Introducing a corpus of conversational stories. Construction and annotation of the *Narrative Corpus*

CHRISTOPH RÜHLEMANN[1] and MATTHEW BROOK O'DONNELL[2]

[1]*University of Munich*
[2]*University of Michigan*

Abstract

Although widely seen as critical both in terms of its frequency and its social significance as a prime means of encoding and perpetuating moral stance and configuring self and identity, conversational narrative has received little attention in corpus linguistics. In this paper we describe the construction and annotation of a corpus that is intended to advance the linguistic theory of this fundamental mode of everyday social interaction: the Narrative Corpus (NC). The NC contains narratives extracted from the demographically-sampled subcorpus of the British National Corpus (BNC) (XML version). It includes more than 500 narratives, socially balanced in terms of participant sex, age, and social class.

We describe the extraction techniques, selection criteria, and sampling methods used in constructing the NC. Further, we describe four levels of annotation implemented in the corpus: speaker (social information on speakers), text (text Ids, title, type of story, type of embedding etc.), textual components (pre-|post-narrative talk, narrative, and narrative-initial|final utterances), and utterance (participation roles, quotatives and reporting modes). A brief rationale is given for each level of annotation, and possible avenues of research facilitated by the annotation are sketched out.

Keywords: *narrative, annotation, participation, textual components, discourse presentation, quotatives, XML, XPath, XQuery*

1. Introduction

Telling a story is part of the DNA of everyday talk. Ochs & Capps (2001: 54), like many others, consider it "a ubiquitous feature of ordinary conversation." The reasons are certainly complex. One aspect is that by telling stories we share experiences. As Schiffrin notes: "The stories that we tell about our own

Corpus Linguistics and Linguistic Theory 8–2 (2012), 313–350 1613-7027/12/0008–0313
DOI 10.1515/cllt-2012-0015 © Walter de Gruyter

Bereitgestellt von | Bayerische Staatsbibliothek
Angemeldet | 194.95.59.195
Heruntergeladen am | 05.11.12 19:08

and others' lives are a pervasive form of text through which we construct, interpret and share experience" (1996: 167). This view of narrative has long been at the heart of narrative enquiry in the tradition of Labov & Waletzky (1967/1997), who had their informants tell danger-of-death stories. By another view, which is more radical and which also allows the inclusion of more mundane and seemingly trivial stories, it is through narrative that we configure self and identity (e.g., Bamberg 2004: 332). This is undoubtedly achieved in intricate ways. One central way is by using stories as carriers of moral stance:

> Everyday narratives of personal experience elaborately encode and perpetuate moral worldviews. Personal narratives generally concern life incidents in which a protagonist has violated social expectations. Recounting the violation and taking a moral stance toward it provide a discursive forum for human beings to clarify, reinforce, or revise what they believe and value. (Ochs & Capps 2001: 46)

To illustrate, consider (1). In this extract, S3 tells a story of how Greg got into the embarrassing situation of being caught out at work as using condoms. Notice how in the utterances marked with arrows that follow the recount of the events, the speakers jointly attempt to come up with an explanation for Greg's behavior to reestablish his moral credibility:

(1) **"Dropped your johnnies"** (KCE-N1)
(Type: T10 [First-person personal narrative] / Embed Level: EC2)

	S3	[singing] [unclear] Oh it was so funny at work today, Greg fell off his chair.
	S1	[laugh]
	S3	Packet of condoms fell out of his pocket [unclear]
	S1	[laugh]
	S3	[laughing] And they were ripped.
	S1	[laugh]
	S3	Ah no he was, he, he wouldn't sit on his chair cos he'd just called me an arsehole and I goes oh sit down [unclear] Greg! I said sit down Gregory and shut up. So he went to sit down but his chair weren't there. All I saw were this pair of legs sticking over the desk and him going aaaaagh!
	S2	And his condoms [unclear]
	S3	And he got up and then one of the girls said hi Greg dropped your johnnies.
	UN	[unclear]
	S3	I've never seen anyone go so red in my life.
→	S2	How old's he?
→	S3	Twenty three. He's married.
→	S1	Well, so?

Bereitgestellt von | Bayerische Staatsbibliothek
Angemeldet | 194.95.59.195
Heruntergeladen am | 05.11.12 19:08

→ S3 He got married si what three months ago.
→ S1 Maybe he doesn't want any children yet.

Given the ubiquity and social significance of narrative, unsurprisingly, in discourse analysis, oral narrative is "one of the most developed areas" (Schiffrin 1984: 314). However, most analyses have traditionally been based either on very small collections of narratives (e.g., Schiffrin 1996), or narratives told by professional narrators (e.g., Leith 1995) or elicited interview-style narratives (e.g., Labov 1972; see Schegloff's 1997 critique) or narratives told in two-party interviews with the researcher being one party (e.g., Gwyn 2000). By comparison, narratives that are strictly *conversational* in the sense that they originate in natural multi-party talk among familiars have been given much less attention. Only recently has conversational narrative come more into focus (e.g., Norrick 2000; Ochs & Kapps 2001; Bamberg 2004; Georgakopoulou 2006a, b). We are aware of only one corpus linguistic approach to conversational narrative, namely Norrick's (2000) *Saarbruecken Corpus of Spoken English* (SCoSE), which is a very small non-annotated corpus. Apart from that, conversational narrative seems *terra incognita* to corpus linguists.

The *Narrative Corpus*, hereafter NC, is intended to fill this gap. This corpus is not only considerably larger (though by corpus linguistic standards still small) but also annotated on various levels to facilitate a broad range of corpus linguistic analyses. The NC is, thus, the first of its kind[1]. Given that the additional layers of annotation are intended to target discourse and pragmatic phenomena, the corpus is an addition to the small class of corpora featuring discourse and pragmatic annotation (cf. Garside et al. 1997; Leech et al. 1997). It is hoped that the use of the NC will advance the linguistic theory of narrative as a primary mode of everyday spoken interaction. This paper aims to outline the construction (Section 2) and the annotation (Section 3) of the NC. In these sections we briefly sketch out possible avenues for future research using the NC. Further, in Section 4 we describe the tools used in annotating the NC and those developed for analyzing the data in the corpus.

2. Corpus construction

2.1 *Data source and sampling*

The NC is a specialized corpus containing narratives extracted from the demographically-sampled subcorpus of the British National Corpus (BNC) (XML version). The demographically-sampled subcorpus (henceforth BNC-C) consists of roughly 4.5 million words and is often referred to as the 'conversational' subcorpus because the transcripts assembled in this subcorpus

Bereitgestellt von | Bayerische Staatsbibliothek
Angemeldet | 194.95.59.195
Heruntergeladen am | 05.11.12 19:08

"consist of casual conversations" (Aston & Burnard 1998: 28; Hoffmann et al. 2008: 32–39; see also Rayson *et al.* 1997 and Biber *et al.* 1999: 133).

We make a distinction between 'files', 'texts', and 'narrratives'. 'File' refers to the source data, that is, the 153 files that make up the BNC-C (each file contains the speech recorded by a specific respondent selected in the sampling process), 'text' to the data assembled in the NC (e.g., text KB7-N1 and KB7-N2 are two distinct texts extracted from file KB7) and, 'narrative' to the story or stories proper contained *within* the texts. Texts in the NC are invariably larger than the narratives they contain. In the corpus, texts include non-narrative stretches of talk leading up to the narratives, as well as non-narrative talk following the narrative. Another reason why texts are invariably larger than the narrative(s) they contain is that stories are frequently responded to by other thematically related stories, thus forming 'narrative chains.' We have retained these chains where possible, allowing up to three narratives in a chain.

2.2 *Data extraction*

In order to retrieve narratives from the BNC-C we primarily used lexical extraction and detailed reading of the files. Searches were conducted in the BNC-C for lexical forms that either the literature or examination of concordance lines suggested were recurrent in narratives. The list of search strings included the following items: 'it was so (funny, weird, etc.)', 'did I tell you', 'reminds me', the interjections 'bloody hell' and 'oh my god', as well as one-word items such as 'anyway', 'suddenly', 'happened', and the lemma REMEMBER. We acknowledge that any use of a search string has the potential to skew the retrieved data. To avoid this danger we did not search for discoursively significant phenomena such as quotatives. In browsing BNC-C files we drew on two observations made in conversation analytical work (e.g., Sacks 1992). First, since stories may include utterances that are longer than utterances in turn-by-turn talk we focussed on utterances of more than 15 words length, that is, a third more than average utterance length (cf. Rayson et al. 1997). Second, we looked for stretches of conversational text in which one speaker occupied every 'third' slot. This technique builds on Sacks's observation that "[f]ormally [a story] can be said to be in the first instance an attempt to control a third slot in talk, from a first" (1992: Vol. II: 18).

The extraction techniques are largely manual and could well be complemented by automatic extraction. At present, however, we are not aware of any established automatic extraction schemes available for conversational narratives. We anticipate that the systematic enquiry into narrative that the NC facilitates will in due course uncover a wealth of structural knowledge of narrative that can inform computer scripts serving to search conversational corpus data for narratives (e.g., Rühlemann & Gries in preparation).

Bereitgestellt von | Bayerische Staatsbibliothek
Angemeldet | 194.95.59.195
Heruntergeladen am | 05.11.12 19:08

2.3 *Selection criteria*

Conversational narrative is a multifaceted and elusive discourse type. This is due to its intricate integration with (non-narrative) conversation and to the broad range of narrative subgenres (cf. Norrick 2000), which range from 'big stories' more traditional research has been concerned with to what has recently been termed 'small stories' (e.g., Georgakopoulou 2006a, b). While acknowledging the typological breadth of stories, we decided for practical purposes to use two critical criteria for the inclusion of narratives in the NC.

The first criterion is the presence of 'exosituational orientation' in the discourse. This relates to linguistic evidence of the fact that stories relate sequences of events that happened in a situation remote from the present, story-telling, situation. In most cases, this exosituational orientation manifests itself in the use of past tense verbs, lexical items that have a clear past time reference (*this morning*, *yesterday*, etc.), reference to locations removed from the location of speaking and referents (typically people) not present in the telling situation. In the case of fantasies–a variety of stories characterized by a projection of future events–the exosituational orientation will typically be realized by future time reference and/or expressions of hypotheticality (e.g., conditional clauses; cf. Norrick 2000: 161). The former type of exosituational orientation, which represents the prototypical one, is exemplified in (2):

(2) S15 Last year, right, I was on holiday in the South of France and I, I'd made friends with these two German people and I'm a bit wary of German people anyway, cos of what's happened and everything. (KPG-N1)

In (2), the incipient narrator steers the discourse towards a situation distinct from the present speech situation by reference to the previous year (*last year*), to what she was doing at the time (*I was on holiday*), where the events occurred (*in the South of France*) and who else was involved (*these two German people*). That is, the scene is set sufficiently clearly for the listener to become aware that a shift in focus is sought by the speaker from the present situation in the here-and-now to a past situation in the there-and-then.

The second required criterion is the presence of what Labov & Waletzky (1967/1997) termed the 'a-then-b' relationship, that is, the use of at least two temporally junctured narrative clauses. This relationship is seen by Labov & Waletzky as the "defining characteristic of narrative" (1967/1997: 15). The a-then-b relationship is illustrated in (3):

(3) S1 yeah no, what happen was the alarm went off in their block or something and they all had to go into a room, I don't know (KPY-N1)

Bereitgestellt von | Bayerische Staatsbibliothek
Angemeldet | 194.95.59.195
Heruntergeladen am | 05.11.12 19:08

Table 1. *Narrative Corpus – basic facts and figures*

No. files	No. texts	No. narratives	No. words
143	279	531	149,520

Here two narrative clauses a and b can be isolated:

a: the alarm went off
b: they all had to go into a room

Finally, we required at least two independent researchers to agree on the narrative status of a text. Texts over whose status as narratives no consensus could be reached were excluded.

To retain the sociological balance that characterizes the BNC-C source texts we attempted to include two texts from each BNC-C file. However, in a number of these files, no narratives were found that met the criteria. As shown in Table 1, we included 279 texts extracted from 143 files (93% of BNC-C). The NC contains 531 narratives and the total word count is almost 150,000.

