Investigating the presentation of speech, writing and thought in spoken British English: A corpus-based approach[1]

Dan McIntyre [a], Carol Bellard-Thomson[b], John Heywood[c], Tony McEnery[c], Elena Semino[c] and Mick Short [c]

[a] Liverpool Hope University College, UK,

[b] University of Kent at Canterbury, UK,

[c] Lancaster University, UK

Abstract

In this paper we describe the Lancaster Speech, Writing and Thought Presentation (SW&TP[2]) Spoken Corpus. We have constructed this corpus to investigate the ways in which speakers present speech, thought and writing in contemporary spoken British English, with the associated aim of comparing our findings with the patterns revealed by the previous Lancaster corpus-based investigation of SW&TP in written texts. We describe the structure of the corpus and the archives from which its composite texts are taken. These are the spoken section of the British National Corpus, and archives currently housed in the Centre for North West Regional Studies (CNWRS) at Lancaster University. We discuss the decisions that we made concerning the selection of suitable extracts from the archives, the re-transcription that was necessary in order to use the original CNWRS archive texts in our corpus, and the problems associated with the original archived transcripts. Having described the sources of our corpus, we move on to consider issues surrounding the mark-up of our data with TEI-conformant SGML, and the problems associated with capturing in electronic form the CNWRS archive material. We then explain the tagging format we adopted in annotating our data for Speech, Writing and Thought Presentation and discuss how this was developed from the earlier version used for tagging written texts. We also discuss some preliminary analyses which point towards fruitful future lines of investigation.

1 Introduction

The presentation of speech and thought has long been of interest to a range of scholars. Recent research in this area has been done by philosophers (Clark and Gerrig 1990), applied linguists (Thompson 1996; Buttny 1997; Baynham and Slembrouck 1999; Myers 1999), conversation analysts (Holt 1999) and psychologists (Ravotas and Berkenkotter 1998). In stylistics, there is a long tradition focussing on speech and thought presentation in written fiction (see, for example, Banfield 1973; McHale 1978; Leech and Short 1981 and Fludernik 1993). One of the most widely accepted frameworks for the description of the phenomenon in this tradition is Leech and Short's (1981) model. Leech and Short proposed parallel scales of speech and thought presentation categories for the novel, arranged on a cline of different degrees of apparent narratorial interference (see Figure 1).

NRA	NRSA	IS	FIS	DS	FDS
NRA	NRTA	IT	FIT	DT	FDT

Figure 1: The cline of speech and thought presentation categories in Leech and Short (1981)

As one moves across the cline from left to right, the categories reflect an increasing lack of apparent narrator 'control' of the report. This results at the extreme right of the scale in the categories of 'free direct' speech or thought, the effect of which is to suggest that what we have in these instances are the words and thoughts of the characters themselves, with no narratorial intervention at all. (The categories themselves are defined below in Section 5.)

Descriptions of speech and thought presentation such as the Leech and Short model have generally been based on a combination of intuition and wide reading experience and have been established and illustrated with carefully selected textual examples, chosen to best illustrate particular phenomena. As a result, existing frameworks have remained untested systematically on large quantities of data. In order to address this issue, in 1994 Short, Semino, Culpeper and Wynne embarked on a corpus-based investigation of speech and thought presentation in written literary and non-literary texts (see Short *et al*. 1996; Semino *et al*. 1997; Wynne *et al*. 1998; Short *et al*. 1999; Short *et al*. 2002; Short 2003 and Semino and Short forthcoming). The aim of this initial project was to test the model of speech and thought presentation described in Leech and Short (1981) against a specially constructed quarter-of-a-million word data-set of fictional and non-fictional narratives to see how robust the framework was and how far it would stand up to exposure to corpus data. Among other things, this project introduced an additional scale, parallel to the speech and thought scales, to take account of writing presentation. In this paper we describe the latest phase of this project, which is to further test and refine the model by investigating the nature of Speech, Writing and Thought Presentation (henceforth SW&TP) in spoken, as opposed to written, data. To this end we have constructed the Lancaster Speech, Writing and Thought Presentation Spoken Corpus. Below we outline in more detail the background to the earlier written project, before going on to describe the spoken corpus and its construction, issues involved in annotation, and the outcomes of some preliminary analyses.

2 A corpus-based approach to SW&TP

2.1 The Lancaster Speech, Writing and Thought Presentation Written Corpus

The Lancaster Speech, Writing and Thought Presentation Written Corpus was built to investigate the nature of SW&TP in written narrative texts. The SW&TP Written Corpus project extended the boundaries of investigation beyond the focus on literary texts in Leech and Short (1981) by including non-literary texts within its remit (see Short et al. 1996). Developed between 1994 and 1997, the corpus is now approximately 260,000 words in size. The relatively small size of this in comparison to most modern electronic corpora is due to the fact that the whole corpus needed to be hand-annotated. It is divided into three narrative genres: (1) prose fiction; (2) newspaper news reports; and (3) (auto)biography. These three genres are then sub-divided into ‘serious’ and ‘popular’ sections. The analysis of the corpus texts resulted in some adjustments to Leech and Short’s earlier model and also revealed the necessity of the parallel scale referred to above to take account of the report of writing (see Wynne et al. 1998; Semino et al. 1999; and Short et al. 1999 for more details).

2.2 The need for a Speech, Writing and Thought Presentation Spoken Corpus

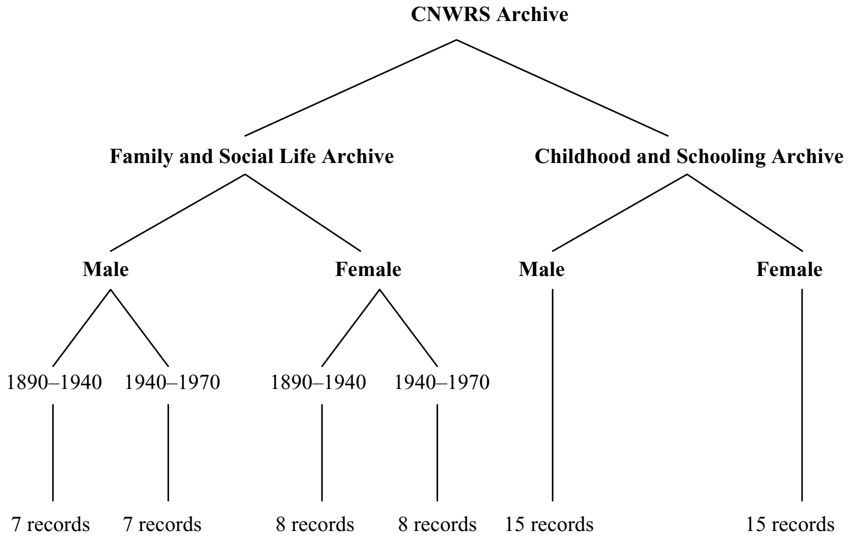
Work on the SW&TP Written corpus raised the question of the extent to which the quantitative and qualitative results that were arrived at would apply to spoken as opposed to written language. The work that has been done on SW&TP in speech has tended to concentrate purely on direct speech (e.g. Baynham 1996), or has analysed qualitatively small amounts of data gathered from very specific contexts (e.g. Hall et al. 1999; Holt 1999). We have attempted to address this issue by constructing a small, balanced corpus of contemporary spoken British English in order to analyse the presentation of speech, thought and writing in spoken data systematically. Our aim is to further test the model of SW&TP originally proposed in Leech and Short (1981) and expanded in the work of Short, Semino and Wynne (e.g. Wynne et al. 1998), in order to arrive at a systematic and comprehensive framework developed through exhaustive analysis of both written and spoken data. For this reason, in building the corpus we decided to explore both elicited and spontaneous speech.

