# Grammatical word class variation within the British National Corpus Sampler

*Paul Rayson, Andrew Wilson and Geoffrey Leech*

UCREL, Lancaster University

## Abstract

*This paper examines the relationship between part-of-speech frequencies and text typology in the British National Corpus Sampler. Four pairwise comparisons of part-of-speech frequencies were made: written language vs. spoken language; informative writing vs. imaginative writing; conversational speech vs. 'task-oriented' speech; and imaginative writing vs. 'task-oriented' speech. The following variation gradient was hypothesized: conversation – task-oriented speech – imaginative writing – informative writing; however, the actual progression was: conversation – imaginative writing – task-oriented speech – informative writing. It thus seems that genre and medium interact in a more complex way than originally hypothesized. However, this conclusion has been made on the basis of broad, pre-existing text types within the BNC, and, in future, the internal structure of these text types may need to be addressed.*

## 1.     Introduction

In this paper, we present some of the findings that have emerged from our frequency study of the British National Corpus (BNC). Over the past year or so, we have been engaged in a substantial analysis of both lexical items and part-of-speech categories in the corpus. Our results have recently been published in the form of a frequency dictionary – *Word frequencies in written and spoken English: based on the British National Corpus* (Leech, Rayson and Wilson, 2001) – which also includes brief discussions of selected word groups as well as reproducing the frequency lists themselves.

Chapter 6 of Leech, Rayson and Wilson (2001) presents frequency lists of part-of-speech categories in a subsample of the British National Corpus known as the BNC Sampler. These lists show to what extent the two different mediums (speech and writing) and four different genres (conversation, task-oriented speech, imaginative writing, and informative writing) that are represented in the corpus vary according to their preferences for employing different parts of speech. This paper provides a study and interpretation of those lists.

## 2.     The tagged BNC Sampler

The BNC as a whole contains approximately one hundred million words of running text (both written and spoken), which has also undergone part-of-speech tagging using the CLAWS suite of programs (see, e.g., Leech, Garside and Bryant 1994). Despite advances in tagging technology, including the use of a Template Tagger to correct

common errors (Fligelstone, Rayson and Smith 1996), the tagging still has an error rate in the region of 2% and therefore, in order to reach 100% accuracy, texts still need to be manually postedited. However, for reasons of scale, it has not been possible to hand-correct the entire tagged BNC. Hence, in order to provide users with a fully accurate tagged corpus, a 2% sample of the entire BNC has been manually postedited. This smaller corpus is known as the BNC Sampler.

In order to achieve maximal accuracy without hand editing, the entire tagged BNC makes use of a set of part-of-speech categories – the C5 tagset – which contains fewer detailed distinctions than previous tagsets that have been used by CLAWS. Also, preferring ambiguity to inaccuracy, it makes use of so-called 'portmanteau tags': these are assigned where there is a very small difference in the likelihood of two possible tags being correct, and in such cases both possible tags are assigned together – for example, past tense *and* past participle of a verb. In contrast to the entire tagged BNC, the tagged BNC Sampler makes use of the most detailed part-of-speech tagset used by CLAWS – the C7 tagset – and does not need to make use of portmanteau tags, since it has been manually edited for full accuracy.

The BNC Sampler, therefore, makes a more reliable data set for an analysis of variation in part-of-speech usage than does the BNC as a whole. Although it is substantially smaller than the entire BNC, this does not necessarily mean that it is less representative for the task in hand. As shown by Biber (1993), the size of corpus needed to be representative is closely linked to the frequency of the item(s) under examination. Since part-of-speech categories are much more frequent than lexical items (as each contains many lexical items) it is possible to work reliably with a much smaller corpus.

## 3.     Methodology

Given two subcorpora of the BNC sampler we wished to compare, we produced a part-of-speech (POS) frequency list for each subcorpus. For each tag in the two frequency lists we calculated the log-likelihood (LL) statistic recommended by Dunning (1993). This was performed by constructing a contingency table as in Table 1.

