The Construction of a Corpus to Investigate the Presentation of Speech, Thought and Writing in Written and Spoken British English[1]

Dan McIntyre, Carol Bellard-Thomson[2], John Heywood,
Tony McEnery, Elena Semino and Mick Short,
Lancaster University, UK

Abstract
In this paper we describe the Lancaster Speech, Thought and Writing Presentation (ST&WP) 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 ST&WP in written texts. We describe the structure of the corpus, the archives from which its composite texts are taken, the decisions that we made concerning the selection of suitable extracts from the archives, and the problems associated with the original archived transcripts. We then move on to consider issues surrounding the mark-up of our data with TEI-conformant SGML, and explain the tagging format we adopted in annotating our data for ST&WP.

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 (Buttny 1997; Thompson 1996; 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 fig. 1).

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

Fig. 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 forthcoming 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, thought and writing presentation (henceforth ST&WP) in spoken, as opposed to written, data. To this end we have constructed the

[1] The research presented in this paper was supported by a grant from the *Arts and Humanities Research Board* (B/B/RG/AN2314/APN12482). We are also grateful to Andrew Hardie and Scott Piao for their help with some technical difficulties.
[2] Carol Bellard-Thomson now works in the School of Languages at the University of Kent, UK.

Lancaster Speech, Thought and Writing 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 ST&WP

2.1 The Lancaster Speech, Thought and Writing Presentation Written Corpus

The Lancaster Speech, Thought and Writing Presentation Written Corpus was built to investigate the nature of ST&WP in written narrative texts. The ST&WP Written Corpus project extended the boundaries of investigation beyond the focus on literary texts in Leech & 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 Semino *et al*. 1999, Short *et al*. 1999 and Wynne *et al*. 1998 for more details).

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

Work on the ST&WP 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 ST&WP in speech has tended to concentrate purely on direct speech, 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 ST&WP 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 ST&WP 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 ST&WP it contains).

The Spoken Corpus is approximately 260,000 words in order to make it comparable in size with the existing ST&WP 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 ST&WP. 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

[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.

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.

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.

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 ST&WP. So, in addition to the reporting verbs picked up in the electronic search, we also looked for further examples of ST&WP 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. In some cases, then, we had to newly transcribe stretches of interaction that had been omitted or simply summarised. We removed anomalous punctuations and corrected misspellings.

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 wav files. We recorded in mono, at 16-bit resolution, in order that the resulting wav 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 ST&WP

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, thought and writing 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 format that we used in annotating our data, it is useful to summarise briefly the categories of ST&WP that we used in analysis. We begin by presenting the category 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 ST&WP categories in the Written and Spoken Corpus projects

Table 1 details the acronyms used to mark instances of ST&WP in the Written Corpus project and their equivalents in the Spoken Corpus

Categories outside the discourse presentation clines			
Written Corpus		*Spoken Corpus*	
Category	*Definition*	*Category*	*Definition*
N	Narration	A	Anything other than ST&WP (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

Table 1 Categories in the ST&WP Written corpus and their equivalents in the Spoken Corpus[6]

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 ST&WP category labels are written in upper-case letters. The ST&WP 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.

5.2 ST&WP 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. Short and Semino (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.

[6] For a full definition of the linguistic criteria of each category, see Wynne *et al*. (1998) and Semino and Short (forthcoming).

- 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 ST&WP, is replaced in the Spoken corpus by A, which, simply standing for '[A]nything other than ST&WP', 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 ST&WP in an effort to provide more information about the forms of ST&WP in our corpus. We assume the default grammatical structure of a stretch of ST&WP 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, thought and writing 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 ST&WP 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 ST&WP

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 function. 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.*

[7] In the early publications arising from the Written Corpus we use a capital 'P' in the acronyms NR{S/T/W}AP to indicate a speech, thought or writing act with an extended topic. We now prefer to use a lowercase 'p' since NR{S/T/W}AP is not a category in its own right, but simply a category variant.

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*

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)
4. 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)
5. Look, they've stuck a sticker in the back *[FDW] cars kill trees*

Free Direct Thought (FDT)
6. 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)
7. [RS] Father said [DS] can my girls come? *[FIS] No they couldn't come*

Free Indirect Writing (FIW)
8. [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)
9. [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)
10. [RS] he said *[IS] it made him happy*

Indirect Writing (IW)
11. [RWr] it was put down on in a book *[IWr] that you'd taken a pair of stockings home*

Indirect Thought (IT)
12. [RT] he thought *[IT] there was nowhere else*

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

[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).

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)
13. [RSA] I just threatened them.

Representation of Writing Act (RWA)
14. [RWA] Vivian voted Conservative

Representation of Thought Act (RTA)
15. [A] I just move some of this stuff out the way, I know, *[RTA] I've had a good idea, a smart idea*

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.

16. I was sitting there *[RV] talking* [A] and they had a drop of wine

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.

17. they had slates *[RNr] and they used to write with a piece of slate*

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

18. [RI] I was frightened to death of him I was really I was frightened to death of 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)
19. [RS] Mrs Hall said [DS] I don't know how you find time to go to your church every morning like this.

Report of writing (RW)
20. [RW] across the certificate he wrote [DW] this man should be in bed

Report of thought (RT)
21. [RT] I decided [IT] I'd like to be an engineer

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.

