More on the language of dialogue in fiction*

Pieter de Haan
University of Nijmegen

1. Introduction

My interest in the language of dialogue in fiction was first aroused when I looked into sentence length (de Haan, 1993) and noticed that the two authors of the crime novels in the Nijmegen corpus were quite different in their use of what in Quirk *et al*. (1985) is called 'reporting clauses', and what in the TOSCA analysis system (Oostdijk, 1993) is referred to as 'reporting utterances'. One of the two authors, for instance, far more often made explicit reference to somebody's speech by means of reporting clauses than the other. The same author also used far more different verbs in these reporting utterances than the other. One of the questions that these observations naturally led to was whether they point to a different attitude of the author to representing dialogue in fiction.

Variation in the use of reporting utterances had also been noticed by Oostdijk (1990), who studied five fiction texts taken from the TOSCA corpus. Oostdijk points out that dialogue in fiction has its own characteristics, which makes it like, but at the same time also different from, spoken language. It is like spoken language, as it clearly reflects the author's attempt to represent spoken conversation. It is clearly unlike spoken language as it is planned, revised, and edited. Oostdijk assumes a continuous scale with planned discourse at one extreme and unplanned discourse at the other, suggesting that dialogue in fiction is found somewhere between these extremes.

Oostdijk (1990) presents an overall picture of five fiction samples from the TOSCA corpus, but suggests that a number of observations point to the existence of idiosyncratic variation among authors. I became more interested in the question whether it is possible to present a coherent picture of this variation, while at the same time characterizing the language of dialogue in fiction more generally. In this article I will restrict myself to a discussion of some preliminary findings and I will try to indicate in which direction further research might be conducted.

It should be borne in mind that the term 'dialogue' in this article (as in Oostdijk's) is used to include any kind of direct speech utterance.

2. The corpus used

The corpus used for the present study was composed of seven fiction texts (see the Appendix). Two texts were taken from the Nijmegen corpus (NIJM01 and NIJM02). These are both crime fiction texts. Five texts were taken from the TOSCA corpus, representing the categories crime (FCRI01), horror (FHOR01 and FHOR02), love & romance (FROM01) and humour (FHUM03). Oostdijk (1990) based her article on the study of the first four of these TOSCA texts, and on one text from the category psychological novel. The corpus used for the present study is composed of the following texts. Full bibliographical references can be found in the Appendix.

1:	NIJM01:	The Mind Readers
2:	NIJM02:	The Bloody Wood
3:	FCRI01:	Carson's Conspiracy
4:	FHOR01:	The Damnation Game
5:	FHOR02:	The Fog
6:	FROM01:	Flight from Bucharest
7:	FHUM03:	The Joke of the Century

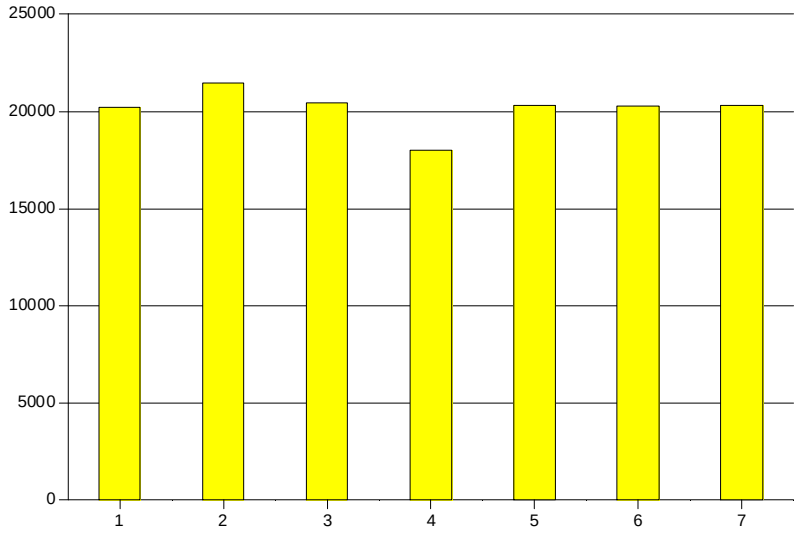

Figure 1: *Number of words per text*

Figure 1 presents the length of the seven corpus texts, expressed in the number of words. It shows that text 2 is slightly longer than the others, whereas text 4 is considerably shorter. Figure 2 presents the length of the seven corpus texts, expressed in the number of sentences.

