Annotating a corpus of Early Modern English writing for categories of discourse presentation

Dan McIntyre and Brian Walker

University of Huddersfield, UK

1 Introduction

This article discusses the process of annotating a small corpus of Early Modern English writing that we have constructed in order to investigate the diachronic development of speech, writing and thought presentation. The work we have done so far is a pilot investigation for a planned larger project. We have constructed a corpus of approximately 40,000 words of Early Modern English (EModE) fiction and news journalism and annotated it for categories of discourse presentation (DP) drawn from a model originally proposed by Leech and Short (1981). This has allowed us to quantify the types of discourse presentation within the corpus and to compare our findings against those from a similarly annotated corpus of Present Day English (PDE) writing (reported in Semino and Short 2004). Our results so far appear to indicate developing stylistic tendencies in fiction and news texts in the Early Modern period, and suggest that it would be profitable to extend the project through the construction of a larger corpus incorporating a greater number of text-types in order to test our hypotheses more rigorously. In this article we concentrate specifically on describing the annotation phase of the project. We discuss the criteria by which we defined the various discourse presentation categories in order to make clear our analytical methodology, as well as the issues we were confronted with in

trying to annotate in a systematic and retrievable way. We conclude with some preliminary results to illustrate the value of this kind of annotation and suggest some hypotheses resulting from this pilot investigation.

2 Discourse presentation

Prototypically, discourse presentation (also known as speech, writing and thought presentation – or SW&TP) refers to the presentation in a posterior discourse context of speech, writing or thought from an anterior discourse context. A person may report the speech, writing and/or thoughts of a third party using a variety of different forms. Hence, the original utterance 'I love corpus linguistics!' may be reported by a third party using any of the following structures (see Table 1 for a description of the categories).

(i) 'I love corpus linguistics!'

(ii) 'I love corpus linguistics!' he said.

(iii) He said that he loved corpus linguistics.

(iv) He loved corpus linguistics!

(v) He expressed his enjoyment.

(vi) He spoke loudly.

Each of the above forms expresses varying degrees of what Leech and Short (1981, 2007) have termed narrator interference, as well as decreasing claims to faithfulness with regard to the reporting of the original utterance. Example (i) expresses the exact words of the original utterance; (ii) includes the exact words plus a reporting clause indicating the

presence of a narrator; (iii) presents the original utterance in an indirect form, with the original speaker's words contained within a subordinate clause, having been subjected to a backshift in tense; (iv) is a free indirect rendering that blends aspects of a narratorial report with a flavour of the original speaker's utterance (in this case, the exclamation mark); (v) reports only the speech act of the original speaker (none of the propositional content of the original utterance can be reconstructed); (vi) reports only the fact that speech occurred.

Examples (i) to (vi) constitute speech presentation, though the same principles apply to the presentation of a third party's writing or thoughts. Discourse presentation can also refer to the presentation of speech, writing or thought in some future discourse context. For example, *He's about to say how much he loves corpus linguistics*. Table 1 shows the categories of discourse presentation that we used in our project:

Speech presentation		**Writing presentation**		**Thought presentation**	
(F)DS	(Free) Direct Speech	(F)DW	Free Direct Writing	(F)DT	Free Direct Thought
FIS	Free Indirect Speech	FIW	Free Indirect Writing	FIT	Free Indirect Thought
IS	Indirect Speech	IW	Indirect Writing	IT	Indirect Thought
NRSA	Narrator's (Re)presentation of a Speech Act	NRWA	Narrator's (Re)presentation of a Writing Act	NRTA	Narrator's (Re)presentation of a Thought Act
NV	Narrator's	NW	Narrator's	NT	Narrator's

	Presentation of Voice		Presentation of Writing		Presentation of Thought
				NI	Internal Narration
NRS	Narrator's Report of Speech	NRW	Narrator's Report of Writing	NRT	Narrator's Report of Thought
N	Narration	N	Narration	N	Narration

Table 1 Speech, writing and thought presentation model based on the description in Short 2007

The categories in Table 1 are those described in Short (2007), itself a development of those presented originally in Leech and Short (1981) and later expanded by Short and a project team at Lancaster University. The model proposed by Leech and Short (1981) suggests that with each move along the cline of discourse categories comes an increased claim to faithfulness with regard to the reporting of the original discourse. (The only deviation from this in the model concerns the Free Direct (FDS/W/T) and Direct (DS/W/T) categories, which are conflated because they represent the same degree of faithfulness to the original). In later conceptions of the model, Short *et al*. dispensed with the notion of discourse report, preferring instead to describe the phenomenon as discourse presentation, as a result of the fact that hypothetical and forward-facing discourse presentation does not involve the report or representation of something already said, written or thought. Nonetheless, the 'R' element (for 'report' or 'representation') has been retained in favour of 'P' (for 'presentation') in some of the acronyms in Table 1 to

avoid confusion with earlier publications on the subject. While we occasionally refer to discourse report or representation, this is only to avoid the confusion that might arise from changing the acronyms in Table 1, and all such references should be taken as referring to the *presentation* of discourse.

