Discourse Patterns in Spoken and Written Corpora

Pragmatics & Beyond New Series

Volume 120

Discourse Patterns in Spoken and Written Corpora
Edited by Karin Aijmer and Anna-Brita Stenström

Discourse Patterns in Spoken and Written Corpora

Edited by

Karin Aijmer

Göteborg University

Anna-Brita Stenström

University of Bergen

John Benjamins Publishing Company

Amsterdam / Philadelphia

Library of Congress Cataloging-in-Publication Data

Discourse Patterns in Spoken and Written Corpora / edited by Karin Aijmer and Anna-Brita Stenström.
p. cm. (Pragmatics & Beyond, New Series, ISSN 0922-842X ; v. 120)
Includes bibliographical references and indexes.
1. Discourse analysis. I. Aijmer, Karin. II. Stenström, Anna-Brita, 1932- III. Series.

P302.D54888 2004
401'.41-dc22 2004041133
ISBN 90 272 5362 5 (Eur.) / 1 58811 506 2 (US) (Hb; alk. paper)

John Benjamins Publishing Co. · P.O. Box 36224 · 1020 ME Amsterdam · The Netherlands
John Benjamins North America · P.O. Box 27519 · Philadelphia PA 19118-0519 · USA

Contents

List of contributors

Karin Aijmer
English Department
Göteborg University
Box 200
SE 405 30 Göteborg
Sweden
karin.aijmer@eng.gu.se

Annalisa Baicchi
Department of Linguistics
University of Pavia
Corso Strada Nuova, 65
I-27100 Pavia
Italy
annalisa.baicchi@unipv.it

Julia Bamford
Dipartimento di Studi Geoeconomici.
Linguistici, Statistici, Storic per l'analysi regionale
Università di Roma 'La Sapienza'
Via Del Castro Laurenziano, 9
00161 Roma
Italy

Marina Bondi
University of Modena and Reggio Emilia
Facoltà di Lettere e Filosofia
Largo S.Eufemia 19
I-41100 Modena
Italy
mbondi@unimore.it

Silvia Bruti
Department of Linguistics
University of Pavia
Corso Strada Nuova, 65
I-27100 Pavia
Italy
s.bruti@unipv.it

Giuliana Diani
Department of Cultural and Language Studies
University of Modena and Reggio Emilia
Largo S. Eufemia, 19
I-41100 Modena
Italy
diani.giuliana@unimo.it

Hilde Hasselgård
Department of British and American Studies
University of Oslo
P. O. Box 1003, Blindern
0315 Oslo
Norway
hilde.hasselgard@iba.uio.no

Gunther Kaltenböck
Department of English
University of Vienna
Spitalgasse 2–4
A-1090 Wien
Austria
gunther.kaltenboeck@univie.ac.at

Anna Mauranen
English Department
School of Modern Languages and
Translation Studies
33014 University of Tampere
Finland
Anna.Mauranen@uta.fi

Christina Samson
Sezione di Lingue e Culture Straniere
Facoltà di Economia
Università degli Studi di Firenze
Sezione di Lingue e Culture Straniere
Università degli Studi di Firenze
Via Curtatone 1
50123 Firenze
Italy
christina.samson@cce.unifi.it

Anna-Brita Stenström
Department of English
University of Bergen
Sydnesplassen 7
N-5007 Bergen
Norway
stenstrom@kristianstad.mail.telia.com

Sanna-Kaisa Tanskanen
Department of English
University of Turku
20014 Turku
Finland
sakata@utu.fi

Kay Wikberg
Department of British and American
Studies
University of Oslo
P. O. Box 1003, Blindern
0315 Oslo
Norway
kay.wikberg@iba.uio.no

Discourse patterns in spoken and written corpora

Karin Aijmer and Anna-Brita Stenström
Göteborg University / University of Bergen

The purpose of the present volume is to bring together a number of empirical studies that use corpora to study discourse patterns in speech and writing. The papers are a selection of those presented in the Section Text and Discourse at the 5th ESSE Conference in Helsinki, 25th–29th August, 2000 with some added papers. The papers represent new trends in the area of text and discourse, characterised by the alliance between text linguistics and areas such as corpus linguistics, genre analysis, literary stylistics and cross-linguistic studies.

Both text linguistics and discourse analysis are concerned with text. But, as Stubbs points out (1983: 9), the terms text and discourse require some comment since their use is confusing. There has been a tendency to use 'text' for the printed record and 'discourse' for spoken texts. This is reflected in the names of the two disciplines text linguistics and discourse analysis. However, it should be kept in mind that there is a great deal of overlap between the disciplines. Brown & Yule (1986: 3) for instance use 'text' as a technical term to refer to the verbal record of a communicative act whether spoken or written.

Does the use of different terminology reflect different perspectives on the same area of research? Given the wide variety of approaches that are concerned with the analysis of text, what do these have in common? To begin with, it is necessary to look at the status and meaning of the terms text, discourse and function in modern linguistic theory.

Background

For the most part of the 20th century, linguists have been concerned with analysing sentences and with linguistic systems rather than the use of language. Chomsky set up as a goal for linguistics to describe the native speaker's competence, i.e. the tacit knowledge of the abstract rules of language formalized as a component consisting

of context-free rewrite rules and rules with transformational power. Made-up sentences were relied on and they hardly ever occurred in a context or cotext. In contrast, the study of discourse goes beyond the sentence and studies texts.

