoun (an idea that we shall return to briefly below).[3] More positively, data presented below will show that this variation turns out to be part of a larger (but still mysterious) pattern. But what we can't do at present is to get beyond speculation by embedding our explanations in an overall theory from which these particular figures would follow.[4] In short, the trends are facts, not artifacts, but they are facts in search of a theory.

Returning to the figures, we can push the analysis further by distinguishing the figures for the 15 individual genres in the two corpora, without distinguishing the subclasses of noun. Table 3 shows both the length of each genre subcorpus (in words) and the percentage of nouns (i.e. common nouns, proper nouns, or pronouns) found in each subcorpus.

	Number of words		% of nouns	
Genre	Brown	LOB	Brown	LOB
A Press: reportage:	88690	89138	42.2	41.2
B Press: editorial:	54505	54447	36.1	35.8
C Press: reviews:	35346	34321	37.0	37.7
D Religion:	34590	34387	34.8	34.9
E Skills, trades, hobbies:	72590	76913	37.2	35.4
F Popular lore:	97223	89090	35.7	35.9
G Belles lettres, biography, essays:	152064	155336	35.5	34.4
H Miscellaneous informational:	62477	60761	37.9	34.5
J Learned & scientific writings:	162211	161900	35.0	33.3
K General fiction:	58380	59204	36.7	35.8
L Mystery and detective:	48208	49145	36.7	35.5
M Science fiction:	12042	12119	35.5	36.3
N Adventure & western:	58416	59391	37.3	36.4
P Romance & love story:	58625	59382	36.5	36.1
R Humor:	18277	18203	36.7	35.6

Table 3. Total words and noun percentage for 15 genres in the Brown and LOB corpora.

What emerges from this table is that the generalization in the title of this note is an oversimplification of a system which is complex but quite regular. We have already seen that the balance between common nouns and pronouns varies widely with genre, but this table shows that the overall percentage for all nouns also varies, between 33% and 42%. These differences among genres

[3] Both of these suggestions come from anonymous readers of an earlier version of this paper.

[4] For a theoretical approach to the question of pronoun/noun variation, however, see recent work by Jean-Marc Dewaele (e.g. 1992). Dewaele proposes that formality is the most important dimension of stylistic variation in word-classes. For instance, he has presented results showing that a more formal context leads to an increase in nouns, articles, adjectives, and prepositions, while a less formal context favors pronouns, adverbs, and verbs. These results derive from studies of genre differences in native Italian, Dutch, and French, and in French interlanguage.

are extremely significant from a statistical point of view, even though the total numbers are now reduced to 'mere' tens of thousands of tokens (chi-square = 1623 for Brown and 1808 for LOB). The differences may be seen as confirmation of the old idea that genres differ in their 'nominality', but such differences as there are turn out to be small when measured in percentage points—just 9% difference between the most nominal and the least nominal styles.

Given that the intergenre differences are significant, what about the generalizations about each genre? Do the figures from Brown and LOB support them? More generally, is it possible to make any generalizations at all in terms of genres which will apply across corpora? The answer seems to be positive: the constructors of these corpora did a good job in selecting their genres, because at least some of the genre collections emerge as sound bases for statistical generalizations. One simple measure of the relationships between two sets of figures such as the Brown and LOB percentage figures for all nouns is a correlation coefficient, *r*. This measures the extent to which one set of figures is predictable from the other, and if *r* is above .514 or below −.514 then the fit between the two sets is significant at the .05 level (which means this degree of fit would be expected by chance in only 5 experiments in every 100). For noun percentages, the correlation between the figures for Brown and for LOB is .817, which is very significant. This suggests that the intergenre differences in the Brown corpus are very similar to those in the LOB corpus.