Discourse presentation as a linguistic phenomenon has been studied from a wide range of academic perspectives, including philosophy (Clark and Gerrig 1990), applied linguistics (Baynham and Slembrouck 1999, Myers 1999), sociology (Holt 1999, Holt and Clift 2006) and psychology (Ravotas and Berkenkotter 1998). Our interest in the phenomenon relates to its stylistic import, hence our use of a discourse presentation model developed from research in linguistic stylistics. Our interest in the diachronic development of the phenomenon is what prompted our study of Early Modern English discourse presentation. Our choice of this period was determined by the fact that this phase of the development of English saw the rise of a standard form of the language as well as an increase in printed texts and literacy. Since the Early Modern period was such a rich era for linguistic development, we reasoned that discourse presentation as a stylistic technique might be used differently from how it is in PDE. There has been some work on the phenomenon from a historical linguistic perspective, though none has used the methodological framework we employ here. Moore (2002), for instance, explores the phenomenon from a qualitative angle, while Jucker (2006), although taking a corpus-based approach, analyses only one text-type (news discourse) and uses an un-annotated corpus. One consequence of this is that Jucker's findings are limited by the structural forms of discourse presentation that it is possible to search for. For example, Jucker does not analyse free indirect discourse, since this is impossible to retrieve through

concordancing; free indirect discourse is only apparent by its context, not by its linguistic form. Our technique of first annotating our data means that we are able to retrieve all instances of all the categories of discourse presentation outlined in Short's (2007) model (see Table 1).

The nature of our project locates it within the growing field of corpus stylistics (see, for example, Semino and Short 2004, Mahlberg 2007, O'Halloran 2007a, 2007b), and particularly historical corpus stylistics (Studer 2008, Culpeper and Kytö 2002, 2006). Our two principle aims are to investigate the forms and functions of discourse presentation in Early Modern English writing and to provide quantitative evidence of the relative frequencies of presentational forms. Long term, a subsidiary aim is to provide a perspective on the history of English that is focused on stylistic developments and goes beyond formal levels of language, thereby contributing to the 'alternative histories of English' advocated in Watts and Trudgill (2002).

3 Corpus construction

Since our aim was to compare the forms and functions of discourse presentation in EModE with those of PDE, the sampling frame for our corpus follows the principles of the Lancaster SW&TP Written Corpus, a 260,000-word corpus of contemporary English writing annotated for the categories of speech, writing and thought presentation outlined in Table 1. The Lancaster corpus comprises equal numbers of 2000-word samples of serious and popular fiction (broadly akin to 'high' and 'low' literature), tabloid and broadsheet news journalism and biography and autobiography. The labour intensive nature of speech, writing and thought presentation annotation meant that we were unable

to construct a corpus of equivalent size, and so we restricted our text-types to just fiction and news reports (needless to say, our quantitative comparisons in section 5 are with the fiction and news sections of the Lancaster corpus only, and we carried out log-likelihood calculations to determine whether differences in tag frequencies between the two corpora were statistically significant). We have around 20,000 words of each text-type, divided equally across fifty-year segments of the Early Modern English period. In defining this time-frame we took the common consensus of historical linguists who date the period from approximately 1500 to approximately 1750, these dates being, respectively, roughly synonymous with Caxton's printing press revolution taking effect and the publication of Johnson's dictionary. (This is not to suggest that these two events had an equal impact on all varieties of English; we are primarily interested in the developing Standard English of the period, on which Caxton and Johnson clearly did have an effect). Tables 2 and 3 outline the content of the fiction and news sections of our corpus, as well as the time periods they are representative of. It is worth noting that our earliest examples of news journalism are somewhat different from PDE newspapers, since the newspaper as a text-type did not evolve until mid-way through the Early Modern period. The newspaper (as we might recognise it) did not exist at the earlier end of our time frame, and news was often in the form of letters or personal accounts, which were printed and distributed on a fairly limited basis. News pamphlets (also called Corantos, or News books) first started appearing towards the end of the 16th Century and became established in the early 17th Century. What is often regarded as the first proper newspaper, *The London Gazette*, did not appear until 1666. Our earliest samples of news journalism are therefore of the contents of letters describing newsworthy events (for example, J1.2 'Hevy news of an

earthquake'), and while the data is not absolutely equivalent to the Lancaster news data, it does afford an opportunity to gain an insight into how the genre develops across the period. A further point to note is that, unlike the Lancaster team, we did not distinguish between serious and popular fiction, since the distinction did not exist in the Early Modern period in quite the same way. Furthermore, sub-dividing our data in this way would not have been a good methodological strategy, since this would have generated raw figures too small to draw reliable conclusions from.