As Stubbs (1983: 12) points out, there has been "a gathering consensus, particularly since the mid-1960s, that some of the basic assumptions of Sausserean-Bloomfieldean-Chomskyan linguistics must be questioned". Such assumptions are for example that language should be studied for itself and that the highest unit of linguistic analysis is the sentence. Even during the heyday of Chomskyan linguistics, we find ideas sharply opposed to those represented in generative grammar in the British linguistic tradition. The importance of text and functions of language in context had been stressed already by Firth (1957); it was inspired by Malinowski's 'context of culture', and work on text and discourse has been taken further in the work by Halliday and by Sinclair. Halliday's theory of language leans towards the functional: "The particular form taken by the grammatical system of language is closely related to the social and personal needs that language is required to serve" (Halliday 1970: 142). Another development, going against the Chomskyan assumption about the superiority of intuitive data, is the corpus-based research by Quirk, Leech, Svartvik and others.

Text linguistics

Historically, text linguistics and discourse analysis represent two different approaches to the study of text and discourse. In Textlinguistics one studies written texts. Much of the work undertaken is concerned with the text as a product 'words on the page' and not as a process (cf. Brown &Yule 1983: 24: the 'text-as-product' view).

The term text linguistics usually refers to work done within a particular European tradition, represented for instance by van Dijk (1972) and by de Beaugrande (e.g. 1980). Linguists in this tradition turned to the text in order to cope with features which a sentence grammar could not handle, such as pronouns, ellipsis, etc. Typical of the text-linguistic approach is also the interest in coherence and cohesion. In Halliday and Hasan's view (Halliday & Hasan 1976: 4), cohesion occurs "where the interpretation of some element in the discourse is dependent on that of another". Cohesion is, for instance, created by reference, repetition, ellipsis, conjunction and lexical organisation.

Another view of text analysis is illustrated by Critical Discourse Analysis, a socially directed application of linguistic analysis. The goal is "to make mechanisms of manipulation, discrimination, prejudice, demagogy, and propaganda explicit and transparent" and to inquire not merely 'how and why' language barriers emerge and exist but also how they 'might be altered or even overcome' (Wodak 1990: 126; quoted from Asher 1994: 4576).

The view of language as a 'social semiotic' — simultaneously socially based and having socially instrumental meanings — shows that a major inspiration behind critical discourse analysis is Halliday (cf. Simon-Vandenbergen 2001: 80). A major proponent is Norman Fairclough (Fairclough 1992, 1995), but critical discourse analysis has also been increasingly recognised by European researchers such as R. Wodak (Vienna) and T. van Dijk (Amsterdam).

Linguistic theory and function

Linguists have had very different attitudes to language functions. Bloomfield (1933) turned his back on the problem by observing that "the statement of meanings is the weak point in language study" (Bloomfield 1933: 140; quoted from Sinclair & Coulthard 1975: 11). Questions of language functions have also been placed on the linguistic agenda as a result of the insights provided within speech act theory. Austin (1962), for example, made a distinction between a sentence and the act it is used to perform, and Searle gave an intentional account of sentence function in terms of felicity conditions associated with illocutionary acts (1969).

A functional approach to language has also been adopted by Halliday (1970, 1994) and by Sinclair (see Sinclair & Coulthard 1975). For Halliday, every text involves a particular context of use. It follows from this that language is organized functionally around particular metafunctions (ideational, textual, interpersonal) that are realised in grammar.

The Hallidayan concept of function is based on an analysis of grammar and is not a discourse notion. What this means is that, in Hallidayan linguistics, there is no need to talk about 'text linguistics' as if it were separated from other branches of linguistics. In this approach, text is an instantiation of the system, which is the potential: "The grammar, then, is at once a grammar of the system and a grammar of the text" (Halliday 1994: xxii, quoted from Simon-Vandenbergen 2001: 80).

Discourse analysis

In the early 1950s, Zellig Harris introduced the term discourse analysis and suggested that the goal of discourse analysis is to discover how discourse differs from random sequences of sentences. Harris only looked at formal patterns within the text. However , "in recent years the idea that a linguistic string (a sentence) can be fully analysed without taking 'context' into account has been seriously questioned" (Brown & Yule 1983: 25). To understand how language is used, we constantly need to refer to context. In discourse analysis some aspect of the context is always taken into account. Context can be interpreted widely as "a world filled with people producing utterances: people who have social, cultural, and personal identities,

knowledge, beliefs, goals and wants, and who interact with one another in various socially and culturally defined situations" (Schiffrin 1994: 364).

Early work in discourse analysis focused primarily on monologue (cf. Hoey 1983). More recently, discourse analysis has established itself particularly in the study of spoken interaction. In addition to work on informal conversation, work has been carried out on discourse which is more structured, such as classroom discourse (Sinclair & Coulthard 1975; Coulthard & Montgomery 1981). In the same descriptive tradition ('the Birmingham school') the discourse features of informal conversation have recently been analysed by Amy Tsui (Tsui 1994).