Another way to compare these intergenre differences in the two corpora is to rank them for their 'nounfulness'—i.e. according to the percentage of word-tokens which are nouns. For example, it is noticeable that press reportage (genre A) is by far the most nounful in both corpora, while learned and scientific writing (genre J) is the least nounful in LOB and the second least nounful in Brown. Now the trouble with simple ranking is that it may reflect differences which are trivially small. For example, genre K is slightly more nounful in Brown than genre L (36.7009% compared with 36.7014%), but slightly less nounful in LOB (35.7864% compared with 35.4990%). These differences are easily explained by chance (chi-square = .0000023 for Brown and .97 for LOB, both well below any significance level), so we should really consider them as equal in rank in both the corpora. If we apply this criterion to the ranking of genres in the two corpora, we get the picture in Table 4, where just two genres

BROWN		LOB	
% NOUNS	GENRE	% NOUNS	GENRE
35.0	J	33.3	J
34.8, 35.5	D, **M**	34.9, 34.5	D, **H**
35.5	G	34.4	G
35.6	F	35.9, 36.3	F, **M**
36.1, 36.5, 36.7	B, P, R	35.8, 36.1, 35.6	B, P, R
37.2, 36.7, 36.7	E, K, L	35.4, 35.8, 35.5	E, K, L
37.3	N	36.4	N
37.1	C	37.7	C
37.9	**H**		
42.2	A	41.2	A

TABLE 4. How 15 genres are ranked for nounfulness in Brown and LOB.

are highlighted: H and M. These are the only two genres whose ranking is significantly different between the two corpora.

What Table 4 shows is that it is indeed possible to make generalizations about genres which apply across corpora; but once again the main point for those of us who are not directly involved in the comparison of genres is that it is possible to generalize about the percentage of word-tokens that we should expect to be nouns in any given text, and that if we know the text's genre we can predict this percentage more accurately.

Just to emphasize our ignorance, let me return to the discussion above about the balance between pronouns and common nouns. I mentioned two suggested explanations which assumed that the difference had something to do with the use of anaphoric pronouns, so let's explore this suggestion. If we subdivide pronouns to distinguish personal pronouns from other kinds, such as compound pronouns (e.g. *someone*), possessive pronouns, reflexive pronouns, and WH pronouns, what differences do we find among genres? If the variation is mainly concerned with the use of anaphoric pronouns, then we might expect personal pronouns, and perhaps reflexive pronouns, to be the only subtypes of pronoun that are linked to genre in this way. But when we look at the figures in Table 5, this expectation is not confirmed.

What Table 5 shows is that there is a significant negative correlation between the use of pronouns and the use of common nouns for ALL types of pronouns, including compound pronouns (which are definitely not anaphoric). This rather surprising finding suggests that the differences among genres have nothing to do with anaphoricity, so we need to look elsewhere for an explanation.

	BROWN	LOB
% ALL PRONOUNS:	−.952	−.976
% COMPOUND:	−.932	−.905
% PERSONAL:	−.923	−.959
% POSSESSIVE:	−.901	−.929
% REFLEXIVE:	−.887	−.864
% EXPLETIVE *there*:	−.605	−.160
% WH:	−.585	−.374

TABLE 5. Correlation coefficients for the percentage of all common nouns and the percentage of pronouns and subclasses of pronoun in Brown and LOB.

All the figures quoted so far have been based on rather large (sub)corpora, measured in terms of thousands of word-tokens. Is that the scale on which we have to project these regularities? Apparently not. Rather similar figures emerge, though with more variability, in texts as short as 100 or 200 words. For example, in the paragraph before this one there are 61 words, of which 21 are nouns, which gives 34%—a figure which corresponds exactly to the expected 33–35% for learned and scientific writing. Here is the paragraph repeated with the nouns highlighted:

> **What Table 5** shows is that **there** is a significant negative **correlation** between the **use** of **pronouns** and the **use** of common **nouns** for ALL **types** of **pronouns**, including compound **pronouns** (**which** are definitely not anaphoric). This rather surprising **finding** suggests that the **differences** among **genres** have **nothing** to do with **anaphoricity**, so **we** need to look **elsewhere** for an **explanation.**

A collection of 27 student projects[5] of about 100 words each reveals roughly similar figures.

It seems, then, that some genres of written English have a typical word-class 'profile' in terms of both nouns and subclasses of nouns. What about other word-classes, other kinds of language, and, indeed, other languages? Table 6 summarizes some data that I have culled from reports on various corpora.