3 Selecting the corpus data

The texts that form our corpus are drawn from two sources: (1) the spoken demographic section of the BNC (World edition); and (2) oral history archives in the Centre for North West Regional Studies (CNWRS) at Lancaster University. Whereas the texts for the written corpus were randomly selected, we deliberately chose spoken texts that appeared to be rich in SW&TP in order to ensure that we had a substantial amount of data to work with (hence we cannot claim that our spoken corpus is representative in terms of the overall amount of SW&TP it contains).

The Spoken Corpus is approximately 260,000 words in order to make it comparable in size with the existing SW&TP corpus. The CNWRS archives and the BNC obviously provide a far larger body of data than we required, and so we opted to select 120 ‘chunks’ (60 from the BNC and 60 from the CNWRS archives) of approximately 2,000 words each (as this was the size of the texts in the written corpus), providing 240,000 words in total. We also decided that the chunks would not be stopped at exactly 2,000 words, but would be allowed to run on a little, to allow each chunk to represent a coherent stretch of conversation, a decision parallel to that made when constructing the written corpus. This gave us the remaining 20,000 words needed to make our corpus approximately 260,000 words in size.

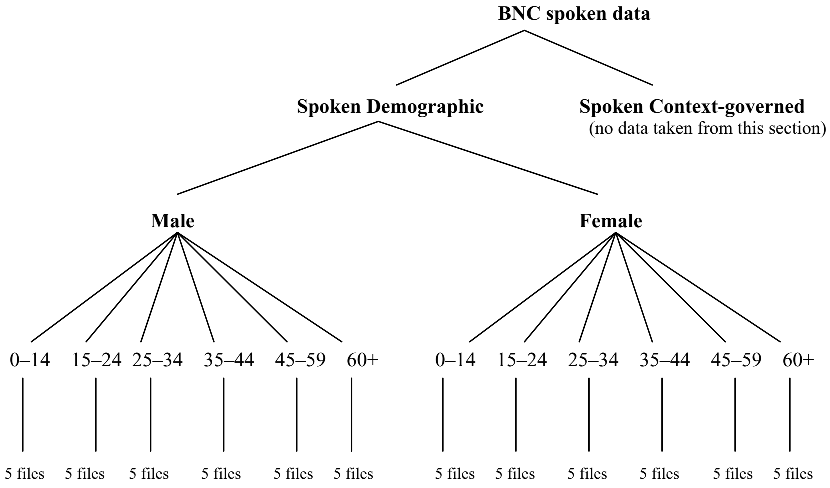
The CNWRS data is drawn from two archives. The ‘Family and Social Life’ archive was compiled from data collected in the 1970s and 1980s by Elizabeth Roberts[3] and Lucinda Beier[4], and consists of 250 hours of interviews, stored on audiocassettes and reel to reel tapes, with accompanying transcripts. We used the transcripts to identify sections rich in SW&TP. The interviewees recall what life was like in Lancaster, Preston or Barrow between the periods 1890–1940 or 1940–1970. The data in the ‘Childhood and Schooling’ archive was collected in the 1980s by Penny Summerfield[5], and consists of approximately 200 hours of interviews on audiocassette, with accompanying transcripts. Again, the interviews are one-to-one, with the interviewees recalling their years spent in education between 1920 and 1950 in Lancaster and Morecambe, Preston, Blackburn, Burnley and Clitheroe. We aimed to balance for male and female interviewees in this data set. Figure 2 shows the number and distribution of CNWRS files in our corpus.



(60 files with a roughly equal balance of male and female speakers from each age-range in each archive; discrepancies in the number of files in the male and female sections of the FASL archive are due to the low quality of sound recordings in the male sections making many of these files unusable.)

Figure 2: Number and distribution of CNWRS texts in the Lancaster SW&TP Spoken Corpus

With regard to the BNC texts, we decided to use only material from the spoken demographic section of the corpus, as this would allow us to contrast spontaneous dialogue with the elicited monologues of the CNWRS archives. Since the BNC data was collected in the early 1990s and the CNWRS data in the 1970s and 1980s, we also left open the possibility of studying diachronic developments in speech. We chose texts from the BNC that cover all age ranges, with an equal division between male and female respondents. We also concentrated solely on face-to-face interaction – we did not use transcripts of radio phone-ins, for example – and we used only those texts which constitute spontaneous, unscripted data. Figure 3 shows the number and distribution of BNC files in our corpus.



(60 files with an equal balance of male and female respondents in each age-range)

Figure 3: Number and distribution of BNC texts in the Lancaster SW&TP Spoken Corpus

After the initial selection of the transcriptions on demographic grounds, we examined each transcript for long turns, on the basis that these were more likely to be narrative turns that would provide a higher density of the kind of features we were interested in. This meant excluding those files which were of a brief question-answer format, or which contained numerous short turns. In addition to this, with the BNC texts we used the BNC Web query facility to search for common discourse reporting verbs. Where the query returned favourable results, we then examined that area of the text in question manually to see if it was likely to yield numerous examples of SW&TP. So, in addition to the reporting verbs picked up in the electronic search, we also looked for further examples of SW&TP in close proximity to these. As with the CNWRS data, each member of the project team then read the texts in order to identify suitable extracts for inclusion within the corpus.

4 Constructing the corpus

The transcripts from the CNWRS archives were initially the most problematic as these had originally been transcribed for an oral history research project, without regard for linguistic transcription conventions. The texts are divided into what the original transcribers deemed to be ‘sentences’, and are punctuated as if they were written texts. We decided to retain this punctuation, except in those cases where it was necessary to re-transcribe a section due to the inaccuracy of the original transcription, or where we had to newly transcribe stretches of interaction that had been omitted or simply summarised. In these cases, the only punctuation that we added were full stops where we felt a sentence boundary would most likely exist in a written form of the interview. We did this to prevent the text from becoming impossibly difficult to read and understand, though punctuation was not used as a criterion when tagging the texts for SW&TP. Anomalous punctuations were removed and misspellings corrected. In the process, it was noted that the transcribers’ idiosyncratic use of inverted commas sometimes indicated the presence of interesting voice quality features related to discourse presentation.