In descriptive discourse analysis, one tries to identify units of different sizes through their functions:

> We are interested in the function of an utterance or part of an utterance in the discourse and thus the sort of questions we ask about an utterance are whether it is intended to evoke a response, whether it is intended to evoke a response, whether it is intended to mark a boundary in the discourse, and so on.
> (Sinclair & Coulthard 1975: 14)

It follows that the grammatical, structural units of clause and sentence are not the most important ones, but there are grounds for postulating units such as lesson and lecture as the highest units in discourse (Sinclair & Coulthard 1975). Other descriptive categories to analyse discourse are for example turn, move and act.

Discourse analysis is one of the most vast but also least defined areas of linguistics. There are differences in terminology capturing different areas of interest. Conversation Analysis is a term used by scholars with an ethnomethodological approach (Sacks, Schegloff & Jefferson 1974; Schegloff 1992, 1997). In Conversation Analysis, the use of conversational data is fundamental, and the focus is on the emergence of discourse organisation and structure. What is said is always a response to what has been said before and has an effect on what comes afterwards.

In present-day linguistics, it is common to use discourse analysis as an umbrella term for all issues that have been dealt with in text and discourse (cf. Stubbs 1983: 10; Östman & Virtanen 1995: 244). This is how the term will be used in this book, which focuses on new topics and trends in text and discourse.

Recent trends in the linguistic study of text and discourse

The use of corpora for text-linguistic purposes

Corpora provide a new and powerful tool for the text linguist. As **Bondi** points out (this volume), text and discourse studies can only be fully developed when closer

analysis of particular instances of communicative events is integrated with quantitative data from wider textual bases.

In recent years, corpora have been increasingly used as a tool for the interpretation of texts. The advantages of corpora are well-known. They provide information about meanings which are not available through intuition and they can be used to study the use of language in different text types; the results are more objective and the research can be replicated (Svartvik 1992: 8ff.). Corpus linguistics has contributed to research in different ways. Large corpora representative of 'general English' such as the British National Corpus, the Cobuild Corpus and the Bank of English have had an impact on the study of texts since they make it possible to use quantitative data to look for the distribution of particular structures and meanings in different text types. Spoken corpora have been mainly conversational (the London-Lund Corpus of spoken English/LLC, the Santa Barbara Corpus of spoken American English, the Bergen Corpus of London teenage language/COLT). Spoken corpora have also been compiled in order to study 'English for Academic Purposes'. For example, the Michigan Corpus of Academic Spoken English (MICASE) contains lectures and other types of types of spoken academic discourse. Recently, corpora have also been used in critical linguistics to study 'stigma key words' in the context of Europescepticism (Teubert 2000) and as a tool for text explication (Sinclair 2001).

The corpora needed for text-analysis may also be tailor-made for the study of particular genres such as journal article abstracts, economic lectures, e-mailing list messages, headlines, titles ('reduced texts'). Such specialized corpora are suitable for investigating the use of micro-features in the text, such as the use of the deictic *here*, *however* or hedges. For example, **Julia Bamford** (this volume) uses a small corpus of academic lectures on economics in English (the Siena Corpus) as well as lecture data from the MICASE Corpus to study deictic terms. In addition, data from parallel corpora and translation corpora has been used to study areas such as information structure (and text organisation; see for instance the articles in *Languages in Contrast* 1999).

The interface between speech and writing

Recently we have seen an increasing concern with how texts are organised differently depending on whether the mode is speech or writing. Biber (1988) has for instance shown that we can apply a variety of techniques to text corpora and identify underlying dimensions of genre variation and variation between speech and writing. A number of studies have looked at differences between speech and writing in the areas of grammar and lexis (see e.g. the anthology by Tottie & Bäcklund 1986). Several articles in this volume show that there are interesting

differences between speech and writing depending on the external circumstances under which we write or speak. For instance, cataphoric (forwards-looking) reference is realised differently, as appears in the contributions by **Baicchi** and **Bruti** (see below).

In recent years personal computers has given rise to a type of dialogic interaction in written texts. **Sanna-Kaisa Tanskanen** shows in her contribution that written dialogue ('e-mail mailing lists') makes use of the same cohesive strategies as dyadic dialogue. It appears that the dyadic conversation represents a situation where the use of cohesion is carried furthest. Mailing-list texts would be situated between the two-party and the three-party conversation but closer to the former than to the latter. The fact that the written dialogue makes use of the same cohesive strategies as are favoured in dyadic dialogue is seen as a strong indication of the collaborative basis of cohesion.

Academic discourse

Academic discourse is a field with potential pedagogical applications. There are many analyses of scientific English reflecting the importance of spoken and written language in science (cf. Stubbs 1983: 18), and the study of the language used within academic disciplines also raises interesting questions about the role of language as constitutive of the discipline. There is a wide variety of variation within academic prose and there are text types which could be regarded as mixtures between several genres and/or modes.

In **Christina Samson's** contribution it is shown that written economic lectures are a mixed genre influenced by research articles as well as the spoken lecture on which they were based. Samson has studied the role of the personal markers *we* and *I* in written academic texts, notably written economics lectures, which are comparable to planned monologues. She argues that writing can be as interactive as speaking, since understanding presupposes collaboration between writer and reader. In order to prove her point she uses a corpus of ten economics lectures, which are all composed in the same way: introduction, middle and conclusion. The choice of personal markers, she says, reflects the way the writer might want to 'involve' the reader in the activity, which presupposes a certain degree of shared knowledge. This, Samson says, is what distinguishes written economics lectures from economics textbooks.