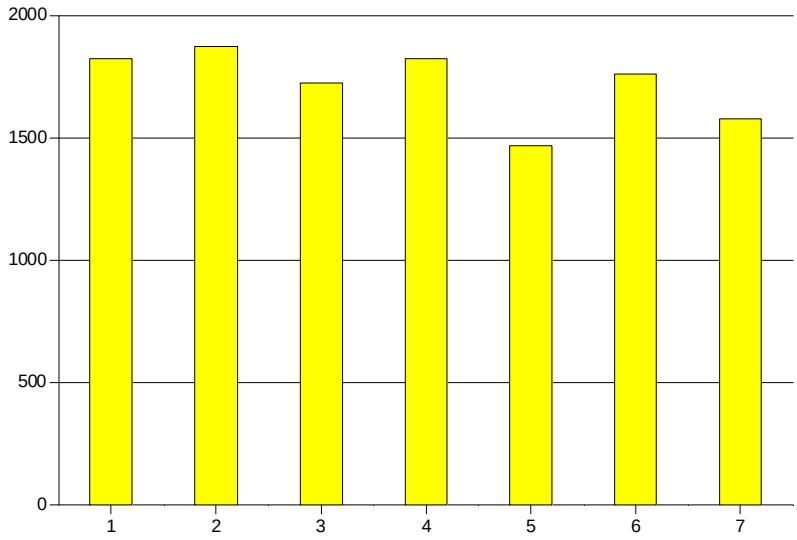

Figure 2: *Number of sentences per text*

It is remarkable to observe that text 4 has a fairly large number of sentences, given its relatively few words. This points to an overall low mean sentence length. I will return to this below. What is the dialogue component (= direct speech component) in each of these texts? As we do not know how sentence length is distributed over the various types of sentence (dialogue or non-dialogue), we must look at this question in two ways:

1 on the basis of the number of **sentences** that make up the dialogue component;

2 on the basis of the number of **words** that make up the dialogue component.

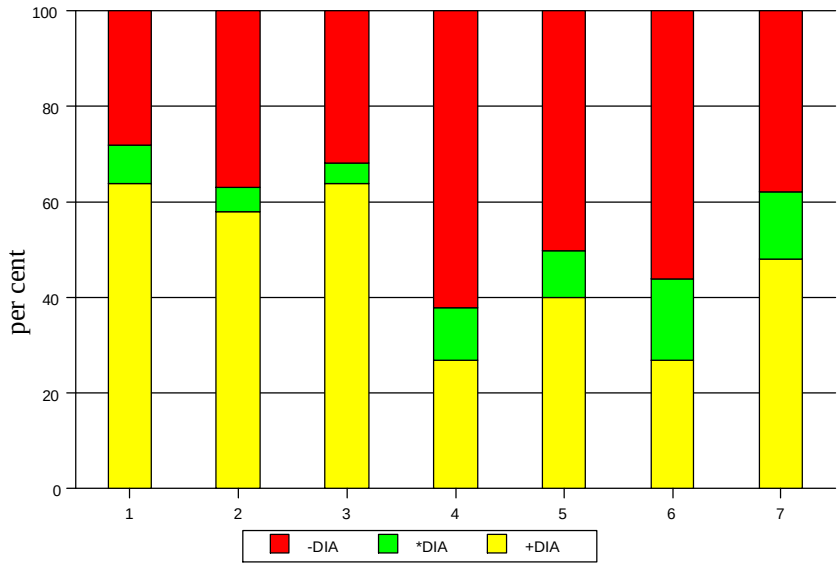

Figure 3: *Dialogue and non-dialogue sentences (proportionally)*

Figure 3 presents the proportion of dialogue and non-dialogue sentences in each of the seven corpus texts. It is, in fact, a breakdown of the numbers in Figure 2 into three categories:

1 +DIA: the sentences that contain only direct speech;

2 *DIA: the sentences that contain both direct speech (reported utterance) and an explicit reporting utterance (as in: *'It's cold', she said*);

3 –DIA: the sentences that contain no direct speech at all, and can be taken to be purely descriptive.

The figures are percentage scores, making a comparison across the seven texts possible. It is obvious that the three crime texts contain far more dialogue than any of the other four texts (roughly 60 per cent of the total number of sentences in each of the three crime texts). It also shows that, in text 6, a relatively large part is taken up by the *DIA category, pointing to a relatively large number of explicit reporting utterances.