EModE Corpus – Prose Fiction sub-section					
Extract No.	**Period**	**Title**	**Word count**	**Author**	**Pubn Date**
PF1.1	1500-1549	The Noble History of King Ponthus	2072	Henry Watson	1511
PF1.2		The Mad Men of Gotham	2002	William Tyndale	1547
PF2.1	1550-1599	The Carde of Fancie	2154	Robert Greene	1584
PF2.2		Arcadia	2022	Philip Sydney	1590
PF3.1	1600-1649	The Blacke Booke	2057	attr. Thomas Middleton	1604
PF3.2		Cloria and Narcissus	2047	attr. Percy Herbert	1653
PF4.1	1650-1699	The blazing-world	2097	Margaret Cavendish	1668
PF4.2		Oroonoko	2073	Aphra Behn	1688
PF5.1	1700-1750	Moll Flanders	1993	Daniel Defoe	1722
PF5.2		Tom Jones	2079	Henry Fielding	1751
Total Words			20596		

Table 2 The composition of the fiction section of the EModE corpus

EModE Corpus – News Report sub-section					
Extrct No.	**Time Period**	**Title**	**Word count**	**Author**	**Pubn Date**
J1.1	1500-1549	An account of the Battle of Flodden	2200	Not Known	1513
J1.2		Hevy newes of an earthquake	825	Not Known	1542
J1.3		A copy of a letter containing certayne newes	1017	Not Known	1549
J2.1	1550-1599	The Spoyle of Antwerpe	2122	George Gascoigne	1576
J2.2		The English Mercurie	1391	Not Known	1588
J3.1	1600-1649	The weeklely Newes	1079	Not Known	1606
J3.2		The courant of newes	1386	Not Known	1620
J3.3		The marchings of Two Regiments	2101	Henry Foster	1643
J4.1	1650-1699	Every Day's Intelligence 1	1019		1653
J4.2		Every Day's Intelligence 2	1013		1653
J4.3		A true designe of the Late Eruption of Mt Etna	2170	Heneage Finch	1669
J5.1	1700-1750	The Flying Post	1107	Not Known	1700
J5.2		London Post	1184	Not Known	1700
J5.3		Country Journal	1876	Not Known	1736
		Total Words	20490		

Table 3 The composition of the news section of the EModE corpus

We collected our data from a variety of sources, using texts that were already in electronic format, as this represented a great time saving and, where possible, checked the electronic version against facsimiles of original publications of the texts. This was to make sure that the later edited version of early texts, which the electronic forms were often drawn from, did not contain, for example, extra or altered punctuation. Our sources included: Early English Books Online (EEBO); the Oxford Text Archive (OTA);

Renascence Editions (University of Oregon); Project Gutenberg; the Lampeter Corpus; the Lancaster Newsbook Corpus; the Corpus of English Dialogues 1560-1760; and a fairly new resource called The Burney Collection, which is a collection of facsimiles of newspapers available from the British Library.

4 The annotation process

The annotation scheme we used was a development of that outlined in McIntyre et al. (2004), which describes the results of a similar project to annotate a corpus of contemporary spoken English for discourse presentation. This involves the application of TEI-conformant XML mark-up that comprises an element *dptag* (*d*iscourse *p*resentation *tag*) and an attribute *cat* (*cat*egory). These are enclosed within angle brackets forming what, in shorthand reference, is called a *tag*. The *cat* attribute consists of fifteen fields into which pre-designated alphanumeric codes are entered detailing the SW&TP categories outlined in Table 1. Each field has a limited number of possible constituents and the combination of constituents from different fields allows for a detailed description of the discourse presentation being marked (these are set out in Table 4).

Field	Possible constituents	Definition of constituent
1	x N F	**N**arrator's; **N**arration; **F**ree
2	x R I D	**R**epresentation; **I**ndirect; **D**irect
3	x S T W V I M	**S**peech; **T**hought; **W**riting; **V**oice; **I**nternal state; **U**se
4	x A	**A**ct
5	x p	To**p**ic
6	x #	**#** = odd/interesting cases
7	x y	discourse summar**y**
8	x g a	**g**rammatical negative; **a**bsence of speech, thought and/or writing
9	x h	**h**ypothetical
10	x i	**i**nferred
11	x q	**q**uote
12	x r	ite**r**ative
13	x v p	interrogati**v**e; im**p**erative
14	x m	no**m**inalisation
15	x 1 2 3 4	**no.s** = DP split into sections

Table 4 Constituents of the fields of the *cat* attribute

The possible constituents designated to the first four fields relate to the major DP categories (outlined in Table 1) and are always capital letters. The constituents designated to the remaining eleven fields relate to DP sub-categories and provide further details about the DP. These are generally lower-case letters, but the hash symbol (#) and numbers are also possible in certain fields. We use *x* as a placeholder and do not mark empty positions following the final attribute value. This means that "cat" attribute

constituents always occur in the same field position, making searches of the annotated corpus for particular DP categories using computer tools more straightforward.

