Summary

Testimony is undeniably a major source of our beliefs about the world, a fact that underscores the centrality of language for our representational capacities. Many maintain that testimony is also a major source of our knowledge of the world, a contention that implies the centrality of language for our felicitous representational capacities. Even granting this contention, however, there is considerable debate about the nature of testimony. Clearly, it involves knowledge, through sensory perception of understood linguistic acts (or their effects), that another has reported that things are thus-and-so. In the reductionist model, it reduces across the board to inference from such knowledge and from knowledge of the reporter's trustworthiness with respect to his or her report. In the anti-reductionist model, no such reduction obtains.

See also: Cognitive Science and Philosophy of Language; Communication, Understanding, and Interpretation: Philosophical Aspects; Epistemology and Language; Representation in Language and Mind; Thought and Language: Philosophical Aspects.

Bibliography

Barnes J (1980). 'Socrates: the jury, Part II.' *Proceedings of the Aristotelian Society, Supplementary Volume 54*, 193–206.

Burge T (1993). 'Content preservation.' *Philosophical Review 107*, 457–488.

Chakrabarti A (1994a). 'Telling as letting know.' In Chakrabarti & Matilal (eds.). 99–124.

Chakrabarti A (1994b). 'Testimony: a philosophical study [Review essays].' *Philosophy and Phenomenological Research 54*, 965–972.

Chakrabarti A & Matilal B K (eds.) (1994). *Knowing from words: Western and Indian philosophical analysis of understanding and testimony.* Dordrecht, The Netherlands: Kluwer.

Coady C A J (1992). *Testimony: a philosophical study.* Oxford: Clarendon.

Fricker E (1987). 'The epistemology of testimony.' *Proceedings of the Aristotelian Society, Supplementary Volume 61*, 57–83.

Fricker E (1994). 'Against gullibility.' In Chakrabarti & Matilal (eds.). 125–162.

Fricker E (1995). 'Telling and trusting: reductionism and anti-reductionism in the epistemology of testimony.' *Mind 104*, 393–411.

Hume D (1995). 'An enquiry concerning human understanding.' [Section X.] In Nidditch P H (ed.) *Enquiries concerning human understanding and concerning the principles of morals*, 3rd edn. Oxford: Clarendon. (Original work published 1777.)

Locke J (1989). *An essay concerning human understanding.* Nidditch P H (ed.). Oxford: Clarendon. (Original work published 1689.)

McDowell J (1994). 'Knowledge by hearsay.' In Chakrabarti & Matilal (eds.). 195–224.

Reid T (1997). *An enquiry into the human mind on the principles of common sense.* Brookes D (ed.). University Park: Pennsylvania State University Press. (Original work published 1764.)

Strawson P F (1994). 'Knowing from words.' In Chakrabarti & Matilal (eds.). 23–27.

Text and Text Analysis

T Sanders, Utrecht University, Utrecht, The Netherlands

J Sanders, Tilburg University, Tilburg, The Netherlands

Communication Through Text and Discourse

People use language to communicate. Language users communicate through discourse. Sometimes, utterances of one word ('John!' 'Okay.' 'Stop!') or one sentence ('I declare the games opened') suffice to get the message across, but usually language users communicate through a connected sequence of minimally two utterances, i.e., discourse. The importance of the discourse level for the study of language and linguistics can hardly be overestimated: "Discourse is what makes us human" (Graesser *et al.*, 1997). It is not surprising, therefore, that the study of text and discourse has become an increasingly important area over the last decades, both in linguistics and psychology.

The term 'discourse' is used as the more general term to refer to both spoken and written language. The term 'text' is generally used to refer to written language. This article focuses on text. Although spoken and written discourse have crucial characteristics in common, the linguistic traditions of the study of written and spoken discourse are very different. 'Monological texts' are traditionally studied in areas such as stylistics, text linguistics, and psycholinguistics, often based on rather specific linguistic analyses and regularly using a quantitative methodology. By contrast, 'dialogical discourse' has long

been the arena of conversation analysis and sociolinguistics, often focused on qualitative interpretations of individual conversations in context. Over the last 10 years, this situation has begun to change. With the growing availability of spoken corpora and the growing insight that the study of spoken and written discourse should be related because they complement each other (Chafe, 1994), the linguistic study of discourse is becoming less and less restricted to one medium. See, for instance, the overview by Ford *et al.* (2001), who relate linguistic subdisciplines such as grammar and the study of conversation.

A text is more than a random set of utterances: it shows connectedness. A central objective of linguists working on the text level is to characterize this connectedness. Linguists have traditionally approached this problem by looking at overt linguistic elements and structures, thereby characterizing it in terms of cohesion (Halliday and Hasan, 1976; *see* **Cohesion and Coherence: Linguistic Approaches**). By this view, connectedness is localized in the text itself because of explicit linguistic clues, such as pronouns referring to earlier mentioned subjects (cohesion type: *reference*), e.g., *he* refers to *bird-watcher* in (1); or conjunctions, such as *because* in (2) (cohesion type: *conjunction*), which express a causal relation.

(1) The bird-watcher had a great day. He observed a kingfisher and a group of 70 cranes.

(2) The bird-watcher had a great day because he observed a kingfisher and a group of 70 cranes.

(3) The bird-watcher had a great day. A kingfisher and a group of 70 cranes were in the area.

Influential as the cohesion approach has been, the interdisciplinary field of text linguistics and discourse studies is nowadays dominated by the 'coherence' approach: the connectedness of text is considered a characteristic of the mental representation rather than of the text itself (*see* **Cohesion and Coherence: Linguistic Approaches** and **Coherence: Psycholinguistic Approach**). The main reason is probably that a sequence of sentences like (1) or (2) is still interpreted as a perfectly normal piece of text if the cohesive elements of reference and conjunction are absent, as in (3). Hence, the connectedness is not dependent on these overt markers. This does not imply, however, that the linguistic elements signaling text coherence are unimportant.