Research articles, lectures, abstracts, etc involve scientific procedures established by the social activity itself and are maintained by members of the professional community. Discourse patterns and discourse markers may also vary across a particular discipline or genre. **Anna Mauranen** compared the use of hedging expressions between the MICASE Corpus and the British National Corpus. A

distinction was made between vagueness hedges (*kind of, sort of, something like that*) and mitigators (*somewhat, a little bit*). When different genres were compared it was shown that the more dialogic genres tended to have more strategic (interactive) than epistemic uses. **Bondi** has shown that connectors may be used differently depending on the specific disciplinary culture. The resulting descriptions differ with regard to the degree of delicacy and depend on whether the focus is on the social activity (the communicative event), the genre (class of social events) or a specific subgenre.

Grammar and discourse

Traditional grammatical analysis stops with the sentence. As we saw above, a major factor in the rise of text linguistics was the fact that texts were needed to supplement existing theories and methods based on the analysis of the sentence. Text was shown to be needed as a unit larger than the sentence to explain grammatical phenomena such as pronouns, tense sequences, connectives. Research in the area of grammar continues to reveal phenomena which can profit from a discourse approach. **Gunther Kaltenböck** focuses in his contribution on the use of non-extraposition in speech and writing, arguing that the communicative function of non-extraposition has not received much attention. The construction is marked insofar as it is much less used than extraposition, especially in the spoken language. A study of its distribution shows that the use of non-extraposition decreases steadily from persuasive writing via academic writing to creative writing and from public to private dialogue; i.e. there is a decrease from formal to informal in either mode. The main reason for its low occurrence in spoken language is said to be due to the extra processing effort required for a subject in initial position, which is contrary both to the 'principle of weight' and 'the light subject constraint'. The author emphasises that the choice of non-extraposition vs. extraposition is related to 'given' and 'new' information in that, unlike extraposition, non-extraposition typically conveys given information, thus contributing to the cohesion of a text.

Coherence and cohesion

In text linguistics one has long been concerned with the principles of connectivity which bind a text together. Eugene Winter (e.g. Winter 1977) defined 'clause relation' as a cognitive process whereby we interpret the meaning of a sentence in the light of its adjoining sentences. Similarly, Halliday & Hasan (1976) are concerned with different resources for text construction and cohesion. Halliday and Hasan's ideas are further elaborated by Martin (1992), using systemic functional grammar to ask questions about text structure. An influential theory used primarily in the computational domain is Mann & Thompson's theory of rhetorical relations

(1988). Text coherence is attributed to rhetorical relations such as contrast and sequence, which are mapped unto schemas rather than structures.

Recently one has also taken an interest in the efficiency and appropriateness of cohesive devices. Baicchi and Bruti deal with cataphoric reference in speech and writing respectively. The focus is on the complexity which results from a momentary gap of information when the reference is cataphoric rather than anaphoric.

Anna-Lisa Baicchi's article is related to an Italian research project aiming at constructing a hierarchical scale of complexity for linguistic and textual phenomena. The article illustrates the interplay between interpretability, complexity and markedness with special emphasis on the cataphoric reference of titles and headlines. Markedness is seen as a gradual concept. In order to show this, Baicchi suggests three basic evaluation criteria: quantity of indexical items in the title, quality of the items, and distance between the cataphoric items in the title and their co-referents in the text. Analysing some titles in terms of transparency, she identifies four different types: totally transparent, partially transparent, symbolically related, and opaque.

Silvia Bruti's article is related to the same project on text complexity as Baicchis's, and this article, too, deals with cataphora, but with the focus on spoken discourse. Bruti bases her analysis on data from the London-Lund Corpus of Spoken English (LLC) and the British National Corpus (BNC). An inventory of cataphoric devices is followed by the analysis of some of these devices in the two corpora. The most frequently used device is the demonstrative pronoun *this*. But also vague expressions, especially *thing*, can have a cataphoric function in conversation, by emphasising what follows and keeping the listener's attention alive. Another device with a similar function is the 'attention-getter' *do you know what I mean*. The discussion about cataphoric complexity, and the parameters involved, shows that cataphora and anaphora are not two opposite textual strategies. The article ends with a brief section on how to calculate the markedness and complexity inherent in cataphoric structures.

Connectors and discourse markers

Hoey (1983: 33) has drawn attention to the fact that there are clues in the surface of discourse making it possible to perceive the structure and making it possible to build an infinite number of discourse patterns. In spoken English we find markers which are not usually considered in grammars and which have essentially interactive functions. Typical markers are *anyway, well, I mean, I think, you know* and many more. These have been shown to segment the discourse flow and to have discourse functions such as changing the topic. The focus of research has also shifted from the functions of markers in local structures to take into account their roles and patterning in specific generic structures, as is clear from **Marina Bondi's** paper.

