2 Data, methods, and tools

2.1 Introduction

This chapter is divided into two major subsections. Section 2.2 describes the construction and annotation of the NC, the corpus of conversational narratives, on which the analyses in this book are based. The account given here is brief; see Rühlemann & O'Donnell (2012) for a more detailed description. In Section 2.3, I outline major methods and tools used in this study. Specifically, I introduce XPath and XQuery, two related query languages which operate on the hierarchical tree structure of XML texts and which were consistently used in generating data for the analytical chapters.

2.2 The Narrative Corpus

2.2.1 Corpus construction

Data source

The NC is a specialized corpus of conversational narratives, all of which were extracted from the demographically sampled subcorpus of the British National Corpus (BNC) (XML version), a large general corpus assembled in the early 1990s in the UK (see Hoffmann *et al.* 2008). The demographically sampled subcorpus, BNC-C henceforth, consists of roughly 4.5 million words and is often referred to as the 'conversational' subcorpus because, as is widely agreed, its 153 files "consist of casual conversations" (Aston & Burnard 1998: 28; see also Rayson *et al.* 1997, Biber *et al.* 1999: 133).

Social balance

In order to ensure social balance, the creators of the BNC employed demographic sampling as is commonly used in public opinion research, that is, informants were selected "on the basis of their age, sex, region, social class, and so on" (McEnery & Wilson 1996: 65). A total of 148 informants were recruited in such a way that roughly equal numbers were achieved in terms of recruits' sex, age group, region, and social stratum.

To maintain this social balance in the NC, the aim was to include and annotate two narratives from each of the 153 texts in the BNC-C. This aim

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 20:18:37, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9781139026987.003

was achieved for the large majority of texts. The social composition of the NC is, indeed, a fair reflection of the composition of the BNC-C; for more detail, see Rühlemann & O'Donnell (2012).

As shown in Table 2.1, there are 500 speakers involved in the narrative components of the texts in the NC, of whom 212 are female and 173 male. The discrepancy is an echo of the small built-in speaker sex bias characteristic of the BNC-C.[1] The bias also comes to the fore in the number of words spoken by men and women: female participants (44,476/56%) talk almost twice as much as male participants do (24,268/31%). The greater verbosity of women in the NC is also reflected in the greater mean number of words: women produce, on average, 210 words, as compared to men's 140 words (for a more detailed discussion and preliminary hypotheses to explain the female overrepresentation in narrative, see Rühlemann & O'Donnell 2012).

As can be seen from Table 2.2, the number of speakers not coded for class (UU) is substantial: 272, representing 54% of all speakers. To judge from the speakers with class information, the distribution of the social classes in the NC too is similar to that of the BNC-C. The numbers of words per participant are reasonably similar across all four classes. As regards absolute numbers of participants per class, speakers coded for social class DE – that is, semiskilled

Table 2.1 *Distribution of male and female participants in the narrative (CNN) components of the NC*

Sex	Number of participants	%	Number of words	%	Mean number of words per participant
Female	212	42	44,476	56	209.79
Male	173	35	24,268	31	140.28
Unknown	115	23	10,141	13	88.18
	500	100	78,885	100	

Table 2.2 *Distribution of narrative participants in the narrative (CNN) component according to social class*

Social class	Number of participants	%	Number of words	%	Number of words per participant
AB	58	12	11,991	15	206.74
C1	77	15	15,239	19	197.91
C2	54	11	11,064	14	204.89
DE	39	8	7,625	10	195.51
UU	272	54	32,966	42	121.20
Total	500	100	78,885	100	

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 20:18:37, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9781139026987.003

Table 2.3 *Distribution of participants in the narrative (CNN) component according to age group*

Age group (years)	Number of participants	%	Number of words	%
0 (0–14)	62	12	6,539	8
1 (15–24)	81	16	12,765	16
2 (25–34)	65	13	11,159	14
3 (35–44)	57	11	13,335	17
4 (45–59)	57	11	12,165	15
5 (60+)	53	11	11,713	15
X or u	125	25	11,209	14
Total	500	100	78,885	100

and unskilled manual workers, e.g., bus drivers, laborers, barmen – seem to be underrepresented: not only are they far less numerous (39/8%) but they also contribute far fewer words (7,625/10%) than speakers coded for the classes AB (11,991/15%), C1 (15,239/19%), and C2 (11,064/14%). This might suggest that, as regards class DE, the NC lost the BNC-C's social balance out of sight. In actual fact, however, DE speakers are similarly underrepresented in the BNC-C (see Hoffmann *et al.* 2008).[2] In other words, the social-class data in the NC are a good match and reflection of the data in the BNC.

A similar distribution can be observed for speakers across the age parameter. Only one age group, viz. children in age group 0, are underrepresented in the NC – just as they are in the BNC-C. All other age groups are reasonably evenly represented.

On the whole, it seems fair to say that the social make-up of the NC is a good fit of the social make-up of the BNC-C and hence that the social balance characteristic of the BNC-C was also achieved for the NC.

Data extraction

Since the BNC contains annotation of parts of speech (POS) only, an automatic retrieval of narratives was not possible. Instead, two main types of extraction techniques were used: lexical extraction and extensive file reading.

Lexical extraction techniques were used by way of conducting searches in the BNC-C for lexical strings that either the literature or inspection of concordance lines suggested were frequent in storytelling. The strings used included, for example, '(there's/was) this guy/bloke/girl/man/woman,' 'it was so (funny, weird, etc.),' 'did I tell you,' 'reminds me,' 'anyway,' 'suddenly,' the interjections 'bloody hell' and 'oh my god,' the lemma REMEMBER and so on.

Also, narratives were extracted by extensive horizontal reading of files in the BNC-C, particularly those files for which the lexical 'hooks' did not help retrieve any candidate texts.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 20:18:37, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9781139026987.003

Selection criteria

Three criteria were critical for conversational passages to be included in the NC as narratives of which all three had to be fulfilled. The first such criterion was 'exosituational orientation.' That is to say that linguistic evidence was found that a speaker was referring to a situation different than the present, storytelling, situation. In personal experience stories, exo-indicators include, for example, past tense verbs, past-time deictics (*this morning*, *yesterday*, etc.), remote-place deictics and reference to referents not present in the telling situation. In fantasies, by contrast, which project future events, exosituational orientation is typically realized by future time reference and/or expressions of hypotheticality. Exosituational orientation in personal experience stories is exemplified in (2.1); exo-indicators are in square brackets:

(2.1) "Rugby injury" (Type: T10/Embed Level: ES)
CPR
(. . .)
CNN
CNI
S1 PNP I [had to take [Tony] [to the casualty] on [Saturday]]?
S3 PRC Why?

In (2.1), the incipient narrator, S1, steers the discourse ostensibly away from the present speech situation by referring to (i) a past event (*had to take Tony to the casualty on Saturday*), (ii) a third-person character (*Tony*), who, upon further reading, turns out to be absent from the present situation, (iii) a remote location (*the casualty*), and (iv) a past weekday (*Saturday*). Thus, a situational frame is established transferring the 'scene' from the here-and-now to a narrative there-and-then.

The second criterion was Labov & Waletzky's (1967/1997) 'a-then-b' relation (see Section 1.4), that is, the use of at least two temporally junctured narrative clauses.[3]

Finally, inter-rater agreement had to be achieved between two independent researchers. Texts were excluded when the raters did not agree.

The texts of a few BNC-C files did not meet the criteria and hence no narratives could be extracted from them. As shown in Table 2.4, a total of 279 texts were finally extracted from 143 BNC-C files. The NC contains 531 narratives and roughly 150,000 words.

Table 2.4 *Narrative Corpus – basic statistics*

Files	Texts	Narratives	Words
143	279	531	149,520

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 20:18:37, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9781139026987.003

Sampling

Since narratives are large linguistic units that unfold discursively, sampling techniques that provide for same-length text samples were seen as inappropriate. Instead, narratives were included as wholes (see Section 2.2.2). Moreover, given that a characteristic of conversational narratives is the use of techniques to ensure coherence between turn-by-turn talk and stories both at the beginning of stories and at story exit points (see Jefferson 1978) both pre-narrative stretches and post-narrative stretches of turn-by-turn talk were also included in the texts. The basic structure of the texts in the NC is thus threefold:

Pre-narrative → Narrative → Post-narrative

As a rule, both the pre-narrative and post-narrative stretches included in the texts were 15 utterances long. In many cases, however, this length could not be achieved, for example, when stories succeed each other there are typically very few non-narrative lines, if any, in between the stories. Also, the non-narrative passages were not calculated over (textual division) boundaries.

2.2.2 *Corpus annotation*

In the following subsections I outline the scheme underlying the annotation of the NC. For pointers to the manifold types of research the annotation may facilitate as well as more technical details relating to annotation procedures see Rühlemann & O'Donnell (2012). Note that the bulk of the annotations were implemented in the narrative components of the texts only. Codings used in narrative components but not in pre-narrative and post-narrative components include: quotatives, reporting modes, and participant status. Annotation of textual components as well as information on the type of embedding and type of narrative concerned the texts as wholes.

All tags were implemented manually; an initial tagged version of a text was proposed by one researcher and checked and, when necessary, revised by another. This inter-tagger collaboration facilitated a learning process allowing the researchers, as the tagging progressed, to revise decisions made earlier on the basis of increased coding competence. What errors and inconsistencies remain are due to the pressures of completing the project on time and with limited funding. It is hoped to correct potential errors in later versions of the NC.