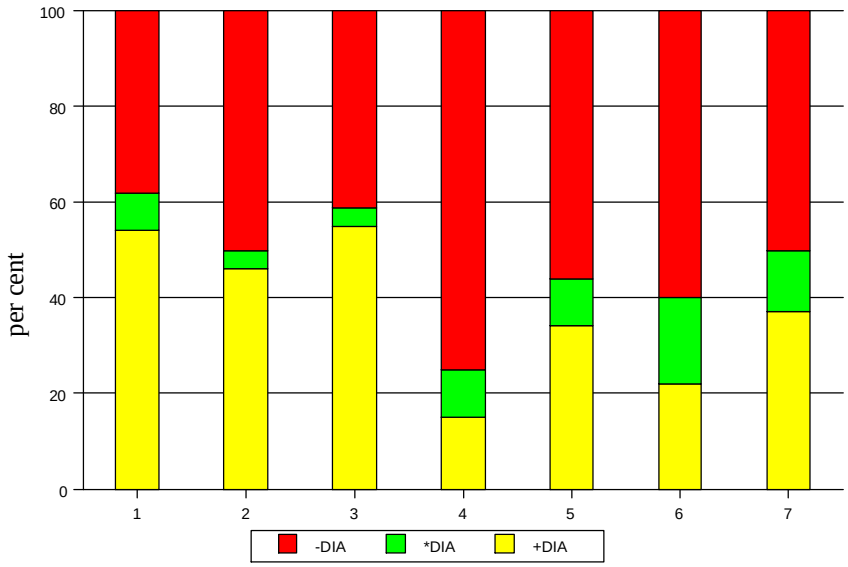

Figure 4: *Dialogue and non-dialogue words (proportionally)*

When we break down the numbers presented in Figure 1 (based on words), we get comparable results (Figure 4), although the differences between dialogue and non-dialogue are less striking for the three crime texts. Whereas in Figure 3 the dialogue component in these texts is shown to be 60 per cent or more, in Figure 4, it is 'only' 50 per cent or less. However, the difference between dialogue and non-dialogue is far more striking in text 4, where the relative part taken up by pure dialogue is now shown to be less than 20 per cent.

3. Sentence length in dialogue passages

These observations would suggest differences in sentence length, particularly between the dialogue sentences and the non-dialogue sentences. In de Haan (1992), these differences are discussed, on the basis of a detailed study of one of the two Nijmegen crime texts. It is shown that on the whole, the dialogue sentences are shorter than the non-dialogue sentences, and that, even in smaller text fragments, there is a good deal of variation in the number of dialogue sentences and, consequently, in

the mean sentence length. The overall mean sentence length in the present corpus of seven fiction texts is 11.7 words. The overall mean sentence length for each of the seven texts is shown in Figure 5:

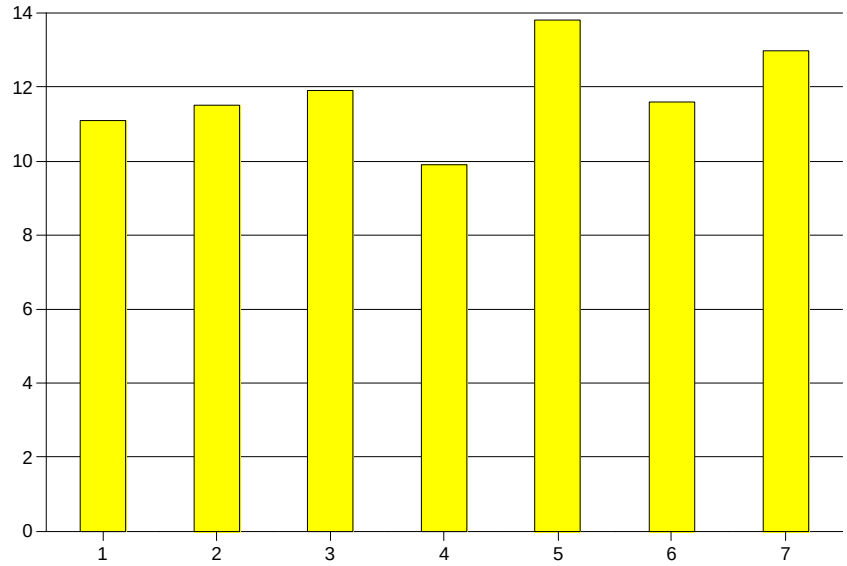

Figure 5: *Mean sentence length*

Figure 5 shows that, interestingly, the two extreme scores, the highest mean score and the lowest mean score, are found in the two horror texts. Breaking these figures down for dialogue and non-dialogue sentences, we get the picture presented in Figure 6.

A number of interesting observations can be made in Figure 6. First of all, with one exception (text 6), the general picture is that dialogue sentences are indeed the shortest on average, and that the non-dialogue sentences are the longest, with the *DIA sentences taking a mid-position.

Secondly, we see that the author of text 4 uses shorter sentences in all categories, and that the author of text 5 uses relatively long sentences in all categories. So the overall difference in mean sentence length between the two horror texts in Figure 5 points to a general characteristic and is clearly not due to a different distribution of length in the different types of sentences.