Bondi highlights the discursive roles of *however* in different parts of journal abstracts from history, economics and sociology. In a broader perspective the aim is to show that there is a close link between the linguistic choices and epistemology in academic disciplines. Connectors such as *however* are seen in terms of their interpersonal meaning: they assume that there is a common ground and contribute to interpersonal or evaluative coherence. It is argued that it is fruitful to analyse *however* and causal connectors in general with reference to the argumentative dimension of shared stereotypic knowledge, and that they are used to present claims and counterclaims. The analysis of the causal connectors was carried out in three steps, starting with a frequency list and key words making it possible to get a better picture of the behaviour of causal connectors in the three disciplines. In the second step the patterns and meanings of causal connectives within the three disciplines were compared. In the final stage a sub-corpus of *however* in historical abstracts was investigated.

Giuliana Diani discusses the pragmatic functions of *I don't know* in examples from the spoken part of the Cobuild Corpus. She argues that the basic semantic meaning of *I don't know* underlies all the pragmatic functions, regardless of whether it is used, for instance, as a mitigating strategy avoiding face threats or as a filler for time, to take two extremes. The first part of the paper deals with the concept of face. This is followed by a discussion of the various functions of *I don't know* in examples from the corpus. The third part of the paper considers the frequent occurrence of *I don't know* with the discourse markers *oh, you know, I mean* and especially *well,* each of which adds a particular pragmatic effect: reinforcement (*oh*), cooperation (*you know*), self-correction (*I mean*), insufficiency (*well*). The author concludes by emphasising the difficulty of providing clear-cut distinctions between different pragmatic functions.

Deixis and non-verbal communication

The interconnectedness of the verbal and visual in communication is a long neglected but rapidly expanding research topic. Deictic links can for example be made both through the spoken and the visual mode. The relationship between the text and the visual are particularly interesting in lectures and conference presentations. **Julia Bamford** uses a small corpus of academic lectures on economics in English (the Siena Corpus) as well as lecture data from the MICASE Corpus to analyse deictic expressions. The Siena Corpus was used to investigate how lecturers use visual materials (graphs, diagrams, maps, etc) and the relation between deictics and gesture.

In the corpus there was a high frequency of occurrences of *here* linked to a gesture. Gestural *here* is always relatively precise and refers to something in the local context. The gesture itself may precede the deictic as seen on the video. In lectures

the majority of examples of *here* were gestural and commented on something visual in the lecture. The gestural *here* is distinguished from other less frequent uses of the adverb. The deictic *here* is also used with vague referents (the symbolic use of deictics). The referent of the symbolic *here* is less precise since it belongs to the common cognitive space of both the speakers and their student audiences. Depending on the context there are several reasons for using the vague *here* such as the wish to create involvement and group-feeling. When *here* has textual function it can indicate a contrast between parts of the text.

Contrastive studies

Contrastive discourse analysis is an area where we can expect more attention in the future. Comparing structures and patterns of texts in different languages has been the subject of much study in contrastive rhetoric (Kaplan 1972, Connor 1987) and genre analysis (Swales 1990). A recent development is to use bilingual corpora for text-linguistic purposes. **Hilde Hasselgård** used Halliday's views on the role of multiple themes to illustrate how we can get additional evidence for the functions of multiple themes by bringing in contrastive data. The data from the English-Norwegian Parallel Corpus and from the Oslo Multilingual Corpus permits the author to see how multiple themes are rendered in Norwegian and German. Since both are V2-languages, translators will have to make priorities as regards the element placed as theme.

A general finding was that the number of elements which could be accommodated as theme was reduced and that the thematic elements were not expressed post-verbally in translation. The investigation further confirmed the hypothesis that the majority of multiple themes contain at least one cohesive tie. As regards the type of cohesive link, Hasselgård found that reference was most frequent, followed by conjunction and lexical cohesion. Another hypothesis which was confirmed was that multiple themes can bring a non-cohesive element into thematic position. When a multiple theme was paragraph-initial, it was generally cohesive and marked continuity rather than a topic-break. A sequence of thematic adverbials on the other hand had an 'ice-breaker' function before a shift in discourse.

The use of corpora to study metaphor

Metaphors have usually been studied on the sentence or clause level. **Kay Wikberg** argues in his paper that it is important to study metaphors in text and analyse their contribution to coherence in discourse. It is shown that most of the metaphors in the corpus investigated are evaluative in some sense and would realise Halliday's interpersonal metafunction. Wikberg shows that multilingual corpora such as the Oslo Multilingual Corpus allow us to study metaphors in authentic texts and their translations. The computer also helps to trace chains of cognitively related meta-

phors and to see the semantic fields that come into play in the metaphorical expressions and their interpretation.

Prospectives for the future

In the editorial of a special issue of the journal *Text* 'Text linguistics at the millennium: Corpus data and missing links' Wilson and Sarangi (2000: 149) write that "we would like to see more corpus-based, descriptive work being undertaken, especially with a special focus on theoretical issues surrounding the organization and consumption of texts in social contexts. Corpus-based studies — which combine quantitative and qualitative studies of language — have already proven to be practically relevant in the area of language teaching and in revisiting the differential norms of spoken and written grammar". It is clear from the contributions to this volume that corpus-based studies have the potential to ask a number of new questions about context, text types, differences between speech and writing, cohesive devices, and stylistic devices such as metaphor.

References

Asher R. E. (ed.)
1994 *Encyclopedia of Language and Linguistics*. Oxford, etc: Pergamon Press.