3. Corpus annotation

In this section we discuss the annotation applied in the NC in terms of the analytical categories and tagset values adopted. Section 4 focuses on the actual implementation of the annotation scheme through the modification of the TEI schema used for BNC-XML files. It also discusses some of the XML-based tools we have used to query the NC.

3.1 *Approach*

Since the NC is the first annotated corpus of conversational narrative we were not able to rely on established practices; rather, the annotation system was largely developed from scratch. However, we did follow previous corpus annotation practice in two areas: (i) annotation of discourse presentation – following categories developed in McIntyre et al.'s (2004) work on discourse presentation in speech–and (ii) research on textual positioning (e.g., Hoey 2005, Hoey & O'Donnell 2008, O'Donnell et al. 2012) in annotating textual components (see Section 3.4.3). Generally, we observed the principle that "[a]nnotation practices should be consensual" (Leech 2005: 21). As a consequence, we did not annotate a number of features that have been given much prominence in narrative research, among them, for example, narrative structure in the Labovian sense (e.g., 1972) and a number of other features on the

Bereitgestellt von | Bayerische Staatsbibliothek
Angemeldet | 194.95.59.195
Heruntergeladen am | 05.11.12 19:08

grounds that they were unlikely to be coded identically by at least two coders (see Section 3.3.3).

3.2 *Tagsets*

The tagsets used in the annotation of the NC were designed in accordance with Leech's (1997) 'standards' for corpus annotation. That is, the labels used are concise (consisting of no more than three characters), perspicuous (easy to interpret)[2], and analyzable (decomposable into their logical parts). Consider, for example, the tagset used for the various forms of the quotative SAY (see also Section 3.3.4.2):

Tagset quotative SAY:

Q	QS	QSB	(quotative SAY base form *say*)
		QSZ	(quotative SAY 3rd-pers. sing. present tense form *says*)
		QSD	(quotative SAY past tense form *said*)
		QSG	(quotative SAY progressive form *saying*)
		QSN	(quotative SAY past participle form *said*)

3.3 *Levels of annotation*

The annotation scheme provides for markup on five different levels: part-of-speech (POS), speaker, text, textual components, and utterance. We assume readers are generally familiar with POS annotation and will not deal with it here. The remaining four levels are briefly explained in the following subsections. A rationale for each level of annotation is given and research questions the annotation may help to address are briefly sketched out.

3.3.1 *Speaker*

In BNC-C all speakers received tags indicating information about their sex, age, social class, region of origin, educational background and so on, where this information was available to the respondent (Hoffmann el al. 2008: 34–36). Any such sociological information is retained in the NC text headers allowing a range of sociolinguistic queries to be formulated. In the interest of space we will not discuss how the data in the NC breaks down in terms of age bands but focus on the distribution according to sex and only note in passing what can be done using the meta-information on speaker social class.

Table 2 shows the distribution of the 500 speakers involved in the narratives (that is, speakers in the CNN component (see 3.3.3) which contains nearly 79,000 words according to sex.

Table 2 shows that there are 212 female participants, 173 male participants, and 115 speakers whose sex is unknown. The overrepresentation of women in

Bereitgestellt von | Bayerische Staatsbibliothek
Angemeldet | 194.95.59.195
Heruntergeladen am | 05.11.12 19:08

Table 2. *Distribution of male and female narrative participants in the CNN components of the NC*

Sex	Number of participants	%	Number of words	%
Female	212	42	44,476	56
Male	173	35	24,268	31
Unknown	115	23	10,079	13
Total	500	100	78,823	100

the NC is due to the small built-in speaker sex bias that is characteristic of the BNC-C. According to Rayson et al. (1997: 135), there are 561 female speakers compared to 536 male speakers. However, the difference is wider in the NC than in the BNC-C. According to chi-squared tests for given probabilities, the difference in numbers of female and male speakers in the BNC-C is insignificant ($p = 0.45$), while the difference in the NC is significant ($p = 0.048$). A 'spill' effect is the unequal number of words spoken by men and women either as narrators or recipients. Female participants (44,476 words / 65%) contribute to storytelling almost twice as many words as male participants do (24,268 words / 35%).[3]

These figures suggest various research paths. A much-studied object of research is the differential amount that men and women talk (for a survey of relevant research, see James & Drakich 1993; Schmid 2003; also Baker 2010: 33–44). The overall consensus seems to be that the amount of talk essentially depends on context of use. The NC can be used to contribute to this line of enquiry. If we assume that the figures presented above are reasonably representative (given the close modeling of the NC on the BNC-C, whose representativeness and sociological balance are widely accepted), they facilitate a number of interesting hypotheses. A crude hypothesis is that men tell fewer stories than women and/or that men's stories are shorter. Another hypothesis is that men's word share is smaller in part because they are less verbose *as recipients* of stories (asking fewer questions and/or producing fewer tokens of listenership). We acknowledge the tentative nature of these hypotheses but still trust they are worth testing.

Among the questions that might be asked of the annotation of social class in the NC we briefly note one. Rayson et al. (1997) found the forms *said* and *says* strikingly high up among the most frequently used words in social classes C2 and DE. In the NC, both *said* and *says* are overwhelmingly used as quotatives introducing direct speech presentation. Rühlemann et al. (2011) demonstrate that the proportions of *said* and *says* used as quotatives (as opposed to non-quotative use) are 93% and 96% respectively. The annotations of speaker class

Bereitgestellt von | Bayerische Staatsbibliothek
Angemeldet | 194.95.59.195
Heruntergeladen am | 05.11.12 19:08

and discourse presentation (cf. Section 3.3.4.2) might then be combined to test the hypothesis that amount of direct discourse presentation is correlated with social class in the sense that the lower social classes C2 and DE use it more frequently than the higher classes AB and C1.

3.3.2 *Text*

Annotation of the NC at the text level is intended to capture characteristics of narratives as discourse units. The annotation includes the narrative's title (given by the researchers), information on the type of embedding (that is, whether the narrative is a stand-alone story or part of a 'narrative chain') and information about narrative subgenre.

(4) `<div title="Bob Marley" embedLevel="EC3" narrativeType="T10">`
(KPG-N1)

In (4), the information captured at `<div>`-level is threefold: 1. the title assigned to the narrative ('Bob Marley'), 2. the value 'EC3' indicates that 'Bob Marley' is the third narrative within a narrative chain, and 3. the value 'T10' identifies the narrative as a 1st person personal experience story.

In defining narrative types, or subgenres, we diverge from previous taxonomies in a number of respects. Firstly, our taxonomy is built around a single criterion, namely the criterion of 'experience', and makes a basic distinction between 'experiencer' – that is, the person who underwent the experience –, and 'type of experience'. We admit two types of 'experiencer', 1st person and 3rd person[4]. That is, a basic distinction is made between stories relating a sequence of events the narrator was involved in and stories describing events the narrator learned about through hearsay. Further, we distinguish between various types of experience: stories can not only relate personal experiences but also recurrent generalized experiences, dreams, fantasies, jokes, and mediated experiences. This latter category reflects the fact that speakers not uncommonly relate the experience of watching a film or reading a book.

Table 3 shows the breakdown of the narratives in the NC according to type of experience and experiencer person (1st or 3rd person). Nearly 80% or 4 in every 5 narratives are first person, which is to be expected in conversation.

The rationale for annotating type of narrative is the observation that "we are probably better off in considering narrative genre as a continuous cline, consisting of many subgenres, each of which may need differential research treatment" (Ervin-Tripp & Küntay 1997: 139).

3.3.3 *Textual components*

Initially, we aimed to annotate elements of narrative structure as identified in Labov & Waletzky (1967/1997), including abstract, orientation, complicating

Bereitgestellt von | Bayerische Staatsbibliothek
Angemeldet | 194.95.59.195
Heruntergeladen am | 05.11.12 19:08

Table 3. *Distribution of the 531 narratives in the NC across narrative types*

Experiencer		1st person	Number of narratives	3rd person	Number of narratives
Type of experience	Personal	T10	337	T30	81
	Generalized	T1G	42	T3G	14
	Dream	T1D	9	T3D	1
	Mediated	T1M	14	T3M	10
	Fantasy	T1F	13	T3F	9
	Joke	–		T3J	1
			415		116

events, resolution, and coda. However, it became clear quickly that the identification of these elements would be largely non-consensual, because Labov & Waletzky's elements are "not always recognizable by traditional narrative-internal criteria" (Ervin-Tripp & Küntay 1997: 133; see also Edwards 1997: 139).

Instead of attempting to annotate structural elements of *narratives* we decided to annotate 'components' of the *texts*. Remember that the texts extracted from the BNC-C include not only the narrative but also stretches of conversational talk preceding and following the narrative proper. As a rule, the length of these non-narrative stretches was limited to 15 utterances both before and after the narrative(s) proper. Moreover, the first utterance and the last utterance within the narratives are tagged. The tagset for text components includes six values:

Tagset Text components:

C	CP	CPR	Pre-narrative conversation
		CPO	Post-narrative conversation
	CN	CNN	Narrative
		CNI	Narrative-initial utterance
		CNF	Narrative-final utterance
		CNI-CNF	Single-utterance narrative

The component structure of the texts in the NC is illustrated in Figure 1 (word counts are given in square brackets). The word counts show that the two non-narrative components CPR and CPO together contain only slightly fewer words (33,001 + 37,696 = 70,697) than the narrative component CNN (78,823). The mix of general conversation and conversational narrative in the NC is thus fairly balanced.

Bereitgestellt von | Bayerische Staatsbibliothek
Angemeldet | 194.95.59.195
Heruntergeladen am | 05.11.12 19:08

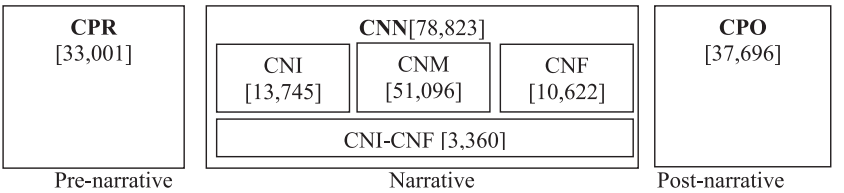


Figure 1. *Componential structure of texts in the NC*

Identification of the five components depends on identification of story boundaries. That is, what precedes or follows a story, and the first or last utterance in a story can only be determined if the beginning and the ending of the story can be discerned. Discriminating between pre-narrative/post-narrative text and narrative text is not always easy. As Ervin-Tripp & Küntay (1997: 133) note, "the onset of conversational stories does not always clearly demarcate the narrative segment from the preceding talk." Therefore, the identification of story boundaries may not always be consensual. However, as evidenced by "considerable disagreement" (Leech 2005: 21) over apparently simple matters, such as defining word classes, complete consensus will never be achieved. The guiding principle in determining story boundaries builds on the first critical criterion underlying the selection of narratives, exosituational orientation (see Section 2.4). We defined a story beginning as that utterance in a conversation in which a shift in orientation could be observed from the present (telling) situation to a non-present (told) situation and to the sequence of events that occurred therein. The reverse, we defined as post-narrative beginning that utterance in the text in which the return to the present situation had been completed. Determining story boundaries thus depends on a clearly defined principle. How this principle is interpreted vis-à-vis actual discourse may vary. However, we trust that our identifications of story boundaries are reasonably consensual. Further, note that while story boundary identification may not be entirely consensual, identification of the six textual components, which all follow automatically from story boundary identification, can be considered uncontroversial.

The rationale for annotating text components is as follows. Annotation of pre-narrative conversation (CPR) and post-narrative conversation (CPO) is included to enable researchers to address two questions critical in research on conversational narrative: 1. how are narratives occasioned in conversation and 2. how are narratives responded to in subsequent talk (e.g., Jefferson 1978)? The annotation of CPR and CPO will allow the empirical analysis of the techniques used to display a relationship between, on the one hand, prior talk and subsequent stories and, on the other, stories and subsequent talk.