**Table 1:** Contingency table for tag frequencies

|                       | SUBCORPUS ONE | SUBCORPUS TWO | TOTAL   |
|-----------------------|---------------|---------------|---------|
| Frequency of tag      | a             | b             | a+b     |
| Frequency of other tags | c-a         | d-b           | c+d-a-b |
| TOTAL                 | c             | d             | c+d     |

Note that the value 'c' corresponds to the number of tags (or words) in subcorpus one, and 'd' corresponds to the number of tags (or words) in subcorpus two (N values). The values 'a' and 'b' are called the observed values (O). We then calculated the expected values (E) according to the following formula:

$$E_i = \frac{N_i \sum_i O_i}{\sum_i N_i}$$

In our case N1 = c, and N2 = d. So, for this tag, E1 = c*(a+b) / (c+d) and E2 = d*(a+b) / (c+d). The calculation for the expected values took account of the size of the two subcorpora, so we did not need to normalise the figures before applying the formula. We then calculated the log-likehood value according to this formula:

$$-2\ln\lambda = 2\sum_i O_i \ln\left(\frac{O_i}{E_i}\right)$$

This equated to calculating LL as follows: LL = 2*((a*ln (a/E1)) + (b*ln (b/E2)))

The tag frequency list was then sorted by the resulting LL values. This gave the effect of placing the largest LL value at the top of the list, representing the tag which had the most significant relative frequency difference between the two subcorpora. In this way, we can see the tags most indicative (or characteristic) of one subcorpus, as compared to the other subcorpus, at the top of the list. The tags which appeared with roughly similar relative frequencies in the two subcorpora appear lower down the list. Given the non-random nature of tags in a text, we are always likely to find frequencies of tags which differ across any two texts, and the higher the frequencies, the more information the statistical test has to work with. For a 2x2 contingency table, the minimum significant LL value is 3.8 (for p<0.05 and 1 d.f.).

An example of the result from this technique can be seen in Table 2 which shows the top 20 tags with the highest LL value for the comparison of the spoken subcorpus with the written subcorpus of the BNC sampler.

The frequencies shown are per million words and rounded to the nearest whole number. It is possible that a tag has a high frequency not because it is widely represented in the language as a whole but because it is 'overused' in a much smaller number of texts, or parts of texts, within the subcorpora. Moreover, this 'overuse' may be due to some factor which was not controlled during the selection of samples for the corpus: for example, the selection of a leisure book about fly-fishing rather than hang-gliding. Important additional information, range and dispersion, was therefore calculated for occurrences in each subcorpus. These showed how widely spread the occurrence of a tag is: whether it is frequent because it occurs in a lot of text samples in the corpus or whether it is frequent because of a very high usage in only a few samples. Frequent tags with high dispersion values may be considered to have high currency in the language as a whole; high frequencies associated with low dispersion values should, in contrast, be treated with caution.

**Table 2:** LL comparison of spoken and written sampler subcorpora

| POS | Spoken Sampler | | | O/U | LL | Written Sampler | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| | **Freq** | **Ra** | **Disp** | | **LL** | **Freq** | **Ra** | **Disp** |
| UH | 31705 | 50 | 0.92 | + | 35177 | 1210 | 37 | 0.74 |
| FU | 23010 | 50 | 0.93 | + | 29404 | 252 | 26 | 0.67 |
| NP1 | 14469 | 50 | 0.94 | - | 19819 | 49089 | 50 | 0.92 |
| PPIS1 | 31379 | 50 | 0.95 | + | 17304 | 6701 | 48 | 0.82 |

| POS | Spoken Sampler | | | O/U | LL | Written Sampler | | |
|---|---|---|---|---|---|---|---|---|
| | Freq | Ra | Disp | | | Freq | Ra | Disp |
| PPY | 26531 | 50 | 0.96 | + | 16894 | 4656 | 48 | 0.81 |
| NN1 | 89193 | 50 | 0.96 | - | 16762 | 152639 | 50 | 0.98 |
| JJ | 36979 | 50 | 0.97 | - | 13395 | 75496 | 50 | 0.97 |
| NN2 | 25432 | 50 | 0.94 | - | 12373 | 57048 | 50 | 0.96 |
| PPH1 | 25899 | 50 | 0.96 | + | 9462 | 8308 | 50 | 0.94 |
| VV0 | 28729 | 50 | 0.97 | + | 8037 | 11113 | 50 | 0.91 |
| AT | 39031 | 50 | 0.95 | - | 7997 | 68200 | 50 | 0.98 |
| IO | 13406 | 50 | 0.93 | - | 7156 | 31039 | 50 | 0.95 |
| VBZ | 26524 | 50 | 0.97 | + | 6472 | 11117 | 50 | 0.95 |
| DD1 | 20945 | 50 | 0.98 | + | 6215 | 7802 | 50 | 0.96 |
| II | 44809 | 50 | 0.97 | - | 6200 | 71619 | 50 | 0.98 |
| XX | 18067 | 50 | 0.95 | + | 6127 | 6135 | 50 | 0.91 |
| VD0 | 7395 | 50 | 0.94 | + | 4875 | 1230 | 50 | 0.83 |
| PPIS2 | 10880 | 50 | 0.93 | + | 4292 | 3277 | 49 | 0.84 |
| VH0 | 9180 | 50 | 0.97 | + | 3356 | 2944 | 50 | 0.94 |
| RR | 42555 | 50 | 0.98 | + | 2928 | 28183 | 50 | 0.97 |