The annotation implemented in the corpus combines the benefits of discourse analysis and corpus linguistics, enabling the analyst to read the texts both 'horizontally' and 'vertically' (see Tognini Bonelli 2010: 19). Manual annotation of narrative texts is based on the annotating linguist's careful horizontal reading of the texts, paying attention to smaller discourse units representing microstructures (such as presented discourse) and larger

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 20:18:37, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9781139026987.003

units relating to the texts' macrostructure (such as narrative boundaries). Each time a unit was identified it was 'captured' by the annotator using its corresponding tag, thus making the unit retrievable for further processing. This horizontal analysis can be seen as a text-linguistic analysis, more specifically in the case of conversational stories as discourse analysis. The fact that this discourse-analytical process was repeated for a large number of texts, amassing ever larger amounts of tags for different types of discourse units, enables a second major type of analysis, viz. the corpus linguistic analysis. This analysis 'reads' the texts and their units vertically, scanning for patterns in the use of the units and generating frequency data which lay the ground for statistical evaluation. The combination of the two methods removes obstacles that inevitably obstruct the analyst's view if either method is applied in isolation, instead allowing a two-dimensional perspective revealing many, and many new, structures and functions of conversational narrative.

Basics

Given the NC's unique status as the first and, as yet, only annotated corpus of conversational narrative in English, it was not possible in large measure to rely on established practices; rather, with the exception of discourse annotation (see McIntyre *et al.* 2004), the annotation scheme needed to be developed from scratch. In devising the annotation scheme, Leech's principle was observed that "[a]nnotation practices should be consensual" (Leech 2005: 21). Accordingly, a number of features that have long been in the spotlight in narrative research were not marked up on the grounds that they were unlikely to be coded identically by at least two coders; among them, for example, narrative structure in Labov's sense (e.g., 1972; see Section 1.6). Further, the annotation scheme follows the guidelines developed by the Text Encoding Initiative (TEI) as an application of the Extensible Markup Language (XML).

Among the many advantages and few disadvantages of TEI-conformant XML, two are of relevance in annotating the NC. A major advantage of XML is, as the name suggests, its extensibility. That is, while all TEI-encoded texts share a set of elements such as <w>, <seg>, , etc., these elements can most conveniently be extended by the researcher adding any attributes and attribute values that are specific to the project at hand. For example, using the <seg> element it was possible to annotate such diverse narrative-specific phenomena as quotatives, discourse presentation, and textual components. On the negative side, TEI-conformant XML works on the assumption that "the data is parsable in such a way that there will be no cross-cutting of elements" (Carruthers 2008: 117). In reality, however, elements may well overlap. For example, discourse presentation encoded as a <seg> element may cut across an intervening or overlapping recipient utterance, that is, a <u> element. Consider:

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 20:18:37, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9781139026987.003

(2.2) "Bob Marley" (Type: T10/Embed Level: EC3)

			Tag structure
1	**<u who='1'>**	(. . .) **<seg>**What! Do you know what you're saying you are, you are some stupid [unclear]**</seg> </u>**	<u> <seg> </seg> </u>
2	**<u who='2'>**	He's a legend, man. **</u>**	<u> </u>
3	**<u who='1'>**	<seg> fool.</seg> You know **</u>**	<u> <seg> </seg> </u>

The word 'fool' in utterance 3 is most likely an integral part of the discourse presentation the speaker, for convenience identified in the excerpt as '1', has begun in utterance 1. However, if the <seg> element which encloses the discourse presentation is extended to include the word 'fool', the <seg> element would overlap with the first <u> element (utterance 1). XML cannot process such overlaps. What XML requires instead is that start and closing tags be embedded within larger tags, as shown in the right-hand column in (2.2); there, the <seg> elements are opened and closed within <u> elements. Therefore, instead of using one <seg> element to capture the single instance of discourse presentation, which would do the linguistic reality more justice, it was necessary to use two separate <seg> elements, one in utterance 1 and another in utterance 3.

XML editor

<oXygen/>XML Editor[4] is a complete cross platform XML editor providing the tools, inter alia, for XML authoring. It also supports browsing, managing, and querying XML databases. A special feature of this editor is that all files are associated with a schema in which the annotations are laid down and which automatically checks XML validity and well-formedness of annotations implemented in the files. <oXygen/> was used for implementing the annotations in the texts and for performing XPath queries, an extremely useful query language for well-formed XML texts such as the NC described in more detail in Section 2.3.2.

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 were intended to be concise, consisting mostly of no more than three characters, perspicuous (that is, easy to interpret), and analyzable (that is, decomposable into their logical parts). Consider, for example, the tag QSZ (for quotative *says*). The tag is read from left to right: the left-hand character designates the attribute (Q stands for quotative); the middle character

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 20:18:37, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9781139026987.003

identifies the attribute (S stands for the verb SAY), whereas the right-hand character identifies the value the attribute is given (Z stands for the third person present tense form). This hierarchical left-to-right ordering is useful in searches using wildcards. For example, a search for QS* finds all forms of the lemma SAY used in quotative function, a search for Q* returns all verbs/forms marked up as quotatives.

The annotation scheme provides for markup on five different levels: text, textual components, utterance, speaker, and part of speech (POS). Readers will be sufficiently familiar with speaker meta-information and POS-tagging. These types of annotation are therefore not addressed in the following. The remaining three levels are briefly explained in the order in which they are located in the XML hierarchy. I begin with annotation related to text type.

Text

Annotation of the NC at text level is intended to grasp characteristics of narratives as textual wholes. In XML, this information is recorded at the level of the element. In the NC, the element records three attributes. The first such attribute is the narrative's 'title,' given by the researchers. In devising the titles it was attempted to use labels which were both concise (consisting of no more than one to four words) and precise (capturing what the story is about). Further, level annotation includes information on the type of embedding of narratives included in the text. Three main types of embedding are distinguished: narratives can stand alone (that is, no response stories were found in the 15-utterance spans following the story), they can be part of a sequence of narratives (what is referred to here as a 'narrative chain'), or they can be contained within a larger narrative ('embedded').

The tagset for embedding includes these values:

Tagset	Embedding	
E	ES	Stand-alone narrative
	EN	Nested narrative
	EC	Narrative chain
	EC1, 2, . . .	First, second, . . . narrative in narrative chain

(2.3) shows the title ("Fathers and daughters") given to text KBE-N2 and its identification as a narrative chain (EC):

(2.3)
(KBE-N2)

Thirdly, level annotation includes information on narrative sub-genre. The annotation of text type in the NC includes not only identification of first-person-experience stories but also a number of 'non-canonical' types of narrative. These include: third-person experience stories, which

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 20:18:37, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9781139026987.003

are "told not from events remembered first-hand but about someone else" (Norrick 2000: 149); 'generalized recurrent experience stories,' that is "narrative-like representation of a recurrent shared past experience in generalized form without reference to any specific instance" (Norrick 2000: 151); dreamtellings (Shanon & Eiferman 1984, Norrick 2000); 'fantasies,' projecting future events or constructing a fictional world (Norrick 2000, Georgakopoulou 2006a, b); 'jokes' (Sacks 1978, Norrick 1994, 2000); and, finally, what I will call 'mediated tellings,' including film retellings and stories about texting and MSN interactions (see Blum-Kulka 1993, Georgakopoulou 2008). The tagset for narrative subgenres includes these values:

Tagset Narrative type:

T	T1	T10	First-person experience narrative
		T1G	First-person generalized recurrent experience story
		T1D	First-person dreamtelling
		T1F	First-person fantasy
		T1M	First-person mediated story
	T3	T30	Third-person experience narrative
		T3G	Third-person generalized recurrent experience story
		T3D	Third-person dream
		T3F	Third-person fantasy
		T3M	Third-person mediated story
		T3J	Joke

As shown in Table 2.5, first-person narrative, in its various incarnations, is by far the most frequent macro-type of stories: in 415 stories, accounting for 78% of all stories, "the teller is the central actor or affected participant" (Norrick 2000: 135). This finding is consistent with findings made on the *Saarbruecken Corpus of Spoken English* (SCoSE), where first-person stories make up the vast majority too (Norrick 2000: 149). The 116 third-person narratives account only for 22% in the corpus.

Table 2.5 *Distribution of the 531 narratives in the NC across narrative types*

	Experiencer	1st person	No. of stories	%	3rd person	No. of stories	%
Experience	**Personal**	T10	337	63	T30	81	15
	Generalized	T1G	42	8	T3G	14	3
	Dream	T1D	9	2	T3D	1	0
	Mediated	T1M	14	3	T3M	10	2
	Fantasy	T1F	13	2	T3F	9	2
	Joke				T3J	1	0
Totals			415	78		116	22

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 20:18:37, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9781139026987.003

Considering next the individual narrative types, the prototypical conversational narrative type is first-person experience story (T10), which reports the teller's actual experience in a concrete anterior situation. The prototypicality of first-person experience is suggested by the sheer dominance of stories of this type in the corpus: the 337 first-person experience stories account for 63% of all stories recorded. The second most frequent type is third-person experience story, with 81 instances accounting for 15%. The third most common narrative type is first-person generalized experience story with 42 instances (8%). All other narrative types account for proportions smaller than 4%.[5]

Textual components

As noted earlier, the texts retrieved from the BNC contain not only the story but also passages preceding and succeeding the story. These three textual macro-components received tags: CPR for pre-narrative component, CNN for narrative component, and CPO for post-narrative component. Moreover, two micro-components were marked up within the narrative component CNN, viz. the utterance launching a story as well as the utterance bringing the story to a close; they were tagged: CNI for narrative-initial and CNF for narrative-final utterance. Another code, CNI-CNF, was used for utterances constituting whole stories, in which narrative-initial and narrative-final utterance merge into a single utterance. In a future version of the NC, the utterances enclosed by the narrative-initial and the narrative-final utterances, referred to as narrative-medial utterances, will be coded as CNM. The present tagset for textual components includes the following codes:

Tagset Narrative components:

C	CP	CPR	Pre-narrative conversation
		CPO	Post-narrative conversation
	CN	CNN	Narrative
		CNI	Narrative-initial utterance
		CNF	Narrative-final utterance
		CNI-CNF	Utterance which is at the same time narrative-initial and narrative-final