Bereitgestellt von | Bayerische Staatsbibliothek
Angemeldet | 194.95.59.195
Heruntergeladen am | 05.11.12 19:08

Annotating narrative-initial (CNI) and narrative-final utterance (CNF) is considered useful for two interrelated reasons. First, these utterances are the loci in which two decisive shifts in participation framework occur: in CNI, participants need to signal the shift from conversational participation to narrative participation (see Section 3.4.4.1) while in CNF, participants accomplish the return to conversational participation patterns. Obviously, these shifts in participation correlate with shifts in the generic framework: from general conversation to narrative and vice versa. Therefore, these boundary utterances are of great interest for researchers working in a conversation-analytical or discourse-analytical tradition interested in how participants manage transitions between different types of participation frameworks and different (sub-)genres. A second reason for annotating the boundary utterances is derived from recent research into textual positioning, which has shed light on the structural associations between lexis and different positions in text (Hoey & O'Donnell 2008; O'Donnell et al. 2012; O'Donnell & Römer In preparation). The investigations of textual positioning so far have concentrated on written data. The componential annotation in the NC may be a useful starting point to explore this phenomenon in spoken data. Given that first and last story utterance are crucial in terms of participation and generic structure, it is highly likely that shifts in that structure will have lexical correlates. Annotation of CNI and CNF facilitates analyses of lexical associations with boundary positions in narratives, thus helping to explore textual colligation in genres of everyday speech. For example, in Rühlemann & O'Donnell (Forthcoming) we reveal an intimate association of 'introductory *this*' (as in *Eh! Our Arthur, Arthur sat there and* ***this*** *girl come to clear the pots away*) with narrative-initial position (i.e., story-first utterance tagged CNI).

3.3.4 *Utterance*

At the utterance level there are two kinds of annotation: 1. pragmatic and 2. stylistic. Utterances are pragmatically annotated by being marked with participant status (Section 3.4.4.1). Annotation at utterance level is stylistic inasmuch as there is systematic markup of discourse presentation. Specifically, two phenomena of discourse presentation are captured: 1. quotatives, that is, verbs serving to introduce Direct or Indirect discourse presentation (Section 3.4.4.2), and 2. presentation modes, that is, meta-information on whether the presentation is Direct, Free Direct, Indirect, Free Indirect and so forth (Section 3.4.4.3).

3.3.4.1 *Utterance I: Participation*

It is frequently noted in the literature that the wide-spread distinction between speaker and hearer in conversation is a gross oversimplification (e.g., Schiffrin 1987: 27). This critique is all the more valid with regard to participation in

Bereitgestellt von | Bayerische Staatsbibliothek
Angemeldet | 194.95.59.195
Heruntergeladen am | 05.11.12 19:08

conversational narrative. Here, the terms 'narrator' and 'recipient' are hypernyms for a broad range of subroles.

We distinguished six major roles, four narrator roles and two recipient roles. The first narrator type includes those narrators whose narration is non-shared in the sense that no reaction whatsoever from the listeners is triggered. We refer to this type as 'Unsupported Narrator' (PNU). However, narrators often do receive support of various sorts from their audience. That is, the narration is no longer done single-handedly but shared between narrator and audience. Three subtypes of shared narration are distinguished. Narrators receive responses that do not contribute to story content in any way but contain only tokens of listenership signalling that the recipients are listening, comprehending, and, potentially, agreeing. Narrators only receiving listenership tokens are referred to as 'Supported Narrators' (PNS). Recipients contributing listenership tokens are referred to as 'Responsive Recipients' (PRR). Another type of recipient contribution is the production of utterances that add to the content or evaluation of the story, for example, by eliciting additional orientation information (where, when, to whom etc. something happened), commenting on individual aspects of the story, challenging the narrator's account or contributing discourse presentation. We refer to this type of recipient as 'Co-constructive Recipient' (PRC). The corresponding narrator role is 'Primary Narrator' (PNP). Finally, narrators may be supported by 'Ratified Co-narrator' (PNC), that is, a participant who has privileged access to the events underlying the story, for example, because he/she was involved in them or a witness to them.

Beside recipients and narrators, we acknowledge two further roles: utterances which are fragmented or unintelligible are classed as 'Unclear Role' (PXX), while speakers who attend to matters completely outside the narrative, for example, to a child crying while another participant is telling a story, are seen as fulfilling a 'Non-narrative Role' (PX0). Building on this participation framework, the participant roles annotated in the NC include the following:

Tagset Participant roles:

P	PN	PNU	Unsupported Narrator
		PNS	Supported Narrator
		PNP	Primary Narrator
		PNC	Ratified Co-narrator
	PR	PRR	Responsive Recipient
		PRC	Co-constructive Recipient
	PX	PX0	Non-narrative Role
		PXX	Unclear Role

Annotation of participant roles can be exploited in a number of ways. Narrative research in the tradition of Labov & Waletzky (1967/1997) and Labov

Bereitgestellt von | Bayerische Staatsbibliothek
Angemeldet | 194.95.59.195
Heruntergeladen am | 05.11.12 19:08

(1972) emphasized the concept of the single teller (Holmes 1997: 94), while more recent research emphasizes the crucial role of participants' contributions to the story (e.g., Ochs & Capps 2001), viewing conversational narrative as an 'interactional achievement' (Schegloff 1997; cf. Rühlemann Forthcoming). The participant break-up provided for in the NC can be used to substantiate this view. Secondly, while research on listenership tokens – that is, the kind of language used by 'Responsive Recipient' (PRR) – has been enormously productive as far as general conversation is concerned (e.g., Kjellmer 2009) we are not aware of any study particularly concerned with back-channel behaviour in narrative. Further, the annotation of participation roles would enable researchers to address important questions such as why it is that many narratives receive audience co-construction of some sort while others do not. By comparing, for example, subcorpora extracted from utterances by PNU, PNS, and PNP it may be possible to discover whether, and how, narrators use language in such a way as to encourage or discourage audience participation.[5]

Another avenue of research is the exploration of gender differences in the co-construction of narratives. One of the very few studies in this field is Holmes (1997), who reports that New Zealand women used more facilitative questions whereas men asked questions that "sometimes disrupted the story" (1997: 95). By using both speaker information and participant annotation, gender-sensitive queries can be formulated to investigate possible differences in the ways that men and women co-construct conversational narratives.

Finally, the NC's participant annotation facilitates analyses of turntaking patterns in narrative. For example, as noted, Sacks (1992) described story-telling as "an attempt [by the main narrator] to control a third slot in talk, from a first" (1992: Vol. II: 18). This is no small claim in that it amounts to the claim that turn order in narrative decisively diverges from ordinary turn order, for which Sacks et al. (1974) postulate the rule that "[t]urn order is not fixed but varies" (Sacks et al. 1974: 701).[6]

3.3.4.2 *Utterance II: Quotatives*

Much research suggests that, in informal talk, a small set of verbs dominate the quotative system across regional varieties of English (e.g., Tagliamote & Hudson 1999: 155; Buchstaller 2002; Barbieri 2007; Macaulay 2001; Winter 2002). Building on this research, the quotative verbs annotated in the NC include SAY, THINK, GO, LIKE (the latter both with and without preceding BE). Moreover, ASK and TELL, which are not infrequent in the NC, were assigned tags as well. Finally, the tag QOO was included for any other quotative. Tagsets for quotatives contain five values for each quotative (except for LIKE without preceding BE and QOO). They are listed below, using quotative GO as illustration:

Bereitgestellt von | Bayerische Staatsbibliothek
Angemeldet | 194.95.59.195
Heruntergeladen am | 05.11.12 19:08

Tagset Quotative GO (tagset for SAY see above)

Q	QG	QGB	(quotative GO base form *go*)
		QGZ	(quotative GO 3rd-pers. sing. pres. tense form *goes*)
		QGD	(quotative GO past tense form *went*)
		QGG	(quotative GO progressive form *going*)
		QGN	(quotative GO past participle form *gone*)

Although there is some agreement as to which verbs have the widest currency across varieties of English, it should be noted that the underlying findings stem primarily from research "based on samples of sociolinguistic interviews or conversational narratives collected in one single location" (Barbieri 2007: 25). Winter (2002), for example, analyses sociolinguistic interviews with 15–16 year-olds in Melbourne. In contrast, there are fewer analyses from corpora of general conversation. Buchstaller (2002), for example, used the Switchboard Corpus and the Santa Barbara Corpus of Spoken English, two American corpora. Little is known about which quotatives are used in conversational narrative in the UK. Therefore, based on the systematic annotation of quotatives provided for in the NC, it can be established which quotatives and which forms are used in the NC and, thus, to discover which quotatives are typical of British narrative (in the early 1990s).

3.3.4.3 *Utterance III: Discourse presentation mode*

Extending McIntyre et al.'s (2004) model of discourse presentation, we distinguish ten presentation modes: Free Direct, Direct, Free Indirect, Indirect, Representation of Speech Act, Representation of Voice with Topic, Representation of Voice, Representation of Use, Reference to Discourse Presentation, and Request for Discourse Presentation. An additional code, MXX, is used for presentations that cannot with certainty be assigned to any of the categories.

The tagset used to capture presentation modes includes the following values:

Tagset for Reporting Mode:

M	MD	MDD	Direct
		MDF	Free Direct
	MI	MII	Indirect
		MIF	Free Indirect
	MN	MSS	Representation of Speech Act
	MV	MVT	Representation of Voice with Topic
		MVV	Representation of Voice
	MU	MUU	Representation of Use
	MR	MRR	Reference to Discourse Presentation
		MRQ	Request for Discourse Presentation
	MX	MXX	Unclear

Bereitgestellt von | Bayerische Staatsbibliothek
Angemeldet | 194.95.59.195
Heruntergeladen am | 05.11.12 19:08

In what follows we illustrate the ten categories. Two examples are given for each category. All (a) examples are invented and all (b) examples are taken from the NC. We take the liberty of using invented examples that share the same context and content hoping that the subtle differences between the modes can be demonstrated more easily. Following Leech & Short's (1981) scalar model of discourse presentation, the examples are ordered in ascending order of what Leech & Short called 'narratorial interference', that is, the ordering begins with those types of presentation in which the presenting speaker is apparently least in control of the discourse presented but closest to what the 'original' discourse may have looked like.

(5) MDF: (a) *Can I have my videos?*
(b) And, no, they had n't done anything racist up until then and I goes ^{QGZ} casually, [^{MDD} do you like Bob Marley?] [$^{\mathbf{MDF}}$ **No, he's a black nigger**.] [$^{\mathbf{MDF}}$ **What! Do you know what you're saying you are, you are some stupid**]
(KPG-N1)

(6) MDD: (a) *Wayne said, can I have my videos?*
(b) I says $^{\mathbf{QSZ}}$ [$^{\mathbf{MDD}}$ **Lindsey if you want us to trust you, you have got to tell the truth**
(KDS-N1)

(7) MIF: (a) *Could he have his videos?*
(b) Ah she does, aye, aye no, and she says $^{\mathbf{QSZ}}$ you know, [^{MDD} answer a few questions] and all that**, [$^{\mathbf{MIF}}$ would I like to take part in a sur survey of how the English language is getting used?**]
(KPD-N1)

(8) MII: (a) *Wayne said, can he have his videos?*
(KE5-N1)
(b) you know this erm, you know this young girl that was killed along Benji Avenue with her mother? Well I thought $^{\mathbf{QTD}}$ [$^{\mathbf{MII}}$ **it was Wendy's daughter**]
(KCP-N1)

(9) MSS: (a) *Wayne wants his videos back.*
(b) Milk up [$^{\mathbf{MSS}}$ **Nicola did n't believe me when I told her**.] She thought $^{\mathbf{QTD}}$ [^{MII} I was playing a joke.]
(KPN-N2)

(10) MVT: (a) *He talked about his videos.*
(b) Well actually on the television I do n't know whether you've noticed when **[$^{\mathbf{MVT}}$ they start talking about**

Bereitgestellt von | Bayerische Staatsbibliothek
Angemeldet | 194.95.59.195
Heruntergeladen am | 05.11.12 19:08

burglaries and things like that] the picture that comes on is Newlands Park you know
(KDY-N1)

(11) MVV: (a) *He was talking to Jean*.
(b) [$^{\mathbf{MVV}}$ **She has talked, we've talked to her yesterday,**] was it yesterday?
(KDS-N1)

(12) MUU: (a) *He said 'my videos'*
(b) PS549> Oh aye, aye. Aye, but . . . you know . . . like you se – she mentioned one in particular, like
PS54D> What?
PS549> [$^{\mathbf{MUU}}$ **the word skeilth]**
PS54D> [$^{\mathbf{MUU}}$ **Skeilth?**]
PS54E> Mm.
(KPD-N1)