In addition to producing electronic copies of the CNWRS transcriptions, we also made copies of their corresponding sound files. We digitised the cassettes using the CoolEdit software package, which allowed us to convert the original tapes to wave files. We recorded in mono, at 16-bit resolution, in order that the resulting wave files should be in a form suitable for later time-alignment with the transcripts.

4.1 Mark-up of the corpus

The 120 files in our corpus are all marked up using TEI (Text Encoding Initiative) conformant SGML (Sperberg-McQueen and Burnard 2001) in order to create a shareable archive, compatible with other corpora and concordancing packages. The SGML mark-up allows the corpus to be searched using concordancing programs such as Wordsmith Tools and SARA. For each file in the corpus we have generated a header containing bibliographical information about the computer file itself (with which it is possible to catalogue the file in a library archive), information about the types of tags that are used in the file and how the encoders resolved any problems that arose during tagging, classificatory and contextual information about the text, and a history of changes made in the development of the electronic version. We have also generated an overall corpus header and a document-type declaration for the corpus files.

5 Annotating the corpus for SW&TP

Having described the structure and composition of our corpus, in this section we explain the system of annotation that we used to tag the files for Speech, Writing and Thought Presentation. To enable us to compare our findings from the Spoken Corpus with those of the Written Corpus project (see Semino et al. 1997), we make use of the system of annotation outlined in Wynne et al. (1998), though with some modifications to take account of the differences between written and spoken data. Before describing the category set and outlining the tagging formalism that we used in annotating our data, it is useful to summarise briefly the categories of SW&TP that we used in analysis. We begin by presenting the categories sets we used in both the Written and the Spoken Corpora and consider the changes that we made to our tag-set as a result of working with spoken data.

5.1 SW&TP categories in the Written and Spoken Corpus projects

Table 1 details the acronyms used to mark instances of SW&TP in the Written Corpus project and their equivalents in the Spoken Corpus.

NRS/T/W and RS/T/W are reporting signals (prototypically reporting clauses), and are not a part of the discourse being presented. They are therefore placed outside the discourse presentational clines. The convention we use is that SW&TP category labels are written in upper-case letters. The SW&TP Written project also developed a set of four additional features that categories might have. These are marked in lower-case to distinguish definitional labels from more minor associated features. The four features are discussed below in Section 5.2.2 as part of the expanded set developed for the Spoken project.

Table 1: Categories in the SW&TP Written corpus and their equivalents in the Spoken Corpus[6]

Categories outside the discourse presentation clines			
Written Corpus		**Spoken Corpus**	
Category	**Definition**	**Category**	**Definition**
N	Narration	A	Anything other than SW&TP (narrative and non-narrative)
		RU	Report of Language Use
NRS	Narrator's Report of Speech	RS	Report of Speech
NRT	Narrator's Report of Thought	RT	Report of Thought
NRW	Narrator's Report of Writing	RW	Report of Writing
Discourse Presentation Categories			
Written Corpus		**Spoken Corpus**	
Category	**Definition**	**Category**	**Definition**
NV	Narrator's Representation of Voice	RV	Representation of Voice
NI	Narrator's Representation of Internal States	RI	Representation of Internal State
NW	Narrator's Representation of Writing	RN	Representation of Writing
NRSA	Narrator's Representation of Speech Act	RSA	Representation of Speech Act
NRTA	Narrator's Representation of Thought Act	RTA	Representation of Thought Act
NRWA	Narrator's Representation of Writing Act	RWA	Representation of Writing Act
NRSAp	Narrator's Representation of Speech Act with Topic	RSAp	Representation of Speech Act with Topic
NRTAp	Narrator's Representation of Thought Act with Topic	RTAp	Representation of Thought Act with Topic
NRWAp	Narrator's Representation of Writing Act with Topic	RWAp	Representation of Writing Act with Topic
IS	Indirect Speech	IS	Indirect Speech
IT	Indirect Thought	IT	Indirect Thought
IW	Indirect Writing	IW	Indirect Writing
FIS	Free Indirect Speech	FIS	Free Indirect Speech
FIT	Free Indirect Thought	FIT	Free Indirect Thought
FIW	Free Indirect Writing	FIW	Free Indirect Writing
DS	Direct Speech	DS	Direct Speech
DT	Direct Thought	DT	Direct Thought
DW	Direct Writing	DW	Direct Writing
FDS	Free Direct Thought	FDS	Free Direct Thought
FDT	Free Direct Thought	FDT	Free Direct Thought
FDW	Free Direct Writing	FDW	Free Direct Writing

5.2 SW&TP categories in the Spoken Corpus project

For the spoken project we began with the tag set in the left half of Table 1 (see Semino et al. 1997 for a discussion of these categories). However, in the course of annotating the spoken data we made various alterations and additions to our categories and their corresponding acronyms, as shown in the right half. The main changes were as follows:

- Leech and Short (1981) initially used the term ‘report’ in the description of NRSA and NRTA. This term was replaced by ‘representation’ in some of the later publications describing the written corpus. Semino and Short (forthcoming: Chapter 1, Section 1.1) now argue for the term ‘presentation’. The arguments for and against these alternative terms are too complex to go into here. However, it will be helpful if we point out that we have retained ‘R’ in our various category acronyms in order to preserve as much annotational continuity as possible and we continue to gloss it as ‘representation’ in order to avoid possible confusion for our various readerships.
- We have dispensed with the N constituent of the categories as a consequence of tagging oral texts. N previously stood for ‘narration’ or ‘narrator’s’, and is not always applicable to non-narrative written data or to spoken data. Hence, what in the written corpus would have been NRSA is in the spoken corpus simply RSA. Likewise, the single N attribute value, which was used in the Written corpus to mark anything not annotated as SW&TP, is replaced in the Spoken corpus by A, which, simply standing for ‘[A]nything other than SW&TP’, comprises both narrative and non-narrative text.
- Dispensing with the N constituent had the knock-on effect of leaving us with the same acronym – RW – to refer both to a reporting clause (or non-clausal equivalent) of writing presentation preceding either the direct or indirect presentation of writing, and the minimal presentation of writing (e.g. ‘I wrote to Eileen’). We therefore needed a different acronym in order to distinguish between the two phenomena. We chose to use RN to refer to the latter, N being the only remaining consonant in the word ‘writiNg’ that is not used elsewhere in the tag-set.
- We have introduced a new tag, RU, to refer to ‘report of language use’. This is used to tag instances where speakers refer to words or expressions, often idiosyncratic, that were habitually used either by groups of people or individuals to refer to particular things. A prototypical example would be ‘So we had a box of [RU] what we called wet day stockings’. Instances of RU are most common (220 out of 247 instances) in the CNWRS texts in our corpus where people are talking about their past lives.
- We have chosen to mark the grammatical structure of instances of SW&TP in an effort to provide more information about the forms of SW&TP in our corpus. We assume the default grammatical structure of a stretch of SW&TP to be declarative and this is not tagged. Imperatives are tagged with ‘p’ and interrogatives with ‘v’. Confusion with the lower case p for ‘topic’ is avoided by their being placed in different positions.