The annotation of the componential structure of texts depends on the identification of story boundaries; these, however, are often not salient (see Ervin-Tripp & Küntay 1997: 133). Determining precisely the junctures between non-narrative conversational text and narrative text may therefore not always be consensual. However, consensus is a rare amenity among linguists; even apparently simple decisions such as what counts as a word or a word class may be issues of debate (see Leech 2005: 21). Complete consensus is therefore not a primary criterion for the validity of annotational decisions. What is indispensable is that the annotations be based on clear, consensual principles. The principle guiding the text segmentation in the NC was exosituational orientation, the criterion underlying the selection of

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 20:18:37, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9781139026987.003

narratives (see Section 2.2.1). That is, the boundary between pre-narrative conversation and narrative was defined as that utterance in which a speaker observably re-adjusts the frame of reference from the present telling situation to a non-present told situation; for illustration, consider example (2.4). The speakers are a group of teenagers of African descent, who have been discussing the role of race in Hollywood films. In the exchange leading up to the utterance tagged as narrative-initial utterance (CNI) the discourse is clearly focussed on *present* states of affairs, as, for example, evidenced by the frequently occurring present tense forms. In CNI, however, although the present tense is retained at first (*Kerry's saying*) and only later exchanged for a past tense form (*Truno was going*), the focus shifts to events in a specific situation in the past, viz. Kerry's and Truno's exchange of opinions on Malcolm X:

(2.4) "Malcolm X" (Type: T30/Embed Level: EC1)

CPR		
(. . .)		
S6		Eddie Murphy films are good though [unclear]
S1		And Whoopie Goldberg and they're brilliant actor
S3		She, she's done some brilliant anti-racist films
S1		yeah they.
S2		But the thing is even people that, even though people are saying, well you've gotta get more black people into it, but if [unclear]
S4		You know, there's, there's a law in America.
S1		But that's, but that's what I mean, they're making, they're saying let's make, when the white people make films black and white people watch them, they're not specifically made for white people, not specifically made for black people, it just happens to be white actors. But when they make a black film, they make a black film for black people to sit and watch.
CNN		
CNI		
S1	PNP	I've got it all on this tape [^{MDF}Malcolm X], Kerry's saying ^{QSG} [^{MDD}Malcolm X is brilliant, he was a black man, he was a nigger] and Truno was going ^{QGG} [^{MDD}shut up] Truno's black too and she was going ^{QGG} [^{MDD}shut up! Malcolm X, stupid black nigger,] do you know what I mean. But because they're black it don't make no difference, do you know what I mean.
S3	PRC	[^{MUU}Cos if we say it]
(. . .)		
(KPG-N1)		

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 20:18:37, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9781139026987.003

Similarly, the first utterance in post-narrative conversation was defined as that utterance in which a speaker effectively re-directs the discourse from the told situation to the present situation. Note that from this decision follows the coding of the immediately preceding utterance as the narrative-final utterance (CNF). Where that utterance was some kind of telling-related listenership token, the next utterance further 'up' which contained story content was defined as narrative-final. To illustrate, in (2.5), the speakers, apparently a group of students, are co-constructing the story of a friend's rugby injury because of which he had to be taken to the casualty. Speaker S1's utterance *well he's er, he says* ^{QSZ} *[*^{MII}*he's done a big X-ray.]* is demonstrably the last utterance adding details to the story content (the following utterance *Mm.* by speaker S3 is merely telling-related and hence not counted as narrative-final). The immediately next utterance by speaker S5 (about the story guide the group is supposed to work through) has no topical connection to the happenings reported in the story; rather, the frame of reference is again the situation in the here and now:

(2.5) "Rugby injury" (Type: T10/Embed Level: ES)

	CNN		
	(. . .)		
	S6	PRC	Has he? What happened to him?
	S1	PNP	He got a knock on
	S3	PXX	[unclear]
	S1	PNP	on Saturday. But I think it was when he was tackling and erm
	S6	PXX	[unclear]?
	S1	PNP	got a cra no he didn't black out. And he says^{QSZ}, but he got like a crack on the side of his head and his hear
	S6	PRR	Oh! On his head?
	S1	PNP	and his hearing went
	S5	PX0	Any questions about that guide?
	S1	PNP	so I had to take him
	S6	PRC	Was it bad that [unclear]?
	S3	PRC	Was it the R V I or the General?
	S1	PNP	No.
	S2	PRC	Was it bad?
	S1	PNP	I'll sho oh I'll show
	S2	PXX	[unclear]?
	S6	PXX	Yes.
	CNF		
→	S1	PNP	well he's er, he says ^{QSZ} [^{MII}he's done a big X-ray.]
	S3	PRR	Mm.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 20:18:37, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9781139026987.003

CPO	
S5	This is your study guide. Work your way through the guide and you're going to answer all the questions on the quiz. As usual, the papers are taken and they're kept in the file for yo reference by your personal tutors. The reason that we're gonna produce biological science quizzes is they're, it's from your feedback evaluations that you don't get enough biological science.
S4	Soon put right.
S7	[laugh]
S5	Not a lot more I can say to you is there?
S6	[laugh]
(. . .)	

In story rounds, the boundary was set in that utterance which effected a re-orientation from the sequence of events occurring in one told situation to the event sequence occurring in another. The principle on which decisions underlying text segmentation rested was thus clear-cut. Once story boundaries were determined, determination of micro-components followed automatically, as first (narrative-initial) utterance and last (narrative-final) utterance; micro-componential annotation can thus, in itself, be considered uncontroversial.[6] The componential structure is shown in Figure 2.1.

The rationale for annotating the major textual components CPR, CNN, and CPO was as follows. Annotation of pre-narrative conversation (CPR) and post-narrative conversation (CPO) was included because the shifts in the generic framework from general conversation to narrative to general conversation along with their repercussions on lexis, turntaking, topic, etc. (see Jefferson 1978) can thus be examined corpus-linguistically.

The main rationale for annotating the micro-components CNI and CNF stems from recent research into textual positioning, revealing structural associations between lexis and position in text, referred to as textual

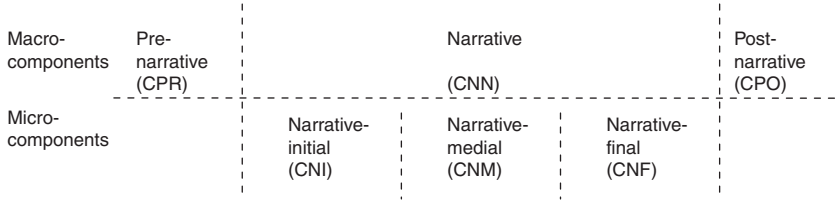


Figure 2.1 Componential structure of texts in the NC

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 20:18:37, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9781139026987.003

colligation (Hoey 2005). While research so far has concentrated on written genres, textual colligation in spoken genres is largely uncharted territory. Since first and last story utterance initiate shifts from one (sub)genre to another, these shifts are highly likely to have lexical correlates; see, for example, the role of introductory *this* in narrative-initial utterances (Rühlemann & O'Donnell forthcoming).

In terms of number of words, the NC's mix between non-narrative conversation and narrative discourse is fairly balanced: combined, the non-narrative components CPR and CPO contain 70,697 words while the narrative component CNN contains 78,823 words. In percentages, the non-narrative parts account for 47% of the whole NC, while the narrative components make up 53% of the corpus.[7]

Participant status

Utterance elements received attribute-value pairings for participant status. Building on the participation framework outlined earlier, the tagset used for participant status includes the following values:

Tagset Participant Roles:

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

For a detailed description of the concepts underlying the codes the reader is referred to Section 1.3. The code PXX is used for utterances that are narrative-related but whose status can not be determined with certainty; the code PXo designates utterances which are related not to the narrative but some event occurring outside of the narrative, e.g., greeting a new participant, admonishing a child, etc. The statistics for participant values are shown in Table 2.6.

As noted, the theme of this study is the co-construction of narrative, that is, the assumption that stories are interactionally achieved. The breakdown of participant roles shown in Table 2.6 lends support to this assumption. Unsupported Narrators (PNU), the label used in the NC which corresponds most closely to Labov & Waletzky's concept of the single teller, have very low shares both in terms of number of words and utterances. PNU utterances contain just 2,316 words, accounting for 3% of all words in the narrative components (CNN). Only 45 utterances (accounting for 1% of all utterances in CNN) are recorded for Unsupported Narrators (PNU). By default, one

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 20:18:37, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9781139026987.003

Table 2.6 *Participant values in narrative (CNN) components*

Participant type	No. of words	% of all words in the NC	No. of utterances	% of all utterances in the NC
PNU	2,316	3	45	1
PNS	10,009	13	288	5
PNP	52,807	67	2,488	42
PNC	3,783	5	450	7
PRR	1,677	2	1,264	21
PRC	6,270	8	919	15
PXo	1,632	2	313	5
PXX	329	0	293	5
Total	78,823		6,060	

PNU utterance constitutes one story. Inspection of contexts, however, shows that some of them get interrupted by listener utterances unrelated to the story, tagged PXo, which splits the story into two or more utterances. Therefore the actual number of *stories* by single tellers is smaller than 45, namely 38. This number accounts for merely 7% of all stories, a proportion reminiscent of the 5% of single-teller stories in Blum-Kulka's (1993: 366) collection of Israeli and American narratives. The next narrator category, Supported Narrator (PNS), gets higher shares: there are 288 PNS-tagged utterances (5% of all utterances in the NC), which include 10,009 words (13% of all words in the NC). By far the largest share is taken up by Primary Narrators (PNP): they have 52,807 words, which account for more than two-thirds (67%) of the word total in the narrative component. Primary Narrators also score highest in terms of number of utterances: they get 2,488 utterances, or 42%, of the utterance total in CNN. The share of Ratified Co-narrators (PNC) is rather small, with 3,783 words (5%) and 450 utterances (7%). As regards recipient types, the number of PRR utterances (1,264) exceeds the number of PRC utterances (919), whereas the reverse relation holds for the total number of words produced: PRC-tagged recipients produce 6,270 words as opposed to just 1,677 words produced by PRR-tagged recipients, suggesting that, unsurprisingly, PRR utterances are far shorter than PRC utterances. Figure 2.2 outlines mean utterance length across participant types.