(13) MRR: (a) *He really said that.*
(b) And when I walked away I thought^{QTD}, [^{MDD}oh fancy saying that] **[$^{\mathbf{MRR}}$ but I'd said it,]** it was too bloody late.
(KD7-N1)

(14) MRQ: (a) *What did he say?*
(b) [$^{\mathbf{MRQ}}$**And what did you say then?]**
(KBY-N2)

At the extreme end of the narratorial intervention cline where the presenter is apparently not in control at all of the discourse presented and where it seems "as if the author has vacated the stage and left it to his characters" (Leech & Short 1981: 334) we find the two direct categories, MDF and MDD, as in (5) and (6). Both modes present speech "in the form in which it is directly manifest to a listener" (Leech & Short 1981: 345). They are distinguishable in that MDD is preceded by a reporting clause (which introduces a first element of narratorial control) whereas MDF is not. The same structural distinction applies to the two indirect categories, MIF and MII, illustrated in (7) and (8). While in (Free) Direct mode presenters purport to give the exact wording of the utterances presented, in (Free) Indirect mode the propositional content of the original speech is specified, "but no claim is made to present the words and structures originally used to utter that proposition" (McIntyre et al. 2004: 61). The categories further include Representation of Speech Act, as in (9). This nomenclature is modelled on McIntyre et al.'s nomenclature; more specifically, however, this mode captures *illocutionary acts*, that is, what is *done in* issuing an utterance (Austin 1962: 99). Further, following McIntyre et al. (2004) we recognize Representation of Voice, as in (11); this type "captures minimal references to speech with no indication of the illocutionary force, let alone the propositional

Bereitgestellt von | Bayerische Staatsbibliothek
Angemeldet | 194.95.59.195
Heruntergeladen am | 05.11.12 19:08

content or form of the utterance (part)" (McIntyre et al. 2004: 62). However, unlike McIntyre et al., we recognize a subtype of MVV, namely Representations of Voice with Topic (MVT), as in (10). These are minimal references to speech that include mention of the topic. Typically, this type includes an *about* phrase. Another type is Representation of Use, as in (12), a meta-linguistic type of discourse presentation used for "mentions of language use, such as the words or expressions habitually used to refer to things, or the way words were spelled or pronounced" (McIntyre et al. 2004: 63). Our final categories are new categories not yet recognized in previous research; they are situated at, and possibly cross, the border line between discourse presentation and narration. We distinguish two such categories: Reference to Discourse Presentation (MRR), as in (13), and Request for Discourse Presentation, as in (14). Instances of MRR are characterized by the fact that the content of a discourse presentation is referred to anaphorically and can hence only be recognized via the preceding context. The type is clearly a borderline phenomenon in that the discourse is not actually presented but only pointed to by referential means. Typically, MRR is realized using referring expressions such as *that* or *it*. Finally, Request for Discourse Presentation (MRQ), as in (14), typically occurs in recipient utterances. Like MRR, MRQ gives no clue as to the form, propositional content or illocutionary force of an anterior discourse but only links to it via reference. Not surprisingly, the question pronoun *what* (a referring expression) is found in all instances of this type in the NC. That is, MRR and MRQ are both references to discourse presentation, MRR in declarative form, MRQ in interrogative form.

The rationale for annotating discourse presentation derives from the central role discourse presentation plays in conversational narrative, both quantitatively – one in four words occurring in the narratives occurs within discourse presentation units (cf. Rühlemann Forthcoming) – and qualitatively in that (Free) Direct discourse presentation (by far the most frequent type in the NC accounting for 66% of all discourse presentation) "is a means by which experience surpasses story to become *drama*" (1986: 312; our emphasis).

The markup of discourse presentation opens up numerous avenues for future research. Little attention has been paid so far to amount of discourse presentation in narratives by men and women. We are aware of only a few relevant studies (e.g., Johnstone 1993; Ferrara & Bell 1995; Barbieri 2007; Harrington 2008), suggesting that women outscore men in terms of use of discourse presentation. Moreover, investigating discourse presentation in terms of positioning across narratives, claims can be tested suggesting that "[t]he climax of a conversational story is often realized in [constructed] dialogue" (Norrick Forthcoming; cf. also Li 1986). Further, if the annotations of discourse presentation and participant role are combined, discourse presentation contributed to a narrative by recipients can be explored (cf. Yule & Mathis 1992: 204).

Bereitgestellt von | Bayerische Staatsbibliothek
Angemeldet | 194.95.59.195
Heruntergeladen am | 05.11.12 19:08

Finally, another possible line of enquiry is looking into 'utterance openers', that is, interjections such as *oh* and *well* used at the onset of quotations, as in *she said* **QSD** *[*[MDD] ***oh*** *you can have a hot cross bun there]*, signalling that speakers are embarking on (Free) Direct discourse presentation (cf. Biber et al. 1999: 1118).

4. Annotation and analytical tools

The discussion of the annotation added to the texts in the NC in the previous section was focused at the conceptual level, describing the theoretical background and justification for the categories used in the annotation. The tagset (or codes) for each of the components of the annotation model and some initial distributional data were also provided. This section considers the annotation model from a more technical perspective with a focus on how the annotation was added to the BNC-XML files from which the NC is extracted. We also discuss some of the XML-based tools we have developed for analyzing the NC data.

4.1 *Implementing the NC annotation model*

Given that the NC is a subset of the BNC-C and built from the BNC-XML version (Burnard 2007) using XML for the NC annotation scheme is a natural choice. Discussions of corpus annotation (e.g. O'Donnell 1999; Leech 2005; McEnery et al. 2008: 29–45; Baker 2010: 15–19, 149–150) and markup frequently present two options: 1. inline or embedded annotation, where tags or SGML/XML elements are added to the textual data in the same file and 2. standoff or stand-alone annotation where textual data is in the base file and different annotation layers are placed in separate files with pointers to relevant spans of text in the base file. Standoff markup is ultimately the better of the two options in terms of the flexibility and extensibility it brings (see the list of advantages in McEnery et al. 2008: 44 and technical advantages discussed in Carletta et al. 2002). One advantage of using standoff markup is that is does not make any changes to the base file and it allows the extra annotation to be redistributed even if the base corpus requires a license. However, standard corpus tools and even those built for working with XML-encoded corpora, such as Xaira (Burnard & Todd 2003) have difficulty working with these kinds of corpus files. When we began the work on the NC we planned to use Xaira to carry out searches. Also, as should be clear from the description of the annotation model in Section 3, many of the categories and values in the model are

Bereitgestellt von | Bayerische Staatsbibliothek
Angemeldet | 194.95.59.195
Heruntergeladen am | 05.11.12 19:08

simply additional attributes added to utterances (Section 3.3.4) already marked in the underlying BNC-XML files. Finally, and most importantly, we have found that embedding the NC model annotation into the BNC XML has allowed us not only to make use of the query functionality in a standard XML database but also to use the XPath query language and its extension, XQuery, which are exceedingly useful tools in XML data retrieval (see Section 4.3). These factors influenced our decision not to develop a standoff annotation scheme for the NC categories but rather to make minimal modifications to the XML Schema provided with the BNC XML version. It should be noted however, that it is a relatively trivial task to extract the additional inline annotation and to create a standoff version of the NC. Once the NC annotation has stabilized and been revised we may explore this option for distribution.

4.2 *An example of the NC annotation*

Returning to the conversational narrative example (1) used at the beginning of the paper, we now focus on the analysis of this text using the NC model and how it is annotated in XML. Figure 2 shows a screenshot of the text formatted in the NC Browser web tool (see Section 4.4). The narrative boundaries are marked by the dotted box with the text components (CNN, CNI, CNF and CPO) shown as inner boxes with solid lines. Each speaker utterance begins with a speaker number and participation role (PNP, PNC, etc.). Quotatives are

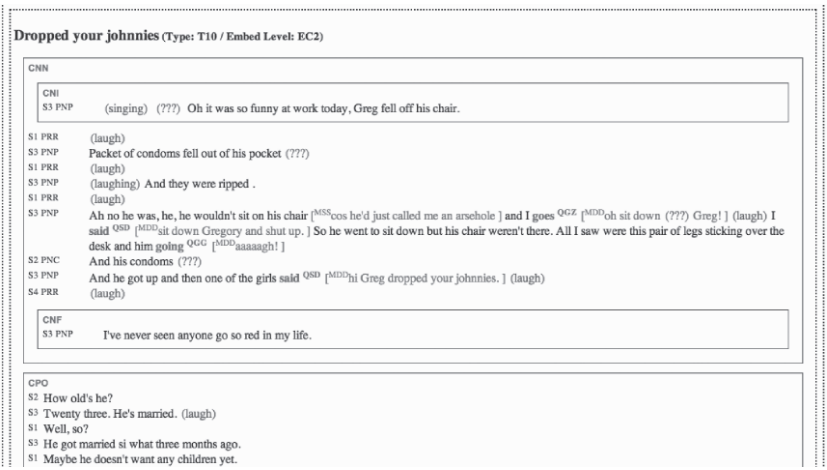


Figure 2. *Formatted version of narrative from file KCE-N1 (Example 1) illustrating NC annotation model*

Bereitgestellt von | Bayerische Staatsbibliothek
Angemeldet | 194.95.59.195
Heruntergeladen am | 05.11.12 19:08

marked in bold type face followed by a superscript tag (QGZ, QGG, QSD, etc.) and stretches of reported discourse demarcated by brackets with an initial superscript tag (MSS, MDD, etc.).

Recall from Section 2.1 that we make a distinction between files, texts and narratives. The example narrative is extracted from the BNC-C file KCE. The KCE file was recorded by a respondent named Helena and contains 24 conversations:

```
<bncDoc xml:id="KCE">
   <teiHeader>
       <fileDesc>
          <titleStmt>
              <title>
               24 conversations recorded by 'Helena'
               (PS0EB) between 12 and 20 March 1992 with
               9 interlocutors, totalling 7370 s-units,
               50776 words, and 5 hours 47 minutes 8
               seconds of recordings.
              </title>
          <titleStmt>
      ...
```

Each of the participants in the conversation are referenced with a code that links to their demographic information, allowing the kind of analysis discussed in Section 3.3.1 (see also Burnard 2005). In the body (`<stext>` element) each conservation is placed within a element:

```
<stext type="CONVRSN">
    ...
    <div n="029501" decls="KCERE006 KCESE006">
        <u who="PS0EG">
            <s n="2467">
                <w c5="AVP" hw="out" pos="ADV">Out </w>
                <w c5="PRP" hw="in" pos="PREP">in </w>
                <w c5="AT0" hw="a" pos="ART">a </w>
                <w c5="NN1" hw="minute"
                pos="SUBST">minute</w> <c c5="PUN">,</c>
                <unclear/> <w c5="VVN" hw="change"
                pos="VERB">changed</w> <c c5="PUN">.</c>
            </s>
        </u>
        <u who="KCEPSUNK">
```

Bereitgestellt von | Bayerische Staatsbibliothek
Angemeldet | 194.95.59.195
Heruntergeladen am | 05.11.12 19:08

```
                <align with="KCELC0T1"/>
                <unclear/><vocal desc="laugh"/>
                <align with="KCELC0T2"/><pause/>
        </u>
        <u who="PS0EG">
            <s n="2468">
                <w c5="AJ0-AV0" hw="quick"
                pos="ADJ">Quick </w> <pause/> <vocal
                desc="laugh"/>
            </s>
        </u>
        ...
    </div>
    ...
</stext>
```

The NC file KCE-N1 is derived from this BNC-C file by leaving the header untouched but removing all content from the <stext> section apart from the 'text' containing the narratives. For KCE-N1 the two narratives are made up of the <s> units numbered 2641 to 2658 and 2659 to 2682. To capture the distinction between a text and a narrative we modified the schema to allow further elements within the children of <stext>. These elements take the three attributes: 1. title, 2. narrativeType and 3. embedLevel discussed in Section 3.3.2, The two narratives in KCE-N1 are part of a narrative chain so have embedLevel values of EC*:

```
<stext type="CONVRSN">
    <div title="Mishaps" embedLevel="EC">
          <div title="Forgetting matches"
              embedLevel="EC1" narrativeType="T10">
              ...
          </div>
          <div title="Dropped your johnnies"
              embedLevel="EC2" narrativeType="T10">
              ...
          </div>
    </div>
</stext>
```