	P	Noun				V	Ad-word		
Corpus		cN	nN	pN	all		Adj	Adv	all
(1) Written English									
all:	12	23	5	9	37	18	7	5	12
informational:	13	25	5	7	37	17	8	5	13
imaginative:	10	18	4	15	37	22	6	7	13
(2) Written Swedish:	12	24	4	11	39	17	8		
(3) New Testament Greek									
all:	8	21		12	32	20	7	5	12
letters:	9	23		10	33	17	8	6	14
narrative:	7	19		13	32	22	6	4	10
(4) Written Welsh:	13	23	3	11	37	16	7	4	11
(5) Spoken English									
broadcasts:	12	>24		>7	>31	14	6	12	18
prepared speeches:	11	>21		>11	>32	19	5	8	13
interviews:	11	>18		>13	>30	21	6	10	16
spontaneous speeches:	9	>18		>15	>32	21	5	9	14
conversations:	8	>15		>16	>31	24	4	11	15
phone-conversation:	7	>14		>17	>31	25	4	11	15
(phone data 1930)		15		22	37	28			12
(6) Children's English									
free play:		11	2	16	29	29			
interview:		15	3	15	33	23			
female:		12	2	16	31	27			
male:		14	3	14	31	26			
6-year-olds:	5	13	2	18	33	26.5			14
8-year-olds:	5	13	2	17	31	26.5			15
10-year-olds:	5	12	2	17	31	26.5			16
12-year-olds:	5	12	2	16	30	26.5			17

Table 6. Word-class token percentages for 6 kinds of corpora covering written English, Swedish, New Testament Greek, and Welsh, and spoken English of adults and children.

The table is laid out so as to juxtapose the columns for prepositions (P) and common nouns (cN), because these tend to covary. (In the table, nN = proper noun.) Within each corpus, subcorpora that are above average for prepositions are also above average for common nouns, and vice versa; and likewise for below-average scores. A similar trend links verbs (V) and pronouns (pN), and these two pairs of parameters appear to be negatively related: every corpus which is high on prepositions and common nouns is low on verbs and pronouns, and vice versa. Biber noted the positive and negative relations among these

[5] Students seem to enjoy counting word classes in texts, and to learn a lot about language in the process.

four variables (1988:102), but it is interesting to see how clearly they emerge even from data as crude as those in Table 6, and how the same pattern is repeated in New Testament Greek.

The corpora are a rather random collection of data that I could process with my limited resources and that happen to have come my way:

(1) Written English. This corpus is made up of the Brown and LOB corpora combined into a single corpus of two million words. The figures for word-classes in the two corpora can be merged because they are virtually identical. The corpora differ by no more than 1% whether they are divided into imaginative and informational subcorpora or not, a difference which is trivial compared with the differences among the subcorpora.

(2) Written Swedish. Ellegård (1978:62) quotes data from an analysis of a million words of Swedish newspaper texts (editorial and reportage) carried out by Sture Allén. He notes 'an almost uncanny similarity' between these figures and the figures which he himself had found in part of the Brown corpus (p. 77).

(3) New Testament Greek. These figures are based on the morphologically parsed UBS text of the entire Greek New Testament.[6] It is interesting to notice how the Brown-LOB difference between informational and imaginative supergenres is parallelled by the difference between the New Testament letters and narrative (the Gospels and Acts). There are far fewer nouns than in written English, which suggests that the 37% norm for writing may vary from language to language.

(4) Written Welsh. This is a tiny corpus of a mere 1500 words of modern written Welsh—500 words each from a novel, a newspaper comment, and a literary review, provided by Aled Jones.

(5) Spoken English. My main source of data on spoken English is the appendix of Biber 1988, which includes figures for the London-Lund corpus, about half a million words of spoken British English (Svartvik & Quirk 1980). The syntactic variables listed include most of the word-classes, but unfortunately Biber's nominal categories are hard to align with those which I used. His figures are consistently several percent lower than mine for the LOB genres, so I assume that he excluded some categories which I included. The missing figures would be enough to bring the total nouns for spoken English up to the same level as for written English, about 37%. This suspicion is confirmed by some extremely crude data reported in French et al. 1930, from a study in which a team of shorthand typists bugged conversations in a public telephone exchange! It is interesting to see how the subcorpora fall into two groups according to whether they are high on prepositions and common nouns or on pronouns and verbs. This difference corresponds to the difference between scripted and unscripted speech.