Although coherence phenomena are of a cognitive nature, their reconstruction is often based on linguistic signals in the text itself. These linguistic expressions are considered 'processing instructions' to language users. For instance, referential expressions, such as pronouns and demonstratives, are used in such a way that interpreters can systematically recover the referential coherence (*see* **Accessibility Theory** and **Discourse Anaphora**). Similarly, connectives (*because*, *however*) and (other) lexical markers of relations, such as cue phrases (*On the one hand, on the other hand*) and signaling phrases (*The problem is ... A solution might be ...*), make the meaning relations between text segments explicit (*see* **Connectives in Text**). In recent years, the relationship between the linguistic surface code, on the one hand, and aspects of the text representation, on the other hand, has become a crucial research issue in the interdisciplinary field of text linguistics and discourse studies (cf. Gernsbacher and Givón, 1995; Sanders and Spooren, 2001; Graesser *et al.*, 2003).

Text

It follows from the discussion above that, in this article, we consider a text to be a monological stretch of written language that shows coherence. The term 'text' derives from the Latin verb *texere* 'to weave' (hence the resemblance between the words 'text' and 'textile'). But what is it that makes a text a text? This question has been at the center of attention of the fields of discourse studies and text linguistics, especially since the 1970s.

Meaning Rather than Form

In the area of syntax – 'sentence analysis' – the principled discussion on the question of whether syntax is an autonomous and purely formal level of representation is still going on, especially with the recent rise of cognitive linguistics (cf. Langacker, 1986; Jackendoff, 1996) (*see also* **Cognitive Linguistics**). At the discourse level such a discussion is nowadays absent. In the pioneering years of text linguistics, scholars like van Dijk (1972) and Petöfi and Rieser (1973) attempted to describe texts as a string of sentences within the framework of generative grammar. Analogous to the way in which sentence grammars described sentences in terms of their constituents, texts were seen as constituted by sentences. In generative grammar, a sentence is the result of rewriting rules of the form: $S \rightarrow NP + VP$.

In 'text grammars,' a text was regarded as consisting of sentences: $T \rightarrow S1 \ldots Sn$. Similarly, the top of hierarchical text representations was formed by a T (for 'text'), analogous to the S for sentence in generative sentence representations. In psychology, so-called 'story grammars' were developed in the late 1970s (Thorndyke, 1977; Rumelhart, 1977). According to such representations, a 'story' consists of a

setting ("Once upon a time, there was a little girl who lived in the woods with her parents. She was called Little Red Riding Hood.") and an 'episode' ("One day, her mother asked her to bring some food to grandmother ...") and, with the help of the same type of rewriting rules, episodes can in turn be represented as a combination of an 'event' ("Why do you have such a big mouth? she asked.") and a 'reaction' ("The wolf jumped out of bed and ate her."):

Story → setting + episode

Episode → event + reaction

Several scholars have argued that the analogy with sentence grammar is not convincing, among them Brown and Yule (1983) and Wilensky (1983):

> ... while our intuition of 'sentencehood' is a clearly linguistic notion, our intuition of 'storiness' most certainly is not [...]. the notion of 'Story' refers to actions, events, goals, or other mental or conceptual objects. In other words, our intuitions about stories are closer to our intuitions about the meanings of sentences than they are about they are about sentences themselves (Wilensky, 1983: 580).

And indeed, ever since Halliday and Hasan (1976), Hobbs (1979), and van Dijk (1977), it is widely accepted that purely formal or syntactic principles play a far smaller role at the discourse level. It is hard, for instance, to make much sense of the idea of a structurally 'well-formed' but semantically anomalous text. There is a consensus that the well-formedness of a discourse is primarily to do with its meaning – more specifically, with the question of whether the meanings of its component segments can be related together to form a coherent message.

What Makes a Text a Text?

What, then, are the crucial characteristics of text? At present, the dominant stance is that 'coherence' explains best the connectedness shown by texts. Coherence is considered a mental phenomenon; it is not an inherent property of a text under consideration. Language users establish coherence by relating the different information units in the text.

Generally speaking, there are two respects in which texts can cohere (Sanders and Spooren, 2001):

- 'Referential coherence': smaller linguistic units (often nominal groups) may relate to the same mental referent throughout the text (*see also* **Discourse Anaphora**); or
- 'Relational coherence': text segments (most often conceived of as clauses) are connected by coherence relations, such as cause-consequence, between them (*see also* **Clause Relations**).

Both coherence phenomena under consideration – referential and relational – have clear linguistic indicators that can be taken as processing instructions. For referential coherence, these are anaphoric devices such as pronouns, and for relational coherence these are connectives and (other) lexical markers of relations.

Ever since the seminal work of linguists such as Chafe (1976) and Prince (1981), both functional and cognitive linguists have argued that the grammar of referential coherence can be shown to play an important role in the mental operations of connecting incoming information to the existing mental representations. For instance, referent NPs are identified as either those that will be important and topical, or as those that will be unimportant and nontopical. Hence, topical referents are persistent in the mental representation of subsequent discourse, whereas the nontopical ones are nonpersistent. In several publications, Ariel (1988, 2001) argued that regularities in grammatical coding should indeed be understood to guide processing. She studied the distribution of anaphoric devices and suggested that zero anaphora and unstressed pronouns cooccur with high 'accessibility' of referents, whereas stressed pronouns and full lexical nouns signal low accessibility. This cooccurrence can easily be understood in terms of cognitive processes of activation: high-accessibility markers signal the **default** choice of continued activation of the current topical referent. Low-accessibility anaphoric devices, such as full NPs or indefinite articles, signal the terminated activation of the current topical referent and the activation of another topic (*see* **Accessibility Theory**).

'Centering theory' (see Walker *et al.*, 1998 for an overview) makes explicit and precise predictions about the referent that is 'in focus' at a certain moment in a discourse. It even predicts that the degree of text coherence is determined by the extent to which it conforms to 'centering constraints.' Given a clause in which referential antecedents are presented, centering theory predicts the likelihood that an antecedent will be a central referent – which is 'in focus' – in the next clause. The salience of a discourse entity is determined by a combination of syntactic, semantic, and pragmatic factors, such as grammatical role (subject or not), expression type (zero, pronoun, or NP), and discourse topic-hood. Several processing studies have demonstrated the 'psychological reality' of linguistic indicators of referential coherence (see Garrod and Sanford, 1994, and Sanford and Garrod, 1994, for an overview; *see also* **Discourse Processing**).