- We expanded the number of additional features.[7]

Below, we explain the acronyms used to refer to the main categories of Speech, Writing and Thought Presentation, via some examples from our corpus. We then describe the acronyms for additional feature constituents. All new additions to our tag set were to cope with particular phenomena we encountered in the spoken data. For ease of interpretation, the tags are represented here in simplified form. The full format is presented at the end.

We now present an explanation of the main categories of SW&TP in 5.2.1, and of possible additional feature constituents in 5.2.2. This is followed by the tagging format we use in 5.3. Our descriptions of the scope of each category and our examples aim to account primarily for central cases, since we do not have the space in this paper to discuss complex and borderline cases.

5.2.1 Main categories of SW&TP

As the three scales are in parallel, to bring out what they have in common we combine their definitions as far as possible. In general, speech and writing presentation categories share many formal features and functions. Thought presentation categories, however, often display different functional properties. We therefore group speech and writing together and place thought last in the following lists.

The Direct Categories (DS, DW and DT)

The direct categories consist of independent clause/s or phrase/s which convey the illocutionary force of speech or writing acts, their propositional content, and which include the deictic features appropriate to the anterior speech, thought or writing event that is being presented. Prototypically, the ‘direct’ categories usually claim to represent the ‘actual words’ used, or to exemplify the kinds of words and expressions typically used. Although Direct Thought is formally similar to Direct Speech and Direct Writing, aspects of the definition of the latter two, such as illocutionary force and ‘actual words’, do not sensibly extend to DT. The following are examples:

Direct Speech (DS)

1. [A] He looked round [RS] and said to all the lot of us lads he said, he said **[DS] I bet you buggers like your fish and chips**

Direct Writing (DW)

2. [RWA] he wrote me this letter [RW] saying erm saying **[DW] I, I realise that there’s been something on your mind recently**

Direct Thought (DT)

3. [RT] I thought **[DT] well I might as well come**
4. [RT] I thought **[DT] oops!**

The Free Direct Categories (FDS, FDW and FDT)

As for DS, DW and DT but without an accompanying RS, RW or RT.[8]

Free Direct Speech (FDS)

5. And I remember al always our Leonard taking me next door and knocking at the door and the **[FDS] I've come to show you our Peggy's new frock**

Free Direct Writing (FDW)

6. Look, they've stuck a sticker in the back **[FDW] cars kill trees**

Free Direct Thought (FDT)

7. I went into the loo **[FDT] it stinks of smoke in here**

The Free Indirect Categories (FIS, FIW, FIT)

The Free Indirect categories are characterised by a mixture of deictic, syntactic and lexical features, some appropriate to current speaker, others to the producer of the anterior speech, writing or thought event that is being presented. They are prototypically realised by an independent clause, but an accompanying RS, RW or RT is sometimes possible.

Free Indirect Speech (FIS)

8. [RS] Father said [DS] can my girls come? **[FIS] No they couldn't come**
9. [RS] and everybody was **[FIS] where was I sort of thing**

Free Indirect Writing (FIW)

10. [RW] Dennis, who had been my boyfriend wrote from Italy where he was stationed, **[FIW] when he came home at Christmas, could we be engaged?**

Free Indirect Thought (FIT)

11. [A] I persisted in getting dressed and immediately went home. I was quite <unclear> by that time, but **[FIT] I wasn't putting up with this garbage I was going home, that was it**

The Indirect Categories (IS, IW and IT)

The Indirect categories consist of a reported clause which is grammatically subordinated to an RS, RW or RT. All deictic features are appropriate to the speaker in the posterior, discourse presenting, situation. Prototypically, the propositional content of the original speech, thought or writing act is specified, but no claim is made to present the words and structures originally used to utter that proposition.

Indirect Speech (IS)
12. [RS] he said **[IS] it made him happy**

Indirect Writing (IW)
13. [RWr] it was put down on in a book **[IWr] that you'd taken a pair of stockings home**
14. [RW] well all that was said was [IW] the had twenty eight days to pay

Indirect Thought (IT)
15. [RT] he thought **[IT] there was nowhere else**
16. [RT] I suddenly realised **[IT] I hadn't got Vicky with me**

Representation of Speech/Writing/Thought Act (RSA/RWA/RTA)
RSAs and RWAs present the illocutionary force of an utterance or text (part) with an optional noun or prepositional phrase indicating the topic, but do not claim to represent the propositional content or the original wording of that content. RTAs are formal equivalents, but the notion of a 'thought act' seems likely to have a much more restricted range than speech or writing acts. More specifically, the notion of 'illocutionary' force in relation to thought acts is problematic, while that of perlocutionary effect associated with speech and writing acts is inapplicable.

Representation of Speech Act (RSA)
17. [RSA] she asked Billy out
18. [RSA] I just threatened them
19. [RSAp] And I told him all rigmarole

Representation of Writing Act (RWA)
20. [RWA] Vivian voted Conservative
21. [RWA] my sister put in for an improvement grant
22. [RWAa] we didn't have to put our religion on the paper

Representation of Thought Act (RTA)
23. [A] I just move some of this stuff out the way, I know, **[RTA] I've had a good idea, a smart idea**
24. I mean **[RTA] you've kind of changed your mind haven't you?**

Representation of Voice/Internal State/Writing (RV/RI/RN)

Representation of Voice (RV)
RV captures minimal references to speech with no indication of the illocutionary force, let alone the propositional content or form of the utterance (part). RVs can present either individual instances of talk or whole Speech Events. As with the RSA category, a reference to a topic may be attached.

25. I was sitting there **[RV] talking** [A] and they had a drop of wine
26. [A] yeah, well **[RVr] they used to teach sort of parrot fashion**
27. [A-RVr] every day did start with a hymn, [RSAr] a prayer, **[RVr] and a reading [RVr] and then we had a scripture lesson [RVr] and by the way a fair bit of that scripture lesson**
28. [A] yes, what about the other masters **[RVp] you were talking a bit about them before**

Representation of Writing (RN)
RN captures minimal references to writing or writing events or to the writing of an instance of a text-type with possibly a minimal reference to topic, but with no indication of the illocutionary force or of the propositional content or linguistic form of the portion of text. RNs can present either individual instances or a series of writing events, or group participation in them.

29. they had slates **[RNr] and they used to write with a piece of slate**
30. [RN] he wrote me this letter
31. We had as well as [RN] taking the School Certificate we also had a religious School certificate that we took

Representation of Internal State (RI)
RI captures references to cognitive or emotional states or processes that do not amount to specific thoughts.

32. [RI] I was frightened to death of him I was really I was frightened to death of him
33. [A]...come on I mean **[RIi] you really liked him [RIi] fancied him [RIi] loved him**

Other categories

Reporting signals (RS, RW and RT)
RS, RW and RT are prototypically represented by a reporting clause associated with a stretch of direct, indirect, and in some cases, free indirect, speech,

thought or writing. As we pointed out in our discussion of Table 2, RS/T/W, as reporting signals, are not a part of the discourse being presented. The RS/RW/RT function is sometimes performed by a noun, adjectival, adverbial or prepositional phrase.