Unsurprisingly, Unsupported Narrators (PNU), who, by definition, tell a story single-handedly, have by far the longest utterances (51.47 words on average). They are followed by Supported Narrators (PNS), who, by definition, receive only minimal telling/teller-related feedback: here, the average utterance length decreases substantially to 34.75 words. Primary narrators, by far the most common type of narrator in the NC, defined as receiving not only minimal telling/teller-related listener feedback but also tale-related responses as well as support from Ratified Co-narrator (PNC), produce, on average, even shorter utterances, with 21.23 words. The decline in utterance

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 20:18:37, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9781139026987.003

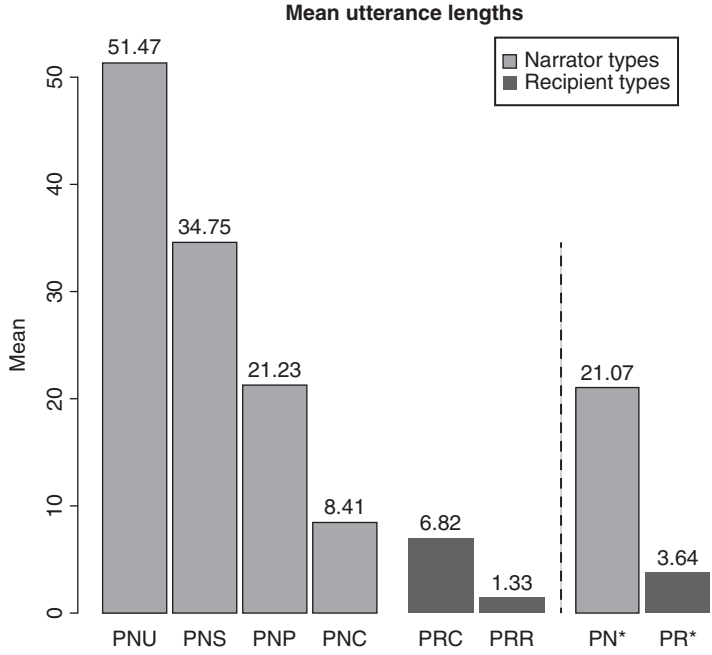


Figure 2.2 Mean utterance length in words across participant roles

length continues to PNC, who, like PNP, can interact with any other narrator type: their utterances are, on average, only 8.41 words long. The differences are (mostly very highly) significant.[8] Recipient responses to storytelling are again much shorter than utterances by any narrator type: minimal, telling/teller-related responses (PRR) are, on average, 1.33 words long, while substantial, tale-related responses (PRC) are 6.82 words long – again a very highly significant difference (see Section 2.3.2). (I will return to the differences in utterance lengths in Section 3.3.)

Discourse presentation

In annotating discourse presentation, two aspects were foregrounded: (i) quotatives, that is, the verbs or other forms used to 'announce' that the discourse forthcoming is not to be attributed to the speaker in the present situation but to a speaker (who may be the same as the present speaker) in a remote situation, and (ii) reporting mode, that is, the stylistic choice made by the speaker in presenting the discourse. Note that quotatives are relevant for two types of reporting mode only, Direct mode and Indirect mode. All other forms of reporting are quotative-free.

There is broad agreement that, in conversation, a small pool of forms dominate the quotative system despite the current "large-scale reorganization

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 20:18:37, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9781139026987.003

of the system" (Buchstaller 2011: 59), which is brought about by newcomers such as BE *like*. The inventory of quotatives comprises four major verbs, viz. SAY, GO, THINK, and BE *like*, largely independently of regional varieties; see, for example, Tagliamonte and Hudson's (1999) list of quotatives for British and Canadian English as well as Buchstaller's (2002) for American English (see also Tannen 1986).[9] For Scottish English a similar top list was observed in Macaulay (2001), while Winter (2002) discusses SAY, GO, and BE *like* (but not THINK) in Australian English. Quotatives in conversation are, then, shared across regional varieties to a very large extent.

Considering this core inventory, it was decided to tag quotative uses of SAY, GO, THINK, and BE *like*. Cursory inspection of the data in the NC suggested that ASK and TELL were of some frequency too, so these two verbs were coded for as well. Additionally, *like* without preceding form of BE was assigned a tag because it is "not uncommon that *like* alone functions as a demarcator between a quotation and the rest of the utterance" (Stenström *et al.* 2002: 1158). Finally, an 'umbrella tag,' QOO, was used for all other forms serving to introduce quoted material.

The tagset for quotatives typically includes five values per quotative, illustrated below using quotative GO; obviously, the tagsets for BE *like*,[10] *like* (without BE), and 'other quotatives' (QOO) diverge from this schema:

Tagset Quotative GO:

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

As regards THINK and other verbs that exhibit ambiguity between quotative introducer and stance marker, as in expressions such as *I think*, *I suppose*, etc.,

Table 2.7 *Quotatives and their tags in the NC*

Quotative	Tag
SAY	QS*
THINK	QT*
GO	QG*
BE like	QB*
Like	QLK
ASK	QA*
TELL	QL*
Other	QOO

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 20:18:37, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9781139026987.003

a restrictive conceptualization was adopted where only instances of genuine quotation referring to thought processes in displaced situations were coded as quotative (see Buchstaller 2011).

The total number of quotative tokens in the NC is 1,921. By far the most important quotative verb, it turns out, is SAY. Its five forms combined have a token frequency of 1,350. Thus, SAY alone accounts for 70% of all quotative forms. By far the most frequent form of the lemma is the past tense form *said*: it occurs 1,005 times, accounting for 74% of all forms of SAY and for 52% of all quotatives. In other words, more than every second quotative in the NC is the form *said*. The second most frequent quotative form of SAY, as well as of all quotative forms, is the present tense form *says*, which occurs 242 times.

The second most frequent quotative lemma is GO with a total of 239 tokens, accounting for 12% of all quotative tokens. Interestingly, contrary to the two most frequent forms *said* and *says* of the lemma SAY, it is not the past tense form *went* which is the most frequent form of the lemma GO but the present tense form *goes*. This finding is in support of Biber *et al.*'s (1999: 1119) claim that, among the quotative uses of GO, "the predominant form is the third person singular present tense *goes*." The third most frequent quotative is THINK, which has a total frequency of 214 tokens. Already far less frequent are TELL (total of 26 tokens) and ASK (16). That is, only three verbs dominate the quotative system in the corpus: SAY, GO, and THINK.[11]

The proportions of the quotative lemmas are given in Figure 2.3.

The annotation scheme used in annotating reporting modes in the NC follows to a large extent that described in McIntyre *et al.* (2004), which, in turn, builds on the seminal model of discourse presentation developed in Leech & Short (1981) and updated in Semino & Short (2004).

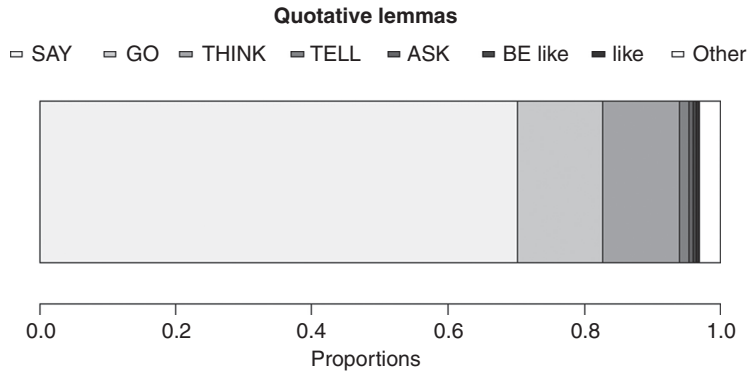


Figure 2.3 Proportions of quotative lemmas

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 20:18:37, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9781139026987.003

McIntyre *et al.*'s model distinguishes six main categories of discourse presentation: Direct, Free Direct, Indirect, Free Indirect, Representation of Speech Act, and Representation of Voice. Additionally, McIntyre *et al.*'s scheme recognizes a seventh, minor, category, Representation of Use (see below). The scheme underlying the NC extends that model in that it recognizes new categories not mentioned elsewhere. The tagset used for reporting mode includes ten values corresponding to ten reporting mode categories. An additional value, MXX, is for cases in which the ambiguity, inherent in "a large part of discourse presentation" (McIntyre & Walker 2011: 114), could not be resolved:

Tagset Reporting mode:

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

In the following, I briefly characterize the ten categories. A more substantial discussion of their properties and functions follows in Section 4.1. The two direct categories MDF and MDD have in common that, by and large, they present discourse as it would appear in natural utterances (but see Section 4.2). A structural distinction between the two is that MDD is part of a bi-clausal structure combining a reporting clause (such as *he said*, *she goes*, etc.) and a reported clause (the quote itself), whereas MDF is a mono-clausal structure, consisting only of the reported clause (see Semino & Short 2004: 11). The same structural distinction applies to the two indirect categories, where only MII (but not MIF) is coupled to a reporting clause. The indirect categories purport less to present the exact form, as the direct categories, but rather the content of the discourse reported (see McIntyre *et al.* 2004: 61). Further, MIF and MII see a decisive shift in perspective: while in the direct categories the perspective is that of the presentee, in the indirect categories it is that of the presenter (Coulmas 1986: 2). The perspectival shift has important consequences on the referential plane: while in the direct categories the deictic center is the presented speaker's, it is the presenting speaker's in the indirect categories; thus, in the latter, all deictic expressions are appropriate to the presenter (for example, a deictic such as *I* used within the indirect report refers indeed to the presenting speaker). Another category which is,

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 20:18:37, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9781139026987.003

like all remaining categories, both mono-clausal and deictically anchored in the presenter's deictic system is Representation of Speech Act (MSS). This type encodes the illocutionary force of the reported utterance. The categories further include Representation of Voice with Topic (MVT) and Representation of Voice (MVV). Both MVT and MVV capture "minimal references to speech with no indication of the illocutionary force, let alone the propositional content or form of the utterance (part)" (McIntyre *et al.* 2004: 629). MVT and MVV are distinguished by the fact that MVT includes mention of what the discourse was about (its topic), whereas MVV only indicates that discourse occurred (but not its topic). A minor category is Representation of Use (MUU), noted in passing in McIntyre *et al.* (2004). It captures "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).[12] The final two categories, Anaphoric Reference to Discourse Presentation (MRR) and Request for Discourse Presentation (MRQ), have not yet been recognized in previous research on discourse presentation. The two categories have in common that the anterior discourse is only being referred to. In MRR, the anterior discourse is referred to anaphorically and in declarative form. That is, the discourse has already been presented in some form in the preceding narrative discourse and is now being taken up again anaphorically, typically by use of referring expressions such as *that* and *it*. In MRQ, by contrast, the anterior discourse has not yet been given in any form; in fact, at the point at which MRQ occurs in narrative it is not even clear whether any anterior discourse can or will be reported. Rather, discourse presentation is only requested, typically, by a recipient. That is, the anterior discourse is referred to in interrogative form. Not surprisingly, all instances in the NC of this category feature the (referring) question pronoun *what*. It goes without saying that this latter category constitutes a borderline phenomenon between discourse presentation and narration or, more precisely, co-narration by the recipient. Its inclusion among the discourse presentation types may thus be questionable.

