

Computer-assisted corpus-based text analysis with TATOE

Melina Alexa and Lothar Rostek

GMD-IPSI, Dolivostrasse 15, 64293 Darmstadt, Germany

KEYWORDS: data-driven, semi-automatic text analysis tool

AFFILIATION: Integrated Publication and Information Systems Institute (IPSI) & German National Research Centre for Information Technology (GMD)

E-MAIL: alexa@darmstadt.gmd.de
rostek@darmstadt.gmd.de

FAX NUMBER: +49 6151 869 818

PHONE NUMBER: +49 6151 869 809

1. Introduction

Data-driven and corpus-based text analysis involves considerably variant tasks and aims. For natural language processing (NLP) applications text analysis practice includes tasks ranging from knowledge acquisition for dictionary construction, grammar development, testing and validation, to discourse analysis for automatic text understanding or/and text generation.

Text analysts in computational linguistics form, thus, a group with mixed objectives. However, although their ultimate aims of analysis may differ, initial exploration, modeling and codification of the data and subsequent extraction of knowledge which may be dependent on the already performed categorization and codification are common tasks. Analysis typically covers more than one level of linguistic description whereby information obtained from one level is used in order to gain a deeper understanding of another and contributing, thus, to a deeper understanding of a single piece of discourse.

In corpus-based NLP, text analysts work with corpora whose size may vary from some thousands of words to tens of millions. The texts can be monolingual or bi- or multi-lingual. When more languages than one are considered the corpora can be parallel or non-parallel. Analysis may concentrate on single words, word collocations, word groups (syntactic or semantic units) clauses, sentences, paragraphs, and so on. Furthermore – and as one of the main reasons for adopting a corpus-based approach to analysis – the focus of analysis is the language use in (real) context.

Computational support is required for the processing of machine-readable monolingual or multi-

lingual corpora, the integration of programs for morphological analysis, Part of Speech (PoS) annotation, lemmatization, etc., extracting statistical information concerning frequency of occurrence of word tokens, word types or lemmas and on-text selection of a word or string of words and presentation of all of its occurrences in context. When it comes to working intellectually with the texts, support is further needed for assigning text segments to categories belonging to schemata which are hierarchically or non-hierarchically structured. This necessitates support for (i) importing a schema which is to be used for codification and classification and (ii) support for defining and constructing one's own schema. Also, functionalities are required which enable merging, modifying and deleting categories of the schema (or schemata) used. Texts may already come annotated (SGML, morphological, syntactic, etc.) and means for integrating and re-using this annotation are essential.

Search for certain linguistic cues, frequency of occurrence information, annotation, representation and retrieval of the annotated information are all tasks which need to be performed in text analysis. Currently, fully automated annotation procedures are possible for some linguistic levels of description, for instance morphological or PoS tagging, but not for all. A way to fill the gap between the knowledge which is acquired in a fully automated manner and knowledge which is still required for text analysis is possible by means of semi-automated “mark up”.

2. Text analysis – the work context

Over the last few years we have performed intensive corpus-based text analysis of corpora belonging to different domains, in different languages and, also, with different analysis objectives:

- *sublanguage analysis*: (Alexa, 1993): analysis of English and Greek job advertisement texts for the purposes of multilingual text generation systems
- *genre and register analysis* (Alexa, 1995): analysis of different text types – artists' biographies and archeological site descriptions – typical of encyclopedic reference works in order to specify contextual (text structure) choices which are to guide text type sensitive automatic generation in the Komet-Penman-MultiLingual (KPML) text generation system (Bateman 1996), (Teich and Bateman 1994)
- *semi-automatic information extraction*: analysis of a corpus of English biography texts in order to first identify prominent and text type specific event structures, i.e. main verbs together with the concepts (their participants and the temporal and spatial information),

parse these events and update the knowledge base used for the *Editor's Workbench* application (for the general context of this work see Rostek *et al.*, 1994 and Rostek and Moehr, 1994)

- *domain specific thesaurus construction*: analysis of press release messages issued by the Deutsche Presse Agentur (DPA) for the construction of a semantic network to support fact extraction from these messages
- *thematic development in encyclopedic reference texts*: exploration, determination and subsequent specification of thematic development preferences typical of English artist biographies and archeological site descriptive texts in order to specify the text type specific thematic choices for automatic text generation.