Austin, John L.
1962 *How to Do Things with Words*. Ed. by J. O. Urmson. London: Oxford University Press.

Beaugrande, Robert de
1980 *Text, Discourse, and Process*. Ablex, Norwood, NJ.

Biber, Douglas
1988 *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Bloomfield, Leonard
1933 *Language*. New York: Henry Holt.

Brown, Gillian and Yule, George
1983 *Discourse Analysis*. Cambridge: Cambridge University Press.

Connor, Ulla
1987 "Argumentative patterns in student essays: Cross-cultural differences". In *Writing across Languages: Analysis of L2 text*, U. Connor and R. B. Kaplan (eds), 57–72. Reading, M. A.: Addison Wesley.

Coulthard, Malcolm and Montgomery, Martin (eds)
1981 *Studies in Discourse Analysis*. London: Routledge & Kegan Paul.

Dijk, Teun A. van
1972 *Some Aspects of Text Grammars*. The Hague: Mouton.

Fairclough, Norman
1992 *Discourse and Social Change.*Cambridge: Polity Press.
1995 *Critical Discourse Analysis.* London: Longman.
Firth, J. R.
1957 "The technique of semantics". In *Papers in Linguistics 1934–51.* London: Oxford University Press.
Halliday, M. A. K.
1970 "Language structure and language function". In *New Horizons in Linguistics,* J. Lyons (ed.), 140–65. Harmondsworh: Penguin.
Halliday, M. A. K.
1994 *An Introduction to Functional Grammar.* 2nd ed. London: Edward Arnold.
Halliday, M. A. K. and Hasan, Ruqaiya
1976 *Cohesion in English.* London: Longman.
Hoey, Michael
1983 *On the Surface of Discourse.* London: George Allen & Unwin.
Kaplan, Robert B.
1972 *The Anatomy of Rhetoric: Prolegomena to a Functional Theory of Rhetoric.* Philadelphia: Center for curriculum development. (Distributed by Feinle & Feinle.)
Mann, William C. and Thompson, Sandra A.
1998 "Rhetorical structure theory: A theory of text organisation". *Text* 8(3): 243–81.
Martin, J. R.
1992 *English Text: System and Structure.* Amsterdam and Philadelphia: John Benjamins.
Östman, Jan-Ola and Virtanen, Tuija
1995 "Discourse analysis". In *Handbook of Pragmatics. Manual.* J. Verschueren, J.-O. Östman and J. Blommaert (eds), 239–53. Amsterdam and Philadelphia: John Benjamins.
Sacks, Harvey, Schegloff, Emanuel and Jefferson, Gail
1974 "A simplest systematics for the organization of turn-taking for conversation". *Language* 50(4): 696–753.
Schegloff, Emanuel
1992 "On talk and its institutional occasions". In *Talk at Work.* P. Drew and J. Heritage (eds), 101–136. Cambridge: Cambridge University Press.
1997 "Whose text? Whose context?" *Discourse and Society* 8(2): 165–187.
Schiffrin, Deborah
1994 *Approaches to Discourse Analysis.* Oxford UK & Cambridge USA: Blackwell.
Searle, John R.
1969 *Speech Acts. An Essay in the Philosophy of Language.* Cambridge: Cambridge University Press.
Simon-Vandenbergen, Anne-Marie
2001 "Analysing text and discourse". *The European English Messenger* X/1:79–85.

Sinclair, John
2001 "A tool for text explication". In *A Wealth of English: Studies in Honour of Göran Kjellmer*, K. Aijmer (ed.), 163–76. [Gothenburg Studies in English 81]. Göteborg. Sweden.
Sinclair, John and Coulthard, Malcolm
1975 *Towards an Analysis of Discourse. The English used by Teachers and Pupils.* Oxford: Oxford University Press.
Stubbs, Michael
1983 *Discourse Analysis. The Sociolinguistic Analysis of Natural Language.* Oxford: Blackwell.
Svartvik, Jan
1992 "Corpus linguistics comes of age". In *Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82 Stockholm, 4–8 August 1991*, J. Svartvik (ed.), 7–13. Berlin: Mouton.
Swales, John
1990 *Genre Analysis: English in Academic and Research Settings.* Cambridge: Cambridge University Press.
Teubert, Wolfgang
2000 "A province of federal superstate, ruled by an unelected bureaucracy: Keyword of the Eurosceptic discourse in Britain". In *Attitudes towards Europe: Language in the Unification Process*, A. Musolff, C. Good, P. Points and R. Wittlinger (eds), 45–86. Aldershot: Ashgate.
Tottie, Gunnel and Bäcklund, Ingegerd (eds)
1987 *English in Speech and Writing. A Symposium.* Stockholm:Almqvist &Wiksell.
Tsui, Amy
1994 *English Conversation.* London: Oxford University Press.
Sarangi, Srikant and Wilson, John
2000 "Editorial". *Text* 20(2): 147–151.
Winter, Eugene O.
1977 "A clause relational approach to English texts: a study of some predictive lexical items in written discourse". *Instructional Science* 6 (1): 1–92.
Wodak, Ruth
1990 "Discourse analysis: Problems, findings, perspectives". *Text* 10: 125–32.