Below is an example of three annotated sentences from the fiction section of the corpus:

> <dptag cat="N"> Here the book dropt from her hand, and a shower of tears ran down into her bosom. In this situation she had continued a minute, when the door opened, and in came Lord Fellamar. Sophia started from her chair at his entrance; </dptag> <dptag cat="NRS"> and his lordship advancing forwards, and making a low bow, said, </dptag> <dptag cat="xDS"> "I am afraid, Miss Western, I break in upon you abruptly." </dptag>

The example shows that the code to mark-up direct speech is 'xDS': the "cat" fields that contain the constituents for direct speech are 2 and 3; field 1 is not required, so is filled with an 'x'; fields 4 to 15 are left blank. Notice that in the example reporting clause (NRS) is tagged. Our annotation policy for this study was to also tag narration and narration phenomena (such as reporting clauses) as well as DP, since this often impacts on the stylistic effect of the DP, as we will show in section 5. It is also the case that our example shows that for every tag that marks the start of a new section of DP or narration phenomena, there is also an end tag (</dptag> in our case) which marks the end of that stretch of DP or narration.

We also indicate instances of embedded discourse presentation using an *e[n]dptag*, where *e* stands for 'embedded' and *n* indicates the level of discourse embedding. An example of this can be seen below:

<dptag cat="xDS"> So yelde I me to you & in to your pryson as your knyght & ye to haue power to doo as of your owne

<e1dptag cat="NRS">& yet he bad me</e1dptag>

<e1dptag cat="xIS">yt I sholde salewe you from hym.</e1dptag>

</dptag>

The example immediately above shows an instance of one level of discourse embedding. More are possible though it is rare to find instances beyond three levels.

It should be recognised that ambiguity is a large part of discourse presentation and we marked this by using ambiguous tags. Consider the following example from the fiction section of the corpus:

But in the other Chappel lined with the Star- stone, she preached Sermons of Comfort to those <u>that repented of their sins</u> […]

Focusing just on the underlined section of the extract, it is unclear whether repenting, in this case, involved a speech act ('I repent of my sins') or some sort of thought process or act, or both. The case, therefore, is genuinely ambiguous, and our annotation scheme reflects this by using a *cat2* attribute to mark an alternative analysis. The *cat2* attribute

follows exactly the same format as the *cat* attribute. Thus, the resulting tagging format for the above example is:

<dptag cat="N"> But in the other Chappel lined with the Star- stone, </dptag>

<dptag cat="NRSAp"> she preached Sermons of Comfort </dptag>

<dptag cat="NRSA" cat2="NRTA"> to those that repented of their sins […] </dptag>

While we do not discuss ambiguous examples in this article, we will return to this issue in future research.

Annotating in this way has a number of methodological advantages. For instance, it forces the analyst to be clear about what constitutes a particular category of discourse presentation. As far as possible, we tagged on the basis of linguistic form (e.g. indirect discourse presentation always involves two clauses, while the NR{S/W/T}A category was used for one clause structures; we discuss this in greater detail below), though we recognised that context often plays a role in determining a particular structure (e.g. free indirect forms). All the texts in the corpus were tagged initially by one of us and then checked by the other and revised in the light of our discussions. A further advantage of this approach is that as our tagging progressed, we were able to revise decisions made earlier in the project on the basis of our increasing experience of identifying the various discourse presentation structures. Annotating also enables the retrieval of problematic structures, for discussion at a later date – for example, ambiguous cases.

In Table 5 we give an example from our corpus, where one exists, of each of the categories set out in Table 1. In each example the DP is underlined. Following this, we discuss the criteria we employed in assigning stretches of text to particular discourse presentation categories.

Discourse presentation					
Speech presentation		**Writing presentation**		**Thought presentation**	
(F)DS	Whan he did com home to his house his wife sayd, where is my Brandiron or trefete.	(F)DW	he began and wrote again -- 'Be mine, with all your poverty.'	(F)DT	'Very well,' thought I;
FIS	the rogues presented each a pistol to them, and bid them deliver, or they would blow the brains out of their heads;	FIW	No occurrence in the corpus	FIT	but Dedalus finding, he could not build his determinations upon these uncertainties, wherein both the safety of the Towne and his own honour, might probably suffer, by reason of the protraction as also the person of the princesse Cloria be

					endangered by his slownesse and neglect,
IS	the Princesse told her, that she had beene lately troubled with a most untoward and fearfull dreame,	IW	Middleton also writes to them out of Holland, that Colonel Dezmond was shipped away …	IT	and he shou'd have been entirely comforted, but for the Thought that she was possess'd by his Grand-father.
NRSA	the wynde also began to blow agayne: wherfore we were glad and lauded and thanked god	NRWA	and comytted unto hym the same by Instruccyon sygned and subscrybed with his owne hande	NRTA	All this, you may be sure, was as I wished,
				NI	his arguments and divisions being so many, that they caused a great confusion in his brain
NV	My lord then made another and a longer speech of the same	NW	The late Parliament having upon their dissolution delivered	NT	filled her imagination with some unprofitable thoughts

	sort		up the Power which they received from his Excellency at their first sitting, by a Writing under their Hands and Seal		