Report of Speech (RS)
34. [RS] Mrs Hall said [DS] I don't know how you find time to go to your church every morning like this
35. [RS] they used to promise [IS] to say so many prayers
36. and sort of looked up at him and **[RS] it was like** [DS] oh hi
37. [RS] I goes [DS] what?

Report of Writing (RW)
38. [RW] across the certificate he wrote [DW] this man should be in bed
39. [RW] but it wasn't saying [IW] I'd made an error *[N.B. the speaker is referring to a bank automat display]*
40. it was said that there was a tombstone there **[RW] with an inscription on it** [DW] Here lies the body of old Tom Thrumb

Report of Thought (RT)
41. [RT] I decided [IT] I'd like to be an engineer
42. Yes well you brought them in **[RT] with great fear** [IT] that your father was going to say something to them
43. [RT] and I used to wonder [IT] what the green van was

Report of Use (RU)
RU captures meta-linguistic mentions of language use, such as the words or expressions habitually used to refer to things, or the ways words were spelled or pronounced.

44. and then you see **[RU] what they called the tacklers** were over the weavers

Anything other than SW&TP (A)
The A tag was applied to all those stretches of text which do not contain any references to speech, thought or writing presentation.

45. [A] Well Mother Monica Mother Mary Monica was the headmistress

5.2.2 SW&TP category features
Of the symbols below, the definitions given here for p (indicating topic), e, h, i, q, and # are those initially developed for the Written Corpus. While we found that e, h, i and # could be applied to the spoken data straightforwardly, the extent

to which p and q could appropriately be applied raised theoretical issues that are currently being investigated. The other symbols were adopted during the annotation of the Spoken Corpus.

p (= topic)
The p suffix marks an extended topic, most commonly of a speech, thought or writing act.

46. [A] Erm I don't ever remember **[RSAp] my mother expressing any interest or desire or wish to have a job**

(= problematic)
The symbol # was used to signal 'problematic' tags that needed further investigation.

47. at ten o'clock at night and <pause> pub was packed **[A-RV#] People singing with the the group**

e (= embedded)
The suffix e marks instances of discoursal embedding where one SW&TP category is embedded discoursally, but not necessarily syntactically, in another.

48. [RV] Joan rang last night [RS] to say [IS] that Reg **[RSe] had asked us [ISe] to go to to see the daffodils**

g (= negative)
The suffix g marks a grammatical negative.

49. [A] And, um, well, I suppose I can't I shouldn't say **[RSApg] but my father would never allow you to go to dances**

a (= absence)
The suffix a signals the marked absence of performance of a speech, thought or writing act.

50. And I never heard once heard my family turn round **[RSa] and say, [DSa] That's my son**

h (= hypothetical)
The suffix h marks an instance of SW&TP that does not present an anterior discourse but "refers" to an event that has not (or not yet) taken place.

51. [RTh] Well if she wants if she wants [ITh] to get rid of it [RShp] ask her [ISh] how much [RTAehv] she wants for it

i (= inferred)
The suffix i signals instances of thought presentation where the reporter did not have direct access to the relevant thoughts.

52. and then, and erm, **[RTi] this woman, receptionist, whatever, obviously thought [DTi] oh well,** [RIei] **he knows the guy**

q (= quotation phenomenon)
The suffix q marks the presence of a direct quotation which is enclosed within a non-direct category of SW&TP and which does not count as a straightforward example of direct speech.

53. [A] I think er I agree with er Tennyson on that. I think **[RWApq] he spoke of Virgil as wielder of the stateliest measure ever moulded by the lips of man**

r (= reiterated)
The suffix r marks an iterated instance of SW&TP.

54. we appear to be the most consistent pub in the area, with er customers and what have you. They all come in and **[RSr] tell us [ISr] we're the busiest [RSr] and I say [DSr] well if we're the busiest, God help those that're the quietest**

v (= interrogative)
The suffix v marks a grammatical interrogative.

55. [RSv] Did they ever say [ISv] why they did it, why they went to view the body and took children

p (= imperative)
The suffix p marks a grammatical imperative. Note that in the tagging format below, p for imperative is differentiated from p for topic by the fact that it appears in a different attribute value slot.

56. [RS] He said er **[DS] [RSep] Tell your mammy [DSep] it'll be alright** [A] and we turned back home

u (= unfinished)
The suffix u signals that the relevant SW&TP category is unfinished.

57. [RS] and I said [DSu] well that was stra

1/2/3 etc
In the Written Corpus, numerals indicate the number of levels of discoursal embedding. In the Spoken Corpus, they are also used to record the number of

repeated adjacent categories represented by one label. The different functions are distinguished by the field in which the numeral occurs (see Section 5.3).

58. Level of Embedding [RT] I felt [IT] I ought [RWAe] to write to him [RT] because I thought [DT] we're both getting old, [RIe] I'd like **[RWAe2] to write [RWeh3] and ask him [IWeh3] [RIeh4] if he remembers his father**

59. Repeated categories [A] He looked round **[RS3] and said to all the lot of us lads he said, he said** [DS] I bet you buggers

5.3 The tagging format

We use the element <sptag> to mark instances of SW&TP in our corpus. Each constituent of the SW&TP categories (detailed in Tables 2 and 3, above) are marked within one of fifteen <sptag> attributes[9] (see the example below). We use 'x' as a placeholder for those attribute value slots that are not filled for a particular SW&TP category. This is done for ease of concordancing. We use an end tag (</sptag>) to mark the end of a particular stretch of SW&TP. Below is an example of an SW&TP tag. This particular example would be used to mark a stretch of hypothetical Free Indirect Speech:

<sptag one="F" two="I" three="S" four="x" five="x" six="x" seven="x" eight="x" nine="h">

Table 2, below, details the allowable values for each of the fifteen attributes: (N.B. We do not mark empty positions that follow the final attribute value for a given SW&TP tag.)

Table 2: Allowable values for each of the fifteen <sptag> attributes

Attribute	Allowable values	Definitions
One	x A F	Anything other than SW&TP; Free
Two	x R I D #	Representation; Indirect; Direct, # interesting example of A
Three	x S T W V I N U	Speech; Thought; Writing; Voice; Internal state; Writing; Use
Four	x A	Act
Five	x p	topic
Six	x # 1 2 3 4	# = odd/interesting cases; numerals = repeated adjacent categories
Seven	x e	embedded
Eight	x g a	grammatical negative; marked absence of SW&TP
Nine	x h	hypothetical
Ten	x i	inferred
Eleven	x q	quotation phenomenon
Twelve	x r	iterative
Thirteen	x v p	interrogative; imperative
Fourteen	x u	unfinished
Fifteen	x 1 2 3 4	numerals = level of embedding

6 Initial results

Our analysis of the corpus is ongoing. Here we present some preliminary quantitative findings based on a comparison between the Spoken Corpus and the Written Corpus of Short et al. (1999), from which some of the differences between SW&TP in the written and spoken data are beginning to emerge. Tables 3, 4 and 5 detail the frequency of occurrence and relative rank orderings of the SW&TP categories in each of the two corpora[10]. The percentages in column A are proportions of the total number of words involved in discourse presentation in the relevant corpus. The percentages in column B express proportions of the total number of words involved in the relevant discourse presentational cline in that corpus. The figures do not include the ‘A’ or ‘N’ categories, the reporting signals ((N)RS, (N)RT and (N)RW) or the RU category, since these do not constitute SW&TP categories in and of themselves. At this stage, all ambiguous tags have also been excluded from the analysis[11]. The final column in each table gives the results of a chi-square test run on the frequency figures to test for the significance of the difference in use of each category between our written and spoken corpora. For this test, which was undertaken with one degree of freedom, the significance level for significance at 0.001 is 10.83 or above. In Tables 3–5, scores statistically significant at the 0.001 level are emboldened.