The conversation elements in the BNC-C files contain just a sequence of utterance, pause and event elements (<u>, <pause>, <vocal>, <event>,

Bereitgestellt von | Bayerische Staatsbibliothek
Angemeldet | 194.95.59.195
Heruntergeladen am | 05.11.12 19:08

<shift> and <trunc>). The textual components discussed in Section 3.3.3 sit between the narrative and these elements in the annotation scheme hierarchy. We opted to use <seg> elements for the textual components and allowed them to nest (see Figure 2). The structure for the narrative in Figure 2 is structured as follows:

```
<div title="Dropped your johnnies" embedLevel="EC2"
narrativeType="T10">
    <seg Components="CNN">
          <seg Components="CNI">
                <u>...</u>
          </seg>
          ...
          <!-- utterance elements -->
          ...
          <seg Components="CNF">
                <u>...</u>
          </seg>
    </seg>
    <seg Components="CPO">
           ...
           <!-- utterance elements -->
           ...
    </seg>
</div>
```

As outlined in Section 3.3.4, at the utterance level the NC model is concerned with three features: 1. participation roles, 2. quotative words / phrases and 3. stretches of text that present anterior discourse. The first of these, participation roles, relates to speakers and their utterances within the narrative so it can be easily handled as an additional attribute on the <u> element with the participant values detailed above.

```
<seg Components="CNN">
    <seg Components="CNI">
      <u who="PS0EG" Participation_roles="PNP">
          ...
      </u>
    </seg>
    <u who="PS0EB" Participation_roles="PRR">
      ...
    </u>
```

Bereitgestellt von | Bayerische Staatsbibliothek
Angemeldet | 194.95.59.195
Heruntergeladen am | 05.11.12 19:08

```
    <u who="PS0EG" Participation_roles="PNP">
      ...
    </u>
    <u who="PS0EB" Participation_roles="PRR">
      ...
    </u>
    ...
</seg>
```

The other two utterance level features can potentially span a number of <w> elements, so again we opted to use a <seg> element for these features. We encountered a problem however because these <seg> elements (particularly those for discourse presentation mode) may cut across from within one <s> unit to another. Consider the utterance in (15) (s-unit boundaries are demarcated by ||):

(15) PNP $_{s1}$|| (*laughing*) (*???*) exploding everywhere wouldn't she! ||
$_{s2}$|| But like, I was **thinking** ^{QTG} [^{MDD}this is gonna be so embarrassing like in P E! ||
$_{s3}$|| (*laughing*) With ha half a bra on!]||
(KCE-N2)

The Direct discourse (MDD) segment begins in s2 and continues to the end of s3. This kind of overlap is illegal in XML where the well-formedness constraint requires all contained elements to be closed before their enclosing (parent) elements (e.g. <x><y></x></y> is illegal, while <x><y></y></x> is 'well formed'; see Leech 2005; Carletta et al. 2005: 453–454 and solutions for overlap in XML in Durusau & O'Donnell 2002 and DeRose 2004). The problem for (15) becomes clearer when the XML is examined:

```
<u who="PS0EC" Participation_roles="PNP">
      <s n="651">
            <shift new="laughing"/> <unclear/> <w c5="VVG"
            hw="explode" pos="VERB">exploding </w> <w c5="AV0"
            hw="everywhere" pos="ADV">everywhere </w>
            <w c5="VM0" hw="would" pos="VERB">would</w>
            <w c5="XX0" hw="not" pos="ADV">n't </w> <w c5="PNP"
            hw="she" pos="PRON">she </w> <shift/>
            <c c5="PUN">!</c>
      </s>
```

Bereitgestellt von | Bayerische Staatsbibliothek
Angemeldet | 194.95.59.195
Heruntergeladen am | 05.11.12 19:08

```
<s n="652">
      <w c5="CJC" hw="but" pos="CONJ">But </w>
      <w c5="AV0" hw="like" pos="ADV">like</w>
      <c c5="PUN">, </c> <w c5="PNP" hw="i" pos="PRON">I
      </w> <w c5="VBD" hw="be" pos="VERB">was </w>
      <seg Quotatives="QTG">
            <w c5="VVG" hw="think" pos="VERB">thinking
            </w>
      </seg>
      <pause/>
      <seg Reporting_modes="MDD">
            <w c5="DT0" hw="this" pos="ADJ">this </w>
            <w c5="VBZ" hw="be" pos="VERB">is </w>
            <w c5="VVG" hw="gon" pos="VERB">gon</w>
            <w c5="TO0" hw="na" pos="PREP">na </w>
            <w c5="VBI" hw="be" pos="VERB">be </w>
            <w c5="AV0" hw="so" pos="ADV">so </w>
            <w c5="AJ0" hw="embarrassing"
            pos="ADJ">embarrassing </w> <w c5="AV0"
            hw="like" pos="ADV">like </w> <w c5="PRP"
            hw="in" pos="PREP">in </w> <w c5="ZZ0"
            hw="p" pos="SUBST">P </w> <w c5="ZZ0" hw="e"
            pos="SUBST">E</w> <c c5="PUN">!</c>
</s>
<s n="653">
            <shift new="laughing"/> <w c5="PRP"
            hw="with" pos="PREP">With </w>
            <trunc> <w c5="UNC" hw="ha" pos="UNC">ha
            </w> </trunc> <w c5="DT0" hw="half"
            pos="ADJ">half </w> <w c5="AT0" hw="a"
            pos="ART">a </w> <w c5="NN1" hw="bra"
            pos="SUBST">bra </w> <w c5="AVP" hw="on"
            pos="ADV">on </w> <shift/> <c c5="PUN">!</c>
      </seg>
</s>
</u>
```

We decided that the s-unit container markup was of little use for the kinds of analysis for which the NC is primarily intended. Indeed the boundaries of s-units are much more fuzzy in spoken discourse. To avoid the overlapping elements problem we switched the <s> container elements in the original

Bereitgestellt von | Bayerische Staatsbibliothek
Angemeldet | 194.95.59.195
Heruntergeladen am | 05.11.12 19:08

BNC-C files to start and end milestone pairs (called 'Trojan' milestones in DeRose 2004) using a simple XSLT stylesheet. The key template is:

```
<xsl:template match="s">
        <s n="{@n}" sID="{@n}"/>
        <xsl:apply-templates/>
        <s eID="{@n}"/>
</xsl:template>
```

This 'flattens' the annotation and removes the conflict between the <s> elements and discourse presentation <seg> elements. The XML fragment for example (15) now looks like this:

```
<u who="PS0EC" Participation_roles="PNP">
      <s n="651" sID="651"/>
      <shift new="laughing"/> <unclear/> <w c5="VVG"
      hw="explode" pos="VERB">exploding </w> <w c5="AV0"
      hw="everywhere" pos="ADV">everywhere </w> <w c5="VM0"
      hw="would" pos="VERB">would</w> <w c5="XX0" hw="not"
      pos="ADV">n't </w> <w c5="PNP" hw="she" pos="PRON">she
      </w> <shift/> <c c5="PUN">!</c>
      <s eID="651"/>
      <s n="652" sID="652"/>
      <w c5="CJC" hw="but" pos="CONJ">But </w> <w c5="AV0"
      hw="like" pos="ADV">like</w> <c c5="PUN">, </c>
      <w c5="PNP" hw="i" pos="PRON">I </w> <w c5="VBD" hw="be"
      pos="VERB">was </w>
      <seg Quotatives="QTG">
            <w c5="VVG" hw="think" pos="VERB">thinking </w>
      </seg>
      <pause/>
      <seg Reporting_modes="MDD">
            <w c5="DT0" hw="this" pos="ADJ">this </w>
            <w c5="VBZ" hw="be" pos="VERB">is </w> <w c5="VVG"
            hw="gon" pos="VERB">gon</w> <w c5="TO0" hw="na"
            pos="PREP">na </w> <w c5="VBI" hw="be"
            pos="VERB">be </w> <w c5="AV0" hw="so" pos="ADV">so
            </w> <w c5="AJ0" hw="embarrassing"
            pos="ADJ">embarrassing </w> <w c5="AV0" hw="like"
            pos="ADV">like </w> <w c5="PRP" hw="in"
            pos="PREP">in </w> <w c5="ZZ0" hw="p" pos="SUBST">P
            </w> <w c5="ZZ0" hw="e" pos="SUBST">E</w>
            <c c5="PUN">!</c>
```

Bereitgestellt von | Bayerische Staatsbibliothek
Angemeldet | 194.95.59.195
Heruntergeladen am | 05.11.12 19:08

```
            <s eID="652"/>
            <s n="653" sID="653"/>
            <shift new="laughing"/> <w c5="PRP" hw="with"
            pos="PREP">With </w>
            <trunc> <w c5="UNC" hw="ha" pos="UNC">ha </w>
            </trunc> <w c5="DT0" hw="half" pos="ADJ">half </w>
            <w c5="AT0" hw="a" pos="ART">a </w> <w c5="NN1"
            hw="bra" pos="SUBST">bra </w> <w c5="AVP" hw="on"
            pos="ADV">on </w> <shift/> <c c5="PUN">!</c>
      </seg>
      <s eID="653"/>
</u>
```

The Appendix contains the full XML using the NC annotation scheme for the narrative section shown in Figure 2. The added elements, attributes and changes from the original BNC-C XML are shown in bold. We found the <oXygen/> XML editor[7] to be a powerful tool for working with the BNC-C files and adding the NC annotation. We modified the bncxml.xsd schema included with the BNC-XML version to allow the elements, attributes and structures discussed above. The annotator opens an unmodified BNC-C file using the XML editor and removes stretches of text that are not part of identified narrative texts and renames the file (e.g. KCE.xml becomes KCE-N1.xml). The modified XML schema file is then associated with the BNC XML file. The file is automatically validated using this schema and <oXygen/> will give the annotator feedback as to where problems exist. It also guides the addition of the textual component elements, utterance attributes and so on.

4.3 Tools

We make use of the eXist XML database[8] to store and index the NC files. eXist supports the XQuery language that allows both for complex queries based on specific XML elements and pattern structures to be carried out and the construction of entire web-applications. The web-based tools developed include: text browsing (see Figure 2), KWIC (see Figure 3), collocation tables (L1–4, R1–4; see Figure 4) and a search function that allows XPath queries. XPath (Clark & DeRose 1999) is a language designed for addressing the elements and attributes of an XML document, that is, it is intended to allow structural traversal and access to the hierarchy created by the XML tags around the text in a document.

Although the use of XML is the recommended state-of-the-art practice in annotating a corpus, particularly when standard XML vocabularies such as the

Bereitgestellt von | Bayerische Staatsbibliothek
Angemeldet | 194.95.59.195
Heruntergeladen am | 05.11.12 19:08

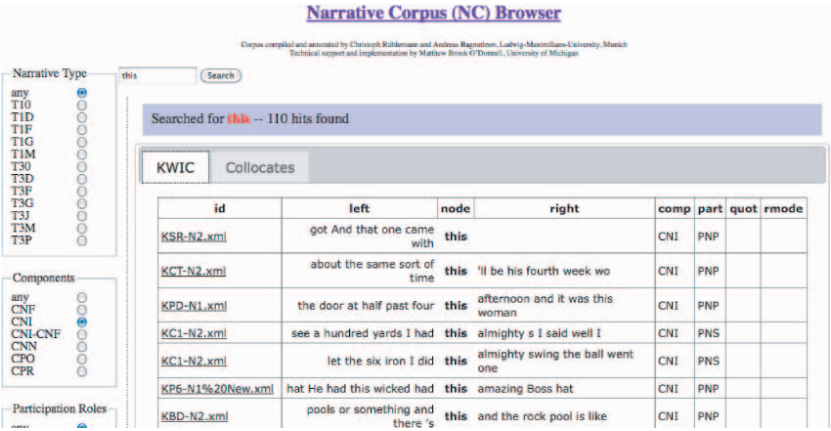
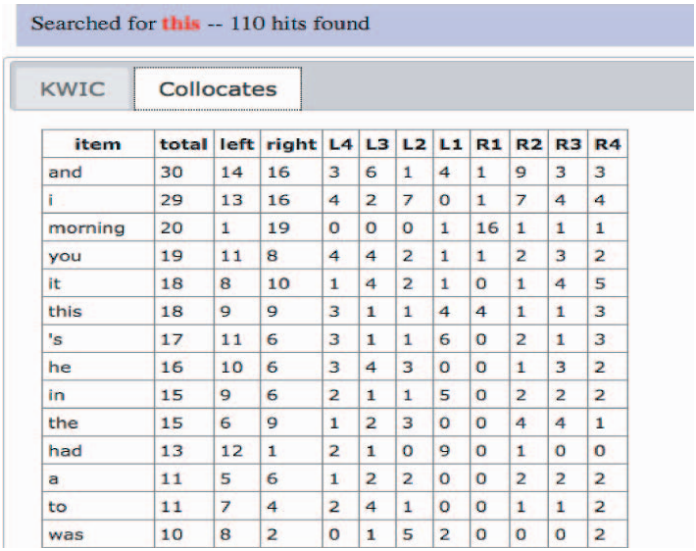