For all the above objectives computational support has been required to structure, compile and edit the text data and to further annotate parts of the text corpora and extract linguistically variant knowledge.

Aiming at supporting such text analysis tasks as the ones listed above we have designed and implemented TATOE, a Text Analysis Tool with Object Encoding, the functionalities of which we present in this paper. Section 3 lists specific requirements for supporting computationally corpus-based text analysis and presents some of the current tools for this task. Section 4 describes TATOE and its main functionalities. Section 5 provides the technical characteristics of the tool. Finally, section 6 presents the future directions of this work and concludes the paper.

3. Requirements for supporting text analysis computationally

The general situation in corpus-based text analysis can be described as one where particular linguistic phenomena observed in the corpus need to be interpreted within a theoretical framework. This interpretation may be an intellectual task typical of scientific work for analysis, understanding and explanation of texts. A tool for semi-automated support analysis, then, needs to provide an integrated and user-friendly working environment whereby at least the following are supported:

- compiling and importing one or more text corpora
- importing an already existing a – hierarchically or non-hierarchically – structured knowledge classification schema
- definition and construction of one's own classification schema
- development or usage of one or more classification schemata for annotation and, thus,

knowledge categorization enabling annotation of text segments (words, groups of words, etc.) according to the schema used

- interactive annotation by mechanisms that are flexible and easy to use on text selections
- enabling work with more than one categorization schemata concurrently, so that the analyst can work separately with different levels of linguistic description, e.g. morphological, lexical, syntactic, etc., and at the same time integrate information from different schemata (for most text analysis tasks integration of different kinds of linguistic knowledge is a requisite)
- the integration of already existing automatic tagging/encoding tools for an initial annotation and enabling work with already annotated corpora, e.g. with SGML, PoS annotation, etc.
- flexibly viewing already encoded text segments; this includes both selecting and arranging according to different criteria (frequency of occurrences, encoded category types, etc.) and presenting information with readable layout styles (fonts, colours)
- concordance list presentations
- calculating different statistics for every word and every annotated text segment and this on the basis of the (hierarchical or non-hierarchical) relations within the categorization schema, e.g. frequency of occurrence of word tokens, word types, schema categories, or demonstration of how these elements are distributed in one or all texts
- providing solutions or mechanisms for exporting both annotated and non-annotated information in different formats, e.g. SGML, for further processing
- multilingual text analysis.

Current available tools and programs which may support the above text analysis requirements can be categorized according to corpus processing tools, general text analysis tools with retrieval capabilities, and annotation tools or programs, i.e. taggers. To give just a flavour of these consider: the DTTool for tagging corpora for anaphora (Aone and Bennett, 1994), the INTEX corpus processing system (Silberztein 1994), the COSMAS corpus storage, maintenance and access system with built-in corpora (al-Wadi, 1994), the WordCruncher, a flexible PC-based text retrieval corpus exploration tool (available from IKS, Bonn), TACT (Bradley, 1990) and Lexa (Hickey, 1992) for general text exploration with presenting information in the form of concordance lists and statistical tables, the knowledge extraction tools KES and GRAAL (as described in (Ogonowski *et*

al. 1994)), the Xtract tool (Smadja and McKeown, 1990) for automatically extracting word collocational information and the WAG-coder (O' Donnell, 1995) for systemic functional linguistics oriented analysis.

The main weakness of the current systems is that no support is offered in an integrated manner where different analysis perspectives can be employed in one working environment and different kinds of information extracted or made available concurrently. Weaknesses usually include one or more of the following: inflexible data model (e.g. the data basis is the whole corpus and the corpus parts such as texts, paragraphs, etc. are not separated), theory-dependence, domain-dependence, applicable only for one level of linguistic analysis, non-linguistically motivated, inability to either import an already existing classification schema or to provide support for consistently and efficiently building one, non-portability, inflexible user interface, lack of multilingual support.