Each subcorpus was divided into 50 roughly equal sized segments, each of about 20,000 words. Sometimes text files were split across two segments. The range and dispersion statistics reflect occurrences of tags in these segments.

Range (Ra) is a simple count of how many segments include the tag in question. Dispersion (Disp) is a statistical coefficient (Juilland's D) of how evenly distributed a tag is across successive segments of the corpus. This is useful, because many segments and texts are made up of a number of smaller, relatively independent units – for example, sectors and stories in newspapers. It may be that, even *within* a text, certain word classes are overused in a given part – e.g. the football-reporting sector of a newspaper. Juilland's D is more sensitive to this degree of variation. It was calculated as follows:

$$D = 1 - \frac{V}{\sqrt{n-1}}$$

where *n* is the number of segments in the subcorpus. The variation coefficient *V* is given by:

$$V = \frac{s}{x}$$

where *x* is the mean sub-frequency of the tag in the subcorpus (i.e. its frequency in each segment averaged) and *s* is the standard deviation of these sub-frequencies. We selected Juilland's D as it has been shown by Lyne (1985) to be the most reliable of the various dispersion coefficients that are available. It varies between 0 and 1, where values closer to 0 show that occurrences are focussed in a small number of segments, and values closer to 1 show a more even distribution across all segments.

The other column in the table (O/U) shows overuse (+) or underuse (-) of a tag in the spoken subcorpus relative to the written subcorpus.

The figures for range and dispersion are not quoted in the paper but are taken into account in our studies of variation.

## 4.     Results

We made four pairwise comparisons of part-of-speech frequencies in the tagged BNC Sampler:

- written language vs. spoken language
- informative writing vs. imaginative writing
- conversational speech vs. 'task-oriented' speech
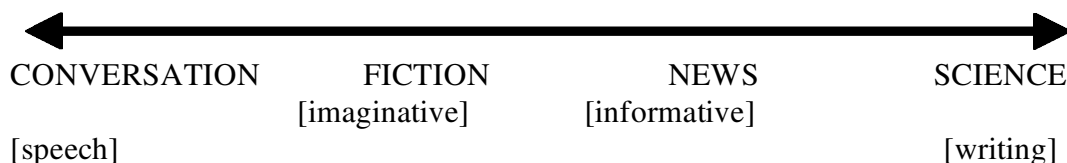- imaginative writing vs. 'task-oriented' speech

As a result of previous studies on both English and other languages, we had strong hypotheses in regard to the first two comparisons.

Hoffmann (1985: 137), for example, presented analyses of various genres of Russian writing, including a sample of imaginative prose and several scientific genres. His figures showed that the proportion of nouns, adjectives and prepositions tended to be higher in the scientific genres, whereas verbs, adverbs and pronouns were more frequent in imaginative prose.

Working on English, Nakamura (1991) used Hayashi's Quantification Method Type III (a procedure somewhat similar to factor analysis) to analyse the tag frequencies in the LOB corpus according to the 15 genre categories within it. The first, and major, axis of his results can be interpreted as a dimension distinguishing the imaginative from the informative genres, with informative genres such as newspaper discourse located closer to the mid-point of that axis than learned and scientific writing. Some of the most characteristic major parts of speech were again nouns, adjectives and prepositions (for the informative genres) and verbs, adverbs and pronouns (for the imaginative genres).