We now turn to (signals of) 'relational coherence.' 'Coherence relations' are often taken to account for the connectedness in readers' cognitive text

representation (cf. Hobbs, 1979; Sanders *et al.*, 1992). They are also termed 'rhetorical relations' (Mann and Thompson, 1988; *see* **Rhetorical Structure Theory**) or 'clause relations' (*see* **Clause Relations**). 'Coherence relations' are meaning relations connecting, at a minimum, two text segments. A defining characteristic for these relations is that the interpretation of the related segments needs to provide more information than is provided by the sum of the segments taken in isolation (Sanders *et al.*, 1992). Examples are relations like 'cause-consequence,' 'list,' and 'problem-solution.' These relations are conceptual and they can, but need not, be made explicit by linguistic markers, so-called connectives (*because*, *so*, *however*, *although*) and lexical cue phrases (*for that reason*, *as a result*, *on the other hand*) (*see* **Connectives in Text**).

In the last decade, much research in relation semantics and pragmatics has focused on the question of how to taxonomize or classify the set of coherence relations (Hovy, 1990; Knott and Dale, 1994; Pander Maat, 1998; Redeker, 1990; Sanders, 1997). The main reason for this interest is the cognitive interpretation of coherence relations: if they are to be considered as cognitive mechanisms underlying discourse interpretation, it is attractive to find out which more general principles are involved in relation interpretation. While work on the hierarchical classification of discourse relations goes back at least as far as Grimes (1975) and Halliday and Hasan (1976), the idea that a small number of reasonably orthogonal primitives is responsible for the differences amongst coherence relations is more recent. Sanders *et al.* (1992) defined the 'relations among the relations,' relying on the intuition that some coherence relations are more alike than others, and that the set of relations can be organized in terms of more primitive notions, such as polarity and causality. Several types of evidence in favor of such an organization were produced, varying from experiments in which text analysts judged relations (Sanders *et al.*, 1992, 1993; Sanders, 1997), to research on the acquisition order of connectives (Evers-Vermeul, 2005) and processing studies indicating how different coherence relations result in different representations (Sanders and Noordman, 2000; *see also* **Connectives in Text**). In such an account of coherence, connectives and other lexical signals are seen as 'processing instructors.' And indeed, experimental studies on the role of connectives and signaling phrases show that these linguistic signals affect the construction of the text representation (cf. Millis and Just, 1994; Noordman and Vonk, 1997).

In sum, it can be concluded that there is compelling evidence, from both linguistic and psycholinguistic studies, in favor of the view that referential and relational coherence are crucial principles, which make a set of sentences a text.

Text Analysis

Now that we have an idea of what a text is, we can define 'text analysis' as the systematic dissection of a textual unity in its constituent parts and the study of those parts in relation to each other. By consequence, text analysis focuses on the linguistic elements present in the text. Texts may be analyzed with different aims and from several perspectives.

A first text-analytic research goal is of a theoretical nature. It concerns the further development of linguistic theory at the discourse level: how are texts structured? There are now several well-established theories that propose mechanisms by which the meaning of individual sentences can be constructed, but the situation with entire texts is different. Text analysis is of crucial importance to the further development of text linguistics.

A second aim is to provide insight into the cognitive processes of reading and writing, or in the text representation that language users have of a text. In reading research, the role of text structure is an important research topic in which text analyses are used to model both the text structure and the representation that readers make of it (see previous paragraph). In writing research, the role of text analysis has received less attention for a long time, even though Bereiter and Scardamalia (1987) argued for the interaction between psychological models and text linguistic research. They pointed to a deficiency in studies of writing and argued that text analysis had a large role to play in discovering the implicit rules of composition.

A third aim is of a computational linguistic nature: the development of computational models of automatic summarization, text generation, and interpretation. Here, the analysis of natural texts should provide the rule system to arrive at such computational models. Although some theories and models discussed in the sections to follow were explicitly developed in the context of such a computational enterprise (such as **Rhetorical Structure Theory**), computational text analyses are not discussed here (*see* **Natural Language Processing: Overview**).

A fourth aim is the evaluation of text quality in the context of written composition and document design. A text analysis can provide the basis for a comparison of similar texts, enabling researchers to compare the writing ability of the authors (Cooper, 1983). In document design, text analysis can predict areas where readers may have difficulties and where

revision is imperative. It is also used to investigate the relationship between text structure and the successful layout of various documents, even multimodal ones (Delin and Bateman, 2002).

From what perspectives do text analysts try to catch the 'meaning' in text? A first division is that between content-oriented and structure-oriented approaches. 'Content-oriented' approaches to text analysis uncover what an individual text is 'about,' either by starting from the smallest building blocks (propositions) or by characterizing texts on a more global level: the topics and subtopics that are covered. 'Structure-oriented' approaches uncover the meaning relations between the textual building blocks, such as causal, contrastive, and additive relations, but also referential relations. Some approaches provide analytic models that allow for a hierarchical representation representing the whole text in such terms.

Content-Oriented Approaches

Micro- and Macrostructure In the context of a psychological model of text processing, Van Dijk and Kintsch (1983) distinguished between three aspects of text representation: 'microstructure,' 'macrostructure,' and 'superstructure' (*see* **Macrostructure**). Superstructures – representing the global structure that is characteristic of a text type – will be discussed in the section on structure-oriented approaches. Micro- and macrostructure concern the **content** of a text. The basic building blocks of these representations are 'propositions,' i.e., a unit of meaning that consists of a predicate and connected arguments. For instance, the proposition underlying sentence (4) would be (4'), where *see* is the predicate and *he* and *kingfisher* are the arguments.