4. Text Analysis Tool with Object Encoding

Based on the variant nature of the work and objectives described in section 2 we have designed TATOE, a tool for a data-driven and multi-layered text analysis. Our intention has been that the tool is language and semantic domain independent, that one or more categorization schema(ta) for encoding can be flexibly developed by the user or imported in the tool, and that the tool is portable between different hardware platforms.

The implementation of TATOE took the requirements listed earlier as its starting point. We designed an object-oriented data model – using the Smalltalk Frame Kit (SFK), an object oriented modeling tool offering a spectrum of features to make model descriptions operational (Fischer and Rostek 1996) – for the representation of the different kinds of information to be stored in a kind of ‘personalized annotation database’. In such a database the text segments are linked to elements of categorization schemata so that the database contents can be seen as a network of information units. We use an object-oriented approach to handle a fine granular network of highly interlinked information units (objects).

A general picture of the tool is given in figure 1.

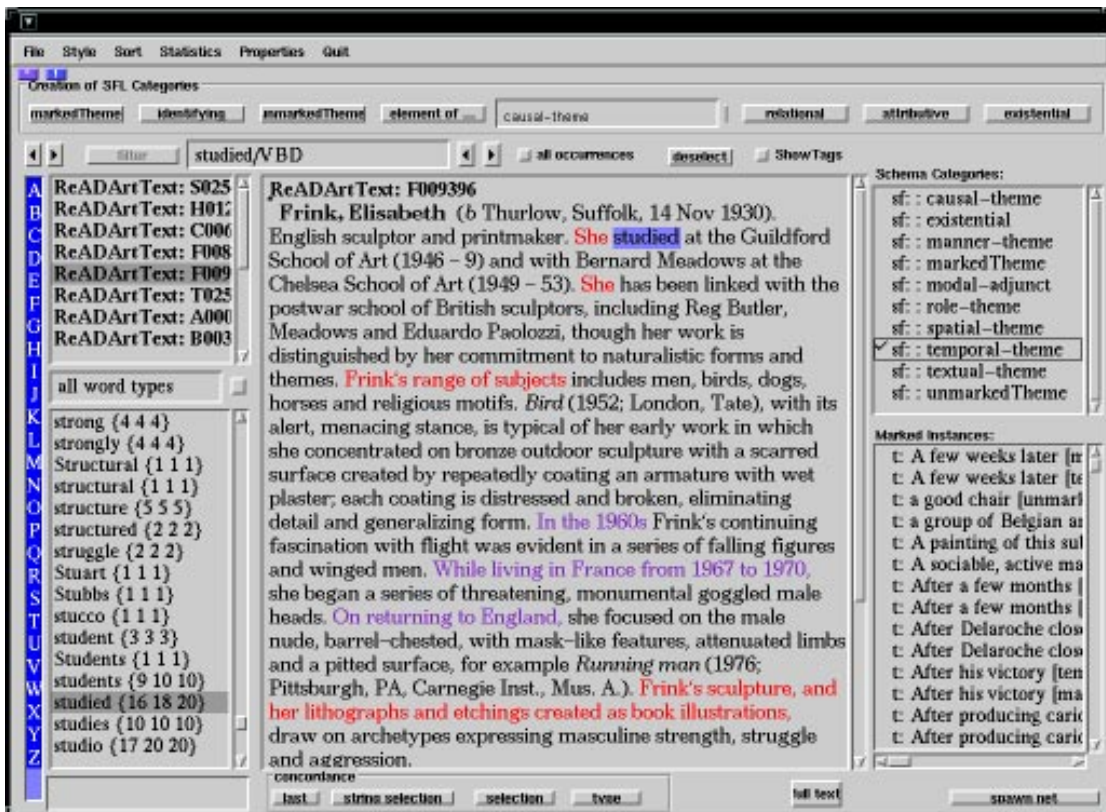


Figure 1: A screen dump of TATOE

The variety of information presented on the screen is arranged in five separate panes in order to keep corpus texts and information about these text(s) side by side. These panes are: the main text pane in the middle, the list of texts at the top left, the word list pane underneath it. On the top right there is the list of categories above the list of marked instances. With the exception of the text list, all the lists contain different information with regards to the words in the corpus, the categories used for annotation and the already annotated instances sorted out in lists.