Table 3: Frequencies of occurrence and rank orderings of speech presentation categories in the written and spoken corpora

Category	**Written Corpus**				**Spoken Corpus**				**Significance**
	Frequency	**A**	**B**	**Rank**	**Frequency**	**A**	**B**	**Rank**	
FDS	927	10.79	15.36	4	191	1.93	3.92	5	**153.80**
DS	2047	23.83	33.92	1	1852	18.97	38.01	1	5.63
FIS	157	1.82	2.60	6	88	0.89	1.80	6	3.39
IS	1114	12.97	18.46	3	588	5.93	12.06	4	**31.54**
(N)RSA	1398	16.27	23.16	2	1305	13.15	26.78	2	6.36
N/RV	391	4.55	6.47	5	848	8.55	17.40	3	**126.53**

Table 4: Frequencies of occurrence and rank orderings of writing presentation categories in the written and spoken corpora

Category	Written Corpus				Spoken Corpus				Significance
	Frequency	A	B	Rank	Frequency	A	B	Rank	
FDW	32	0.37	6.36	5	115	1.16	14.35	3	**10.76**
DW	109	1.26	21.66	2	88	0.89	10.98	4	**14.34**
FIW	32	0.37	6.36	5	25	0.25	3.12	6	4.56
IW	74	0.86	14.71	3	45	0.45	5.61	5	**17.20**
(N)RWA	215	2.50	42.74	1	350	3.53	43.69	1	0.04
NW/RN	41)	0.47	8.15	4	178	1.79	22.22	2	**22.38**

Table 5: Frequencies of occurrence and rank orderings of thought presentation categories in the written and spoken corpora

Category	Written Corpus				Spoken Corpus				Significance
	Frequency	A	B	Rank	Frequency	A	B	Rank	
FDT	69	0.80	3.36	5	5	0.05	0.17	6	**48.31**
DT	38	0.44	1.85	6	175	1.76	6.00	4	**28.41**
FIT	275	3.20	13.40	2	10	0.10	0.34	5	**210.04**
IT	201	1.20	9.79	3	748	7.54	25.66	2	**93.11**
(N)RTA	114	1.32	5.55	4	396	3.99	13.58	3	**44.38**
NI/RI	1355	15.77	66.03	1	1581	15.93	54.23	1	**16.14**

6.1 Ranking and quantitative 'norms' in SW&TP

Before briefly considering the various statistically significant differences shown in the tables, it is first necessary to consider an apparent difference that is in fact not significant – rank ordering. In Tables 3–5, the rank orderings of the different categories between speech, thought and writing representation differ slightly. For example, FDT is ranked 5th in Tables 4 and 5, while it is ranked 4th in Table 3. Similarly, within the tables there are differences in rankings between spoken and written data. For example, within Table 5, the FIT category is ranked 2nd in speech yet 3rd in writing. In order to discover whether these slight differences were significant, we carried out a Wilcoxon signed ranks test (see Oakes 1998: 21). However, the test revealed that the differences in ranks observable in the tables above were not statistically significant. This finding is relevant to the issue of the quantitative 'norms' for each of the three modes of presentation in the two corpora.

Leech and Short (1981) originally suggested a fundamental difference between what counts as the ‘norm’ for speech as opposed to thought presentation, on the basis of what they call ‘the semantics of reporting’ (Leech and Short 1981: 345). Because speech is a physical, public phenomenon, Leech and Short proposed that the ‘norm’ for speech presentation is DS, since “it is the mode which represents speech in the form in which it is directly manifest to a listener” (Leech and Short 1981: 345). In contrast, they saw IT as the ‘norm’ for thought presentation, since thought is a private phenomenon, which, outside fictional conventions, is only directly accessible to the person experiencing the thoughts, and which is not necessarily articulated in verbal form (see also Halliday 1994: 250ff. for a similar view).

Tables 3 and 5 enable us to check the quantitative validity of Leech and Short’s qualitative observations as far as our two corpora are concerned, while Table 4 makes it possible to see how writing presentation compares with the other two modes of presentation. The lack of statistically significant differences in the rank ordering of categories already suggests that the ‘norms’ for each of the three modes of presentation are similar across our Written and Spoken corpora. More specifically, Table 3 shows that the quantitative ‘norm’ for speech presentation in our two corpora is the same as Leech and Short’s ‘semantic’ norm, namely DS. The dominance of DS over other speech presentation categories is even more marked if we consider that FDS is best seen as a variant of DS, so that the separate figures in the first two rows of Tables 3–5 could in fact be combined (see Short et al. 1996, Semino et al. 1997, Semino and Short forthcoming: Chapter 6, Section 6.5.2).

In contrast, Table 4 shows that the quantitative ‘norm’ for writing presentation in the two corpora is not at the direct end of the scale, but rather NRWA. In their analysis of the Written Corpus, Semino and Short (forthcoming: Chapter 5, Section 5.2.6) suggest that this may be due to the fact that (i) DW and FDW do not have the same effects of immediacy and dramatization as DS and FDS, and (ii) the use of DW and FDW is subject to greater faithfulness constraints than that of DS and FDS: when the original ‘text’ is written rather than spoken, it is more likely that misquotations will be noticed, so that, other things being equal, reporters may be less willing to take liberties with DW and FDW than with DS and FDS. A more detailed qualitative analysis of our Spoken Corpus will be necessary, however, before we can comment on how well these kinds of observations apply to the Spoken Corpus.