(4) he sees a kingfisher
(4') (see (he, kingfisher))

The microstructure is a network of propositions like these that represents the textual information in a bottom-up fashion, sentence by sentence. Building on earlier work, van Dijk and Kintsch (1983) presented an influential model of text comprehension, which predicted the information recalled best by readers. For the purpose of text analysis, it is important to focus on another component of the Van Dijk and Kintsch model: macrostructure. On the basis of the microstructure or 'text base,' a macrostructure can be built – an abstract representation of the global meaning structure that would reflect the gist of the text (*see* **Macrostructure**). This is achieved by applying macro-rules to the detailed meaning representation of the microstructure. 'Deletion,' 'generalization,' and 'construction' are such macrorules, which produce macro-propositions: the main ideas in the text (see especially van Dijk, 1980). This idea of producing the macrostructure on the basis of the details of the microstructure is certainly appealing. The results of some experimental processing studies seem to show that macrostructures can predict recall and summarization results: Propositions present in the macrostructure are remembered better than propositions that are 'only' present in the microstructure (Graesser, 1981). Arguably, the theoretical and empirical status of this part of the van Dijk and Kintsch theory is less clear than the microstructure part. This was probably a result of the fact that macrorules were underspecified. In addition, it is not always easy to identify linguistic signals of macropropositions at the surface level of the text, even though titles, headings, abstracts, and topical sentences are mentioned as signalling macropropositional ideas. In recent years, Kintsch (1998) and others have argued that macrostructures can be derived from texts by using 'latent semantic analysis' (*see* **Latent Semantic Analysis**). Here, the meaning of sentences is represented by a vector in a high-dimensional semantic space. Vectors that relate most to the rest of the text can be identified as macropropositions.

Theme and Thematics 'Thematics' is the interdisciplinary study of 'about-ness' in text (*see* **Thematics**). The notion of 'theme' refers to the main idea or topic of the text. For instance, a text can be about a kingfisher or about an ornithologist having a great day. The study of theme has been popular in literary studies. Thanks to the involvement of text linguistics and stylistics, the study of linguistic cues that create thematic meaning has become increasingly important (Louwerse and Van Peer, 2002). For instance, formulations and stylistic figures also emphasize the thematic meaning of a text.

However, regular aspects of formulation, such as the linear order of the information in clauses and sentences, can also contribute to the identification of the theme. A typical linguistic aspect studied in more detail is the way in which the first position in a clause has a special textual status. The terminology is somewhat confusing here, because linguists refer to the information provided in this position with the term 'theme,' whereas any information following this local theme is called 'rheme' (*see* **Theme in Text**). The opening positions of clauses often contain information that guides the reader in constructing a picture of the text as a whole. In linguistics, and especially in systemic functional grammar, sequences of theme–rheme are studied, resulting in patterns of thematic development.

Structure-Oriented Approaches

Most linguistic methods of text analysis focus on the general properties of text structure, abstracting away from the specific content of individual texts. Accounts of text structure usually pay attention to

1. the meaning of the left-right relations between text segments, where the analysis is based on relational and referential coherence; and
2. the hierarchical structure of the text, which accounts for the intuition that the information that is ordered higher in a tree-like representation is more important than the lower information.

Superstructure van Dijk and Kintsch's (1983) model included micro- and macrostructures, which resulted in a representation of the text content, as was discussed above. The third element in their model is the 'superstructure,' which "provides a kind of overall functional syntax for the semantic macrostructures" (van Dijk and Kintsch, 1983: 242). It is the conventional, hierarchical form in which the content of the macrostructure is presented. An example of such a superstructure is that of the type 'news discourse,' in which superstructural categories are distinguished, for example, *headlines, lead, context, event*. Superstructural categories are typically of a global nature in that they organize larger chunks of text rather than consecutive sentences. In addition, a superstructure analysis proceeds top-down: it starts from the highest text level. Superstructures for several other conventional text types were developed, among them the 'Experimental article.' There seems to be a clear parallel here with text type and genre: it would seem logical to expect that stereotypical text types can be characterized in terms of a superstructure (*see* **Genre and Genre Analysis**). Therefore, a text analysis in terms of superstructures is text type-specific by definition.

Clause Relations, Coherence Relations, and Discourse Patterns By contrast, a text analysis based on clause or coherence relations would be generally applicable, independent of text types. It proceeds bottom-up, starting from consecutive clauses. One common relation is called 'problem-solution' or 'solutionhood' (*see* **Problem-Solution Patterns**). See examples (5) and (6).

(5) I'm hungry. Let's go to the Fuji Gardens.

(6) What if you're having to clean floppy drive heads too often? Ask for Syncom diskettes, with burnished Ectype coating and dust absorbing jacket liners.

Mann and Thompson (1986, 1988) treated solutionhood as simply one of the relations, where others have argued that solutionhood was more complex than that (Grimes, 1975; Hoey, 1983; Sanders *et al.*, 1993): "Both of the plots of fairy tales and the writings of scientists are built on a response pattern. The first part gives a problem and the second the solution" (Grimes, 1975: 211). On the basis of clause relations, more complex structures can be built: a 'discourse pattern' (Hoey, 1983) or a 'response pattern' (Grimes, 1975). Hoey (1983) argued that a recurrent combination of clause relations can organize a substantial text fragment, or even a whole text. See the illustrating example from Hoey (1983: 35):

(7) (i) I was on sentry duty.
(ii) I saw the enemy approaching.
(iii) I opened fire.
(iv) I beat off the attack.

Hoey provided several paraphrase tests to recognize the clause relations on which the pattern is based: 'instrument-achievement' with '(iii) thereby (iv),' 'by (iii) ... ing,' and '(iii) by this means (iv)' (Hoey, 1983: 39–41); and 'cause-consequence' 'because (ii), (iii)' and '(ii) therefore (iii)' (Hoey, 1983: 41–42). Paraphrase tests like these are often a great help for inexperienced text analysts, who find it hard to determine the exact relationship expressed between text segments.

This heuristic to identify discourse patterns is an outstanding example of a text-analytic method in the field of clause and coherence relations. The research in this field discussed earlier in this section has probably been more important for the identification of coherence relations and for the theoretical issues discussed earlier (the nature of coherence, taxonomies of relations, the linguistic expression and processing of relations). However, a very important account has not been discussed so far: rhetorical structure theory.