The main characteristics of TATOE and its functionalities are:

- **Category set up and maintenance:** The categories to be used for the classification are usually defined prior to the beginning of the analysis, depending on the analysis objectives. It is possible to either import an already available categorization schema or for the analysts themselves to define a categorization schema. More than one schemata can be defined and used concurrently. This enables multi-layered text analysis (lexical, grammatical, etc.) A categorization schema in TATOE can be modified by adding, deleting, renaming and merging categories. This holds for schemata which have been developed within TATOE as well as for those imported. For our analysis tasks we have used Brill's rule-based PoS tagger (Brill, 1992) for the

English texts and imported the tagging categorization (together with the annotations) as a separate categorization schema. We have also imported the merged upper model (Henschel and Bateman, 1994) linguistic ontology as an additional classification schema.

- **Searching and selecting:** Searching for a word is done by selecting the appropriate item from a word list with all the corpus word types. Searching for a text is performed by selecting the relevant text from the corpus text list. Selection is performed either by choosing one of the items in the word list or by selecting straight on the text. On text selection is also possible when viewing a concordance lists.
- **Concordance lists:** Together with a 'classic' concordance list there is also the possibility for category concordance lists. Category concordance lists include all the words which have been already assigned to a schema category. Figure 2 shows such a concordance list where all annotated instances for the category 'artist' are presented (in a black and white graph the highlighted instances are shown in white). Rather than using positioning to indicated the selected items in the concordance list in TATOE colour is used. The reason and advantage of this is the possibility to justify the left and right context of the selected item 'on the fly'.