Figure 3. *Web-based concordance search for the NC that allows targeted searches within certain narratives, text components, participation roles and so on.*



item	total	left	right	L4	L3	L2	L1	R1	R2	R3	R4
and	30	14	16	3	6	1	4	1	9	3	3
i	29	13	16	4	2	7	0	1	7	4	4
morning	20	1	19	0	0	0	1	16	1	1	1
you	19	11	8	4	4	2	1	1	2	3	2
it	18	8	10	1	4	2	1	0	1	4	5
this	18	9	9	3	1	1	4	4	1	1	3
's	17	11	6	3	1	1	6	0	2	1	3
he	16	10	6	3	4	3	0	0	1	3	2
in	15	9	6	2	1	1	5	0	2	2	2
the	15	6	9	1	2	3	0	0	4	4	1
had	13	12	1	2	1	0	9	0	1	0	0
a	11	5	6	1	2	2	0	0	2	2	2
to	11	7	4	2	4	1	0	0	1	1	2
was	10	8	2	0	1	5	2	0	0	0	2

Figure 4. *Collocate table for targeted searches of NC.*

TEI are utilized (e.g. McEnery et al. 2006: 22–28), there exists very little support for fully searching XML in the standard corpus tools. Many concordance and frequency tools will read files containing XML markup and either exclude the markup from the analysis or have primitive means of using a specific tag or

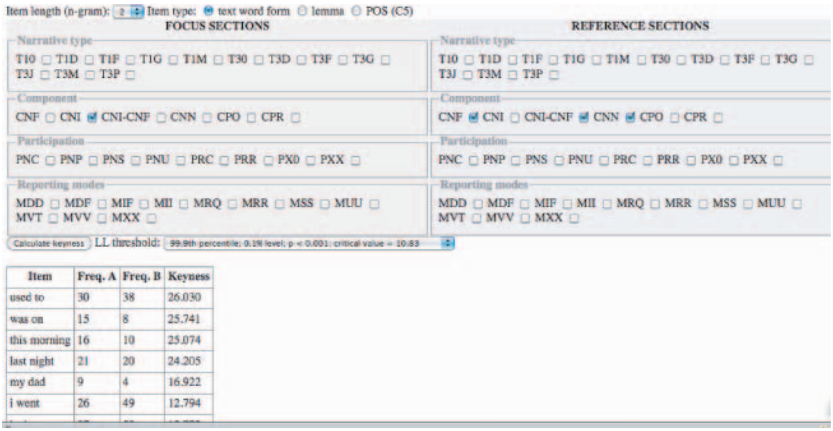
Bereitgestellt von | Bayerische Staatsbibliothek
Angemeldet | 194.95.59.195
Heruntergeladen am | 05.11.12 19:08

tag combination to restrict processing (e.g. the Tags functionality in WordSmith Tools [Scott 2010]). Other tools such as BNCweb can be said to be 'XML aware' in that they include tags in the indexing process and allow certain types of contextual queries (Hoffmann et al. 2008: 230–235). The Xaira tool (Burnard & Todd 2003) that is packaged with the BNC XML and BNC XML Sampler corpora is an exception. There are a number of reasons why the proper treatment of processing of XML documents is a challenge for corpus tool designers. The syntactical rules of XML are quite restrictive, for instance the constraints on element nesting discussed above in Section 4.2, and the representation of an XML document (or DOM = Document Object Model) built during the parsing process is quite memory intensive. Further, at least in Version 1.0 of the XPath language there is very limited support for dealing with strings and textual data, which is of course at the heart of most corpus analysis. However, there are numerous available tools and software libraries designed to work with XML documents and that provide XPath query support. We have found that XPath queries are particularly useful for exploring discourse patterning in the NC. Simple examples of XPath expressions include:

- `//u[count(descendant::seg[starts-with(@Quotatives, 'QS')])=1)]`
  counts utterances with one instance of quotative SAY each
- `//seg[@Components='CNN' and descendant::w[position()>= 100]]`
  selects all narratives with more than 100 words
- `//u[starts-with(@Participation_roles,'PN') and descendant::seg[@Reporting_modes='MDF']]`
  finds all narrator utterances which contain Free Direct discourse presentation[9]

In addition to the XML database and XPath/XQuery queries we wanted to be able to create frequency lists of various subcorpora (item length: 1–4, item type: word form, lemma, POS) and to carry out keyness analysis of various subcorpora (item length: 1–4, item type: word form, lemma, POS; see Figure 5). These kinds of lexical and frequency list based functions are distinctly suboptimal when implemented using XPath/XQuery. So we created a relational database view of the annotated NC following the pattern described in O'Donnell & Römer (In preparation), where each row represents a word (`<w>` element) in the corpus with a certain amount of right and left context and a series of values for textual component, participant roles, reporting modes and so on as additional fields. This approach essentially 'flattens' the XML hierarchy but allows for the highly efficient construction of frequency lists of words, lemma, POS tags and n-grams grouped by these values.

Bereitgestellt von | Bayerische Staatsbibliothek
Angemeldet | 194.95.59.195
Heruntergeladen am | 05.11.12 19:08



Item	Freq. A	Freq. B	Keyness
used to	30	38	26.030
was on	15	8	25.741
this morning	16	10	25.074
last night	21	20	24.205
my dad	9	4	16.922
i went	26	49	12.794

Figure 5. *Keyness comparison tool for NC.*

5. Concluding remarks

This paper has reported on the construction and annotation of a corpus of conversational narrative in British English. A distinguishing feature of the corpus is that it enlarges the as yet small group of corpora with discourse and pragmatic annotation.

The aim in this paper has been to demonstrate the great potential the NC offers for narrative research from a broad range of linguistic disciplines including not only corpus linguistics but also pragmatics, conversation analysis, discourse analysis, and sociolinguistics.

Specifically, it is worth emphasizing that the rich discourse annotation provided in the NC may enable investigations impossible in raw-text, POS-tagged, and parsed corpora. Unlike these 'traditional' corpus types, which enable lexically driven analyses examining mainly surface phenomena, the various layers of discourse annotation in the NC can be combined in novel ways thus enabling new approaches to the study of discourse and pragmatics. For example, in Rühlemann et al. (2011) we investigate how paralinguistic features interact with discourse presentation. By performing analyses of the co-occurrence of silent pauses (which have XML tags) with Direct and Indirect reporting modes, tagged MDD and MII, we discovered that silent pauses exhibit a clear tendency to mark the onset of Direct discourse presentation. That is, what was revealed by means of the annotation is the statistical co-occurrence of one discourse phenomenon (silent pauses) with another discourse phenomenon (discourse presentation). Many more discourse associations of this type can be

Bereitgestellt von | Bayerische Staatsbibliothek
Angemeldet | 194.95.59.195
Heruntergeladen am | 05.11.12 19:08

investigated using the markup available in the NC (cf. Rühlemann Forthcoming). We suggest that this annotation-driven corpus linguistics has great potential for the examination of discourse and pragmatic phenomena that lie beyond the lexical surface level.

It is hoped that the corpus will thus advance the linguistic theory of conversational narrative and will be used by other researchers interested in narrative, a primordial site of everyday social interaction.

Appendix

(KCE-N1)

```
<div title="Dropped your johnnies" embedLevel="EC2"
                                   narrativeType="T10">
  <seg Components="CNN">
    <seg Components="CNI">
      <u who="PS0EG" Participation_roles="PNP">
        <s n="2659" sID="2659"/>
        <shift new="singing"/><unclear/><shift/>
        <w c5="ITJ" hw="oh" pos="INTERJ">Oh </w> <w c5="PNP" hw="it" pos="PRON">it
        </w> <w c5="VBD" hw="be" pos="VERB">was </w> <w c5="AV0" hw="so" pos="ADV">so
        </w> <w c5="AJ0" hw="funny" pos="ADJ">funny </w> <w c5="PRP" hw="at"
        pos="PREP">at </w> <w c5="NN1" hw="work" pos="SUBST">work </w> <w c5="AV0"
        hw="today" pos="ADV">today</w> <c c5="PUN">, </c><w c5="NP0" hw="greg"
        pos="SUBST">Greg </w> <w c5="VVD" hw="fall" pos="VERB">fell </w> <w c5="PRP-
        AVP" hw="off" pos="PREP">off </w> <w c5="DPS" hw="he" pos="PRON">his </w>
        <w c5="NN1" hw="chair" pos="SUBST">chair</w> <c c5="PUN">.</c>
        <s eID="2659"/>
      </u>
    </seg>
    <u who="PS0EB" Participation_roles="PRR">
      <vocal desc="laugh"/>
    </u>
    <u who="PS0EG" Participation_roles="PNP">
      <s n="2660" sID="2660"/>
      <w c5="NN1" hw="packet" pos="SUBST">Packet </w> <w c5="PRF" hw="of"
      pos="PREP">of </w> <w c5="NN2" hw="condom" pos="SUBST">condoms </w> <w c5="VVD"
      hw="fall" pos="VERB">fell </w> <mw c5="PRP"><w c5="AVP" hw="out" pos="ADV">out
      </w> <w c5="PRF" hw="of" pos="PREP">of </w> </mw><w c5="DPS" hw="he"
      pos="PRON">his </w> <w c5="NN1" hw="pocket" pos="SUBST">pocket </w>
      <align with="KCELC0UU"/><unclear/><align with="KCELC0UV"/>
      <s eID="2660"/>
    </u>
    <u who="PS0EB" Participation_roles="PRR">
      <align with="KCELC0UU"/>
      <vocal desc="laugh"/><align with="KCELC0UV"/>
    </u>
    <u who="PS0EG" Participation_roles="PNP">
```