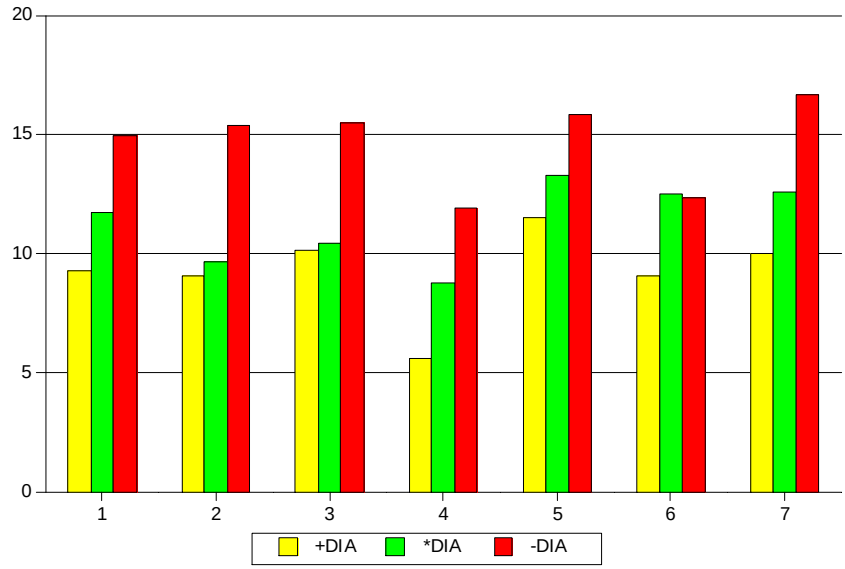

Figure 6: *Mean length of the various types of sentence*

In the third place, we see that the patterns presented in texts 2 and 3 are virtually the same. This may point to an idiosyncrasy, as texts 2 and 3 were both written by the same author (Michael Innes). None of the other authors displays the same extreme difference in mean length (more than five words) between the DIA sentences on the one hand, and the +DIA and *DIA sentences on the other. Likewise, no other author displays so much similarity in the mean length scores of the +DIA and *DIA sentences (the difference is less than one word).

The difference in mean length between the +DIA and the *DIA sentences is not entirely due to the presence of the reporting utterances in the *DIA sentences. This is shown in Figure 7, which breaks down the mean length of the *DIA sentences into the mean length of the reported utterances (in other words: the dialogue parts of these sentences) and the reporting utterances. In order to put these figures in their proper perspective, they are presented together with the figures for the +DIA and –DIA sentences.

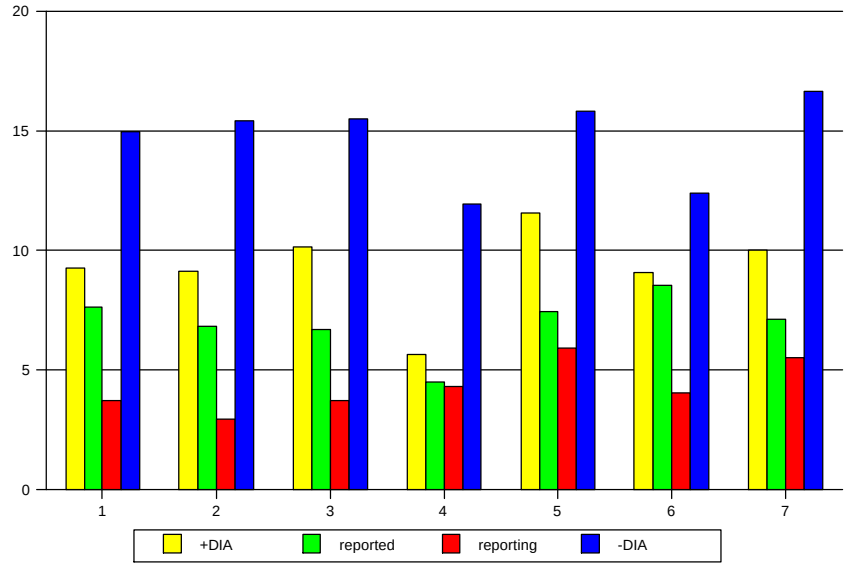

Figure 7: *Mean length of the various types of sentence*

Figure 7 clearly shows that the dialogue parts of the *DIA sentences are far shorter, on average, than the +DIA sentences. This may be due either to a balancing principle, by which authors are somehow restrained from making the difference in length between reporting and reported utterances too great, or to a more general restraining principle, which simply 'allows' authors fewer words to complete a sentence by means of a reported utterance, simply because part of the sentence is already 'reserved' for the reporting utterance. Given the fact that the *DIA sentences are on the whole shorter than the –DIA sentences (see Figure 6), this restraining principle would seem to place stricter limits, for most authors, on *DIA sentences than on –DIA sentences.

4. The use of reporting verbs

In this section I will discuss a few characteristics of the sentences with explicit reporting utterances, and then I will make a few suggestions for further research. Although I had originally taken the *DIA sentences as a separate group mainly in order to make a straightforward comparison between the other two categories possible, I soon realised that this group

deserved attention in its own right, if only because they shared the unusual feature of having a reporting utterance and a reported utterance. The first thing I looked into was the actual reporting verb used. It will not come as a surprise that the single most common verb is *said*, i.e. the past tense of the verb *say*. In almost two thirds of all the cases this was the verb used. In all, 111 different reporting verbs were used in the reporting utterances in this corpus, 56 of them occurring only once. Table 1 lists the twenty most frequent reporting verbs, in decreasing order of frequency. They all occur at least five times in this corpus. Among them, they make up slightly over 85 per cent of all the reporting verbs in the seven texts in this corpus.