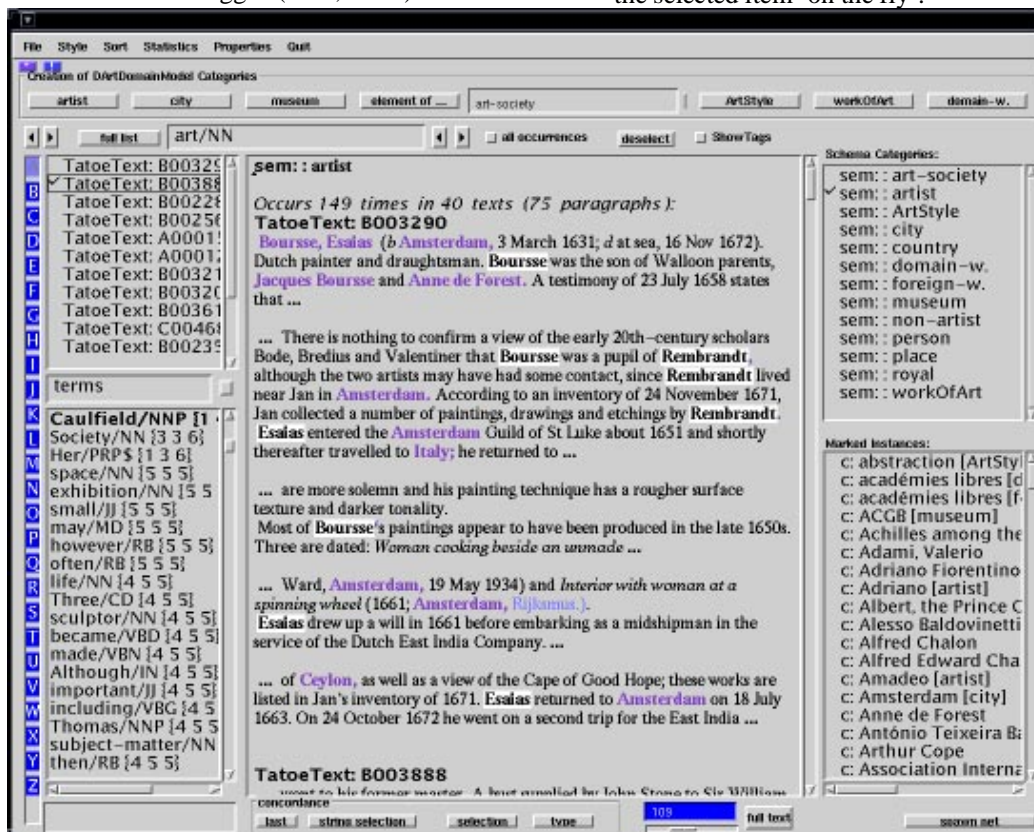


Figure 2: Type concordance list of all instances belonging to a selected schema category

- **Annotating:** annotation is possible for either a single occurrence of a selected item or all its (identical) occurrences in the corpus. Mark up of a selected item is possible either from the full text display or from the concordance display. The marked up items are then highlighted (in colour) so that they are distinguished from the rest of the text items which have not been annotated yet.
- **Qualitative and quantitative information:** upon selection of a marked up instance or a schema category TATOE provides qualitative information, e.g. relations such as broader, part-of, etc. holding between instances or between single schema categories. Selection of an item from the word list or the list of all schema categories provides quantitative information, e.g. frequency of occurrence information or total number of marked instances.
- **Viewing:** Performed, or already existing, annotation can be viewed either in the form of a list of all annotated items or as a list of occurrences in their context, i.e. KWIC lists, or as statistical tables for the already annotated data (with each category together with the total number of tokens or of types assigned to it) or as a graph with all the relations between the categories used represented in a semantic network form. An example of the latter is given in figure 3, where part of the upper model categorization schema is shown graphically.