Part I

Cohesion and coherence

The cataphoric indexicality of titles*

Annalisa Baicchi
University of Pavia

"*Lost Illusion* is the undisclosed title of every novel"
André Maurois

1. Introduction

In this paper, some issues related to the complex nature of titles are discussed in connection with information encoding and text processing. The paper is part of a much wider research project, co-funded by the Italian Ministry of Education and the University of Pisa (http://www.humnet.unipi.it/citatal), which aims at arranging linguistic and textual phenomena along a hierarchical scale of complexity. The project also aims at identifying specific criteria for a theoretically based definition of complexity. Language and text complexity is motivated syntactically, lexically and pragmatically, but the notion of complexity is still too general and not always based on theory, since it is also arrived at through empirical evidence, and intuition. Complexity differs from difficulty and the two notions should be kept distinct, at least from a theoretical point of view, although they can be correlated when information processing and discourse understanding are taken into account. In the present approach, interpretability, complexity and markedness are three faces of the same object and their interplay is what this paper will try to illustrate. The specific phenomenon investigated is the marked status of titles in terms of *phoricity*. Markedness is provisionally intended to be a way to assess its complexity.

2. The retrieval of data

A corpus designed for allowing queries about *Titlelogy* has been assembled with recourse to the "Online Books Page" (http://digital.library.upenn.edu) and to the

"English Server" (http://eserver.org). The "Online Books Page" website is hosted by the University of Pennsylvania Library, founded and edited by Jonh Mark Ockerbloom, a digital library planner and researcher at that University. The website facilitates access to books that are freely readable over the Internet. The index includes more than 20,000 works in various formats and genres written in or translated into English. The index of individual titles includes books and definitive collections (e.g. literary works from ancient to modern world literatures, popularising books, handbooks), and major serials (i.e. magazines, newspapers, journals, and the like). The main criterion for works to be listed in the Online Books is to be listed as books or serials in the online catalogue of a major library such as the Library of Congress. The "English Server" website, founded in 1900 and hosted by the Iowa State University, stores almost 32,000 works belonging to forty-four collections on different topics like world literatures of any genre and period, magazines, and journals, but also design, multimedia, contemporary art, and current political and social issues.

3. The *phoricity* of titles

The referential function of titles can be exophoric and endophoric reflecting the fact that titles may at the same time refer to entities in the outside world and to entities present in the text base. Exophoric function includes semantic reference, indexical reference (reference to the writer's attitude), and intertextual reference. Endophoric titles refer to the text and have intratextual function. They are cataphoric from the perspective of the receiver and anaphoric from the perspective of the text producer. When titles are mainly exophoric, that is, refer to the general context, they are less *complex* in terms of interpretability since they can also rely on the receiver's world knowledge. In contrast, strictly endophoric titles may require exclusive recourse to the text for their interpretability. Complexity in the latter case is definable according to various parameters, which will be illustrated presently. Let us consider the following examples:

(1) The President of the U. S. meets the Pope
(2) Access to Web negated
(3) He wrote after she died

In (1) the referents for both the first and second noun phrase are easily accessed, thanks to the knowledge of the world that a reader is supposed to have, whereas in (2) the receiver's interest is naturally focused on who the actor and patient of the action are. In the latter case only recourse to the text (a newspaper article) will help

the receiver recover the information required. In (3), the headline clearly relies on previous information, provided by the same or other newspapers. In any case, it directs the reader to the text for filling the two referential pro-forms. In (2) and (3) especially, the title is clearly defined as cataphoric (on the part of the receiver).

Cataphora is a marked phenomenon as compared to anaphora, since it represents a gap of information, and, in semiotic terms, a signans with deficient or no signatum. My interest is especially in identifying criteria and parameters for measuring the cataphoric markedness of various types of titles. Markedness is here intended as a scalar concept (see Section 6).

The cataphoric quality of titles derives from their being indexical signantia, whose signata have to be retrieved from the text base. Although titles are external to the text, they are related to it in terms of contiguity, which follows from their being indexical.[1]

The notion of index can be better explained with recourse to Ch.S. Peirce's description of the nature of signs:

> Every sign has, actually or virtually, what we may call a *Precept* of explanation according to which it is to be understood as a sort of emanation, so to speak, its Object. If the Sign be an Icon, a Scholastic might say that the 'species' of the Object emanating from it found its matter in the Icon. If the Sign be an Index, we may think of it as a fragment torn away from the Object, the two in their Existence being one whole or a part of such whole. (Peirce 1965: 2.230).

Applying the Peircean notion to our discussion, the title may be viewed as representing the *fragment torn away* from the text. The fragment is supposed to retain traces of its primordial unity with the object, or, abandoning the metaphor, titles are expected to contain elements similar to or congruent with the text content. My aim is to investigate the relation between title and text. I will analyse various types of titles in terms of their indexical (in)efficiency, which is tantamount to saying that I will evaluate their cataphoric markedness, and the correlated complexity.

