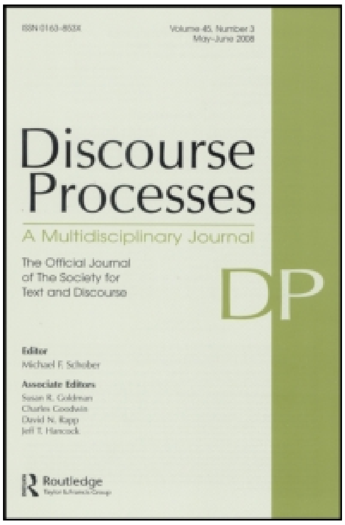

ISSN 0163-853X
Volume 45, Number 3
May-June 2008

Discourse Processes
A Multidisciplinary Journal
The Official Journal of The Society for Text and Discourse
DP

Editor
Michael F. Schober

Associate Editors
Susan R. Goldman
Charles Goodwin
David N. Rapp
Jeff T. Hancock

Routledge
Taylor & Francis Group

Discourse Processes
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t775653637

Surface Cues of Content and Tenor in Texts

Luuk Lagerwerf; Wilbert Spooren; Liesbeth Degand

Online publication date: 08 June 2010

To cite this Article Lagerwerf, Luuk , Spooren, Wilbert and Degand, Liesbeth(2006) 'Surface Cues of Content and Tenor in Texts', Discourse Processes, 41: 2, 111 — 116

To link to this Article: DOI: 10.1207/s15326950dp4102_1

URL: http://dx.doi.org/10.1207/s15326950dp4102_1

DISCOURSE PROCESSES, *41*(2), 111–116

INTRODUCTION

Surface Cues of Content and Tenor in Texts

This special issue of *Discourse Processes* contains a selection of articles from the workshop Multidisciplinary Approaches to Discourse (MAD03)—a biennial workshop bringing together researchers from various disciplines and with a mutual interest in the study of discourse. The 2003 edition's aim was to tackle the issue of how to analyze content and tenor of texts. This topic has a background in various disciplines, which were all represented at the workshop: content analysis, discourse psychology, and computational and cognitive–functional linguistics. A number of articles addressed questions concerning the cues signaling a text's content and tenor and the kind of information or effect that is conveyed this way. Four manuscripts evolving from these articles were selected for this issue. They are examples of the various approaches of discourse analysis making use of text corpora. Different statistical and computational techniques are used to analyze surface cues that signal content or tenor in texts. In this introduction, we present a short overview of current topics in corpus analysis as a tool for discourse analysis. We will show how the four contributions to this special issue represent recent developments.

CORPUS ANALYSIS AS A TOOL FOR DISCOURSE ANALYSIS

Corpus analysis is a method of linguistic analysis based on naturally occurring samples of texts and spoken discourse (corpora). Corpora have become a standard element of the linguist's toolkit. Their function is obvious: If linguists want to study actual language use, they need to look at concrete instances of that use to come up with new hypotheses, to increase the reliability of their analyses, and to test already available hypotheses.

The advent of sophisticated computer technologies has made it feasible to perform large-scale systematic research of large bodies of text on specific linguistic properties. Large corpora of text have been assembled for linguistic research purposes, as well as methods to retrieve the relevant linguistic information from these texts (see Biber, Conrad, & Reppen, 1998, for an example of this kind of work). It follows that traditional sentence analysis tends to be replaced by discourse analysis for various reasons: Robust linguistic models need to be able to cope with the complexities of discourse and with naturally occurring examples usually found in discourse. The complexity of linguistic phenomena in discourses makes it inconvenient and implausible to make up discourses for analytic purposes.

In all areas of linguistic research, corpora are being used. Some influential examples are studies of cross-language variation (Degand, 2001; Granger, Lerot, & Petch-Tyson, 2003), cross-genre stylistic variation (Conrad & Biber, 2001), cross-linguistic comparison of conceptual metaphor (Cameron & Low, 1999), and language learning (Tomassello, 2003).

Tools for corpus analysis are not applied in linguistics exclusively. An example from social psychology is the *Linguistic Inquiry and Word Count Dictionary*, which is used to detect personality differences between people on the basis of tenor differences in their language use (Pennebaker, Francis, & Booth, 2001; Pennebaker & King, 1999).

By using techniques of tagging parts of speech in electronic corpora, it is possible to make abstractions over linguistic categories. Developments in rule-based and probabilistic methods of recognizing and annotating linguistic elements have made tagging faster and more extensive (Brill, 1995; van Halteren, Daelemans, & Zavrel, 2001). The representations that part-of-speech taggers produce are very useful for discourse analysis. It has become possible to search systematically for those surface cues that signal discourse phenomena. These increasing possibilities make it more interesting to study the way in which surface cues may determine discourse characteristics.

This special issue is dedicated to the study of some of these surface cues. The question that unites the articles is what discourse functions these surface cues serve. The articles differ in their choice of surface cues and discourse function. Together they provide an overview of current approaches in discourse analysis.