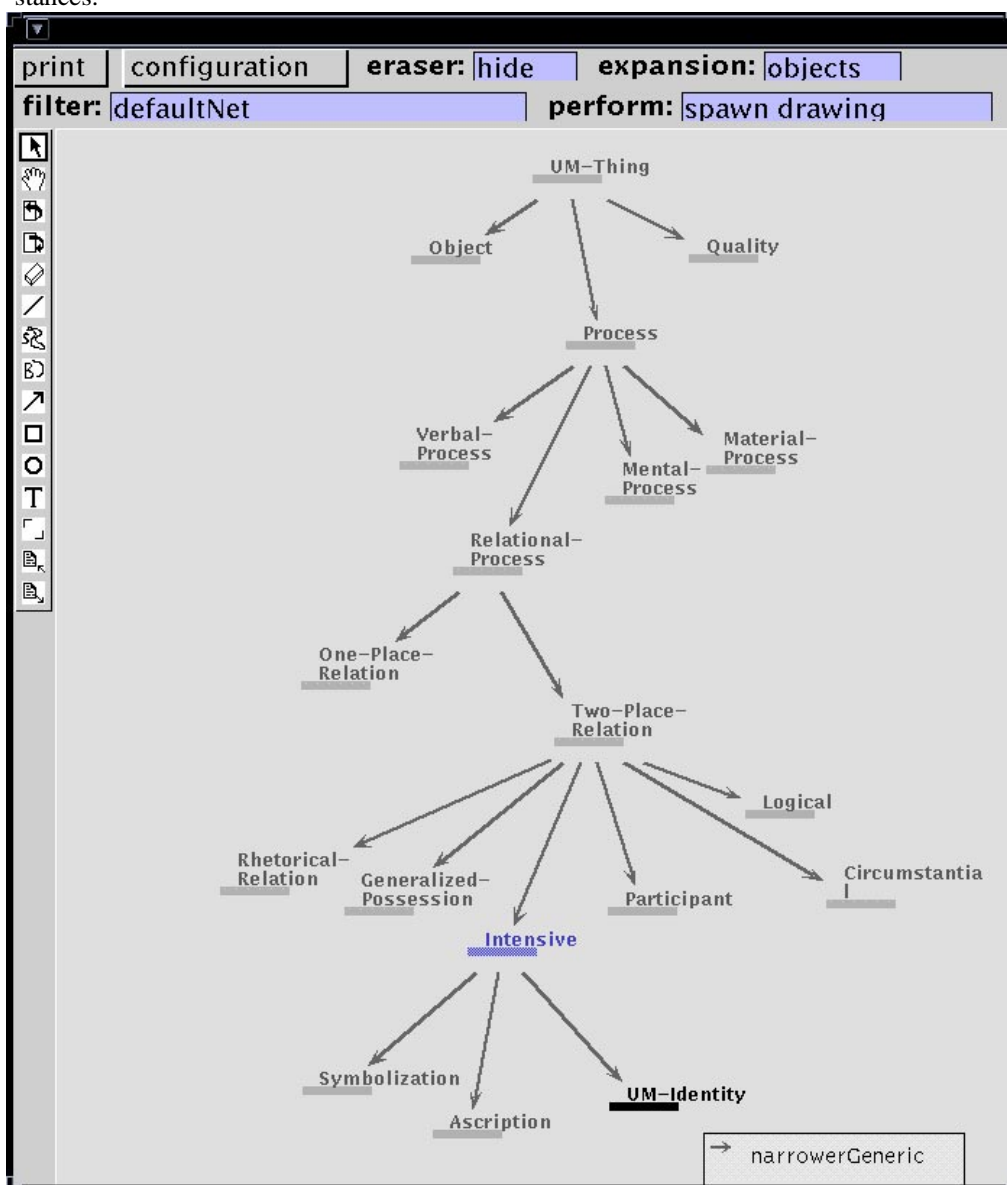


Figure 3: Graphical presentation of categorization schema categories in TATOE

- **Checking and correcting:** In TATOE consistency is checked by selecting a text segment and viewing all its instances in a concordance list. If the intention was that they were all assigned the same category and this has not happened, then the analyst can mark up the remaining instances accordingly. Alternatively, if some of the text segments should have been assigned a different category than the one they are shown to belong to, these can be deselected and then selected again for the category intended. Correction follows the selection and deselection procedure, when the marking up is concerned.
- **Frequency information:** Statistical information is supplied in a number of ways: (i) word frequency lists which are sorted alphabetically, or according to the highest frequency of all occurrences in the whole corpus or to the number of texts or paragraphs one word occurs (note that frequency of occurrence as shown in the word list of figure 1, is sorted according to all occurrences of a word in the whole corpus indicated by the far right number, frequency of occurrence according to the total number of paragraphs a word occurs, indicated by the middle number, and frequency of occurrence according to the total of corpus texts a word occurs, indicated by the far left number), (ii) frequency list of all the categories used for the encoding, (iii) frequency tables of all schema categories (for each schema categorization) together with the total of the text segments encoded for each one of them.

Technical characteristics

TATOE is implemented in the Smalltalk-based programming environment VisualWorks (from ParcPlace). It can be used on various hardware platforms, e.g. Suns, Sparcstations, PCs and Macs.

Future work and conclusions

Currently TATOE is a prototype and its scaling up for handling efficiently very large amounts of textual data is one of the most important tasks. We plan to test its usability for more applications involving corpus-based text analysis. We expect that new applications will result in refining the TATOE's data model to accommodate the specific needs of these applications.

Our intention is to develop a tool which can be used for multilingual corpus-based text analysis and we therefore plan to test its usability for a wide range of languages apart from the ones it has already been used for.

Regarding the functionalities of TATOE, we are working on adding search capabilities of complex

formal patterns for extraction of collocational information in a more flexible way.