The ten reporting mode categories are exemplified in (2.6) to (2.15):

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

(2.7) MDD: I says QSZ [^{MDD} Lindsey if you want us to trust you, you have got to tell the truth] (KDS-N1)

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

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 20:18:37, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9781139026987.003

(2.9) MII: you know this erm, you know this young girl that was killed along Benji Avenue with her mother ? Well I **thought** [QTD] [**MII it was Wendy 's daughter**] (KCP-N1)

(2.10) MUU: PS549> Oh aye, aye. Aye, but . . . you know . . . like you se – [**MUU she mentioned one in particular**], like
PS54D> What?
PS549> [**MUU the word skeilth**]
PS54D> [**MUU Skeilth?**]
PS54E> Mm. (KPD-N1)

(2.11) MSS: Milk up [**MSS Nicola did n't believe me when I told her**.] She thought [QTD] [MII I was playing a joke .] (KPN-N2)

(2.12) MVT: Well actually on the television I do n't know whether you've noticed when **[MVT they start talking about burglaries and things like that]** the picture that comes on is Newlands Park you know (KDY-N1)

(2.13) MVV: [MVV She has talked, we've talked to her yesterday,] was it yesterday ? (KDS-N1)

(2.14) MRR: and he said [QSD] [MDDif you ain't gone by then] he said [QSD] [MDDI'll smash your fucking face in,] [**MRR that's exactly what he said to her**.] (KC1-N1)

(2.15) MRQ: [laughing] [**MRQ Well what did Mrs [name] say ?**] (KBW-N2)

A total of 2,712 reporting mode values have been assigned altogether. Direct discourse presentation, tagged MDD, is by far the most frequent reporting mode in the NC. Figure 2.4 lists the observed frequencies and the respective percentage values of all ten tags used for reporting modes.

The two direct modes, Direct (MDD) (58%) and Free Direct (MDF) (8%), combined account for two-thirds (66%) of all discourse presentation units in the NC, whereas the shares for the two indirect modes, Indirect (MII) (11%) and Free Indirect (MIF) (1%), and the third-ranked Representation of Speech Act (MSS) (9%) are far smaller. The 'norm' in conversational narrative, then, is for discourse presentation to be direct (see Leech & Short 1981). Further, Figure 2.4 shows that the four most frequent types (MDD, MII, MSS, and MDF) taken together account for 85% of all discourse presentation. As will be argued in Section 4.1, these four categories can be said to support varied degrees of retrievability of the utterance reported (the fifth category which facilitates retrievability, viz. MIF, is ranked eighth). This suggests that in conversational narrative, a basic concern in presenting discourse is in reporting with sufficient *immediacy*, offering the

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 20:18:37, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9781139026987.003

N	Reporting mode	Freq
1	MDD	1,573
2	MII	292
3	MSS	240
4	MDF	211
5	MVV	164
6	MVT	70
7	MUU	61
8	MIF	35
9	MRR	34
10	MRQ	14
11	MXX	32
Total		2,726

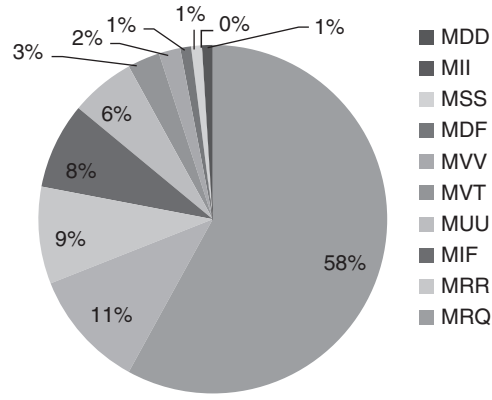


Figure 2.4 Frequencies and proportions of reporting mode types in narrative (CNN); MDD: Direct, MDF: Free Direct, MII: Indirect, MIF: Free Indirect, MSS: Representation of Speech Act, MVT: Representation of Voice with Topic, MVV: Representation of Voice, MUU: Representation of Use, MRR: Anaphoric Reference to Discourse Presentation, MRQ: Request for Discourse Presentation

recipient (varied degrees of) access to the utterance reported. The notion of immediacy in reporting is discussed in Section 4.1.

How do these proportions compare to proportions reported in related research? One point of reference is Mayes (1990), who investigated direct and indirect speech reports occurring in more than 22 hours of taped and transcribed conversation. If, in the NC, the instances of Direct and Free Direct and, on the other hand, Indirect and Free Indirect are combined and compared, the relative proportions thus obtained *exactly* reflect the proportions obtained in Mayes (1990): 85% for (Free) Direct and 15% for (Free) Indirect. Looking further beyond the direct and indirect modes, the only comparable corpus (I am aware of) which is consistently annotated for speech (and also thought) presentation is the *Lancaster Speech, Writing and Thought Presentation Spoken Corpus* (see McIntyre *et al.* 2004), a small, balanced corpus of contemporary spoken English drawn from two sources: the conversational subcorpus of the BNC and oral history archives from Lancaster University containing elicited interviews. While, as noted above, discourse presentation in the NC is annotated for ten categories, McIntyre *et al.* (2004) use six categories: Direct, Free Direct, Indirect, Free Indirect, Representation of Speech Act, and Representation of Voice. The proportions found in the LSWTPSC and the NC are juxtaposed in Figure 2.5.

As is shown in Figure 2.5, Direct mode (MDD) is the most common discourse presentation type in both corpora; however, in the NC it accounts

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 20:18:37, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9781139026987.003

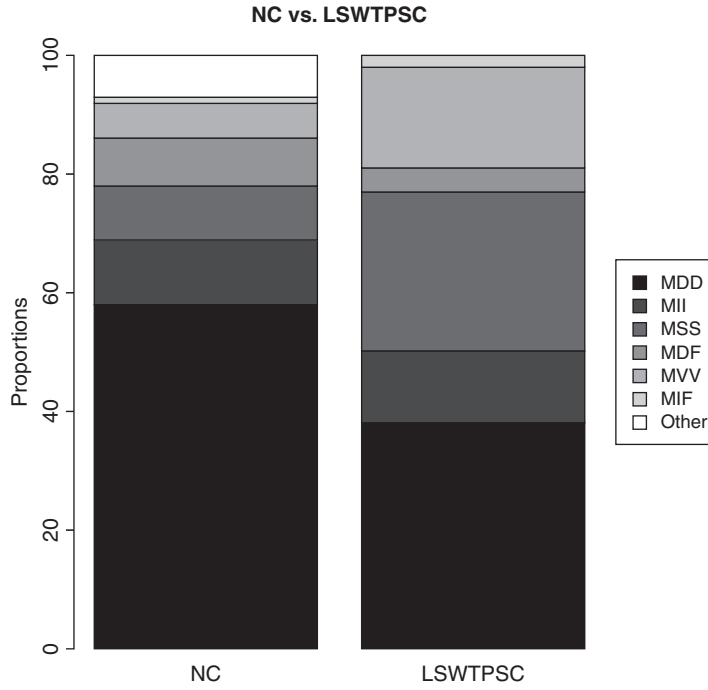


Figure 2.5 Proportions of six speech presentation categories in the LSWTPSC and the NC; the category 'Other' covers the proportions of MVT, MUU, MRR, and MRQ, which are not included in the LSWTPSC

for 58% and in the LSWTPSC it accounts for 38%. The NC also scores higher on Free Direct mode (MDF): here, the percentage value is 8% in the NC as opposed to 4% in the LSWTPSC. So, clearly, the (Free) Direct modes are overrepresented in the NC as compared to the LSWTPSC. On the other hand, the values obtained for Indirect mode (MII) are almost identical in the two corpora with 12% in LSWTPSC and 11% in the NC. Also, the proportions for Free Indirect mode (MIF) are very similar: 2% in LSWTPSC and 1% in the NC. While (Free) Direct is overrepresented in the NC and (Free) Indirect is comparably frequent in the corpora, MSS and MVV are clearly underrepresented in the NC: here they account for 9% (MSS) and 6% (MVV) respectively, while in the LSWTPSC they obtain proportions three times as high (27% for MSS and 17% for MVV). The differences between the two corpora are hence major, affecting mostly the direct categories, which are more important in the NC, and both MSS and MVV, which are more frequent in the LSWTPSC. It appears that the differences can be explained with the distinct generic compositions of the

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 20:18:37, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9781139026987.003

two corpora. As noted above, the LSWTPSC contains data from BNC-C, as does the NC, but the conversational data in the LSWTPSC are by no means narrative data alone, as in the NC, but present a good cross-section of all types of conversational subgenres. Second, the LSWTPSC also includes data from oral history interviews, a genre which is decidedly not conversational but much more formalized and restricted in terms of (fixed) role assignment (informant and researcher) and interaction type (interview questions from researcher to interviewee).