Rhetorical Structure Theory In the 1980s and 1990s, Mann and Thompson (see especially Mann and Thompson, 1988) presented 'rhetorical structure theory' (RST), a functional theory of text organization developed in the context of linguistics and cognitive science (*see* **Rhetorical Structure Theory**). At the heart of RST are the so-called 'rhetorical relations,' similar to clause or coherence relations, and including relations like 'cause,' 'elaboration,' and 'evidence.' The relations are defined in terms of conditions on the nucleus (the most important segment in a relation), on the satellite (which depends on the nucleus), and their combination, and in terms of the effect on the reader. Relations are identified between adjacent text segments (e.g., clauses) up to the top level of the

text. The top level of an RST tree organizes the text as a whole: a relationship that dominates the total text structure.

Rhetorical structure theory has proven to be a very useful analytic tool. One of its benefits is that it allows for a complete analysis of any text type: expository, argumentative, or narrative. The system has been applied to many real-life texts, among them newspaper articles, advertisements, and fundraising letters (Mann and Thompson, 1992). As a rule, an RST analysis starts with an inspection of the entire text. The analysis does not proceed in a fixed way; it proceeds bottom-up (from relations between clauses to the level of the text) or top-down (the other way around) or follows both routes (Mann *et al.*, 1992). The analysis results in a hierarchical structure that encompasses the entire text and has a label attached to each of its branches.

Although RST defines rhetorical relations in a fairly exact way, the assignment of a label is ultimately based on observed 'plausibility.' Four general constraints are the guidelines: 'completedness,' 'connectedness,' 'uniqueness,' and 'adjacency' (Mann and Thompson, 1988: 248–249). How the analysis actually proceeds is left to the intuitions of the analyst and is, in the end, a matter of text interpretation. Still, it has been shown that RST can be applied with a reasonable amount of consensus by expert text analysts (Den Ouden, 2004) and to a certain extent, RST analyses can even be produced automatically (Marcu, 2000).

Procedural Text Analysis Rhetorical structure theory requires a fair amount of text interpretation based on the analysts' overview of the text as a whole. This overview situation may not reflect the way in which writers produce texts. Spontaneously produced texts, especially, are the result of a more incremental process. Sanders and van Wijk (1996) developed 'procedures for incremental structure analysis' (PISA), which incorporates both ideas about written text production and insights from the text analytical literature, especially with respect to hierarchical aspects of text structure. The two texts in example (8) are taken from the PISA corpus. They were written by 12-year-old boys in response to a request to explain who Saint Nicholas is to someone who knows nothing about the subject. The texts are translated from Dutch in a rather literal way, preserving the original punctuation. For the analysis, texts were divided into segments, roughly corresponding to clauses. The hierarchical structures of the texts in (8a and 8b) are shown in **Figures 1** and **2**.

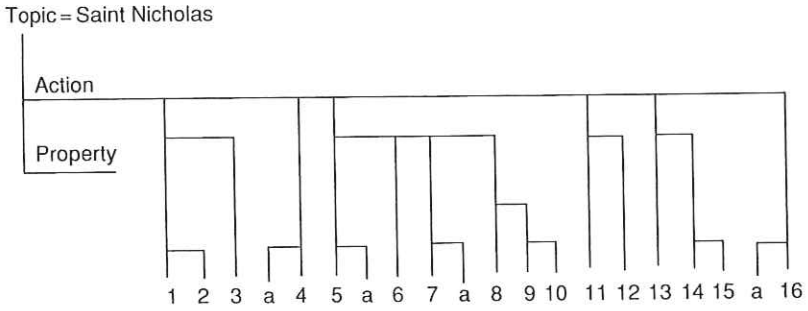

Figure 1 Hierarchical structure of an explanatory text, dominated by a sequence of actions. (Reproduced from Sanders T & van Wijk C (1996). 'PISA – a procedure for analyzing the structure of explanatory texts.' *Text, 16(1)*, 91–132 with permission by Elsevier.)

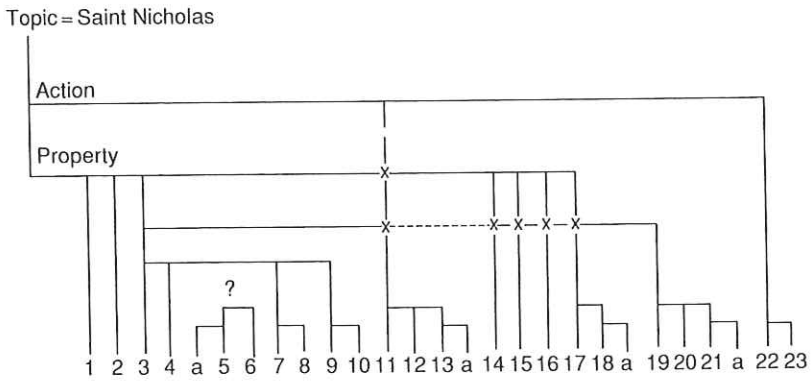

Figure 2 Hierarchical structure of an explanatory text, dominated by referential coherence. (Reproduced from Sanders T & van Wijk C (1996). 'PISA – a procedure for analyzing the structure of explanatory texts.' *Text, 16(1)*, 91–132 with permission by Elsevier.)

(8a) (1) Every year Saint Nicholas comes (2) that is on the 4th of December (3) That day he comes by steamboat (4a) When he arrives in the Netherlands (4) then everyone waves to him (5) At night (5a) when it is pitch-dark (5) Saint Nicholas rides over the roofs with Black Peter (6) and throws lots of presents through the chimney (7) while the children sing a song (7a) such as: Sinterklaas kapoentje ... (8) Saint Nicholas gives lots of presents (9) but he always gets something in return (10) That is either a carrot for the horse or a bit of water, also for the horse (11) On the fifth of December Saint Nicholas really has his birthday (12) on that day he brings the presents (13) and then he leaves again. (14) He also looks into the red book with the cross on it. (15) There it says whether you have been naughty or not (16a) When Saint Nicholas leaves again (16) the children sing Bye bye Saint Nicholas.