Cataphora, as compared to anaphora, is an example of marked indexicality. Whereas anaphora is an efficient index since the retrieval of its co-referent, and therefore co-interpretability, is immediate, cataphora is a less efficient index because its co-interpretability is delayed, sometimes even to the end of the reading process. Still, relative (in)efficiency can be measured and various aspects of it defined. A more efficient index allows easier retrieval of the object (and is less marked and therefore less complex). Retrieval is strongly dependent on explicitness (of the title), which is the first variable of interest to defining the markedness of cataphoric reference. Such titles as Eco's *The Name of the Rose* will be interpreted only through the use of complex inferences. This type of title, that Eco labels *evocative titles*, does not anticipate anything and, rather, may be misleading before

the reading process starts, and not easily connected to a signatum during or even after the reading process. Other variables, modifying the indexical status of titles, are the remoteness of the co-referents in the text, the number of signantia in the title that need to be filled with their respective signata, (dis)similarity between referential expressions in title and text, etc.

Complexity of titles is to be envisaged as: (1) a low degree of explicitness, that is, more difficult access to the referent; (2) delay in the co-interpretation; (3) few elements of contiguity between title and text base; (4) many elements to be filled with recourse to text items. These aspects are partly quantitative (number of cataphoric items to be made interpretable and of congruent items between title and text, quantity of text separating the cataphoric items in the title and their co-referents in the text), and partly qualitative (semantic transparency, i.e. access to referent, semantic load of lexis in terms of connotations, associations, etc.).

My analysis of titles attributes values on three different scales: indexical efficiency, markedness, and complexity. Phenomena will be shown to co-vary along these three scales, which will indicate, on the one pole, indexical efficiency, minimal markedness, minimal complexity, and, on the other pole, the opposite values. Due to the limited size of this paper, a thorough analysis is impossible, but I expect that my suggestions will also apply to a larger database.

To summarise, in order to arrange titles along the scales, I propose to utilize the following basic evaluation criteria: (1) number of indexical items contained in the title, (2) quality of the items in terms of semantic transparency, and (3) distance between the cataphoric items in the title and their co-referents in the text. We will see that each criterion subsumes some other criteria.

3.1 How titles are viewed in the literature

To begin with, I will report some discussions about the nature of titles to be found in the relevant literature, with the aim of defining a frame of reference and providing background knowledge. My own approach is, however, different from the majority of these treatments.

Hoek (1972, 1973) defines the title as an artificial object dependent on its reception. It is interpreted, more or less arbitrarily, by readers, critics, or bibliographers on the basis of the layout of the book cover or the frontispiece. He suggests a bipartite distinction based on the position of the two constituents that are separated by a comma: what comes before the comma he labels 'title', and what comes next 'subtitle'.

Duchet (1973, 1979) considers Hoek's proposal too vague, and suggests a more articulated labelling. As an example he considers the title *Zadig ou la Destinée, histoire orientale*, and proposes to name *Zadig* the title, *ou la Destinée* the second title, and

what comes after the comma, *histoire orientale*, the subtitle. Hoek (1981), analysing the same example, considers *Zadig* to be the title, *ou la Destinée* to be the secondary title, and *histoire orientale* to be the subtitle. Genette (1987) summarizes the debate, and makes a proposal, whose goal — he claims — is not a matter of labelling, but of identifying the constitutive elements of a title. In my opinion, Genette's proposal does no more than introduce a different terminology and he does not shed new light on the question. The example is the same as before, but this time the new labels are title for *Zadig*, subtitle for *ou la Destinée*, and generic indication for *histoire orientale*. As the author admits, the generic indication is a rather heterogeneous ingredient since it has to be defined in functional terms, whereas the title and the subtitle are defined in formal terms. The generic indication is a more autonomous paratextual element that can have a varying influence on the other two, depending on which element of the entitling the reader appends it to: if appended to the title *Zadig*, the generic indication *histoire orientale* may be interpreted as an attribute for the protagonist; if appended to the subtitle *ou la Destinée*, it may be interpreted as an attribute relevant to the whole plot of the novel. In Genette's view, the entitling act is reduced to a structure formed by a title plus subtitle.

Rey-Debove (1978) defines titles as metalinguistic proper names in that they are means for designating the text. On the syntagmatic axis, a title is in apposition to the following text, whereas, on the paradigmatic axis, it is a synonym that may be a substitute for the text since it signifies, in a focalised and abbreviated way, the same thing signified by the text. With reference to Rey-Debove, Marello (1992) suggests that the terms *apposition* and *synonym* should be replaced by the two terms *anaphora* and *cataphora*, borrowed from text linguistics (see also Lyons 1977 (vol. 2, ch.16.2), Weinrich 1993). These terms were first used by Hoek (1981), who, as pointed out by Marello (1992), indicated that the relation between title and text is one of cataphoric expansion or anaphoric contraction, depending on the perspective. Cataphoric expansion occurs when, starting from a given title (i.e. *macrostructure* in our terminology) a cotext, development, or comment (i.e. *microstructure*) is provided. Anaphoric contraction is the reversed process, that is, when, starting from a given topic (*microstructure*), a summary (*macrostructure*) is produced.

My own approach, as I have made clear, concentrates on the cataphoric expansion scheme, which takes the receiver's perspective and points forwards in the text.

3.2 Functions of the title

According to Grivel (1973), critics are agreed on recognizing three main functions for titles: they can (1) identify the work, (2) designate its content, and (3) evaluate it. Hoek (1981) accepts these three functions and integrates them into his own