To conclude this section, the NC is a specialized corpus not only because it targets narratives and their surrounding conversational contexts but also because it is richly annotated for various discourse features. It thus adds to the small set of discourse-annotated corpora (see Garside *et al.* 1997).

2.3 Methods and tools

2.3.1 Methods

Keyness analysis

To fully exploit the rich annotational texture in the NC, a broad variety of corpus-linguistic methods are used. They are by no means mutually exclusive but conveniently complement each other. Most of them have been widely used and are well documented, among them: concordance line analysis (e.g., Hunston 2002: chapter 3), which facilitates essentially qualitative types of manual analysis, and predominantly quantitative methods such as word-lists (see Scott and Tribble 2006: chapter 2), including most notably frequency-ordered lists and keyness analysis (Scott 1997, Scott and Tribble 2006). Keyness refers to "items of unusual frequency in comparison with a reference corpus of a suitable kind" (Scott and Tribble 2006: 55). Analyzing keyness has the potential of revealing two key aspects of a text/text collection: what it is about (aboutness indication) and by what it is typified stylistically (stylistic indication) (see Scott and Tribble 2006: chapter 4). Keyness analysis is a particularly versatile method which can not only establish key words, but also key lemmas, key n-grams, key POS-tags and key semantic domains (see Rayson 2008), a methodological versatility that can be almost fully exploited using the NC analysis tools (see below). A specific type of keyness analysis is intra-textual keyness analysis (O'Donnell *et al.* 2012). The hallmark of this method is that key items are calculated for a subcorpus/subset of a text/text type (narratives in this case) as compared to another subcorpus/subset of the same text/text type (for example, key POS items in (Free) Direct reports compared against (Free) Indirect reports; see Section 4.2.1).

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 20:18:37, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9781139026987.003

Statistical inference

Given that this study aims to provide insights on co-construction gained from quantitative analyses of the narratives in the NC, particular emphasis is placed on hypothesis and significance testing. The intention is to report results that are meaningful not only for the NC, which is, after all, just a minute sample of the population of all conversational narratives in English, but for that very population as a whole. It is, however, essential to define that population more exactly. The population of which the NC is a (hopefully representative)[13] sample is *not* the total of all stories told, being told, and to be told in English, which is potentially infinite. That population is not known to anybody; it might even be said not to exist as such (containing as it does future stories). The population of which the NC is a likeness (in a nutshell) is much smaller (but still frighteningly large): it is the total of stories told by people living in a particular place on the globe at a particular point in time – namely the British Isles in the last decade of the twentieth century (the time the data for the BNC-C were collected). That is, the tests I am going to conduct are intended to examine whether the findings made on the NC can be generalized from the microcosm of the corpus to the macrocosm of telling stories in English conversation in Britain during the 1990s.[14] Almost two decades have since gone by, the millennium has changed – has storytelling in Britain changed as well (such that observations that could be generalized in the 1990s may no longer be generally relevant)? To address this question a representative corpus of storytelling in the early twenty-first century would be required. It does not exist, so the question cannot be answered with certainty. What is certain is that English has changed since the BNC-C was created. A striking example at the level of the lexicon is the quotative BE *like*, which had hardly entered British English in the first half of the 1990s and is accordingly virtually absent from the BNC but widespread today (see endnote 11). Grammatical change too has occurred; see, for example, the continued decline in modal usage (e.g., Leech *et al.* 2012). Storytelling in conversation, however, is essentially neither a lexical nor a grammatical phenomenon but a discourse phenomenon with crucial interactional underpinnings (such as recipient design, specific participation patterns, evaluation in its various forms, etc.). These may have their roots in interaction principles "that are at source independent of variations in language and culture" (Levinson 2006: 40; see also Schegloff & Lerner 2009). They are therefore unlikely to change in a 20-year span. What will be presented here as significant findings is hence likely to have remained significant into the early twenty-first century.

How is the generalization from sample to population achieved? The answer is, by using inferential statistics. Its aim is "to infer from the properties of a part the likely properties of the whole" (Woods *et al.* 1986: 49). The principle at the heart of inferential statistics is what Stefanowitsch (2005)

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 20:18:37, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9781139026987.003

termed the 'expected-frequency epiphany': "[o]bserved frequencies of occurrence must be evaluated against their expected frequencies of occurrence before they become relevant facts for scientific analysis" (Stefanowitsch 2005: 296). That is, the frequencies a researcher can observe are merely a starting point. What is additionally needed if a researcher wishes to make statements about the population is a second type of frequency, viz. expected frequencies. These are the frequencies that are expected in case the null hypothesis H_o is true. If, for instance, the observed frequencies differ a lot from the expected ones, then it is justified to say that H_o is not true and to accept the alternative hypothesis H_1. Expected frequencies, however, cannot be found in a corpus, they can only be calculated, on the basis of the observed frequencies, using appropriate statistical procedures. Although quantification undoubtedly lies at the heart of corpus linguistics – Gries (2009a: 11), for example, takes the extreme but arguable position that "the only thing corpora can provide is information on frequencies" – the uptake of inferential statistics into corpus linguistic research has been somewhat slow. Those corpus studies that do incorporate statistical tests only rarely state the hypotheses they are testing: neither the null hypotheses nor the alternative hypotheses that the analyses are tailored to are fully spelled out. It thus remains somewhat vague what precisely is being tested. To avoid this vagueness all the major hypotheses to be tested will be stated in explicit terms in this study.

Textual colligation

Finally, in a number of case studies, positional analyses will be carried out. They build on recent work which has demonstrated the close association of lexis with positions in texts (e.g., Hoey 2005, Scott & Tribble 2006, Hoey & O'Donnell 2008, Mahlberg & O'Donnell 2008). In Hoey's theory of 'lexical priming,' the claim is made that "every word is primed to occur in, or avoid, certain positions within the discourse" (Hoey 2005:13). This claim is referred to as 'textual colligation.' Also, textual positioning is "likely to be genre and even subgenre-specific" (Hoey & O'Donnell 2008: 300).

Research into text-lexis associations is an intriguing new area straddling the fields of lexical studies and discourse analysis. Except for Römer & O'Donnell (2010), who have examined (student) academic writing, work on textual positioning has tended to pivot on (sub)genres of news reportage. To the best of my knowledge, spoken data have not yet been examined in terms of textual positioning at all.

2.3.2 Tools

In this section, I describe the three tools key in this study: the NC analysis tools, the programming language R, and XPath and XQuery, two related

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 20:18:37, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9781139026987.003

query languages which have huge analytical potential for XML-tagged corpora.

NC analysis tools

Many analyses presented in this book build on tools developed by Matthew Brook O'Donnell at the University of Michigan. The tools he developed especially for the NC include: text browsing, KWIC, collocate tables (L1–4, R1–4), frequency lists of various subcorpora (item length: 1–4, item type: word form, lemma, POS), keyness analysis of various subcorpora (item length: 1–4, item type: word form, lemma, POS), distributional analysis across components, and XPath and XQuery queries (see below) allowing extraction and quantification of XML structures (for more details, see Rühlemann & O'Donnell 2012).

R

The tool used to perform significance tests of various sorts is R, a programming language and environment for statistical computing and graphics. While it has become standard among statisticians, its uptake in corpus linguistics is still in its early stages. For a not too technical introduction to R for corpus linguists, see Gries (2009a); for a more comprehensive and more advanced introduction to R for linguists of all denominations see Gries (2009b); an even more advanced introduction for linguists is Baayen (2008). Readable introductions for non-linguists are Dalgaard (2008) and, at a more advanced level, Crawley (2007). One of the major plusses of R is its very powerful graphical capabilities, of which only a small proportion can be exploited in this study (see for example, Mittal 2011).

One particularly useful graphical representation is the boxplot (see Benjamini 1988, Frigge *et al.* 1989). It will be used on a number of occasions in this book and is therefore introduced briefly in the following. The usefulness of the boxplot stems from the fact that it summarizes 'at a glance' several pieces of crucial information related to the location and spread of data as well as its potential skewness (see Crawley 2007: 155). A boxplot consists of five graphical components: a box, a bold line across the box, notches within the box, whiskers below and above the box, and, potentially, empty circles below and above the whiskers. The box represents the interquartile range (IQR), that is, the range from the first to the third quartile of the data (roughly, the 50% of the data around the median). The bold horizontal line depicts the median, a robust measure of the central tendency of the data (the middle value of a sorted distribution; see Woods *et al.* 1986: 30–32). The notches on either side of the median line indicate the 95% confidence interval within which the 'true' median is expected to fall (note that notches are an optional feature; for example, the boxplots in Section 3.3.3 are drawn without notches). Also, if the notches of two or more boxes do *not* overlap (that is,

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 20:18:37, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9781139026987.003

if they do not cover the same range on the y-axis), they "are likely to prove to have significantly different medians" (Crawley 2007: 157). The vertical dashed lines connecting the boxes and the whiskers "extend to maximally 1.5 times the interquartile range" (Baayen 2008: 30). The whiskers can be understood as 'fences' separating "the outermost observations that are not extreme enough to be flagged as outside by an exploratory rule of thumb" (Frigge *et al*. 1989: 51) from outliers defined as "data points with values that are surprisingly large or small given all data points considered jointly" (Baayen 2008: 27). Finally, the empty circles represent individual data points falling outside the fences, thus designating potential outliers. For illustration consider boxplots showing the above-mentioned lengths of utterances by the two recipient types, PRR and PRC.

In Figure 2.6, the boxplots are presented in two different panels. While the left panel shows the distributions of the lengths of utterances by the recipient types PRR and PRC entirely (with all outliers included), the right panel presents the boxes in greater detail to permit a clearer view of the boxes and the notches.