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

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

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

5.2.2 ST&WP 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.

24. [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.

25. 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 ST&WP category is embedded discoursally, but not necessarily syntactically, in another.

26. [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.

27. [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.

28. 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 ST&WP that does not present an anterior discourse but "refers" to an event that has not (or not yet) taken place.

29. [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.

30. 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 ST&WP and which does not count as a straightforward example of direct speech.

31. [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 ST&WP.

32. 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.

33. [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.

34. [RS] He said er *[DS] [RSep] Tell your mammy [DSep] it'll be alright* [A] and we turned back home [NWRS 177]

u (= unfinished)
The suffix u signals that the relevant ST&WP category is unfinished.

35. [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 5.3).

Level of Embedding
36. [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*

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

5.3 The tagging format
We use the element <sptag> to mark instances of ST&WP. Each constituent of the ST&WP categories are marked within one of fifteen <sptag> attributes. We use 'x' as a placeholder for those slots that are not filled for a particular ST&WP category. This is done for ease of concordancing. We use an end tag (</sptag>) to mark the end of a particular stretch of ST&WP. Below is an example of an ST&WP 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:

Attribute	Allowable values	Definitions
One	x A F	Anything other than ST&WP; 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 ST&WP
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

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

6. Conclusion

Our project, so far, has demonstrated 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 applicable to spoken data, with few modifications. Our work on the Spoken Corpus would seem to confirm the robustness of the model of ST&WP that we are using.

We are currently carrying out quantitative analyses of the various different ST&WP categories in the Spoken Corpus in order to determine the distribution and frequency of these. We aim to compare our quantitative findings with those for the Written Corpus, in order to consider any differences in ST&WP between the spoken and written data. In addition we are also carrying out qualitative analyses of the corpus data in order to try and explain more fully our statistical findings. We will report on these issues in future publications.

References

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

Baynham, M, Slembrouck, S 1999 Speech representation and institutional discourse. *Text* 194: 439-57.

Buttny, R 1997 Reported speech in talking race on campus. *Human Communication Research* 234: 477-506.

Clark, H H, Gerrig, R J 1990 Quotations as demonstrations. *Language* 66: 764-805.

Fludernik, M 1993 *The Fictions of Language and the Languages of Fiction: The Linguistic Representation of Speech and Consciousness*. London, Routledge.

Hall, C, Sarangi, S, Slembrouck, S 1999 Speech representation and the categorization of the client in social work discourse. *Text* 194: 539-70.

Holt, E 1999 Just gassing: an analysis of direct reported speech in a conversation between employees of a gas supply company. *Text* 194: 505-37.

Leech, G N, Short, M H 1981 *Style in Fiction*. London, Longman.

McHale, B 1978 Free indirect discourse: a survey of recent accounts. *Poetics and Theory of Literature* 3: 235-87.

Myers, G 1999 Unspoken speech: hypothetical reported discourse and the rhetoric of everyday talk. *Text* 194: 571-90.

Ravotas, D, Berkenkotter, C 1998 Voices in the text: the uses of reported speech in a psychotherapists notes and initial assessments. *Text* 182: 211-39.

Semino, E, Short, M forthcoming *Corpus Stylistics: A Corpus-based Study of Speech, Thought and Writing Presentation in Fictional and Non-fictional Narratives* provisional title. London, Routledge.

Semino, E, Short, M, Culpeper, J 1997 Using a corpus to test and refine a model of speech and thought presentation. *Poetics* 25: 17-43.

Semino, E, Short, M, Wynne, M 1999 Hypothetical words and thoughts in contemporary British narratives. *Narrative* 73: 307-34.

Short, M 1988 Speech presentation, the novel and the press. In van Peer, W (ed), *The Taming of the Text*. London/New York, Routledge, pp 61-81.

Short, M forthcoming A corpus-based approach to speech, thought and writing presentation. in Wilson, A, Rayson, P, McEnery, A (eds), *Corpus Linguistics by the Lune: A Festschrift for Geoffrey Leech*. Frankfurt/Main, Peter Lang.

Short, M, Semino, E, Culpeper, J 1996 Using a corpus for stylistics research: speech and thought presentation. In Short, M, Thomas, J (eds), *Using Corpora in Language Research*. London, Longman, pp 110-31.

Short, M, Semino, E, Wynne, M 2002 Revisiting the notion of faithfulness in discourse presentation using a corpus approach. *Language and Literature* 114: 325-55.

Short, M, Wynne, M, Semino, E 1999 Reading reports: discourse presentation in a corpus of narratives, with special reference to news reports. In Diller H J, Gert Stratmann, E O (eds), *English via Various Media*. Heidelberg, Universitatsverlag C Winter, pp 39-66.

Sperberg-McQueen, C M, Burnard, L (eds) 2001 *TEI P4: Guidelines for Electronic Text Encoding and Interchange*. Oxford-Providence-Charlottesville-Bergen, The TEI Consortium.

Thompson, G 1996 Voices in the text: discourse perspectives on language reports. *Applied Linguistics* 174: 501-30.

Wynne, M, Short, M, Semino, E 1998 A corpus-based investigation of speech, thought and writing presentation in English narrative texts. In Renouf, A (ed), *Explorations in Corpus Linguistics*. Amsterdam, Rodopi, pp 231-45.