To Nakamura's study can also be added the detailed corpus-based study by Biber et al. (1999). Working with a larger corpus of more recent English, Biber et al.'s findings again mirrored studies such as Nakamura's: nouns, adjectives and prepositions were shown to be more frequent in informative genres, and verbs, pronouns, and adverbs more frequent in imaginative genres. Moreover, Biber et al. also found a similar gradient to Nakamura, with news writing taking an intermediate position between scientific writing and fiction on the informative-imaginative scale. However, an important advance of Biber et al.'s study, made possible by the composition of their corpus, was the inclusion of conversational speech. Compared with informative writing, this showed a part-of-speech frequency profile somewhat similar to that of imaginative writing, but, when compared with imaginative writing, the trends were magnified: for instance, pronouns were yet more frequent in conversation than in fiction.

On the basis of these previous studies, therefore, we might hypothesise that there exists a part-of-speech variation gradient showing a gradual progression from informal, conversational speech at one extreme through more formal speech towards increasingly formal written genres at the other extreme:

◄──────────────────────────────────────────────────►

| CONVERSATION | FICTION | NEWS | SCIENCE |
|---|---|---|---|
| | [imaginative] | [informative] | |

[speech]                                                            [writing]

One comparison that has not yet been made concerns what the BNC compilers termed context-governed speech. (We have chosen in our work to call this task-oriented speech.) This is speech which is closely tied to particular activity types: for example, medical consultations, committee meetings, sermons, lectures, court proceedings, etc. On the whole, we might characterise such speech as more public, and hence, probably, more formal, although in many instances it will still involve substantial amounts of relatively spontaneous speaking as well as prepared reading.

If our hypothesis about a single variation gradient is correct, we would expect task-oriented speech to be positioned somewhere between conversational speech and imaginative writing. We therefore made two comparisons using task-oriented speech. Firstly, we compared it with conversational speech, to ascertain whether task-oriented speech appeared to differ from conversation in a similar direction to that in which writing differs from speech overall; and, secondly, we compared it with imaginative writing, to see whether it differed from that genre in the same direction as speech when compared with writing. Our detailed findings on the major part-of-speech categories for each of our four comparisons are discussed in the following paragraphs. Table 3 summarises these findings by showing overused POS categories in each of the four comparisons.

**Table 3:** Overused POS categories within four comparisons

| | Wri. | Spoken | Inf. | Imag. | Conv | Task O. | Imag. | Task O. |
|---|---|---|---|---|---|---|---|---|
| **Conj.** | CC | CS but | CC | CS but | CS but | CC | as than | CC CS |
| **Noun** | N* | | N* | | NP | N* NPM* | NN1 | NN2 |
| **Verb** | VV* | VM aux V VVGK | VVN | VM VV* | V* | | VVD | V* VVN |
| **Pron.** | PNQ* PNX1 | P* | PNX1 | P* | P* | we us | he/she him/her | they |
| **Adverb** | RRR RRT | R* | RRR RRT REX | R* | RR RL RT | R* | RL | R* |
| **Adject.** | J* | | J* | | | J* | J* | |
| **Prep.** | I* | | I* | | | I* | I* IW | IF IO |
| **Article Determ.** | A* | DA1 DB* | A* | DA1 DB* | DA1 DB* | A* | A* | DA* |

## 4.1 Conjunctions

In terms of the 'basic' categories of coordinating conjunction (mainly representing *and*) and subordinating conjunction, the significant differences between spoken and written language were in the opposite direction to that predicted by early studies on medium variation (e.g. Blankenship 1962 and Kroll 1977): spoken language contained significantly more subordinating conjunctions than written language (with the exception of *that*), and written language contained significantly more coordinating conjunctions than speech. This, however, corroborated the more recent findings of Biber et al. (1999: 81). Biber et al. also found that *but*, in contrast to *and*, was favoured by speech more than writing, and this was again corroborated by our study (*but* being a special category of coordinating conjunction in the C7 tagset). Furthermore, again as found by Biber et al., coordinating conjunctions were more frequent in informative (as opposed to imaginative) writing, with the exception of *but*, which was more common in imaginative writing. A reverse trend in the case of subordinators was again evident. In comparing conversational and task-oriented speech, the pattern was similar to that between speech and writing: conversation tended to have more subordinators and task-oriented speech tended to have more coordinators (again apart from *but*). However, a different pattern emerged in comparing task-oriented speech with imaginative writing: most conjunctions (both coordinating *and* subordinating) showed a preference for task-oriented speech, with the exception of the subordinators *as* and *than*. The coordinator *but* did not show a statistically significant difference.