(8b) (1) Saint Nicholas is an old man (2) He has a white-grey beard (3) He has a steamboat with a lot of little black men (4) They are called Black Peter (5a) When it is the fifth of December (5) the time has come. (6) The day of Saint Nicholas is a Big Festival. (7) The Peters have a birch. (8) A birch is a bundle of swishing branches. (9) They also have a sack (10) that is a kind of wool and a shape. (11) The children in all villages and towns got ginger-nuts. (12) And ginger nuts are four-sided blocks, with sugar. (13) That is candy (13a) that children like. (14) Well, Saint Nicholas is an old man, with white hair. (15) He wears a red robe. (16) And wears a sort of shirt. (17) He also has a whitish grey. (18) A grey is a horse. (18a) (noble animal) (19) Just now I was talking about a steamboat. (20) A steamboat is a big ship, (21) a steamboat often has got a big funnel (21a) from which the steam comes. (22) The children sing songs on the fifth of December, (23) that is because there is a big feast.

The texts in (8) illustrate two different ways of arranging information in an explanatory text. The first one follows the **temporal order** of events: first he arrives, then, at night, he rides over the roofs. On December 5, he has his birthday and the children get presents. And then he leaves and the children sing to him. By contrast, the writer of the second text solves the problem of explanation by focusing on the topic of the text. He mentions all kinds of **properties** of Saint Nicholas in a rather associative way: He is old, has a white beard, a steamboat, etc.

Clearly, these two texts differ in their global structure. The first one is dominated by an 'action-line,' a temporal sequence of actions, and the second one is dominated by a 'property-line,' a list of characteristics of the topic. Text structures like these may reflect the way in which the writer has organized the information during the production of the text. For example, the first text can be produced by running through episodic memory; the actions or events are mentioned in temporal succession and the text ends with the closing of an episode. The second text, on the other hand, is probably produced by searching semantic memory in an associative way. It resembles brainstorming; it lists information related to the topic, which is potentially relevant to explain it.

In a series of publications, it has been argued that the product of this text-analytical method is cognitively interpretable because it correlates with pause time distribution during writing (Schilperoord, 1996; Sanders and Schilperoord, 2005) and explains writing development (van der Pool, 1995), sentence-combining results, and problems during writing (van Wijk and Sanders, 1999). The generalizability of an incremental and procedural approach like this is limited; it has specifically been shown to be successful for spontaneously written explanatory texts and judicial letters that were produced without much planning.

Conclusion and Further Research

There are several interesting developments for the research agenda in the years to come. Before we go into detail, a general methodological remark seems in order. Text analyses of corpora of natural language texts have a crucial role to play in text linguistics and discourse studies, because the development of theoretical models of discourse phenomena needs to proceed in interaction with the study of the (sometimes very complex) reality of natural language in use (cf. Emmott, 1997).

Let us now focus on some specific issues that follow from our analysis of the state-of-the-art in the preceding sections. A first important issue is the linguistics/text linguistics interface. There are clear rapprochements between grammarians, (formal) semanticists, and pragmaticists on the one hand and text linguists on the other hand (Sanders and Spooren, in press). Questions to be asked are: what is the relationship between information structuring at the sentence level and at the discourse level? How do factors such as tense, aspect, and perspective influence discourse connections (Lascarides and Asher, 1993; Oversteegen, 1997)? For instance, discourse segments denoting events that have taken place in the past (*The bird-watcher saw a small blue bird near the river. It was a kingfisher*) will typically be connected by coherence relation of the content type, whereas segments in the present/future, which contain many evaluations or

other subjective elements (*Here is that small blue bird again. It must be a kingfisher*), are prototypically connected by epistemic or argumentative relations (*see* **Connectives in Text** and **Evaluation in Text**). This correlation, in turn, should be studied in connection with issues like perspective and subjectivity (Sanders and Redeker, 1996; Pander Maat and Sanders, 2001).

A second obvious issue is the relationship between the principles of relational and referential coherence. Clearly, the two types of principles both provide language users with signals during text interpretation. Theses signal are taken as instructions for how to construct coherence. Therefore, the principles will operate in parallel, and they will influence each other. The question is: How do they interact? Consider a simple example.

(9) John congratulated Pete on his excellent play.
(a) He had scored a goal.
(b) He scored a goal.

At least two factors are relevant for the solution of the anaphor *he* in (a/b): the aspect of the sentence, and the possible coherence relations that can be inferred between sentences. Part (9a) has perfect tense, and at the discourse level, the interpretation of one coherence relation is obvious – namely the backward causal relation 'consequence-cause.' The tense of (9b) is imperfect, and at the discourse level several coherence relations can exist, including 'temporal sequence' (of events) and 'enumeration/list' (of events in the game). Hence, the resolution of the anaphor-antecedent relation seems to be related to these two factors. In (9a) *he* must refer to Pete; in (9b), both antecedents are possible: John or Pete. How do aspect and the coherence relation interact in the process of anaphor resolution? And: Is the anaphor resolved as a consequence of the interpretation of the coherence relation? Questions like these were already addressed in the seminal work of Hobbs (1979) and recently taken up again in a challenging way by Kehler (2002). Text analysis of natural texts has a large role to play here: How often do ambiguities like these actually show up in text? What are the heuristics apparently used by language users?

A third issue is the further characterization of genres and text types in terms of their text structure. Genre and text type are both frequently used concepts (*see* **Genre and Genre Analysis**) that are often not defined in articulate text-internal characteristics (see Virtanen, 1992). Now that text-analytic models like RST are available and the theory of different types of coherence relations has matured, it is high time that structural analysis of real-life corpus texts show whether text types differ systematically in their text structure. In a first corpus study (Sanders, 1997), such a correlation was indeed found. 'Informative texts' (in which the writer's goal is to inform the reader about something) were compared to 'expressive texts' (in which the writer's goal is to express his or her feelings and attitudes) and 'persuasive texts' (in which the writer's goal is to persuade the reader of something). It was shown that persuasive texts were indeed dominated by more subjective relations, used by the writer to put forward the argument, whereas encyclopedic texts were shown to be informative because their structure was dominated by more objective relations, in which the writer simply described the content area. The realization of this type of text-analytic work on a larger scale would make notions of text type more concrete, but it also provides an example of the way in which text structural characteristics could be operationalized for the further study of language use, on a par with many stylistic text characteristics.