The situation with thought presentation is rather more complex. Table 5 suggests that the quantitative ‘norm’ in both corpora is the least direct category, NI/RI (which was not part of Leech and Short’s 1981 model). However, Semino and Short (forthcoming: Chapter 5, Section 5.3.5) show how a detailed analysis of NI in the Written Corpus questions whether NI is properly seen as a thought presentation category, as opposed to a sub-type of narration. We do not have the space here to discuss the ways in which NI is unlike other thought presentation categories, and also unlike the most minimal categories of both speech and writing presentation (NV/RV and NW/RN respectively). For a detailed discussion, see Semino and Short forthcoming: Chapter 5, Section 5.3.5. If, for a moment, we reconsider Table 5 without including NI/RI among the thought presentation categories, IT would turn out to be the quantitative 'norm' in our spoken corpus, thereby adding weight to Leech and Short's proposal. In the Written Corpus, FIT appears to be the most frequent thought presentation category (after NI), but in this case the overall figure of 275 occurrences is somewhat deceptive. Of the total 275 instances of FIT in the Written Corpus, 230 occur in the fiction section, 45 in the (auto)biography section, and none in the press section. As a consequence, if we exclude NI, the quantitative 'norm' for thought presentation in the Written Corpus is FIT in the fiction section and IT in the press and (auto)biography sections (see Semino and Short forthcoming: Chapter 5, Sections 5.3, 5.3.2). Further research is needed in order to determine more accurately the status of NI/RI. What we can conclude, however, from our initial quantitative results is that the quantitative 'norms' in SW&TP appear to be remarkably similar across the two corpora.

6.2 Preliminary observations on the differences between the two corpora

Returning to the significance tests shown in Tables 3–5 above, it is clear that there are some statistically significant differences in the data we have annotated. The most general pattern that emerges is that, in spite of the similarities in relation to quantitative 'norms', there are some considerable differences in SW&TP in written as opposed to spoken data. The results of the pair-wise comparison, using a chi-square test, of the number of occurrences of different categories in the spoken and written data in Tables 3–5 above show that this is usually significant statistically. 13 of the 18 comparisons shown in these tables are significant at the 0.001 level.

We are not yet in a position to comment in detail on these statistically significant differences, or to attempt explanations. However, within the pattern of general difference, there are some points of note. Firstly, the quality of the observed significant difference varies – some categories are present significantly more frequently in the spoken data (N/RV, FDW, NRW/RN, DT, IT, (N)RTA, N/RI), while others appear significantly more often in the written data (FDS[12], IS, DW, IW, FDT, FIT). Secondly, the thought presentation categories show a much more marked variance between spoken and written data than do the speech and writing presentation categories. Unlike the speech and writing presentation categories, the difference between spoken and written data gives rise to a significantly different distribution across written and spoken data for all of the categories of thought presentation annotated in the corpus. The result of this is that the FI{S/W/T} and N/R{S/W/T}A categories, which do not yield significant differences for the speech and writing scales across the spoken and written data, do generate significant differences on the thought scale. This second observation led us to explore the differences between thought presentation as opposed to speech and writing presentation. In order to do so, we undertook a factor analysis based upon the observed differences of the distributions of the categories in speech, thought and writing across the spoken and written data. Our aim in doing so was to discover how marked were the observations that we had made regarding the differences between the thought categories and the speech and writing categories.

When compared through a factor analysis using the frequency of the categories across spoken and written data, the presentation of speech and writing cluster together (with values of 732.54 and 751.22 respectively), while the presentation of thought clusters separately (with a score of 112.78). This confirms some of the findings that arose from the analysis of the Written Corpus. Semino and Short (forthcoming: Chapter 5, Section 5.4) note that, apart from a few exceptions, speech and writing presentation are remarkably similar in terms of the forms and functions of individual categories. In contrast, the thought presentation categories tend to be quite different in use, function and form from their counterparts for speech and writing presentation (for a detailed discussion see Semino and Short forthcoming: Chapter 5, Section 5.3.6). This is largely due to the fact that speech and writing are both modes of ostensible communication leading to the physical production of ‘discourse’, while thought is a private and often non-verbal phenomenon.

A detailed qualitative analysis of thought presentation in the two corpora will be necessary before we can begin to explain why the various categories differ significantly in their frequencies across the two corpora. The need for such detailed work can be illustrated if we consider the row relating to FIT in Table 5, which could lead to mistaken conclusions. Overall, there is a very highly significant difference in the frequency of FIT in the Spoken Corpus as opposed to the Written Corpus. However, as we mentioned earlier, 230 out of 275 instances of FIT in the Written Corpus occur in the fiction section. As a consequence, we do not appear to be faced with a straightforward contrast between spoken and written data as far as FIT is concerned, but rather a contrast between fiction and non-fiction. Indeed, FIT has often been associated with fictional narratives, and particularly with contemporary fiction (our fiction data dates from the latter part of the 20th century), for reasons that we do not have the space to present here (but see Cohn 1978 and Fludernik 1993, among others, for suggestions).

7 Conclusion

The figures from the Spoken Corpus lend some quantitative weight to Leech and Short's (1981) claims concerning the presentational norms for the speech and thought scales (DS and IT respectively), and suggest in addition that the quantitative norm for the presentation of writing in the two corpora appears to be (N)RWA (though in semantic terms the Leech and Short argument for DS as the speech presentation norm would also seem to apply to DW writing presentation). Perhaps the main overall finding of the project so far, though, is that the model of speech, thought (and later) writing presentation suggested by Leech and Short (1981) and developed by Short, Semino and Wynne in their work on the Written Corpus is itself applicable to spoken data, with few modifications. Our work on the Spoken Corpus to date would seem to confirm the robustness of the model of SW&TP that we are using.

The next stage in our research will be to carry out quantitative analyses of the various different SW&TP categories in the Spoken Corpus, in order to determine the distribution of the various category features appended to the main tags, and the significance of any correlations that emerge. In addition we will begin the qualitative analysis of the corpus in order to try and explain more fully our statistical findings. The demographic data included in the individual file headers also allows for the possibility of examining the distribution of categories according to the sex of the respondent. In addition, our choice of data will also allow us to consider the differences between SW&TP in elicited and spontaneous dialogue, and in story and non-story text, thus providing further insights into the nature of SW&TP in spoken language.

Notes

1. The research presented in this paper was supported by a grant from the *Arts and Humanities Research Board* (B/B/RG/AN2314/APN12482). The authors are also grateful to Andrew Hardie and Scott Piao for their help with some technical difficulties, and to Damon Berridge, of the Centre for Applied Statistics at Lancaster University, for assistance with the statistical analyses.