Bereitgestellt von | Bayerische Staatsbibliothek
Angemeldet | 194.95.59.195
Heruntergeladen am | 05.11.12 19:08

```
  <s n="2661" sID="2661"/>
  <shift new="laughing"/>
  <w c5="CJC" hw="and" pos="CONJ">And </w> <w c5="PNP" hw="they" pos="PRON">they
  </w> <w c5="VBD" hw="be" pos="VERB">were </w> <w c5="VVN-AJ0" hw="rip"
  pos="VERB">ripped </w> <shift/><c c5="PUN">.</c>
  <s eID="2661"/>
</u>
<u who="PS0EB" Participation_roles="PRR">
  <vocal desc="laugh"/>
</u>
<u who="PS0EG" Participation_roles="PNP">
  <s n="2662" sID="2662"/>
  <w c5="ITJ" hw="ah" pos="INTERJ">Ah </w> <w c5="ITJ" hw="no" pos="INTERJ">no
  </w> <w c5="PNP" hw="he" pos="PRON">he </w> <w c5="VBD" hw="be"
  pos="VERB">was</w> <c c5="PUN">, </c><w c5="PNP" hw="he" pos="PRON">he</w>
  <c c5="PUN">, </c><w c5="PNP" hw="he" pos="PRON">he </w> <w c5="VM0" hw="would"
  pos="VERB">would</w> <w c5="XX0" hw="not" pos="ADV">n't </w> <w c5="VVI"
  hw="sit" pos="VERB">sit </w> <w c5="PRP-AVP" hw="on" pos="PREP">on </w>
  <w c5="DPS" hw="he" pos="PRON">his </w> <w c5="NN1" hw="chair" pos="SUBST">chair
  </w>
  <seg Reporting_modes="MSS">
    <w c5="CJS" hw="cos" pos="CONJ">cos </w> <w c5="PNP" hw="he" pos="PRON">he</w>
    <w c5="VHD" hw="have" pos="VERB">'d </w> <w c5="AV0" hw="just" pos="ADV">just
    </w> <w c5="VVN" hw="call" pos="VERB">called </w> <w c5="PNP" hw="i"
    pos="PRON">me </w> <w c5="AT0" hw="an" pos="ART">an </w> <w c5="NN1"
    hw="arsehole" pos="SUBST">arsehole </w>
  </seg>
  <pause/><w c5="CJC" hw="and" pos="CONJ">and </w> <w c5="PNP" hw="i" pos="PRON">I
  </w>
  <seg Quotatives="QGZ">
    <w c5="VVZ" hw="go" pos="VERB">goes </w>
  </seg>
  <seg Reporting_modes="MDD">
    <w c5="ITJ" hw="oh" pos="INTERJ">oh </w> <w c5="VVB" hw="sit" pos="VERB">sit
    </w> <w c5="AVP-PRP" hw="down" pos="ADV">down </w> <unclear/><w c5="NP0"
    hw="greg" pos="SUBST">Greg</w> <c c5="PUN">!</c>
  </seg>
  <s eID="2662"/>
  <s n="2663" sID="2663"/>
    <vocal desc="laugh"/>
    <w c5="PNP" hw="i" pos="PRON">I </w>
  <seg Quotatives="QSD">
    <w c5="VVD" hw="say" pos="VERB">said </w>
  </seg>
  <seg Reporting_modes="MDD">
    <w c5="VVB" hw="sit" pos="VERB">sit </w> <w c5="AVP-PRP" hw="down"
    pos="ADV">down </w> <w c5="NP0" hw="gregory" pos="SUBST">Gregory </w>
    <w c5="CJC" hw="and" pos="CONJ">and </w> <w c5="VVB" hw="shut"
    pos="VERB">shut </w> <w c5="AVP" hw="up" pos="ADV">up</w> <c c5="PUN">.</c>
  </seg>
  <s eID="2663"/>
  <s n="2664" sID="2664"/>
    <w c5="AV0" hw="so" pos="ADV">So </w> <w c5="PNP" hw="he" pos="PRON">he </w>
```

Bereitgestellt von | Bayerische Staatsbibliothek
Angemeldet | 194.95.59.195
Heruntergeladen am | 05.11.12 19:08

```
    <w c5="VVD" hw="go" pos="VERB">went </w> <w c5="TO0" hw="to" pos="PREP">to
    </w> <w c5="VVI" hw="sit" pos="VERB">sit </w> <w c5="AVP" hw="down"
    pos="ADV">down </w> <w c5="CJC" hw="but" pos="CONJ">but </w> <w c5="DPS"
    hw="he" pos="PRON">his </w> <w c5="NN1" hw="chair" pos="SUBST">chair </w>
    <w c5="VBD" hw="be" pos="VERB">were</w> <w c5="XX0" hw="not" pos="ADV">n't
    </w> <w c5="AV0" hw="there" pos="ADV">there</w> <c c5="PUN">.</c>
  <s eID="2664"/>
  <s n="2665" sID="2665"/>
    <pause/><w c5="DT0" hw="all" pos="ADJ">All </w> <w c5="PNP" hw="i"
    pos="PRON">I </w> <w c5="VVD" hw="see" pos="VERB">saw </w> <w c5="VBD" hw="be"
    pos="VERB">were </w> <w c5="DT0" hw="this" pos="ADJ">this </w> <w c5="NN0"
    hw="pair" pos="SUBST">pair </w> <w c5="PRF" hw="of" pos="PREP">of </w>
    <w c5="NN2" hw="leg" pos="SUBST">legs </w> <w c5="VVG" hw="stick"
    pos="VERB">sticking </w> <w c5="PRP-AVP" hw="over" pos="PREP">over </w>
    <w c5="AT0" hw="the" pos="ART">the </w> <w c5="NN1" hw="desk"
    pos="SUBST">desk </w> <w c5="CJC" hw="and" pos="CONJ">and </w> <w c5="PNP"
    hw="he" pos="PRON">him </w>
  <seg Quotatives="QGG">
    <w c5="VVG" hw="go" pos="VERB">going </w>
  </seg>
  <seg Reporting_modes="MDD">
    <w c5="NN1-VVB" hw="aaaaagh" pos="SUBST">aaaaagh</w> <c c5="PUN">!</c>
  </seg>
  <s eID="2665"/>
</u>
<u who="PS0EF" Participation_roles="PNC">
  <s n="2666" sID="2666"/>
  <w c5="CJC" hw="and" pos="CONJ">And </w> <w c5="DPS" hw="he" pos="PRON">his
  </w> <w c5="NN2" hw="condom" pos="SUBST">condoms </w> <align with="KCELC0UW"/>
  <unclear/><align with="KCELC0UX"/>
  <s eID="2666"/>
</u>
<u who="PS0EG" Participation_roles="PNP">
  <s n="2667" sID="2667"/>
  <align with="KCELC0UW"/>
  <w c5="CJC" hw="and" pos="CONJ">And </w> <w c5="PNP" hw="he" pos="PRON">he </w>
  <w c5="VVD" hw="get" pos="VERB">got </w> <w c5="AVP" hw="up" pos="ADV">up </w>
  <align with="KCELC0UX"/><w c5="CJC" hw="and" pos="CONJ">and </w> <w c5="AV0"
  hw="then" pos="ADV">then </w> <w c5="CRD" hw="one" pos="ADJ">one </w> <w
  c5="PRF" hw="of" pos="PREP">of </w> <w c5="AT0" hw="the" pos="ART">the </w>
  <w c5="NN2" hw="girl" pos="SUBST">girls </w>
  <seg Quotatives="QSD">
    <w c5="VVD" hw="say" pos="VERB">said </w>
  </seg>
  <pause/>
  <seg Reporting_modes="MDD">
    <w c5="ITJ" hw="hi" pos="INTERJ">hi </w> <w c5="NP0" hw="greg"
    pos="SUBST">Greg </w> <pause/><w c5="VVD" hw="drop" pos="VERB">dropped </w>
    <w c5="DPS" hw="you" pos="PRON">your </w> <w c5="NN2" hw="johnny"
    pos="SUBST">johnnies</w> <c c5="PUN">.</c>
  </seg>
  <vocal desc="laugh"/><align with="KCELC0V0"/>
  <s eID="2667"/>
```

Bereitgestellt von | Bayerische Staatsbibliothek
Angemeldet | 194.95.59.195
Heruntergeladen am | 05.11.12 19:08

```
    </u>
    <u who="KCEPSUNK" Participation_roles="PRR">
      <align with="KCELC0UY"/><vocal desc="laugh"/><align with="KCELC0V0"/>
    </u>
    <seg Components="CNF">
      <u who="PS0EG" Participation_roles="PNP">
        <s n="2668" sID="2668"/>
        <w c5="PNP" hw="i" pos="PRON">I</w> <w c5="VHB" hw="have" pos="VERB">'ve </w>
        <w c5="AV0" hw="never" pos="ADV">never </w> <w c5="VVN" hw="see"
        pos="VERB">seen </w> <w c5="PNI" hw="anyone" pos="PRON">anyone </w>
        <w c5="VVI" hw="go" pos="VERB">go </w> <w c5="AV0" hw="so" pos="ADV">so </w>
        <w c5="AJ0" hw="red" pos="ADJ">red </w> <w c5="PRP" hw="in" pos="PREP">in </w>
        <w c5="DPS" hw="i" pos="PRON">my </w> <w c5="NN1" hw="life" pos="SUBST">life
        </w> <c c5="PUN">.</c>
        <s eID="2668"/>
      </u>
    </seg>
  </seg>
</div>
```

Bionotes

Chris Rühlemann is the author of *Conversation in Context: A Corpus-driven Approach* (Continuum, 2007) and *Narrative in English Conversation* (Cambridge University Press, forthcoming). He has published on different topics relating to conversational English in edited collections and journals such as *Applied Linguistics*, the *ICAME Journal*, the *International Journal of Corpus Linguistics*, and the *Journal of English Linguistics*. His main interests are in corpus linguistics, pragmatics and sociolinguistics. He is currently completing his professorial thesis (Habilitationsschrift) and co-editing (with Karin Aijmer) the *Handbook of Corpus Pragmatics* (Cambridge University Press, forthcoming), covering the fast-growing field of pragmatic studies based on corpus methods. E-mail: chrisruehlemann@googlemail.com

Matthew Brook O'Donnell is the lab manager and a researcher in the Communications Neuroscience Lab at the University of Michigan. He was previously a research fellow in the UM English Language Institute and involved in a range of projects including MICASE and MICUSP, contributing expertise in corpus compilation, annotation and the development of computational tools for analysis. His research interests include the integration of corpus and psycholinguistic methods, the study of language acquisition in terms of lexical associations and usage-based theories, as well as the application of techniques from machine learning and natural language processing to corpus linguistic tools and methods. E-mail: mbod@umich.edu

Bereitgestellt von | Bayerische Staatsbibliothek
Angemeldet | 194.95.59.195
Heruntergeladen am | 05.11.12 19:08

Notes

1. A small number of corpora of narrative seem to have come into existence. We are aware of three such corpora. An online (but not freely available) corpus consisting of more than 1,000 literary narrative works from medieval to modern times is the *Corpus de la littérature narrative du Moyen Âge au 20e siècle* (cf. http://www.usc.edu/libraries/databases/records/database.php?db=NIR). An analysis of a 150,00 word corpus sampled from 30 Spanish narrative works from the 20th century is presented in Irizarry (1990). The only spoken and thus more comparable corpus (we are aware of) is Carruther's (2008) corpus of French 'new story-telling' (néo-contage), a type of narrative performed publicly to eclectic audiences.
2. One of the tags which is, as an anonymous reviewer rightly noted, not overly perspicuous is 'T10', a label used to designate first-person experience stories. The tag may be replaced by a more transparent one in an updated version of the corpus.
3. Note the figures in Rayson et al. (1997: 135) for the whole BNC-C, where the discrepancy in verbosity is less marked (women: 2593452 words / 60%; men: 1714443 words / 40%).
4. Second-person stories, in which the teller relates what 'you' experienced, did not occur in the data.
5. See Norrick's (2008) work on narrator strategies for stimulating and modulating recipient response and, based on the NC, Rühlemann (Forthcoming) for a logistic regression model for response encouragement.
6. For analyses of narrative turn order, and turn size, patterns using the NC annotation, see Rühlemann (Forthcoming) and Rühlemann & Gries (In preparation).
7. See http://www.oxygenxml.com/.
8. See http://exist-db.org.
9. See Rühlemann (Forthcoming) for more complex XPath expressions and XQuery scripts applied to the NC.

References

Austin, John L. 1962. *How to do things with words.* 2nd ed. Cambridge/MA: Harvard University Press.

Aston, Guy and Lou Burnard. 1998. *The BNC handbook. Exploring the British National Corpus with SARA.* Edinburgh: Edinburgh University Press.

Baker, Paul. 2010. *Sociolinguistics and corpus linguistics*. Edinburgh: Edinburgh University Press.

Bamberg, Michael. 2004. Form and function of slut bashing in male identity constructions in 15-year-olds. *Human Development* 47: 331–353.

Barbieri, Federica. 2007. Older men and younger women. A corpus-based study of quotative use in American English. *English World-Wide* 28(1): 23–45.

Biber, Doug, Stig Johansson, Geoffrey Leech, Susan Conrad and Edward Finegan. 1999. *Longman grammar of spoken and written English. Harlow*: Pearson Education Limited.

Buchstaller, Isabelle. 2002. *He goes* and *I'm like*: The new quotatives re-visited. Internet Proceedings of the University of Edinburgh Postgraduate Conference 1–20.

Burnard, Lou. 2005. What is metadata and why do we need it? In Martin Wynne (ed.), *Developing Linguistic Corpora. A Guide to Good Practice,* 30–46*.* Oxford: Oxbow.

Burnard, Lou. 2007. *Reference guide for the British National Corpus (XML edition)*. (URL http://www.natcorp.ox.ac.uk/docs/URG/ accessed May 2011)

Bereitgestellt von | Bayerische Staatsbibliothek
Angemeldet | 194.95.59.195
Heruntergeladen am | 05.11.12 19:08

Burnard, Lou and Tony Todd. 2003. Xaira: An XML aware tool for corpus searching. In Dawn Archer, Paul Rayson, Andrew Wilson and Tony McEnery (eds.). *Proceedings of Corpus Linguistics 2003*, 142–144. Lancaster University.