Table 1: The most frequent reporting verbs in seven fiction texts

	verb	#	pct	cum.pct
1	said	743	64.4	64.4
2	asked	80	6.9	71.3
3	replied	22	1.9	73.2
4	went on	17	1.5	74.7
5	whispered	17	1.5	76.2
6	cried	15	1.3	77.5
7	murmured	15	1.3	78.8
8	added	12	1.0	79.8
9	told	10	.9	80.7
10	muttered	9	.8	81.5
11	shouted	8	.7	82.1
12	suggested	8	.7	82.8
13	explained	7	.6	83.4
14	began	6	.5	84.0
15	demanded	6	.5	84.5
16	gasped	6	.5	85.0
17	was	6	.5	85.5
18	was saying	6	.5	86.0
19	continued	5	.4	86.5
20	remarked	5	.4	86.9

It is interesting to observe that, while all the authors use the verb *said* a great deal (but see below), the next frequent verb (*asked*) is not used at all by the author of the first text (NIJM01). This is the only author who does not use this verb. At the same time, the four occurrences of the verb *inquired* are all found in this particular text. This may point to an author-specific feature.

An example of the verb *was* in a reporting utterance is found in the following sentence, taken from FHOR02.

> 'Where's Carys?' was Marty's first question.

When we look at the distribution in the seven texts we see considerable variation in the proportion of reporting utterances in which *said* is used. This is shown in Figure 8:

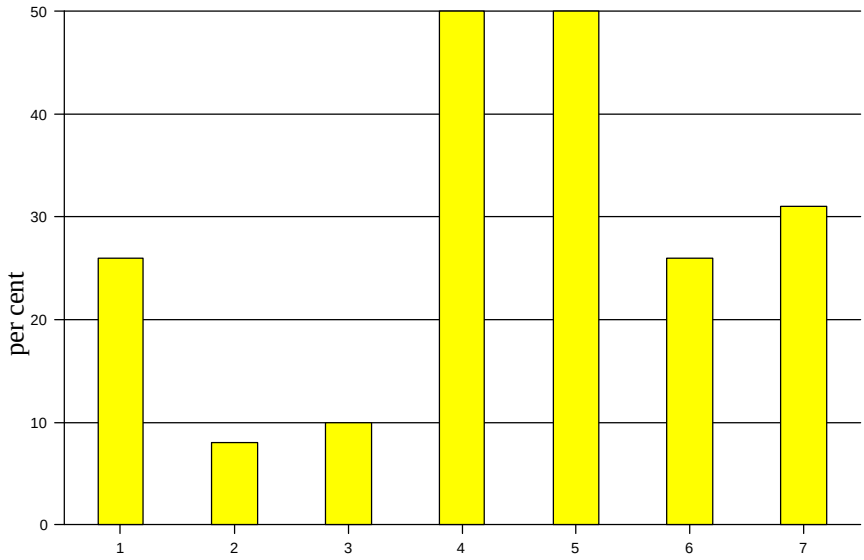

Figure 8: *Said vs. other reporting verbs*

The author of text 6 uses *said* in over 80 per cent of the reporting utterances. Now, bearing in mind that this was the author who (1) used quite a number of reporting utterances, and (2), contrary to the other

six, had *DIA sentences which were actually longer, on average, than the pure DIA sentences, the question arises whether this points to some idiosyncrasy on the part of this author with respect to precisely the *DIA sentences. That question is not so easy to answer at this point, and should be looked into in greater detail.

First consider Figure 9, which shows how many different reporting verbs each of the authors uses:

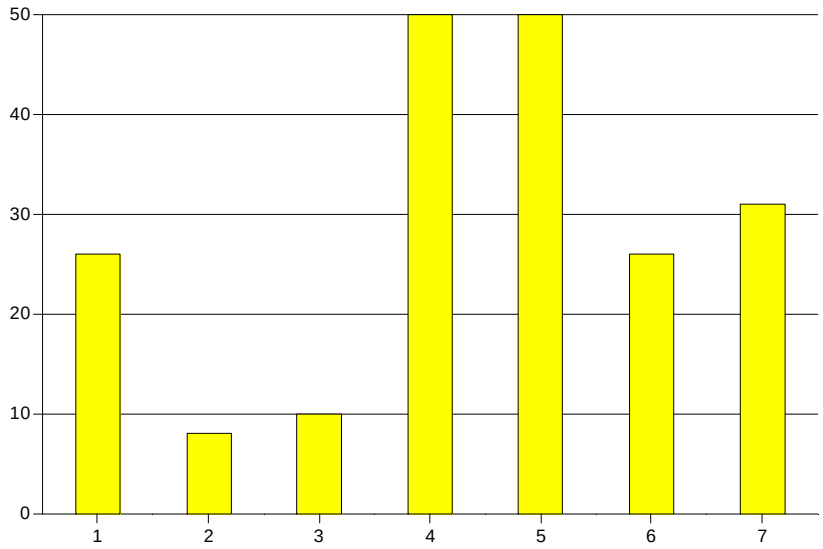

Figure 9: *Number of different reporting verbs per text*

Figure 9 shows that Michael Innes (2 and 3) uses extremely few different reporting verbs, whereas the two authors of the horror texts (4 and 5) each uses as many as 50. However, this does not say anything about the relative distribution of these 50 verbs. Many of these verbs, as was pointed out above, occur only once. And, of course, we have to make allowances for the fact that Innes uses very few explicit reporting utterances in the first place (see Figure 4). His technique of representing dialogue could be found to be quite different from that of the others. Consider the following two passages, both taken from NIJM02:

> ‘She loved the kingfishers.’ Charles Martineau’s voice was not quite under control. ‘But as we were saying last night, they have taken their departure,’
>
> ‘Fell ... is he here?’ For the moment it was the only thing Appleby found to say.
> ‘Fell?’ It was almost meaninglessly that Charles Martineau seemed to repeat the word.