## 4.2 Nouns

As found by previous studies, nouns tended, on the whole, to be more frequent in writing (as opposed to speech) and informative (as opposed to imaginative) writing. Our study also found them to be more common in task-oriented (as opposed to conversational) speech, although, within the class of nouns, conversation had more proper nouns than task-oriented speech (the exception being the names of months). The comparison between imaginative writing and task-oriented speech was, however, more ambivalent. Around half of the noun tags showing significant differences (i.e., 9 out of 16) showed a preference for task-oriented speech and the other half a preference for imaginative writing. This included a clear split within the class of common nouns: singular common nouns were preferred in imaginative writing and plural common nouns in task-oriented speech.

## 4.3 Verbs

Our findings on verbs again broadly supported previous studies. In the comparison of imaginative and informative writing, the modal verbs, and most forms of lexical verbs, were more common in imaginative writing. The exception here in the case of lexical verbs was the past participle (VVN), which was commoner in informative writing. This finding almost certainly reflects a greater use of the passive in informative genres (cf. Biber et al. 1999: 477; Svartvik 1966). Other verb forms preferred by informative writing were: *being*, *be* (as infinitive), *been* (again a past participle), *are*, and *has*. The

comparison of conversation and task-oriented speech presented a very similar picture, with the majority of verbal forms showing a preference for conversation. The exceptions list was likewise similar to the previous one: it contained the past participle of lexical verbs, together with *were*, *being*, *be* (infinitive), *been*, and *are*. The overall comparison of speech and writing was rather more complicated. The modal verbs, and most of the auxiliary verbs (though not necessarily in an auxiliary function), were preferred in speech, the exceptions being *had* (past tense), *having*; *be* (finite base form), *were*, *was*, *being*, *be* (infinitive), and *been*. Speech also preferred the base and infinitive forms of lexical verbs, as well as the catenative *-ing* participle (e.g. *going to*). Other forms of lexical verb, however, tended to be more frequent in writing. The comparison of task-oriented speech and imaginative writing was again complex. Most of the verb forms listed above were frequent in task-oriented speech. However, there were a number of exceptions, which included the simple past-tense forms of both lexical verbs and auxiliaries as well as the present and past participles of *be*, the present participle of lexical verbs, and the third person singular of lexical verbs. Interestingly, the past participle of lexical verbs, which was preferred in writing, was, in this comparison, preferred in the spoken text type.

### 4.4    Pronouns

Again confirming the findings of previous studies, pronouns were generally more common in speech, conversational (as opposed to task-oriented) speech, and imaginative writing (when contrasted with informative writing). There were, however, some exceptions to this general trend. In the comparison of speech and writing, the WH (i.e. relative and interrogative) pronouns were more common in writing, as were the reflexives. The same exceptions applied in the comparison of conversational versus task-oriented speech, where, additionally, the first-person plural pronouns (*we/us*) were more common in task-oriented speech than in conversation, probably reflecting more speech on behalf of groups and organisations rather than on behalf of the individual speaker. In the comparison of imaginative and informative writing, the only exception was the reflexive indefinite pronoun (*oneself*). But the comparison of task-oriented speech and imaginative writing once again introduced a more mixed picture, and, as with nouns, there was an approximately 50:50 split in the direction of preference for those pronouns showing a significant difference. For example, *he* and *she* were preferred by imaginative writing, whereas *they* was preferred in task-oriented speech. However, although *him* and *her* were also strongly preferred in imaginative writing, *them* was not equally preferred in task-oriented speech (the preference was in the direction of imaginative writing, but fell just short of statistical significance).

### 4.5    Adverbs

Confirming once again the previous studies cited above, in the comparison of spoken and written language, most adverbs showing a significant difference were more frequent in the spoken language. The exceptions were the comparative and superlative forms, which were all significantly more frequent in written language, as well as those introducing appositive constructions (such as *e.g.* or *namely*). The picture was broadly

similar in the comparison of imaginative and informative writing (the adverbs being more common in imaginative writing); the comparatives, superlatives and appositives were again exceptions, although the comparative general adverb was preferred in imaginative writing. In the comparison of conversation and task-oriented speech, the trend was towards task-oriented speech. The exceptions in this case were the base form general adverb, the locative and temporal adverbs, and the particle: this suggests a greater use of spatial and temporal deixis in conversation, as well as more phrasal verbs, both hypotheses being supported by Biber et al.'s (1999) detailed grammatical analysis. In the comparison of task-oriented speech and imaginative writing, most adverbs were preferred in the spoken genre, apart from locative adverbs, particles, superlative degree adverbs, catenative prepositional adverbs, and comparative general adverbs. In this case, then, the appositives, preferred in written language and informative writing, were preferred in task-oriented speech.