Table 5 EModE corpus examples of SW&TP categories

Wherever possible, we used linguistic form to guide our tagging. Having clear criteria for deciding between one DP category and another was particularly important for direct and indirect discourse forms. The example for (F)DS in Table 5 shows that quotation marks were not always used to mark direct speech. However, the example is clearly one of direct speech because there is (i) a reporting clause that introduces the speech; (ii) a shift to present tense; and (iii) a shift in deixis that is appropriate to the original speaker, marked by the pronoun. Indirect discourse consists of a reporting clause and a subordinate reported clause, along with a corresponding back-shift in tense. The important criteria here is that the reported discourse must be in a separate clause, which can be finite or non-finite. Table 5 shows a prototypical example involving a reporting clause and a subordinate reported clause signalled by the subordinating conjunction *that*. Non-prototypical but fairly common forms are those where the subordinating conjunction is elided, for example:

<dptag cat="NRT"> for now she fear'd </dptag>

<dptag cat="xIT"> the Storm wou'd fall on the Prince; </dptag>

Additionally, the reported clause can also be marked by an infinitive verb (underlined below), for example:

<dptag cat="NRS"> I chargde them </dptag>

<dptag cat="xIS"> to stay and watch the house belowe, </dptag>

Without these criteria it can sometimes be difficult to distinguish between indirect discourse and Narrator's Report of Speech, Writing or Thought act (NR{S/W/T}A), particularly when there is a topic specified, as the following two examples demonstrate:

<dptag cat="N"> and hearing some one sighing in the other Room, she pass'd on, and found the Prince in that deplorable Condition, </dptag>
<dptag cat="NRTAp"> which she thought needed her Aid: </dptag>

<dptag cat="NRSAp"> when presently I demaunded of this Leiuetenant the place of his abode, and when hee last heard of him </dptag>

The above examples demonstrate the use of a DP sub-category *p* to indicate topic. Using the formal criteria described above helps to distinguish between propositional content and topic, which can sometimes be problematic.

While we endeavoured to tag on form, some DP categories, particularly free indirect examples, also require consideration of the wider context. The examples of FIS

and FIT in Table 5 (no examples of FIW occur in our corpus) demonstrate this. The FIS has tense and pronouns appropriate to an indirect form, but there is no reporting clause introducing the discourse; hence this was distinguished on formal grounds. The FIT, however, is a more difficult case. We tagged this as free indirect thought because the preceding clauses introduce Dedalus' internal state and indicate his thought process. Consequently, we decided that the underlined section of the example is not simply narration but relates to the propositional content of his thoughts, containing some flavour of the original discourse.

5 Results

Analysis of the corpus is ongoing and here we present some of our initial quantitative findings, along with some qualitative analysis, in order to demonstrate the usefulness of the kind of annotation we have undertaken.

Figure 1 shows the overall distribution of speech, writing and thought presentation in both the EModE and PDE data:

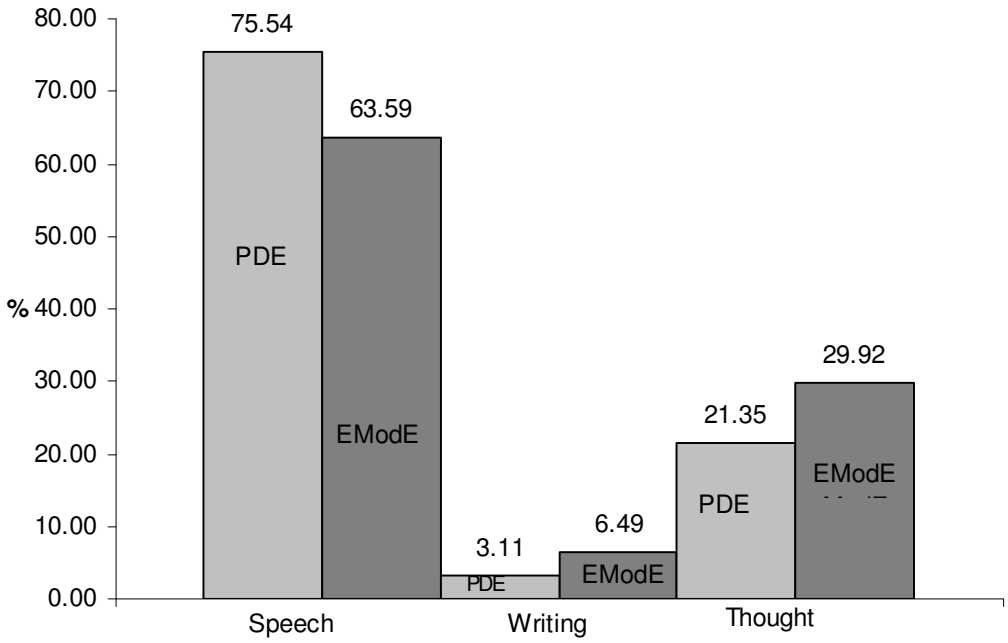

Figure 1 Overall distribution of speech, writing and thought presentation in the EModE and PDE data

What we can observe from this graph is that the overall distribution of discourse presentation in the EModE data follows that of the PDE corpus; that is, speech presentation dominates, followed by thought and writing presentation. While the histogram suggest that there is more thought and writing presentation in the EModE data than in PDE, log-likelihood tests show that these are not significant differences. Initially then, the distribution of SW&TP in EModE is the same as for PDE. However, we begin to see differences when we consider the distribution of individual categories on the speech, writing and thought presentation clines. Tables 6, 7 and 8 show, respectively, the frequency of categories of speech, writing and thought presentation in the corpus

compared against the PDE data. The percentages are based on number of tags as opposed to number of words within each category. We provide a fuller discussion of these results in McIntyre and Walker (forthcoming) and here concentrate particularly on what the results reveal about writing presentation. Log-likelihood (LL) figures in bold indicate statistically significant differences between the two corpora.

Category	**PDE**				**EModE**				**LL**
	No. of tags	**% of total**	**% of cline**	**Rank**	**Tag Freq**	**% of all DP**	**% of cline**	**Rank**	
(F)DS	2339	40.38	53.45	1	275	27.89	43.87	1	**37.12**
FIS	90	1.55	2.06	5	11	1.12	1.75	4	1.17
IS	784	13.53	17.92	3	120	12.17	19.14	2	1.20
NRSA	918	15.85	20.98	2	109	11.05	17.38	4	13.89
NV	245	4.23	5.60	4	112	11.36	17.86	3	**64.73**
Totals	4376	75.54	100.00		627	63.59	100.00		5.79

Table 6 Frequencies of instances of speech presentation (number of speech tags) and rank orderings

Category	PDE				EModE Corpus				LL
	No. of tags	% of all DP	% of cline	Rank	Tag Freq	% of all DP	% of cline	Rank	
(F)DW	43	0.74	23.89	2	17	1.72	26.56	2	6.96
FIW	12	0.21	6.67	5	0	0.00	0.00	5	6.96
IW	30	0.52	16.67	3	36	2.64	40.63	1	**19.68**
NRWA	82	1.42	45.56	1	10	1.01	15.63	4	10.38
NW	13	0.22	7.22	4	11	1.12	17.19	3	5.26
Totals	180	3.11	100.00		74	6.49	100.00		0.59

Table 7 Frequencies of instances of writing presentation (number of writing tags) and rank orderings

Category	PDE				EModE Corpus				LL
	No. of tags	% of all DP	% of cline	Rank	Tag Freq	% of all DP	% of cline	Rank	
(F)DT	84	1.45	6.79	4	5	0.51	1.70	5	15.00
FIT	230	3.97	18.59	2	1	0.10	0.34	6	**51.08**
IT	119	2.05	9.62	3	66	6.69	22.37	2	**45.13**
NRTA	71	1.23	5.74	5	39	3.96	13.22	3	**28.39**
NT					11	1.12	3.73	4	4.38
NI	733	12.65	59.26	1	173	17.55	58.64	1	1.69
Totals	1237	21.35	100.00		295	29.92	100.00		12.34

Table 8 Frequencies of instances of thought presentation (number of thought tags) and rank orderings

If we focus on writing presentation in EModE, it is clear that the foregrounded category is Indirect Writing. This is statistically over-used in the EModE data when compared to the PDE data. The beginnings of one potential explanation for this can be found in the fact that the majority of the Indirect Writing presentation in the EModE corpus (34 out of 36 examples) occurs in the news journalism data. Below is a concordance of all the instances of indirect writing presentation in this sub-section of the corpus:

itch; the substance of which was, **<IW>** that the charge of filling up, the fixing of posts
remarkable, they write from thence, **<IW>** that his majesty one day took three wild boars,
e for the said county, threatening, **<IW>** that in case he procceded any farther in taxing
'd on the spot. They write from Lynn, **<IW>** that on Sunday se'nnight they had such a viole
Marquis de Monti has lately wrote to the magistrates of Dantzick, **<IW>** that they may soon
the captain of which vessel reports, **<IW>** that two Maltese men of war have taken the Adm
We have like wise advice from Genoa, **<IW>** that a ship belonging to Majorca is arrived in
he has, as they write from Vienna, **<IW>** settled the succession
own for good of the publick, and the honour of that mighty empire, **<IW>** he has, as they
wound. We have an account, **<IW>** that one Mons. Munier, who has lived in England the
pril next ensuing: They also tell us, **<IW>** that the Reform of the troops, which was actua
Our letters from Tournay, of the 30th past, say, **<IW>** that an arrest of the council state
Our letters this day from Brussels say, **<IW>** that the burgers, who have fled from their h
Our Advices from Copenhagen say **<IW>** they were busy there fitting out a squadron of men
We have advice from Moscow **<IW>** that his Czarish Majesty had disbanded a great many
Our letters from Paris make mention, **<IW>** as if the Pope, who had been relapsed, were re
We have advice from Lubeck, **<IW>** that 5 ships were lately cast away on the coast of
They tell us from Stetin, **<IW>** that the Governour General Mellin had, by Placaet,
We have an account from Lysland, **<IW>** that they are busy levying a tax there, which is to
We have advice from Warsaw, **<IW>** that, pursuant to the accommodation made with the E
Last Sunday Publication was made throughout the kingdom, **<IW>** that the Month of February
k Dec 24 Our letters from Poland say, **<IW>** that Prince Alexander Sobietzki, designed to g
her her collar. The port-letters say, **<IW>** that the Mary of London was put into Plymouth,
Our acounts from most of the provinces of this kingdom say, **<IW>** that there's nothing but
ordering the landtgrave of Hess d'Armstadt **<IW>** to forbear his hostilities against the
Middleton also writes to them out of Holland, **<IW>** that Colonel Dezmond was shipped away
t a Letter to Glencarn , assuring him **<IW>** that the K. of France , and Denmark , the Duke

tters from the Hague , it is written, **<IW>** that the French Ambassador there,
ther in men nor money, desiring him **<IW>** to be with in what he formerly Promised unto the
of their Garrisons, and they wonder **<IW>** that he is able to send them no aid , neither in
unto the sayd Generall, advising him **<IW>** that the sayd Ile of Lantore did belong unto t
brought Advice into Plymouth, **<IW>** that he had descried the Spanish Armado near the Li
Capt. Fleming, who had beene ordered **<IW>** to cruize in the Chops of the Channell, for Di
certen requestes, as he termed them) **<IW>** to remedye the grieffes of the Devonshirmen,

What is particularly interesting about these examples is the reporting clause that precedes the discourse presentation. In each case, considerable emphasis is placed on identifying the source of the report that follows. Thus, we have clauses such as 'it is written', 'the port-letters say', 'they tell us from Stetin', 'we have advice from Lubeck' and 'they write from Lynn'. It appears, then, that there is a concern among EModE writers of news reports to make clear that the report of news is taken from a identifiable source, rather than being, say, conjecture on the part of the writer. That these reporting clauses should be followed by indirect writing presentation is perhaps explained by the fact that an indirect report allows for the reconstruction of the exact words of the original writer. This seems appropriate when so much emphasis is placed on accounting for the source of the story. In effect, the indirect category makes it clear that the news report is a *re*presentation of an original source, as opposed to, say, a summary report. It does this by presenting the original writer's words in an alternative format but one which also allows the reader recourse to the words and structures of the original discourse. Conboy (2007) makes the point that EModE journalism relied heavily on written reports, though he makes no suggestion as to why these reports were presented in the forms that they were:

> Throughout the eighteenth century, news often comprised the contents of letters received, conveying both opinion and information, and the language reflected the letter-writing style of the time.
>
> > We have a report here, but we hope without foundation, that his Majesty's frigate Minerva was not lost on the back of the Isle of Wight on Friday last night last, when it really blew a hurricane. (*London Evening Post*, 31 December to 3 January 1764)
>
> Newspapers depended on such reports for their own content, together with letters from readers to fill their pages. Communication and distribution technologies available at the time meant that maintaining a regular flow of news was a problem. It meant that the language of the reports which were in regular supply could be more elaborate.
>
> (Conboy 2007: 6-7)