These two passages illustrate a way of placing dialogue in a descriptive context, indicating who utters the reported utterances without actually using so-called reporting utterances. In order to get a complete picture, it is necessary to examine all the references to dialogue, whether inside or outside reporting utterances. However, this is outside the scope of the present article. In the meantime, a fairer picture of the use of reporting verbs is presented if we relate the number of different reporting verbs to that of the number of reporting utterances. This is done in Figure 10. The scores in Figure 10 are obtained by dividing the number of different reporting verbs by the total number of reporting utterances in each text.

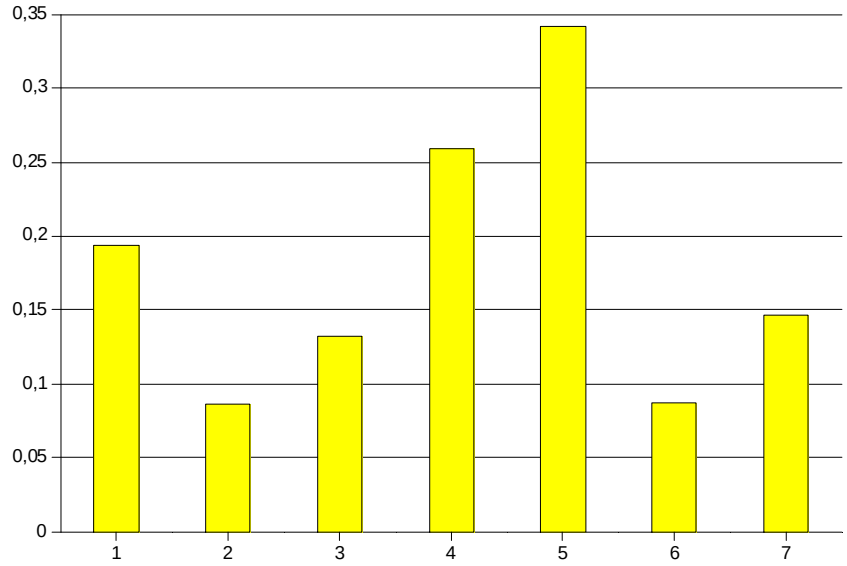

Figure 10: *Variation in reporting verbs*

Now we see that Innes is not so different, really, from the authors of texts 6 and 7. The two horror texts, however, retain a high score, indicating that these two authors are indeed more varied in their use of reporting verbs than the others.

Of course, the reporting utterances cannot be characterized on the basis of the reporting verbs alone. An author can do a lot by means of the structure of the clauses in which s/he uses the reporting verbs. Assuming that reporting utterances have, at least, the dual role of (1) indicating who is talking and (2) indicating the circumstances under which, or the manner in which, the direct speech is uttered (cf. also Oostdijk, 1990), it could, for instance, be hypothesized that an author might be more inclined to qualify the more or less 'neutral' verb *said* by means of adverbials, whereas an author who has more variation in the actual reporting verb might be less inclined to do so. I will return to this point below. The actual figures for reporting clause patterns in absolute terms are presented in Figure 11:

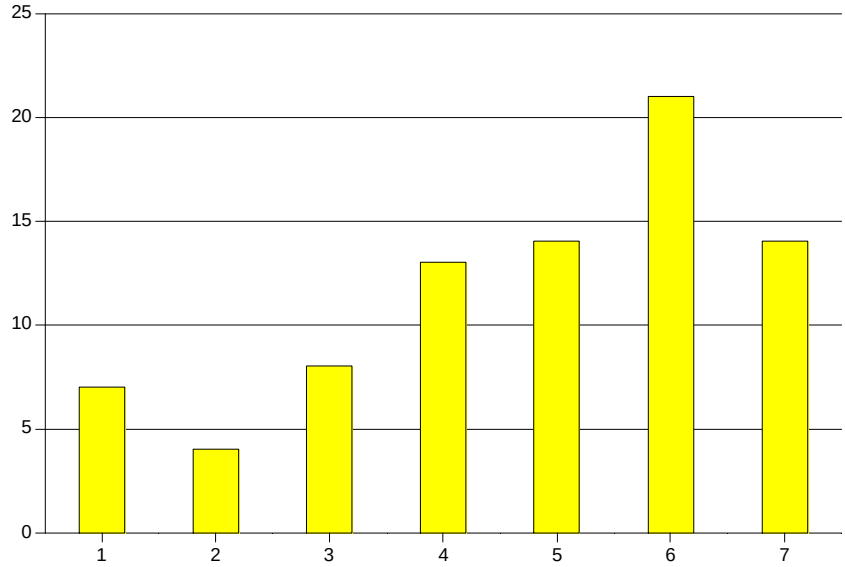

Figure 11: *Number of different patterns in reporting utterances*

We can see that Innes (2 and 3), again, has a fairly low score here. We also see that the two horror texts do not stand out any longer, and that text 6 turns out to emerge at the top, with more than twenty different syntactic patterns in the reporting utterances. However, it is the same with the clause patterns in the reporting utterances as with the reporting verbs themselves: many of them occur only once. So when we relate the actual number of different patterns to that of the total number of reporting utterances per text, we see a completely different picture (Figure 12).

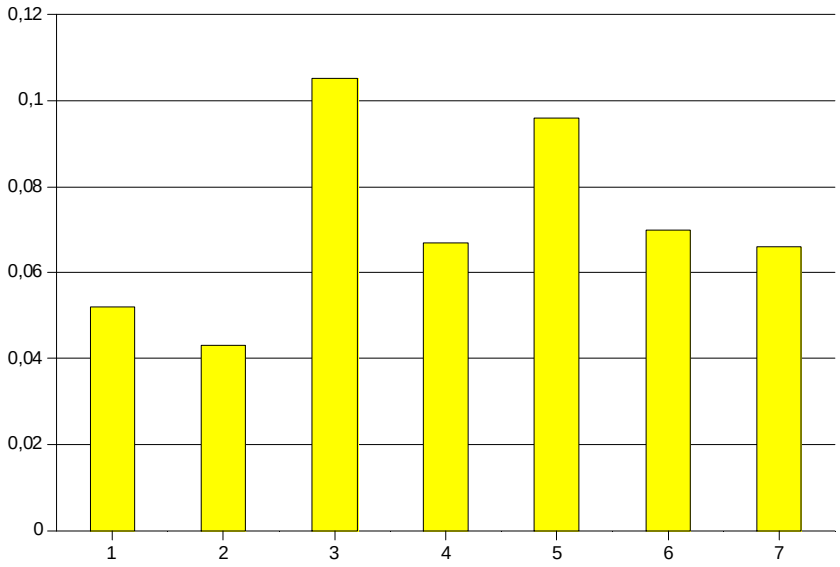

Figure 12: *Variation in reporting utterance patterns*

Now we see that Innes scores very high in text 3 (FCRI01), which is due to the fact that the eight different patterns he uses are found in a total of less than 80 reporting utterances. What we do not know at this point is what the proportion of each of the different patterns is. It was suggested above that there might be an inverse relationship between the variation in reporting verbs and the variation in reporting clause patterns. In other words: an author who uses the 'simple' reporting verb *said* more often may be inclined to use qualifying adverbs or adverbials more often. If we look at the picture presented for the proportion taken

by the verb *said* in the seven texts (Figure 8), we see that text 6 uses *said* in over 80 per cent of the reporting utterances, while text 5 has *said* in less than 40 per cent.

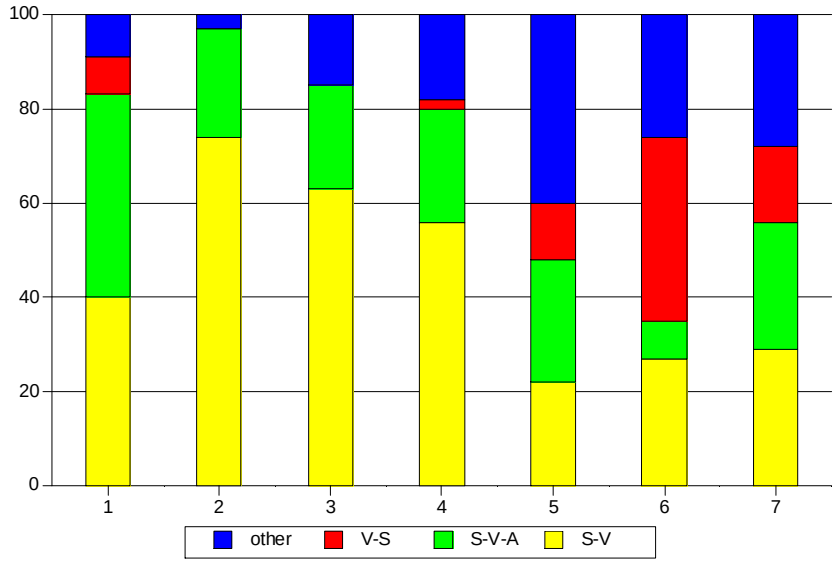

Figure 13: *The most common patterns in reporting utterances*

When we now look at the proportion taken up by the 'basic clause pattern' (S–V) in the seven texts (Figure 13), we see that the only text for which a possible inverse relationship appears to hold is text 6 (over 80 per cent *said*, less than 30 per cent S–V). Note that 'S–V' means exactly: one subject, one verb, in that order, and nothing else. All the other texts show the opposite: the smaller the relative part taken by *said*, the smaller the relative part taken by S–V. On closer examination we can see that, actually, a large proportion of the so-called variation in text 6 is taken up by the pattern V–S, which could arguably be called just a positional variant of the basic pattern S–V. What is more important in this respect is the fact that in this text there are relatively few patterns involving adverbials.