## 4.6    Adjectives

In all comparisons, adjectives showed straightforward and expected preferences. All types of adjective were significantly more common in, respectively, writing, informative (as opposed to imaginative) writing, and task-oriented (as opposed to conversational) speech. The only exception was a slight preference for the catenative adjective (e.g. *able to/willing to*) in speech as opposed to writing, which, however, was not significant at the p < 0.05 level (LL = 1.2). For once, the comparison of task-oriented speech and imaginative writing straightforwardly followed the hypothesised direction: all adjectives (again with the exception of catenative) were preferred in the written genre.

## 4.7    Prepositions

In three of the comparisons, the usage patterns of prepositions supported previous findings: they were more frequent in writing, task-oriented speech (as opposed to conversation), and informative (as opposed to imaginative) writing. However, there was again a 50:50 split in the comparison of imaginative writing (which showed a greater use of *with/without* and general prepositions) and task-oriented speech (which showed a greater use of *of* and *for*). The preference for *of* in task-oriented speech is suggestive of more complex, postmodified noun phrases.
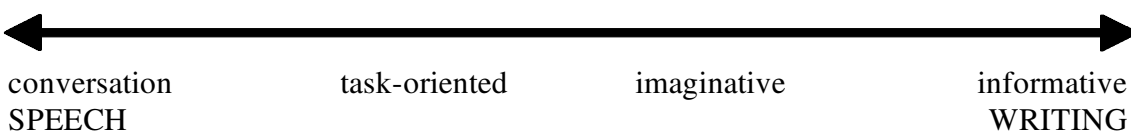
## 4.8    Articles and determiners.

As expected, articles showed a preference for writing, informative (as opposed to imaginative) writing, and task-oriented (as opposed to conversational) speech: this is obviously tied to the higher frequency of nouns in these genres. Determiners were somewhat harder to interpret, as in CLAWS tagging many are ambiguous between pronominal and determiner functions. However, the singular determiner and the pre-determiner certainly showed preferences in the same direction as pronouns (i.e. towards speech, imaginative writing, and conversation). In the comparison of task-

oriented speech and imaginative writing, the singular article and pre-determiner did not show significant differences. The general article showed a preference for imaginative writing, and the singular and plural determiners both showed a preference for task-oriented speech. In this case, therefore, the trends were approximately as hypothesised, with the written genre behaving like writing and the spoken genre behaving like speech.

To summarise, then, these findings, like those of Biber et al. (1999), show a close correlation between the preferences on three of the dimensions of variation. Thus, the parts of speech that are more common in spoken language are also more common in conversation and in imaginative writing, whereas those parts of speech that are more common in writing are also more common in task-oriented speech and in informative writing. However, the findings also clearly refute the hypothesis that both genre variation (e.g., imaginative vs. informative) and medium variation (i.e., speech vs. writing) can be represented on a single gradient going from conversational speech to informative writing. This is because task-oriented speech, when compared with imaginative writing, has an unstable pattern of preferences: sometimes it shows more similarity to written language and informative writing than it does to spoken language and conversation, and sometimes the reverse is the case. For example, with adjectives, we get the following gradient:
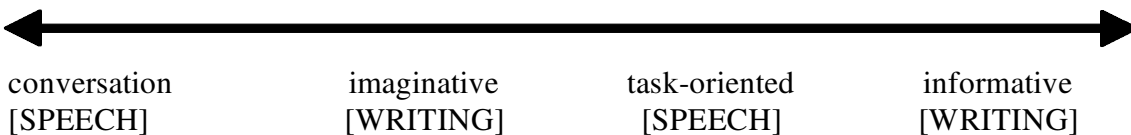
LEAST ADJECTIVES                                          MOST ADJECTIVES

⬅️➡️

conversation            task-oriented           imaginative            informative
SPEECH                                                                 WRITING

but with certain key part-of-speech subcategories (e.g. past participles) we get a gradient like the following:

LEAST PAST PARTICIPLES                      MOST PAST PARTICIPLES

⬅️➡️

conversation            imaginative           task-oriented           informative
[SPEECH]                [WRITING]              [SPEECH]                [WRITING]

and, in several cases (such as with prepositions), no overall direction of preference can be ascertained.