Carletta, Jean, David McKelvie, Amy Isard, Andrea Mengel, Marion Klein and Moten Baun Møller. 2002. In Geoffrey Sampson and Diana McCarthy (eds.). *Corpus linguistics: Readings in a Widening Discipline*, 449–459. London: Continuum.

Carruthers, Janice. 2008. Annotating an oral corpus using the Text Encoding Initiative. Methodology, problems, solutions. *Journal of French Language Studies* 18: 103–119.

Clark, James and Steve DeRose. 1999. XML *Path Language (XPath) Version 1.0*. (available online at http://www.w3.org/TR/xpath [accessed May 2011]).

DeRose, Steven. 2004. Markup Overlap: A Review and a Horse. *Proceedings of Extreme Markup Languages 2004*. Montreal.

Durusau, Patrick and Matthew Brook O'Donnell. 2002. Just-In-Time-Trees (JITTs): Next Step in the Evolution of Markup? *InterCharge: Newsletter of the International SGML/XML Users' Group* 8(4): 21–26.

Edwards, Derek. 1997. Structure and function in the analysis of everyday narratives. *Journal of Narrative and Life History* 7(1–4): 207–215.

Ervin-Tripp, Susan M. and Aylin Küntay. 1997. The occasioning and structure of conversational stories. In Talmy Givón (ed.), *Conversation: Cognitive, communicative and social perspectives,* 133–166. Amsterdam/Philadelphia: John Benjamins.

Ferrara, Kathleen & Barbara Bell. 1995. Sociolinguistic variation and discourse function of constructed dialogue introducers: The case of *be + like*. *American Speech* 70(3): 265–290.

Garside, Roger, Geoffrey Leech and Tony McEnery (eds.) 1997. *Corpus annotation. Linguistic information from computer text corpora.* London/New York: Longman.

Garside, Roger, Steve Fligelstone, and Simon Botley. 1997. Discourse annotation: anaphoric relations in corpora. In Roger Garside, Geoffrey Leech & Tony McEnery (eds.), *Corpus annotation. Linguistic information from computer text corpora,* 66–84. London/New York: Longman.

Georgakopoulou, Alexandra. 2006a. Thinking big with small stories in narrative and identity analysis. *Narrative Inquiry* 16(1): 122–130.

Georgakopoulou, Alexandra. 2006b. The other side of the story: towards a narrative analysis of narrative-in-interaction. *Discourse Studies* 8(2): 235–257.

Gwyn, Richard. 2000. "Really unreal": Narrative evaluation and the objectification of experience. *Narrative Inquiry* 10(2): 313–340.

Harrington, Kate. 2008. Perpetuating difference? Corpus linguistics and the gendering of reported dialogue. In Kate Harrington, Lia Litosseliti, Helen Sauntson & Jane Sunderland (eds.), *Gender and language research methodologies*, 85–102. Basingstoke: Palgrave McMillan.

Hoey, Michael. 2005. *Lexical priming. A new theory of words and language*. London/New York: Routledge.

Hoey, Michael and Matthew B. O'Donnell. 2008. Lexicography, grammar, and textual position. *International Journal of Lexicography* 21(3): 293–309.

Hoffmann, Stefan, Stefan Evert, Nick Smith, David Lee and Ylva Berglund Prytz. 2008. *Corpus Linguistics with the BNCweb – A Practical Guide.* Frankfurt am Main: Peter Lang.

Holmes, Janet. 1997. Struggling beyond Labov and Waletzky. *Journal of Narrative and Life History* 7(1–4): 91–96.

Holmes, Janet & Maria Stubbe. 1997. Good listeners: Gender differences in New Zealand conversation. *Women and Language* 20(2): 7–14.

Ide, Nancy and Keith Suderman. 2006. Merging Layered Annotations. *Proceedings of Merging and Layering Linguistic Information*, Workshop held in conjunction with LREC 2006, Genoa, Italy.

Irizarry, Estelle. 1990. Stylistic analysis of a corpus of twentieth-century Spanish narrative. *Computers and the Humanities* 24(4): 265–274.

Bereitgestellt von | Bayerische Staatsbibliothek
Angemeldet | 194.95.59.195
Heruntergeladen am | 05.11.12 19:08

James, Deborah & Janice Drakich. 1993. Understanding gender differences in amount of talk: A critical review of research. In Deborah Tannen (ed.), *Gender and Conversational Interaction*, 281–312. New York/Oxford: Oxford University Press.

Jefferson, Gail. 1978. Discourse analysis and narrative. In: Jim Schenkein (ed.), *Studies in the organization of conversational interaction*, 219–248. New York: Academic Press.

Johnstone, Barabara. 1993. Community and contest: Midwestern men and women creating their worlds in conversational storytelling. In Deborah Tannen (ed.), *Gender and conversational interaction*. New York/Oxford: Oxford University Press, 63–80.

Kjellmer, Göran. 2009. Where do we backchannel? On the use of mm, mhm, uh huh and such like. *International Journal of Corpus Linguistics* 14(1): 81–112.

Labov, William. 1972. *Language in the Inner City*. Oxford: Basil Blackwell.

Labov, William and Joshua Waletzky. 1967/1997. Narrative analysis: Oral versions of personal experience. In June Helms (ed.), *Essays on the verbal and visual arts,* 12–44. Seattle: University of Washington Press. (reprinted in *Journal of Narrative and Life History* 7: 3–38).

Leech, Geoffrey & Mick Short. 1981. *Style in Fiction.* London/New York: Longman.

Leech, Geoffrey. 1997. Introducing corpus annotation. In Roger Garside, Geoffrey Leech, and Tony McEnery (eds.), *Corpus Annotation. Linguistic Information from Computer Text Corpora,* 1–18. London/New York: Longman.

Leech, Geoffrey. 2005. Adding Linguistic Annotation. In Martin Wynne (ed.), *Developing Linguistic Corpora. A Guide to Good Practice,* 17–29. Oxford: Oxbow.

Leech, Geoffrey, Tony McEnery, and Martin Wynne. 1997. Further levels of annotation. In Robert Garside, Geoffrey Leech, and Tony McEnery (eds.), *Corpus annotation. Linguistic information from computer text corpora,* 85–101. London/New York: Longman.

Leith, Dick. 1995. Tense variation as a performance feature in a Scottish folktale. *Language in Society* 24(1): 53–77.

Li, Charles L. 1986. Direct and indirect speech: A functional study. In: Florian Coulmas (ed.), *Direct and indirect speech*. Berlin: Mouton de Gruyter, pp. 29–45.

Macaulay, Ronald. 2001. "You're like why not?" The quotative expressions of Glasgow adolescents. *Journal of Sociolinguistics* 5(1): 3–21.

McEnery, Tony, Richard Xiao and Yukio Tono. 2006. *Corpus-based language studies: An advanced resource book*. London: Routledge.

McIntyre, Dan, Carol Bellard-Thomson, John Heywood, Tony McEnery, Elena Semino & Mick Short. 2004. Investigating the presentation of speech, writing and thought in spoken British English: A corpus-based approach. *ICAME Journal* 28: 49–76.

Norrick, Neal R. 2000. *Conversational narrative. Storytelling in everyday talk.* Amsterdam/Philadelphia: John Benjamins.

Norrick, Neal R. 2008. Negotiating the reception of stories in conversation. Teller strategies for modulating response. *Narrative Inquiry* 18(1): 131–151.

Norrick, Neal R. (Forthcoming). Interjections. In: Karin Aijmer & Christoph Rühlemann (eds.) The Cambridge Handbook of Corpus Pragmatics. Cambridge: Cambridge University Press.

Ochs, Elinor and Lisa Capps. 2001. *Living narrative. Creating lives in everyday storytelling.* Cambridge/MA: Harvard University Press.

O'Donnell, Matthew Brook. 1999. The Use of Annotated Corpora for New Testament Discourse Analysis: A Survey of Current Practice and Future Prospects. In Stanley E. Porter and Jeffrey T. Reed (eds.), *Discourse Analysis and the New Testament: Results and Applications*, 71–117. JSNTSup, 170; Sheffield: Sheffield Academic Press.

O'Donnell, Matthew Brook, Mike Scott, Michaela Mahlberg and Michael Hoey. 2012. Exploring Text-initial Concgrams in a Newspaper Corpus. *Corpus Linguistics and Linguistic Theory* 8(1): 73–101.

Bereitgestellt von | Bayerische Staatsbibliothek
Angemeldet | 194.95.59.195
Heruntergeladen am | 05.11.12 19:08

O'Donnell, Matthew Brook and Ute Römer. In preparation. Investigating the interaction between phraseological items and textual position.

Rayson, Paul, Geoffrey Leech and Mary Hodges. 1997. Social differentiation in the use of English vocabulary: Some analyses of the conversational component of the British National Corpus. *International Journal of Corpus Linguistics* 2(1): 133–152.

Rühlemann, Christoph, Andrej Bagoutdinov and Matthew Brook O'Donnell. 2011. Windows on the mind: Pauses in conversational narrative. *International Journal of Corpus Linguistics* 16(2): 198–230.

Rühlemann, Christoph. 2007. *Conversation in context: A corpus-driven approach*. London: Continuum.

Rühlemann, Christoph. Forthcoming. *Narrative in English conversation. A corpus analysis of storytelling as an interactional achievement*. Cambridge: Cambridge University Press.

Rühlemann, Christoph & Matthew Brook O'Donnell. Forthcoming. Deixis. In Karin Aijmer & Christoph Rühlemann (eds.), *Handbook of corpus pragmatics*. Cambridge: Cambridge University Press.

Rühlemann, Christoph & Stefan Th. Gries. In preparation. Turn-trigrams: Using turntaking patterns as a means for automatic identification of narratives in conversation.

Ryave, Alan L. 1978. On the achievement of a series of stories. In Jim Schenkein (ed.), *Studies in the organization of conversational interaction*, 113–132. New York: Academic Press.

Sacks, Harvey. 1992. *Lectures on conversation. Vols. I & II.* Cambridge: Blackwell.

Schegloff, Emanuel A. 1997. "Narrative Analysis" thirty years later. *Journal of Narrative Inquiry and Life History* 7(1–4): 97–106.

Schiffrin, Deborah. 1981. Tense variation in narrative. *Language* 57(1): 45–61.

Schiffrin, Deborah. 1984. How a story says what it means and does. *Text* 4(4): 313–346.

Schiffrin, Deborah. 1987. *Discourse markers.* Cambridge: Cambridge University Press.

Schiffrin, Deborah. 1996. Narrative as self-portrait: Sociolinguistic construction of identity. *Language in Society* 25(1): 167–203.

Schmid, Hans- Jörg. 2003. Do women and men really live in different cultures? Evidence from the BNC. In Andrew Wilson, Paul Rayson & Tony McEnery (eds), *Corpus Linguistics by the Lune. A Festschrift for Geoffrey Leech,* 185–221. Frankfurt am Main: Peter Lang.

Scott, Mike. 2010. *WordSmith Tools Version 5.0*. Lexical Analysis Software, Liverpool.

Tagliamonte, Sali & Rachel Hudson. 1999. *Be like* et al. beyond America: The quotative system in British and Canadian youth. *Journal of Sociolinguistics* 3(2): 147–172.

Tannen, Deborah. 1986. Introducing constructed dialogue in Greek and American conversational and literary narrative. In Florian Coulmas (ed.), *Direct and indirect speech,* 311–332. Berlin/New York/Amsterdam: Mouton de Gruyter.

Tannen, Deborah. 1988. Hearing voices in conversation, fiction and mixed genres. In Tannen, D. (ed.), *Linguistics in context: Connecting observation and understanding,* 89–113. Norwood, NJ: Ablex.

Winter, Joanne. 2002. Discourse quotatives in Australian English: Adolescents performing voices. *Australian Journal of Linguistics* 22(1): 5–21.

Yule, George and Terrie Mathis. 1992. The role of staging and constructed dialogue in establishing speakers topic. *Linguistics* 30: 199–215.

Bereitgestellt von | Bayerische Staatsbibliothek
Angemeldet | 194.95.59.195
Heruntergeladen am | 05.11.12 19:08