A fourth and final issue concerns the role of text analysis in text evaluation and document design. Many teachers believe that the best and the worst essays written in class differ in organization. The best one is structured clearly, whereas the worst one is hard to follow. Traditionally, there are few results from research to underpin observations like these. However, this situation has recently improved. For instance, children's explanatory texts showing continuity might be judged better than texts that show discontinuities (Sanders and van Wijk, 1996; van Wijk and Sanders, 1999).There are at least two cognitive reasons to link structure and judgments about text quality: texts are easier to understand without such discontinuities, and discontinuities often point to a lack of text planning during writing (Sanders and Schilperoord, 2005).

The use of text analysis in document design is particularly promising because it not only appears valuable in the study of 'classical' text structure, but it is also a useful basis to investigate the matching of text structure, content, and layout, including visual images (Delin and Bateman, 2002). This type of work shows the way to the text analysis of the 21st century: that of multimodal documents.

See also: Accessibility Theory; Clause Relations; Cognitive Linguistics; Coherence: Psycholinguistic Approach; Cohesion and Coherence: Linguistic Approaches; Connectives in Text; Discourse Anaphora; Discourse Processing; Evaluation in Text; Generative Grammar; Genre and Genre Analysis; Latent Semantic Analysis; Macrostructure; Natural Language Processing: Overview; Problem-Solution Patterns; Rhetorical Structure Theory; Thematics; Theme in Text.

Bibliography

Ariel M (1988). 'Referring and accessibility.' *Journal of Linguistics 24*, 65–87.

Ariel M (2001). 'Accessibility theory: an overview.' In Sanders T, Schilperoord J & Spooren W (eds.) *Text representation: linguistic and psycholinguistic aspects.* Amsterdam: Benjamins. 29–87.

Bereiter C & Scardamalia M (1987). *The psychology of written composition.* Hillsdale NJ: Erlbaum.

Brown G & Yule G (1983). *Discourse analysis.* Cambridge: Cambridge University Press.

Chafe W L (1976). 'Givenness, contrastiveness, definiteness, subjects, topics and points of view.' In Li C (ed.) *Subject and topic.* New York: Academic Press. 25–55.

Chafe W L (1994). *Discourse, consciousness, and time. The flow and displacement of conscious experience in speaking and writing.* Chicago: Chicago University Press.

Cooper C (1983). 'Procedures for describing written texts.' In Mosenthal P, Tamor L & Walmsley S (eds.) *Research on writing: principles and methods.* New York: Longman. 287–313.

Delin J & Bateman J (2002). 'Describing and critiquing multimodal documents.' *Document Design 3*, 141–155.

Den Ouden J N (2004). Prosodic realizations of text structure. Ph.D. diss., Tilburg University.

Emmott C (1997). *Narrative comprehension: a discourse perspective.* Oxford: Clarendon Press.

Evers-Vermeul J (2005). Connections between form and function of Dutch connectives. Change and acquisition as windows on form-function relations. Ph.D. diss., Utrecht Institute of Linguistics OTS, Universiteit Utrecht.

Ford C E, Fox B A & Thompson S A (eds.) (2001). *The language of turn and sequence.* Oxford: Oxford University Press.

Garrod S C & Sanford A J (1994). 'Resolving sentences in a discourse context: How discourse representation affects language understanding.' In Gernsbacher M A (ed.) *Handbook of psycholinguistics.* San Diego: Academic Press. 675–698.

Gernsbacher M A & Givón T (eds.) (1995). *Coherence in spontaneous text.* Amsterdam: Benjamins.

Graesser A C (1981). *Prose comprehension beyond the word.* New York: Springer Verlag.

Graesser A C, Gernsbacher M A & Goldman S (eds.) (2003). *Handbook of discourse processes.* Mahwah, NJ: Erlbaum.

Graesser A C, Millis K K & Zwaan R A (1997). 'Discourse comprehension.' In Spence J, Darley J & Foss D (eds.) *Annual Review of Psychology 48.* Palo Alto, CA: Annual Reviews Inc. 163–189.

Grimes J (1975). *The thread of discourse.* The Hague: Mouton.

Halliday M A K & Hasan R (1976). *Cohesion in English.* London: Longman.

Hobbs J R (1979). 'Coherence and coreference.' *Cognitive Science 3*, 67–90.

Hoey M (1983). *On the surface of discourse.* London: George Allen & Unwin.

Hovy E H (1990). 'Parsimonious and profligate approaches to the question of discourse structure relations.' *Proceedings of the 5th International Workshop on Natural Language Generation.*

Jackendoff R (1996). 'Conceptual semantics and cognitive linguistics.' *Cognitive Linguistics 7(1)*, 93–129.

Kehler A (2002). *Coherence, reference and the theory of grammar.* Chicago: University of Chicago Press.

Kintsch W (1998). *Comprehension. A paradigm for cognition.* Cambridge: Cambridge University Press.

Knott A & Dale R (1994). 'Using linguistic phenomena to motivate a set of coherence relations.' *Discourse Processes 18*, 35–62.

Langacker R (1986). 'An introduction to cognitive grammar.' *Cognitive Science 10*, 1–40.

Lascarides A & Asher N (1993). 'Temporal interpretation, discourse relations and common sense entailment.' *Linguistics and Philosophy 16(5)*, 437–493.

Louwerse M & van Peer W (eds.) (2002). *Thematics: interdisciplinary studies.* Amsterdam: Benjamins.

Mann W C & Thompson S A (1986). 'Relational propositions in discourse.' *Discourse Processes 9*, 57–90.

Mann W C & Thompson S A (1988). 'Rhetorical structure theory: toward a functional theory of text organization. *Text 8*, 243–281.

Mann W C & Thompson S A (eds.) (1992). *Discourse description. Diverse analyses of a fund-raising text.* Amsterdam: Benjamins.

Mann W C, Matthiessen C M I M & Thompson S A (1992). 'Rhetorical structure theory and text analysis.' In Mann W C & Thompson S A (eds.). 39–78.