## 5.    Conclusion

Our results suggest that genre and medium may interact in a much more complex way than hypothesised, and particularly that certain spoken genres may show a greater trend towards the overall norms of written language than do some of the actual written genres.

However, it has to be emphasised that this conclusion has been made on the basis of rather broad, pre-existing genre categories within the BNC, and it is possible

that we need, in future studies, to re-think the way in which some text categories are made up and used in medium and genre variation studies. For instance, one reason why imaginative writing turns out to be more speech-like than other written genres is almost certainly that it contains a higher proportion of direct speech quotation. Arguably, therefore, we are not strictly comparing speech and writing, but, at least in part, speech and speech (albeit invented speech). Nevertheless, dialogue is an integral part of fictional prose and its inclusion is therefore valid.

In contrast, some text categories may involve artificial inconsistencies. In particular, it may be that the category of task-oriented (or context-governed) speech, which showed such an ambiguous position when compared with conversation and imaginative writing, is ill-defined. For example, it contains rather formal, and frequently prepared, monologues (such as sermons) as well as less formal, more spontaneous, and interactive discourse (such as meetings). It is thus quite probable that, strictly speaking, this category contains examples of written informative language that are simply, or largely, read out, as well as spontaneous discourse that is genuinely spoken language. In the same context, it is worth noting that Biber's (1988) first dimension of variation – which Lee (2000), for example, found to be the only stable and replicable dimension – is **involved** versus **informative**, not speech versus writing: even thinking introspectively, we have both primarily involved (e.g. dialogue) and primarily informative (e.g. monologue) texts represented within the task-oriented category of the BNC. Thus, if we were to separate out, for instance, the dialogue and monologue components of the category, it is conceivable that our initial hypothesis of a single gradient of variation could still hold.

## References

Biber, Douglas (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press

Biber, Douglas (1993). 'Representativeness in corpus design'. *Literary and Linguistic Computing* 8: 243–57.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad, Edward Finegan (1999). *Longman grammar of spoken and written English*. London: Longman.

Blankenship, Jane (1962). 'A linguistic analysis of oral and written style'. *The Quarterly Journal of Speech* 48: 419–422.

Dunning, Ted. (1993). 'Accurate methods for the statistics of surprise and coincidence'. *Computational Linguistics*, 19.1, March 1993: 61–74.

Fligelstone, Steven, Paul Rayson and Nicholas Smith (1996). 'Template analysis: bridging the gap between grammar and the lexicon'. In Jenny Thomas and Mick Short (eds), *Using corpora for language research*. London: Longman. 181–207

Hoffmann, Lothar (1985). *Kommunikationsmittel Fachsprache: eine Einführung*. 2nd ed. Tübingen: Günter Narr.

Kroll, Barbara (1977). 'Ways communicators encode propositions in written and spoken English. A look at subordination and coordination'. In Elinor Keenan and Tina Bennett (eds) *Discourse across time and space*. Southern California Occasional Papers in Linguistics, No. 5. Los Angeles, CA: University of Southern California. 69–108.

Lee, Yong Wey David (2000). *Modelling variation in spoken and written language: the multi-dimensional approach revisited*. Ph.D. thesis, Lancaster University.

Leech, Geoffrey, Roger Garside and Michael Bryant (1994). 'The large-scale grammatical tagging of text: Experience with the British National Corpus'. In Nelleke Oostdijk and Pieter de Haan (eds) *Corpus-based research into language*. Amsterdam: Rodopi. 47–64.

Leech, Geoffrey, Paul Rayson and Andrew Wilson (2001). *Word frequencies in written and spoken English: based on the British National Corpus*. London: Longman.

Lyne, Anthony (1985). *The vocabulary of French business correspondence*. Geneva: Slatkine.

Nakamura, Junsaku (1991). 'The relationship among genres in the LOB corpus based upon the distribution of grammatical tags'. *JACET Bulletin* 22: 44–74.

Svartvik, Jan (1966). *On voice in the English verb*. The Hague: Mouton.