[6] I received these figures from James Tauber (personal communication, 1993, and see now Tauber 1994). Tauber used a parsed text prepared by the University of Pennsylvania's Center for the Computer Analysis of Text; that morphologically analyzed text is based in turn on a 1975 United Bible Society (UBS) text and a 1981 analysis by Barbara Friberg & Timothy Friberg.

(6) CHILDREN'S ENGLISH. These data are from the Polytechnic of Wales corpus (Fawcett & Perkins 1980).[7] This is a 70,000-word collection of speech produced under controlled conditions by 120 children of 4 ages, both sexes, and 5 social classes in South Wales. (All the children were monolingual English speakers.) Some of the speech was produced by groups of 3 children building a house of Lego, and some came from interviews. The figures amply justify the careful design of the corpus, as they allow a comparison of the effects of age, sex, social class, and situation (play or interview). The main influence is situation, which produces enormous differences that apply absolutely consistently to every class, sex, and age: interviews produce more nouns overall, and in particular more common and proper nouns, while play produces more verbs and pronouns. Sex and age both have some effect as well, completely independently of other distinctions: females use more pronouns and fewer common nouns, and the proportion of all nouns (especially pronouns) increases steadily with age. Oddly enough, the figures for adults are more similar to those for 6-year-olds than to those for 12-year-olds. Social class shows no consistent influence, but 8- and 10-year-olds from the lowest class consistently used the highest proportion of nouns regardless of sex and situation.

In conclusion, there seem to be regularities in language of which most of us have been completely unaware—regularities which involve the statistical probability of any randomly selected word belonging to a particular word-class. At present we cannot explain these regularities, but they are a challenge that our grandchildren may (possibly) be able to meet.

REFERENCES

BIBER, DOUGLAS. 1988. Variation across speech and writing. 1988. Cambridge: Cambridge University Press.

——. 1989. A typology of English texts. Linguistics 27.3–43.

CARROLL, JOHN B. 1960. Vectors of prose style. In Sebeok, 283–92.

DEWAELE, JEAN-MARC. 1992. How to measure formality of speech? A model of synchronic variation. Proceedings of the second conference of the European Second Language Association, Jyvaskyla, Finland.

ELLEGÅRD, ALVAR. 1978. The syntactic structure of English texts. A computer-based study of four kinds of text in the Brown University corpus. Gothenburg: Acta Universitatis Gothoburgensis.

FAWCETT, ROBIN, and MICHAEL PERKINS. 1980. Child language transcripts 6–12 (4 vols.). Pontypridd: Dept of Behavioural and Communication Studies, Polytechnic of Wales.

FRANCIS, W. NELSON, and HENRY KUČERA. 1982. Frequency analysis of English usage: Lexicon and grammar. Boston: Houghton Mifflin.

FRENCH, NORMAN; CHARLES CARTER; and WALTER KOENIG. 1930. The words and sounds of telephone conversations. The Bell System Technical Journal 9.290–324.

JOHANSSON, STIG, and KNUT HOFLAND. 1989. Frequency analysis of English vocabulary and grammar, based on the LOB Corpus, I: Tag frequencies and word frequencies. Oxford: Clarendon Press.

LABOV, WILLIAM. 1972. Language in the inner city. Philadelphia: University of Pennsylvania Press.

[7] The figures quoted by age are derived from the tagged corpus provided to me by Clive Souter, and the others were extracted specially for me by Mike Day and Robin Fawcett.

MILROY, LESLEY. 1987. Observing and analysing natural language. Oxford: Blackwell.

SANKOFF, DAVID (ed.) 1978. Linguistic variation: Models and methods. New York: Academic Press.

SEBEOK, THOMAS A. (ed.) 1960. Style in language. Cambridge, MA: MIT Press.

SVARTVIK, JAN, and RANDOLPH QUIRK (eds.) 1980. A corpus of English conversation. Lund: Gleerup.

TAUBER, JAMES K. 1994. New Testament statistical data. Perth: University of Western Australia, MS.

WELLS, RULON. 1960. Nominal and verbal style. In Sebeok, 213–20.

Department of Phonetics and Linguistics
University College London
Gower Street
London WC1E 6BT, UK
[uclyrah@ucl.ac.uk]

[Received 30 June 1993;
accepted 17 October 1993;
final version received 10 January 1994.]