Marcu D (2000). *The theory and practice of discourse parsing and summarization.* Boston, MA: MIT Press.

Millis K K & Just M A (1994). 'The influence of connectives on sentence comprehension.' *Journal of Memory and Language 33*, 128–147.

Noordman L G M & Vonk W (1997). 'The different functions of a conjunction in constructing a representation of the discourse.' In Fayol M & Costermans J (eds.) *Processing interclausal relationships in production and comprehension of text.* Hillsdale, NJ: Erlbaum. 75–93.

Oversteegen E (1997). 'On the pragmatic nature of causal and contrastive connectives.' *Discourse Processes 24*, 51–85.

Pander Maat H (1998). 'The classification of negative coherence relations and connectives.' *Journal of Pragmatics 30*, 177–204.

Pander Maat H & Sanders T (2001). 'Subjectivity in causal connectives: An empirical study of language in use.' *Cognitive Linguistics 12(3)*, 247–273.

Petöfi J & Rieser H (1973). *Studies in text grammar.* Dordrecht: Reidel.

Prince E (1981). 'Toward a taxonomy of given-new information.' In Cole P (ed.) *Radical pragmatics.* New York: Academic Press. 223–255.

Redeker G (1990). 'Ideational and pragmatic markers of discourse structure.' *Journal of Pragmatics 14*, 305–319.

Rumelhart D E (1977). 'Understanding and summarizing brief stories.' In LaBerge D & Samuels S J (eds.) *Basic*

processes in reading: Perception and comprehension. Hillsdale, NJ: Erlbaum. 265–303.
Sanders J & Redeker G (1996). 'Perspective and the representation of speech and thought in narrative discourse.' In Fauconnier G & Sweetser E (eds.) *Spaces, worlds and grammars.* Chicago: University of Chicago Press. 290–317.
Sanders T (1997). 'Semantic and pragmatic sources of coherence: on the categorization of coherence relations in context.' *Discourse Processes 24,* 119–147.
Sanders T & Spooren W (2001). 'Text representation as an interface between language and its users.' In Sanders T, Schilperoord J & Spooren W (eds.) *Text representation: linguistic and psycholinguistic aspects.* Amsterdam: Benjamins. 1–25.
Sanders T & Spooren W (in press). 'Discourse and text structure.' In Geeraerts D & Cuykens H (eds.) *Handbook of cognitive linguistics.* Oxford: Oxford University Press.
Sanders T & van Wijk C (1996). 'PISA – a procedure for analyzing the structure of explanatory texts.' *Text 16(1),* 91–132.
Sanders T, Spooren W & Noordman L (1992). 'Toward a taxonomy of coherence relations.' *Discourse Processes 15,* 1–35.
Sanders T, Spooren W & Noordman L (1993). 'Coherence relations in a cognitive theory of discourse representation.' *Cognitive Linguistics 4,* 93–133.
Sanders T J M & Noordman L G M (2000). 'The role of coherence relations and their linguistic markers in text processing.' *Discourse Processes 29,* 37–60.
Sanders T J M & Schilperoord J (2005). 'Text structure as a window on the cognition of writing; How text analysis provides insights in writing products and writing processes.' MacArthur C, Graham S & Fitzgerald J (eds.) *Handbook of writing research.* New York: Guilford Press.
Sanford A J & Garrod S C (1994). 'Selective processes in text understanding.' In Gernsbacher M A (ed.) *Handbook of psycholinguistics.* San Diego: Academic Press. 699–720.
Schilperoord J (1996). *It's about time. Temporal aspects of cognitive processes in text production.* Amsterdam: Rodopi.
Thorndyke P W (1977). 'Cognitive structure in comprehension and memory of narrative discourse.' *Cognitive Psychology 9,* 77–110.
van Dijk T (1972). *Some aspects of text grammars. A study in theoretical linguistics and poetics.* The Hague: Mouton.
van Dijk T (1977). *Text and context. Explorations in the semantics and pragmatics of discourse.* New York: Longman.
van Dijk T (1980). *Macrostructures.* Hillsdale, NJ: Erlbaum.
van Dijk T & Kintsch W (1983). *Strategies of discourse comprehension.* New York: Academic Press.
van der Pool E (1995). Writing as a conceptual process. A text-analytical study of developmental aspects. Ph.D. diss., Tilburg University.
van Wijk C & Sanders T (1999). 'Identifying writing strategies through text analysis.' *Written Communication 1,* 52–76.
Virtanen T (1992). 'Issues in text typology: narrative – a 'basic' type of text?' *Text 12,* 293–310.
Walker M, Joshi A K & Prince E F (1998). *Centering theory in discourse.* Oxford: Clarendon Press.
Wilensky R (1983). 'Story grammars and story points.' *The Behavioural and Brain Sciences 6,* 579–623.

Text Formatting

D M Berry, University of Waterloo, Waterloo, Ontario, Canada

Introduction

The *text formatting* problem, to be solved by software residing on a computer, is to take a sequence of words stored in an input file and arrange them in the same order on as many pages as are needed, subject to the user's commands and his choices on a variety of constraints and options. The pages can be printed on paper at any time, and usually the user is able to see on the computer's screen approximately how the pages will appear when printed. The approximation is usually accurate to the resolution of the computer's screen.

The results the user gets depend on the commands he gives and on the value of the choice he makes for each constraint or option. These commands are given and choices are made by directly or indirectly (see the later discussion on direct manipulation) inserting additional characters, often called *markup*, at appropriate points in the input file. Later in the article, the commands are summarized and a full list of constraints and options is given. In the mean time, the constraints and options are collectively referred to as 'choices.' Telling a formatter to apply a particular option is called *turning the option on* and telling the formatter not to apply a particular option is called *turning the option off.*

For concreteness, the text that you are now reading was formatted in lines that are 3.154 inches (80.1 millimeters) wide, in the Sabon family of typefaces at the 10 point size on lines spaced at 12 points. The text is strictly left to right and is left- and right-justified with hyphenation turned on. The commands issued were mainly for introducing the title, introducing the author's name, introducing sections, introducing