2. Readers of Short, Semino et al.'s earlier work on SW&TP in written narratives will notice that we have changed the order of the letters of the acronym in this article. Previously we referred to ST&WP rather than SW&TP. This change has come about partly as a result of our work on the corpus of speech and partly because of recent work on our earlier corpus of written narratives. We have made the change because our research suggests that the speech and writing presentation scales are rather close together in terms of how they operate, whereas the thought presentation scale is rather different. The alignment of the forms and functions of the various presentational categories on the speech presentation and writing presentation scales is fairly close (although not excact). This is because both scales are related to the presentation of independently observable linguistic communication. Thought, on the other hand is not independently observable, not communicative in the way that speech and writing are, and is at most only partly realised through linguistic means. Hence the thought presentation scale is based on a rather inexact analogy with the speech and writing presentation scales, and this means that the forms and functions of the categories on that scale will differ widely from those on the other two scales. This issue is discussed in some detail in Semino and Short (forthcoming: Chapters 3, 6 and 9) and will be the subject of a paper which is currently in preparation.
3. Emeritus Reader in History, Lancaster University.
4. Professor in the departments of History and Political Science at Illinois State University.
5. Professor of Modern History at the University of Manchester.
6. For a full definition of the linguistic criteria of each category, see Wynne et al. (1998) and Semino and Short (forthcoming).
7. In the SW&TP Written corpus, a 'p' was initially added to the NRSA, NRTA and NRWA tags to form a sub-category, indicating a speech, thought or writing act with an extended topic expressed through an extensive noun phrase or participial phrase. This was largely in response to the frequent occurrence of such extended topics in journalistic prose. While the early publications arising from the written corpus use a capital P in the relevant acronyms (NRSAP, NRTAP, NRWA), Semino and Short (forthcoming, Chapter 3, Section 3.2.1, Note 2) point out that the p suffix is better seen as indicating variants of the larger categories, and therefore use a lower case p suffix (NRSAp, NRTAp, NRWAp). In the Spoken corpus, topics are most frequently indexed by short phrases or anaphoric pronouns. This provides further evidence that the p does not mark categorial status. As a consequence we treat p simply as an additional feature, represented by a lower

case letter that can also be appended to other categories of SW&TP. In particular, we have encountered topics attached to minimal forms of SW&TP, such as RVp.

8. The analysis of the written corpus provided support for Short's (1988) proposal that FDS should be seen as a variant of DS, and also suggested that the same applies to FDT and FDW (see Semino et al. 1997 and Semino and Short forthcoming: Chapter 6, Section 6.5.2).
9. Because it proved difficult to find labels that would accurately summarise the variety of values allowable for each of the attributes, we prefer to use numeric identifiers.
10. At the time of writing, the Spoken Corpus has not yet been fully double-checked and revised by the project team. The figures for the Spoken Corpus are therefore still provisional. Nevertheless, we do not expect our further analytical revisions to result in major changes to the figures we present here.
11. Ambiguous categories currently comprise 13.38 per cent of the total, of which 7.63 per cent are ambiguous with A, and 5.75 per cent represent ambiguities either between categories on one scale or across scales. In the process of doing final checks with the sound files, these figures are likely to reduce. However, a considerable amount of ambiguity is to be expected, given the nature of SW&TP phenomena (see Short et al. 1996; Semino et al. 1997).
12. In Tables 3–5, where the frequencies for the Free Direct and Direct categories are calculated separately, the only comparison between the corpora which is not statistically significant is that for DS. When the categories are combined, however, the difference in frequency of FDS/DS between the spoken and written data becomes statistically significant: the result of a chi-square test run on the combined number of occurrences for DS/FDS in the Written Corpus (2,974 instances) vs. the Spoken Corpus (2,043) is 173.77.

References

Banfield, Anne. 1973. Narrative style and the grammar of direct and indirect speech. *Foundations of Language* 10: 1–39.

Baynham, Mike. 1996. Direct speech: What's it doing in non-narrative discourse? *Journal of Pragmatics* 25: 61–81.

Baynham, Mike and Stef Slembrouck. 1999. Speech representation and institutional discourse. *Text* 19.4: 439–57.

Buttny, Richard. 1997. Reported speech in talking race on campus. *Human Communication Research* 23.4: 477–506.

Clark, Herbert H. and Richard J. Gerrig. 1990. Quotations as demonstrations. *Language* 66: 764–805.

Cohn, Dorrit. 1978. *Transparent minds: Narrative modes for presenting consciousness in fiction*. Princeton NJ: Princeton University Press.

Fludernik, Monika. 1993. *The fictions of language and the languages of fiction: The linguistic representation of speech and consciousness*. London: Routledge.

Hall, Christopher, Srikant Sarangi and Stef Slembrouck. 1999. Speech representation and the categorization of the client in social work discourse. *Text* 19.4: 539–70.

Halliday, M. A. K. 1994. *An introduction to functional grammar.* 2nd edition. London: Edward Arnold.

Holt, Elizabeth. 1999. Just gassing: An analysis of direct reported speech in a conversation between employees of a gas supply company. *Text* 19.4: 505–37.

Leech, Geoffrey N. and Michael H. Short. 1981. *Style in fiction*. London: Longman.

McHale, Brian. 1978. Free indirect discourse: A survey of recent accounts. *Poetics and Theory of Literature* 3: 235–87.

Myers, Greg. 1999. Unspoken speech: Hypothetical reported discourse and the rhetoric of everyday talk. *Text* 19.4: 571–90.

Oakes, Michael. 1998. *Statistics for corpus linguistics*. Edinburgh: Edinburgh University Press.

Ravotas, Doris and Carol Berkenkotter. 1998. Voices in the text: The uses of reported speech in a psychotherapist's notes and initial assessments. *Text* 18.2: 211–39.

Semino, Elena and Mick Short. Forthcoming. *Corpus stylistics: Speech, writing and thought presentation in a corpus of English writing*. London: Routledge.

Semino, Elena, Mick Short and Jonathan Culpeper. 1997. Using a corpus to test and refine a model of speech and thought presentation. *Poetics* 25: 17–43.

Semino, Elena, Mick Short and Martin Wynne. 1999. Hypothetical words and thoughts in contemporary British narratives. *Narrative* 73: 307–34.

Short, Mick. 1988. Speech presentation, the novel and the press. In W. van Peer (ed.). *The taming of the text*, 61–81. London: Routledge.

Short, Mick. 2003. A corpus-based approach to speech, thought and writing presentation. In A. Wilson, P. Rayson and A. McEnery (eds.). *Corpus linguistics by the Lune: A festschrift for Geoffrey Leech*, 241–71. Frankfurt: Peter Lang.

Short, Mick, Elena Semino and Jonathan Culpeper. 1996. Using a corpus for stylistics research: Speech and thought presentation. In M. Short and J. Thomas (eds.). *Using corpora in language research*, 110–31. London: Longman.

Short, Mick, Elena Semino and Martin Wynne. 2002. Revisiting the notion of faithfulness in discourse presentation using a corpus approach. *Language and Literature* 114: 325–55.

Short, Mick, Martin Wynne and Elena Semino. 1999. Reading reports: Discourse presentation in a corpus of narratives, with special reference to news reports. In H. J. Diller and E. O. Gert Stratmann (eds.). *English via various media*, 39–66. Heidelberg: Universitätsverlag C. Winter.

Sperberg-McQueen, C. M. and Lou Burnard (eds.). 2001. *TEI P4: Guidelines for electronic text encoding and interchange*. Oxford-Providence-Charlottesville-Bergen: TEI Consortium.

Thompson, Geoff. 1996. Voices in the text: Discourse perspectives on language reports. *Applied Linguistics* 174: 501–30.

Wynne, Martin, Mick Short and Elena Semino. 1998. A corpus-based investigation of speech, thought and writing presentation in English narrative texts. In A. Renouf (ed.). *Explorations in corpus linguistics*, 231–45. Amsterdam: Rodopi.