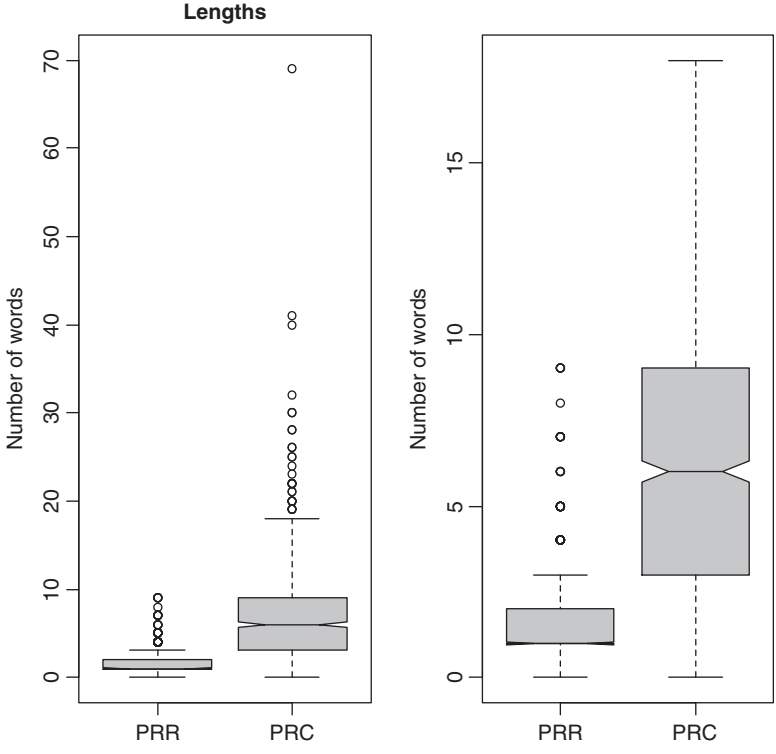


Figure 2.6 Boxplots of lengths of PRR and PRC responses

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 20:18:37, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9781139026987.003

In the left panel, it can be seen that the dispersion of the lengths of PRC responses is much greater than the dispersion of the lengths of PRR responses. This greater dispersion, and hence greater heterogeneity, is suggested, not only by the larger range of outliers (indicated by the empty circles), but also by the greater distance between the whiskers (the dashed vertical lines) and the larger size of the box. This skewness shows up even more clearly in the right panel. While, for PRC responses, the upper part of the box (above the median line) and the lower part are reasonably symmetrical, there exists no lower part (below the median line) for PRR – the box is decidedly asymmetrical. Also, the median (indicated in the bold horizontal line in the box) for PRC is located much higher on the y-axis than for PRR, suggesting that, on average, PRC utterances tend to be much longer than PRR utterances. Further, the notches (indicating the 95% confidence intervals) for PRR and PRC are far from overlapping. This is prima facie evidence suggesting that the difference in median length is significant (see Gries 2009a: 205). (Indeed, according to a Wilcoxon rank sum test, the difference is very highly significant: p-value < 2.2e-16.)

In the following, I introduce XPath and XQuery, two immensely powerful tools for the analysis of XML texts.

Xpath and XQuery

The XML Path Language, XPath, is a metalanguage which allows the navigation of XML documents (see Clark & DeRose 1999), such as the texts assembled in the BNC and the NC. To understand how XPath works an understanding of how XML texts are structured is helpful.

XML documents are characterized by the fact that any element in them is part of a hierarchical tree structure. XML elements can be empty but, more often, they have content, for example, a particular word form. For instance, the narrative component (CNN) of text KP1-N2 starts with the following words: *That was the week when Peter came*. The individual words are all defined as <w>-elements, receiving both a start tag (<w>) and an end tag (</w>):

(2.16) <w>That </w>
<w>was </w>
<w>the </w>
<w>week </w>
<w>when </w>
<w>Peter </w>
<w>came </w>

However, XML elements may be further defined by attributes, which, in turn, have values. The same passage from KP1-N2 with all attributes and values given in the NC reads as follows:

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 20:18:37, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9781139026987.003

(2.17)
```
<w c5="DT0" hw="that" pos="ADJ">That </w>
<w c5="VBD" hw="be" pos="VERB">was </w>
<w c5="AT0" hw="the" pos="ART">the </w>
<w c5="NN1" hw="week" pos="SUBST">week </w>
<w c5="AVQ" hw="when" pos="ADV">when </w>
<w c5="NP0" hw="peter" pos="SUBST">Peter </w>
<w c5="VVD" hw="come" pos="VERB">came </w>
```

All seven <w> elements in the excerpt have three attributes: *c5*, which stands for Claws5, the tagset used in automatically assigning part of speech (POS) tags to each word in the BNC; *hw* for the headword (or lemma) under which individual word forms can be subsumed; and *pos*, a simpler tagset which contains far fewer POS tags than the Claws5 system. Each of the three attributes has a value. The first word *That*, for example, has the value *DT0* on the *c5* attribute, which means that Claws5 identified the word *That* as a 'general determiner.' Since the word *that* has no morphological variation, the headword (*hw*) value is the same as the word form itself. Compare the next word *was*: the value it has on the *c5* attribute is *VBD*, a label used for the past tense form (D, taken from the most common past tense ending *–ed*) of the verb (V) BE (B) and the *hw* value is, accordingly, *be*.

XML documents are not simple strings of <w> elements. Just as leaves are parts of branches, which are parts of the tree, so XML documents are hierarchically ordered in the sense that sets of elements can depend on higher-order sets of elements, which, in turn, can themselves be dependent on yet higher sets of elements. Cf. example (2.18), which gives the same excerpt as above with its immediately surrounding annotations; dependency levels are numbered:

(2.18)

```
1 <div title="Mining accident" embedLevel="ES" narrativeType="T10">
2         <seg Components="CNN">
3                 <seg Components="CNI">
4                         <u who="PS50U" Participation_roles="PNP">
5                                 <w c5="DT0" hw="that" pos="ADJ">That </w>
5                                 <w c5="VBD" hw="be" pos="VERB">was </w>
5                                 <w c5="AT0" hw="the" pos="ART">the </w>
5                                 <w c5="NN1" hw="week" pos="SUBST">week </w>
5                                 <w c5="AVQ" hw="when" pos="ADV">when </w>
5                                 <w c5="NP0" hw="peter" pos="SUBST">Peter </w>
5                                 <w c5="VVD" hw="come" pos="VERB">came </w>
                          </u>
                  </seg>
          </seg>
  </div>
```

The highest-order element in the excerpt (level 1) is the (from 'division') element, which brackets the whole text; see the end tag .

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 20:18:37, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9781139026987.003

It has three attributes and accordingly, three values: the *title* attribute, whose value is *Mining accident*, and the attribute *embedLevel*, which states whether the narrative contained in the text is part of a narrative chain or a stand-alone story; here the value is *ES*, denoting that the text contains a single story. The third attribute is *narrativeType*; its value is *T10*, the label used for first-person narratives (see Section 2.2.2). At the next lower level (level 2) we find a <seg> (from 'segment') element whose attribute *Components* has the value *CNN*, that is, this <seg> identifies the beginning of a narrative component. One level further down (level 3), there is a second <seg> element; it has the same attribute *Components* but the value is *CNI* (for narrative-initial utterance). Yet another dependency level further down (level 4) we find a <u> (utterance) element. It has two attributes: *who* and *Participation_roles*. The *who* attribute acts like a passport defining the speaker in terms of sex, age, and a number of other social categories; the value is *PS50U*, which the text header identifies as a 70-year-old housewife speaking the dialect of Northwest Midlands. The *Participation_roles* attribute registers information about the narrative role the speaker of the utterance is taking on in the course of the incipient narration; the value here is *PNP* (Primary Narrator). Finally, at the lowest dependency level (level 5), there are all the <w> elements. An important term that comes into play at this stage is 'node': elements and attributes (including their values) are referred to as nodes.[15]

XPath operates on that node tree in such a way that any given (series of) nodes can be defined via what are called 'axes.' So, what we have so far referred to as dependency levels 1–4 would be referred to in XPath parlance as the axes 1–4. XPath axes "tell the XPath processor which 'direction' to head in as it navigates around the hierarchical tree of nodes" (Watt 2002: 82). To express the various levels of superordination and subordination between the axes, the XPath syntax uses a terminology which is borrowed from human family relations. Thus, in (2.18) above, the first two axes stand in a parent-child relationship: axis 1 is the parent of axis 2, or the reverse, axis 2 is the child of axis 1. The first and all subsequent axes stand in an ancestor-descendant relationship: axis 1 is the ancestor of axes 2–4, while axes 2–4 are descendants of axis 1. Looking at axis 4 only, the Xpath syntax defines all seven <w> elements, which are on the same 'footing' with one another, as siblings.

If we assume, for illustrative purposes, that (2.18) were a complete XML document[16] we could make use of the knowledge of the structure of XML documents and of the XPath syntax to formulate 'paths' taking us to any node in it.

For example, the simple XPath `//div` locates the text printed in (2.18). This is, of course, not surprising in that the document has only one element. If the document contained more narratives it would contain more elements and, accordingly, the XPath `//div` would find all these elements. If, in that longer document, we were specifically interested in just the text called *Mining accident*, we could exploit the text's title attribute

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 20:18:37, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9781139026987.003

using this XPath: //div[@title='Mining accident']. Since no other text in the NC has the same title value, this XPath would unambiguously identify the text *Mining accident*. Another simple XPath would be //u: this finds any utterance in the document. Since we have only one <u> element, it would find the utterance in (2.18). If the document were longer containing more utterances assigned to different participant roles, the XPath //u [@Participation_roles='PNP'] would find all utterances issued by a Primary Narrator. If, in the same longer document, we were interested in precisely the utterance given in (2.18), the XPath //u[@who='PS50U' and parent::seg[@Components='CNI']] would locate the utterance in question by looking out for an utterance which is produced by speaker *PS50U* and whose parent element is a <seg> element whose *Components* attribute has the value *CNI*.