DISCOURSE ANALYSIS

In both linguistics and discourse psychology, discourse analysis has become a prominent part of the work. In psychology, discourse analysis is used to study how a reader goes through the process of converting linguistic symbols into knowledge (Graesser, Millis, & Zwaan, 1997; Kintsch, 1998; van Dijk & Kintsch, 1983). Textual elements such as connectives function as cues to build a hierarchical structure of a text (Mann & Thompson, 1988), and they help in building a coherent representation (Sanders, Spooren, & Noordman, 1993). Moreover, connectives are used to indicate interactive effects of written discourse, such as the degree of a writer's subjectivity toward the content of what is expressed (Halliday, 1985; Langacker, 1985; Pander Maat & Degand, 2001; for spoken discourse, see Schiffrin, 1987). Connectives thus serve as surface cues modeling content (hierarchical and coherent representations) as well as tenor (subjectivity).

Parallel to the work on discourse representation, computational models have been developed to represent the content of a text (Gardent & Webber, 2001; Lagerwerf, 1998; Polanyi, 1988; Prüst, Scha, & van den Berg, 1994). These models make use of formal theories of discourse representation (Asher & Lascarides, 2003; Beaver, 2001; Kamp & Reyle, 1993). There also has been a substantial body of computational work on the link between the intention of a writer and the production of text (Grosz & Sidner, 1986; Hovy, Lavid, Maier, Mittal, & Paris, 1992; Matthiessen & Bateman, 1991). These theories model the tenor of a text. As in the discourse psychological and text linguistic work, all of these computational approaches study the properties of surface cues as the provider of essential information with which to build their models.

A recent line of work is based on the co-occurrence of words and statistical analysis (Bod & Scha, 1996). One of the more sophisticated frameworks is Latent Semantic Analysis (LSA; Landauer, Foltz, & Laham, 1998). In this framework, a semantic space is built on the basis of a specific statistical analysis of all word–context combinations in a text corpus. LSA possibly mimics the cognitive processes that take place during language comprehension (Landauer & Dumais, 1997), including the processing of text coherence (Foltz, Kintsch, & Landauer, 1998). In this line of work, corpus analysis and discourse analysis are integrated. Surface cues are, in fact, all words in their contexts, without distinction, and frequencies of words are the initial measures. In a second stage, when semantic spaces have been built, distances between specific surface cues can be used to make discourse analytic inferences.

The combination of corpus analysis and discourse analysis enables researchers to build specific computer applications. Numerous computer applications of LSA exist (Foltz, 2005). A comprehensive application addressing the automatic determination of the readability of texts, using LSA cohesion measures as well as well-known readability formulas and other measures, is Coh-Metrix (Graesser, McNamara, Louwerse, & Cai, 2004).

In this issue, each article represents one of these approaches in discourse analysis. The articles exemplify discourse analysis in various forms, with the aid of different kinds of corpus analysis, including both quantitative and qualitative methods. They share the purpose of analyzing how surface cues indicate and model content and tenor of texts.

TEXTUAL SURFACE CUES OF CONTENT AND TENOR

The four contributions in this issue vary in the kind of information that surface cues give: density of information, subjectivity, causality, and interclausal versus intraclausal discourse representation. They also vary in method of corpus analysis: descriptive and hypothesis-testing statistical techniques, automated statistical techniques to falsify hypotheses, and in-depth qualitative analyses of selected examples. Together they exemplify the wide variety of discourse analysis in corpus research.

Dorit Ravid and Ruth Berman show how written narratives, compared with spoken narratives about the same event from the same narrator, contain much less nonreferential material, making the representation of information more dense and providing less explicit clues about the representation. They compared spoken and written corpora on the use of specific surface cues. Their contribution represents the discourse psychological approach.

Mirna Pit shows how subjectivity is expressed differently for several causal connectives by analyzing how these connectives interact with other subjectivity indicators in a corpus analysis. A corpus of newspaper items was analyzed systematically, and a linguistic approach to discourse analysis was followed in this article.

Yves Bestgen, Liesbeth Degand, and Wilbert Spooren used automated techniques like LSA and Thematic Text Analysis to test hypotheses about the subjectifying properties of certain causal connectives. They used large amounts of texts as input for their analyses. Their information retrieval approach was used to test discourse analytic hypotheses.

Michael Grabski and Manfred Stede studied the different occurrences of the German preposition–connective *bei* and analyzed its function for discourse representation. They based their analysis on selected examples from several corpora. Their work can be placed in the context of computational discourse analysis.

With this issue, we intend to present to the community one version of the state of the art with respect to discourse analysis in corpus research. We hope that this volume will contribute to that aim. Other approaches of text research that came forth from MAD03 were published in a special issue of *Information Design Journal + Document Design* (see Foltz, 2005).

We thank the following people for their role in the reviewing process:

Nadjet Bouayad, *Universitat Pompeu Fabra*
Wallace Chafe, *University of California at Santa Barbara*
Lucile Chanquoy, *Université de Nice Sophia Antipolis*
Peter Foltz, *New Mexico State University*
Alistair Gill, *University of Edinburgh*
Michael Grabski, *Technical University of Berlin*
Eduard Hovy, *University of Southern California*
Walter Kintsch, *University of Colorado*
Emiel Krahmer, *University of Tilburg*
Ronald Langacker, *University of California, San Diego*
Max Louwerse, *University of Memphis*

Luuk Lagerwerf
Wilbert Spooren
Liesbeth Degand
Co-Editors

REFERENCES

Asher, N., & Lascarides, A. (2003). *Logics of conversation.* Cambridge, England: Cambridge University Press.