5. Summary, conclusion and suggestions for further research

These are just a few preliminary findings. I hope that the foregoing discussion has made it clear that Oostdijk's (1990) claim about author-specific characteristics certainly deserves attention, but that it is by no means clear that idiosyncratic behaviour shows itself in the same, or even in similar areas in different authors.

A comparative study into author-specific characteristics of dialogue might include an attempt at answering the question how authors keep dialogue going on — with much or little explicit reference to participants, and if so, by means of many or few reporting utterances. We have seen that Innes does not use many *DIA sentences, but does refer to speakers and circumstances of conversation in between the pure dialogue sentences.

A comparative study of spoken conversation on the one hand and dialogue in fiction on the other might include such factors as:

- lexical density;
- length of utterances (which will be difficult to compare, as the analysis of real spoken data relies on the way it has been transcribed; cf. Stenström, 1994);
- syntactic patterns (which may be equally difficult to study, as spoken language may turn out to contain a far wider variety of unconventional clause patterns than we have hitherto assumed — although Oostdijk (1990) has pointed at a couple of features of spoken language that do occur in dialogue in fiction, such as hesitations, omission of operator in questions, topicalization, and ellipsis, but these may turn out to occur less often than in real conversation);
- the way in which turn taking is signalled (cf. Stenström, 1994).

With respect to the continuous scale suggested by Oostdijk (1990), something along these lines has been put forward, e.g. in Biber and Finegan (1986), who show that there are, in fact, several dimensions along which fiction takes a sort of middle position between more formal writing on the one hand, and face-to-face conversation on the other. However, it has yet to be shown how the dialogue vs. non-dialogue parts in fiction score on these dimensions.

What I would like to see is a study along the lines of more recent work by Biber and Finegan (1994), who have looked at text-internal characteristics of medical texts, and who found very clear differences between introduction, methods, results and discussion sections. Something

similar, it would seem to me, might be done in fiction texts, for dialogue vs. non-dialogue passages.

Correspondence

Department of English
P.O. Box 9103
NL-6500 HD Nijmegen
The Netherlands
E-mail: dehaan@let.kun.nl

Note

* This is a slightly revised version of the paper read at ICAME 16, in Toronto, in May 1995.

References

Biber, Douglas, and Edward Finegan. 1986. An initial typology of English text types. In *Corpus linguistics II*, ed. by Jan Aarts and Willem Meijs. 19–46. Amsterdam: Rodopi.

Biber, Douglas, and Edward Finegan. 1994. Intra-textual variation within medical research articles. In *Corpus-based research into language*, ed. by Nelleke Oostdijk and Pieter de Haan. 201–221. Amsterdam & Atlanta, GA: Rodopi.

de Haan, Pieter. 1992. The optimum corpus sample size? In *New directions in English language corpora*, ed. by Gerhard Leitner. 3–19. Berlin-New York: Mouton de Gruyter.

de Haan, Pieter. 1993. Sentence length in running text. In *Corpus-based computational linguistics*, ed. by Clive Souter and Eric Atwell. 147–161. Amsterdam: Rodopi.

Oostdijk, Nelleke. 1990. The language of dialogue in fiction. *Literary and Linguistic Computing* 5: 235–241.

Oostdijk, Nelleke. 1993. *Corpus linguistics and the automatic analysis of English*. Amsterdam: Rodopi.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech and Jan Svartvik. 1985. *A comprehensive grammar of the English language*. London & New York: Longman.

Stenström, Anna-Brita. 1994. *An introduction to spoken interaction*. London & New York: Longman.

Appendix — the seven fiction texts used

Allingham, Margery. 1965. *The Mind Readers*. (NIJM01). Edition used: Penguin Books, 1968.

Barker, Clive. 1985. *The Damnation Game*. (FHOR01). Edition used: Sphere Books limited, 1986.

Herbert, James. 1975. *The Fog*. (FHOR02). Edition used: New English Library, 1986.

Hughes, David. 1985. *The Joke of the Century*. (FHUM03). Edition used: Taplinger Publishing company, 1986. (originally published under the title *But for Bunter*).

Innes, Michael. 1966. *The Bloody Wood*. (NIJM02). Edition used: Penguin Books, 1968.

Innes, Michael. 1984. *Carson's Conspiracy*. (FCRI01). Edition used: Penguin Books, 1986.

Stevens, R. 1977. *Flight from Bucharest*. (FROM01). Edition used: Fontana Books, 1978.