At present, we are investigating exporting the annotated information according to Text Encoding Initiative guidelines. This has to take into account that annotated text segments may often overlap.

We have presented a tool for corpus-based text analysis. Such analysis often requires to allow for the possibility to view the data from different perspectives. TATOE enables importing and maintaining and developing different categorization schemata for annotation, and the utilization of the different categorization structures for calculating a variety of cumulative statistical information.

References

- Alexa, Melpomeni (1993): Corpus-based sub-language analysis for a multilingual text generation system. Ph.D. Thesis, CCL, UMIST, Manchester, 1993.
- Alexa, Melina (1995): Making principled selections: a methodology for register analysis and description for text generation. 22nd International Systemic-Functional Congress, Beijing, China, July 1995.
- al-Wadi, Doris (1994): COSMAS Benutzerhandbuch. XII/2785. Institut für deutsche Sprache, Mannheim, Germany, 1994.
- Aone, Chinatsu and Scott W. Bennett (1994): Discourse tagging tool and discourse tagged multilingual corpora. In Proceedings of International Workshop on Sharable Natural Language Resources (SNLR), Nara, Japan, 1994.
- Bateman, John A. (1996): KPML: The KOMET-Penman (Multilingual) Development Environment, Release 0.9. Technical report, Institut für Integrierte Publikations- und Informationssysteme (IPSI), GMD, Darmstadt, March 1996.
- Bradley, John (1990): TACT: User's Guide. Technical report, University of Toronto, 1990. Version 1.2.
- Brill, Eric (1992): A simple rule-based part of speech tagger. Proceedings of the Third Conference on Applied Natural Language Processing, 1992.
- Fischer, Dietrich and Lothar Rostek. (1996): SFK: a Smalltalk Frame Kit. Technical report, GMD/Institut für Integrierte Publikations- und Informationssysteme, Darmstadt, 1996.
- Henschel, Renate and John Bateman (1994): The merged upper model: a linguistic ontology for German and English. Proceedings of 25th International Conference on Computational Linguistics (COLING 94), Kyoto, Japan, August 1994.
- Hickey, Raymond (1992): Lexa, corpus proces-

- ing software. Technical report. Report series: the Norwegian Computing Centre for the Humanities, 1992. Vol. 57, 58, 59.
- O'Donnell, Mick (1995): From corpus to coding: semi-automating the acquisition of linguistic features. Proceedings of the AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation, California, March 27–29 1995. Stanford University.
- Ogonowski, Antoine, Marie Luce Herviou and Eva Dauphin: Tools for extracting and structuring knowledge from texts. Proceedings of the 15th International Conference on Computational Linguistics (COLING 94), vol. II, pp. 1049 – 1053, Kyoto, Japan, 1994.
- Rostek, Lothar and Wiebke Möhr (1994): An editor's workbench for an art history reference work. Proceedings of the ACM European Conference on Hypermedia Technology, Edinburgh, ACM, New York, NY 1994.
- Rostek, Lothar, Wiebke Moehr and Dietrich Fischer (1994): Weaving a Web: the structure and creation of an object network representing an electronic reference network. Proceedings of Electronic Publishing (EP) '94, pp. 495 –506, 1994. Special issue of the International Journal of Electronic publishing – organization, dissemination and design, Volume 6(4).
- Silberztein, Max D. (1994): Intex: a corpus processing system. Proceedings of the 15th International Conference on Computational Linguistics (COLING 94), vol II, pp. 579 – 583, Kyoto, Japan, 1994.
- Smadja, Frank A. and Kathleen M. McKeown (1990): Automatically extracting and representing collocations for language generation. Proceedings of the 28th Conference of the Association for Computational Linguistics, ACL-90, pp. 252 – 259, Pittsburgh, 1990.
- Teich, Elke and John A. Bateman (1994): Towards an application of text generation in an integrated publication system. Proceedings of the Seventh International Workshop on Natural Language Generation, Kennebunkport, Maine, USA, June 21–24, 1994.