Beaver, D. I. (2001). *Presupposition and assertion in dynamic semantics*. Chicago: University of Chicago Press.

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use.* Cambridge, England: Cambridge University Press.

Bod, L. W. M., & Scha, J. H. (1996). *Data oriented language processing: An overview.* Amsterdam: ILLC.

Brill, E. (1995). Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. *Computational Linguistics, 21*(4), 543–565.

Cameron, L., & Low, G. (Eds.). (1999). *Researching and applying metaphor.* Cambridge, England: Cambridge University Press.

Conrad, S., & Biber, D. (Eds.). (2001). *Variation in English: Multi-dimensional studies*. London: Longman.

Degand, L. (2001). *Form and function of causation. A theoretical and empirical investigation of causal constructions in Dutch*. Leuven, Belgium: Peeters.

Foltz, P. W. (2005). Automated content processing of spoken and written discourse: Text coherence, essays, and team analyses. *Information Design Journal + Document Design, 13*(1), 5–13.

Foltz, P. W., Kintsch, W., & Landauer, T. K. (1998). The measurement of textual coherence with latent semantic analysis. *Discourse processes, 25*(2–3), 285–307.

Gardent, C., & Webber, B. (2001). Towards the use of automated reasoning in discourse disambiguation. *Journal of Logic, Language and Information, 10*(4), 487–509.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers, 36(2)*, 193–202.

Graesser, A. C., Millis, K. K., & Zwaan, R. A. (1997). Discourse comprehension. *Annual Review of Psychology, 48,* 163–189.

Granger, S., Lerot, J., & Petch-Tyson, S. (Eds.). (2003). *Corpus-based approaches to contrastive linguistics and translation studies*. Amsterdam: Rodopi.

Grosz, B. J., & Sidner, C. L. (1986). Attention, intentions, and the structure of discourse. *Computational Linguistics, 12,* 175–204.

Halliday, M. A. K. (1985). *An introduction to functional grammar.* London: Edward Arnold.

Hovy, E., Lavid, J., Maier, E., Mittal, V., & Paris, C. (1992, April). Employing resources in a new text planner architecture. In *Proceedings of the 6th International Workshop on Natural Language Generation.* Workshop conducted in Castel Ivano, Trento, Italy.

Kamp, H., & Reyle, U. (1993). *From discourse to logic. Introduction to modeltheoretic semantics of natural language, formal logic and discourse representation theory.* Dordrecht, The Netherlands: Kluwer.

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

Lagerwerf, L. (1998). *Causal connectives have presuppositions. Effects on discourse structure and coherence.* Utrecht, The Netherlands: LOT.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: the Latent Semantic Analysis theory of acquisition, induction and representation of knowledge. *Psychological Review, 104,* 211–240.

Landauer, T. K, Foltz, P. W., & Laham, D. (1998). An introduction to Latent Semantic Analysis. *Discourse Processes, 25,* 259–284.

Langacker, R. W. (1985). Observations and speculations on subjectivity. In J. Haiman (Ed.), *Iconicity in Syntax* (pp. 109–150). Amsterdam: Benjamins.

Mann, W. C., & Thompson, S. A. (1988). Rhetorical structure theory: Toward a functional theory of text organization. *Text, 8*(3), 243–281.

Matthiessen, C. M. I. M., & Bateman, J. A. (1991). *Text Generation and systemic-functional linguistics; experiences from English and Japanese*. London: Pinter.

Pander Maat, H., & Degand, L. (2001). Scaling causal relations and connectives in terms of speaker involvement. *Cognitive linguistics, 12*(3), 211–246.

Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). *Linguistic inquiry and word count dictionary: LIWC 2001*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Pennebaker, J. W., & King, L. A. (1999). Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology, 6,* 1296–1312.

Polanyi, L. (1988). A formal model of the structure of discourse. *Journal of Pragmatics, 12,* 601–638.

Prüst, H., Scha, R., & van den Berg, M. (1994). Discourse grammar and verb phrase anaphora. *Linguistics and Philosophy, 17,* 261–327.

Sanders, T. J. M., Spooren, W. P. M., & Noordman, L. G. M. (1993). Coherence relations in a cognitive theory of discourse representation. *Cognitive Linguistics, 4,* 93–133.

Schiffrin, D. (1987). *Discourse markers*. Cambridge, England: Cambridge University Press.

Tomassello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Cambridge, MA: Harvard University Press.

van Dijk, T. A., & Kintsch, W. (1983). *Strategies of discourse comprehension*. Orlando, FL: Academic.

van Halteren, H., Daelemans, W., Zavrel, J. (2001) Improving accuracy in word class tagging through the combination of machine learning systems. *Computational Linguistics, 27*(2), 199–229.