If our interest were in the two verbs in (2.18), *was* and *came*, we might formulate this XPath: //w[@pos='VERB' and ancestor::u [@who='PS50U']]. This tells the processor to look for any simple POS tag that has the value *VERB* and whose ancestor axis is an utterance issued by the speaker whose speaker ID is assigned the value *PS50U*. By contrast, the XPath //u[@Participation_roles='PNP' and descendant::w[8]] would not find anything in the document in (2.18) because the XPath specifies that the utterance must have at least eight <w> elements, whereas (2.18) has only seven.

So far we have been examining a single (constructed) document consisting of four distinct axes and just a few words. The underlying structure of the text in (2.18) can be represented thus:

 → <seg> → <seg> → <u> → <w>

This structure is simplistic. The texts and their annotations in the NC are invariably more complex. The texts not only consist of many more words but also of many more annotations, which may even be embedded within each other. To begin, the majority of the texts in the NC contain more than one narrative. As a result, one element is used for the whole text and, one level further down, there may occur as many subordinated elements as there are narratives in the text. Further, <seg> elements were used to encapsulate both textual components as well as quotatives and reporting mode types. Discourse presentation is a feature of utterances. Therefore <seg> elements denoting quotatives and reporting modes must occur within <u> elements. Finally, embedded narratives (that is, narratives told by a recipient before the current narrator has finished his/her narrative) are given a element too, which 'intrudes' into the sequence of utterances issued by the current narrator. Thus, the structure of texts in the NC is often more likely to look like this:

 → → <seg> → <seg> → <u> → <w> → <seg> → → <u> → <w> . . .etc.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 20:18:37, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9781139026987.003

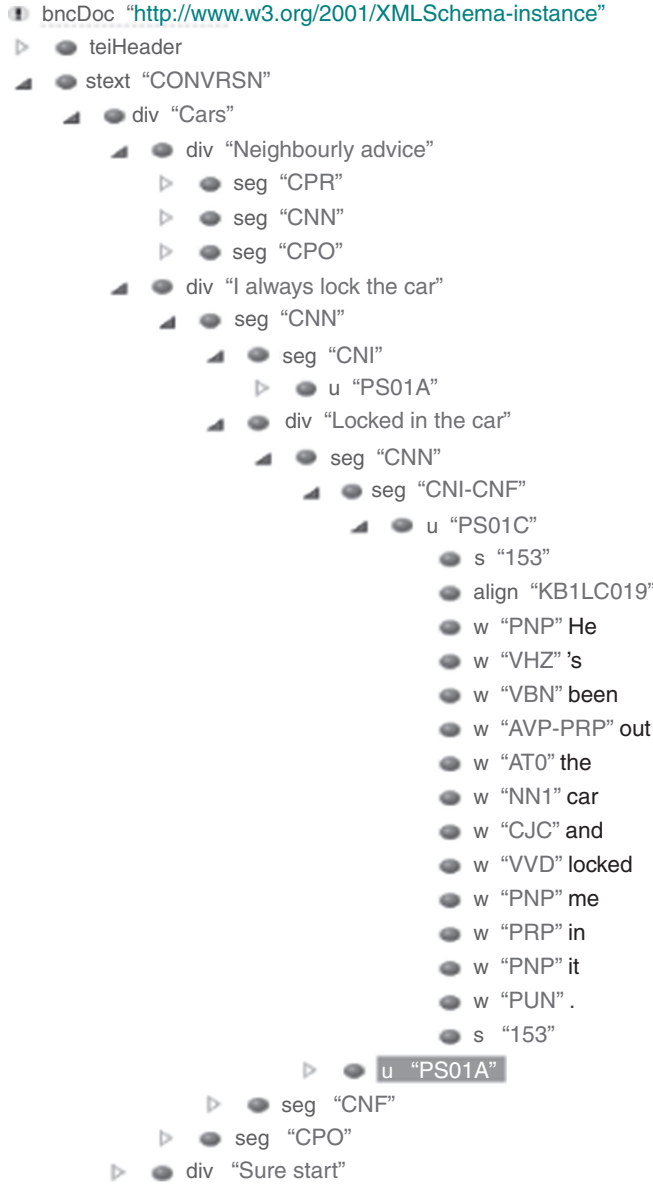


Figure 2.7 Tree structure of text KB1-N1

To illustrate the intricate embedding, Figure 2.7 shows the tree structure of text KB1-N1 entitled *Cars*. The text is enclosed in a element, the highest-order element embracing all other nodes in the document. This first element contains three narratives, each of which is assigned a

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 20:18:37, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9781139026987.003

element of its own. The first such subordinate element has the title value *Neighbourly advice*. As most elements, this first subordinate is superordinate to three major text components, *CPR*, *CNN*, and *CPO*, each enclosed in a <seg> element. The second subordinate , whose title value is *I always lock the car*, does not contain a pre-narrative component (*CPR*). It starts with the main narrative component *CNN*, the first element of which denotes the *CNI* (narrative-initial utterance) component. Immediately following the *CNI* <seg> and the <u> element it encloses is yet another element entitled *Locked in the car*. This interrupts, as it were, the chain of utterances making up the main narrative in the superordinate entitled *I always lock the car*. It is an example of an embedded narrative. Like many such embedded narratives, it consists of only one utterance, that is, the narrative-initial and the narrative-final utterances are collapsed into a single utterance. This single utterance is enclosed in a <seg> element whose *Components* value is *CNI-CNF*. After that 'interlude,' the main narrative is resumed and brought to a close in the narrative-final utterance in the *CNF* <seg>, which is subordinate to the *CNN* <seg>. The *CNF* <seg> is followed by *CNN*'s sibling <seg> *CPO*, the post-narrative component. As soon as the *CPO* <seg> is completed, a new element occurs containing the third story in the story chain.

The levels of embedding of annotations in the NC, which are only sketched out crudely (for more detail on this, see Rühlemann & O'Donnell 2012), are numerous and complex. Finding 'paths' that, where necessary, circumvent or, where desirable, operate on these embeddings is not easy. It is nonetheless feasible and, what is more, revealing. Because of the density of annotation provided in the NC, XPath can retrieve almost any set of data researchers might be interested in.

An extension of XPath is XQuery, which supports the same functions and operators as XPath (for an introduction, see Walmsley 2007).[17] The difference lies in the increased flexibility of XQuery: while Xpath is restricted in terms of the output it can return (namely either frequencies or text but not both at the same time), users of XQuery can program their own scripts to get much more diverse and complex output. This is because XQuery supports the use of what are called for loops. The advantage of using loops is that not only can a specific *type* of node set be specified, which is also possible using XPath, but diverse data can be retrieved for each individual instantiation of that type. Take for example the node set Responsive Recipient (PRR) utterances. Suppose we are interested in the length of PRR utterances in narratives that contain exactly one such PRR utterance. If we use XPath only, we can only determine the total number of words and the total number of such PRR utterances. Dividing the former by the latter the average number of words in the PRR units can be determined. As is well known, though, a danger of averages is that they may harbor important variation. Grasping that

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 20:18:37, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9781139026987.003

variation, however, is essential for many tests of statistical significance. Thus, averages alone are of limited use. Using XQuery, we can go beyond these limitations: here individualized data can be retrieved. That is, instead of asking, 'How many words are there in PRR utterances altogether?', we can ask, 'How many words are there in each and every single PRR utterance?' One possible XQuery script to address this question would be the following:

```
for $seg in //seg[@Components='CNN' and
        count(descendant::u[@Participation_roles='PRR'])=1]
let $n_words_CNN := count($seg//w)
let $n_words_PRR := count($seg//u[@Participation_roles='PRR']//w)
let $text_PRR := $seg//u[@Participation_roles='PRR']
return
concat($n_words_CNN, ' ', $n_words_PRR,' ', normalize-space(string-
        join($text_PRR//w/text(),' ')))
```

The for loop in line 1 instructs the processor to loop over those CNN components in which the number of PRR units (which are 'descendants' of CNN components) is 1. The let commands in lines 2, 3, and 4 define parameters for which data are to be pulled: let $n_words_CNN counts the number of words in each CNN component in the node set; let $n_words_PRR counts the number of words in each PRR unit, and let $text_PRR instructs the software to retrieve the words contained in the PRR units. The return command in line 5, finally, defines the output to be obtained: in this case, the three parameters targeted in the let clauses. The first ten lines of the output look like (2.19) (where no text is printed, the PRR response consists of non-verbal behavior, such as laughter):

(2.19)

$n_words_CNN	$n_words_PRR	$text_PRR
0	1	yes
12	1	yes
191	1	yeah
82	1	blimey
78	0	
45	1	mm
145	1	yeah
24	0	
25	1	yeah
35	2	oh no

The detailed information generated by the XQuery script enables a wide range of quantitative analyses and graphical representations: for example, dispersion measures can be calculated (standard deviation, interquartile range, etc.), the distribution of the data can be compared to similar

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 20:18:37, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9781139026987.003

distributions (using t-tests or Wilcoxon rank sum tests), the distribution could be visualized, for instance, in boxplots.

The XPath and XQuery technologies are not yet as widely used in corpus linguistics as they could be. This is arguably due to the fact that the voice of those who advocate the use of text-only corpora (that is, without even POS annotation) is still widely heard. It is also because the structure of POS annotated corpora is far less complex so that the necessity to resort to sophisticated technologies such as XPath and XQuery has not been felt so far. Further, the number of corpora whose annotations reach beyond POS tagging and whose analysis would, then, most benefit from XPath and XQuery, is still very limited. I suspect that the usefulness of XPath and XQuery will be recognized by more corpus linguists as the benefit of working with discourse-annotated corpora is more fully recognized and more such densely annotated corpora are made available.

In this book, Xpath and XQuery are the key tools used in selecting data for further processing. The query script used for the analysis in Section 3.2 will be given in the main body of the text while most other query scripts can be found in endnotes. Thus, readers interested in this technology can not only verify the precision, or lack thereof, with which the present author used XPath and XQuery to extract relevant data but also familiarize themselves with its syntax and appreciate its benefits. The author hopes that this book will contribute to a better recognition and fuller exploitation of XPath and XQuery in corpus linguistic research.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 20:18:37, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9781139026987.003