Conboy's chosen example is similar in structural terms to those in our corpus in that the reporting clause identifies the source of the report which is then presented in an indirect form. However, when it comes to explaining this kind of structure, Conboy seems to be suggesting that the style of newspaper reports was in part due to a need to fill up space. This seems counter-intuitive, since an easier way to fill up space than establishing a more long-winded written style would have been to print in larger type. We suggest another possibility; that indirect presentational forms are more dominant in the EModE data because of a desire to be seen to *re*present the news. This would accord with the relative

absence of freedom of the press in this period, and the necessity of avoiding overtly critical comment in news writing (taking care to indicate that a report is based on information in another source, and presenting that source in such a way that the original discourse is recoverable, is one way of implicitly claiming no responsibility for the content of the report; it is also noteworthy that most of the news report is anonymous). The systematic annotation and analysis of both reported and reporting clauses allows us to note this as a pattern in the news reporting of the time. We can also note that the quantitative norm of NRWA (Narrator's Representation of Writing) for Present Day English has the function of summarising more than reporting. Indirect Writing, on the other hand, is the closest we can get to the original discourse while still allowing the reporting of this from a different viewpoint. This, we suggest, may be indicative of the developing nature of the news report genre from report to summary, in effect, a move towards the narrator end of the discourse presentation cline. We might further speculate that the use of IW in EModE news continues in a written form the word of mouth tradition from which the transmission of news grew. More research would be needed to validate these hypotheses.

6 Conclusion

This brief article is intended as a record of the decisions we made during the tagging of our pilot corpus, and as an example of what can be gained through stylistic annotation. Our pilot investigation has already generated a number of hypotheses which might be tested further in a larger project. Stylistic annotation thus offers the possibility of moving

us closer towards being able to make generalisations about stylistic development over time.

References

Baynham, M. and Slembrouck, S. (1999) 'Speech representation and institutional discourse', *Text* 19(4): 439-57.

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

Conboy, M. (2007) *The Language of the News*. London: Routledge.

Culpeper, J. and Kytö, M. (2006) '"Good, good indeed, the best that ere I heard": exploring lexical repetitions in the Corpus of English Dialogues, 1560-1760', in Taavitsainen, I., Härmä, J. and Korhonen, J. (eds) *Dialogic Language Use / Dimensions du Dialogisme / Dialogischer Sprachgebrauch* (Mémoires de la Société Néophilologique de Helsinki, Vol. 66), pp. 69-85. Helsinki: Société Néophilologique.

Culpeper, J. and Kytö, M. (2002) 'Lexical bundles in Early Modern English: a window into the speech-related language of the past', in Fanego, T., Méndez-Naya, B. and Seoane, E. (eds) *Sounds, Words, Texts, Change. Selected Papers from the Eleventh International Conference on English Historical Linguistics (11 ICEHL)*, pp. 45-63. Amsterdam: John Benjamins.

Holt, E. (1999) 'Just gassing: an analysis of direct reported speech in a conversation between employees of a gas supply company', *Text* 19(4): 505-37.

Holt, L. and Clift, R. (eds) (2006) *Reporting Talk: Reported Speech in Interaction*. Cambridge: Cambridge University Press.

Jucker, A. H. (2006) '"but 'tis believed that...": speech and thought presentation in Early English newspapers', in Brownlees, N. (ed.) *News Discourse in Early Modern Britain. Selected Papers of CHINED 2004*, pp. 105-25. Bern: Peter Lang.

Leech, G. and Short, M. (1981) *Style in Fiction*. London: Longman.

Leech, G. and Short, M. (2007) *Style in Fiction*. 2nd edn. London: Pearson Education.

Mahlberg, M. (2007) 'Clusters, key clusters and local textual functions in Dickens', *Corpora* 2(1): 1-31.

McIntyre, D., Bellard-Thomson, C., Heywood, J., McEnery, A., Semino, E. and Short, M. (2004) 'Investigating the presentation of speech, writing and thought in spoken British English: a corpus-based approach' *ICAME Journal* 28: 49-76.

McIntyre, D. and Walker, B. (forthcoming) 'Discourse presentation in Early Modern English writing: a preliminary corpus-based analysis.'

Moore, C. (2002) 'Reporting direct speech in Early Modern slander depositions', in Minkova, D. and Stockwell, R. (eds) *Studies in the History of the English Language: A Millennial Perspective*, pp. 399-416. Berlin: Mouton de Gruyter.

Myers, G. (1999) 'Unspoken speech: hypothetical reported discourse and the rhetoric of everyday talk', *Text* 19(4): 571-90.

O'Halloran, K. A (2007a) 'The subconscious in James Joyce's 'Eveline': a corpus stylistic analysis which chews on the 'Fish hook'', *Language and Literature* 16(3): 227-44.

O'Halloran, K. A. (2007b) 'Corpus-assisted literary evaluation', *Corpora* 2(1): 33-63.

Ravotas, D. and Berkenkotter, C. (1998) 'Voices in the text: the uses of reported speech in a psychotherapist's notes and initial assessments', *Text* 18(2): 211-39.

Semino, E. and Short, M. (2004) *Corpus Stylistics: Speech, Writing and Thought Presentation in a Corpus of English Writing*. London: Routledge.

Short, M. (2007) 'Thought Presentation 25 Years on', *Style* 41(2): 227-57.

Studer, P. (2008) *Historical Corpus Stylistics: Media, Technology and Change*. London: Continuum.