Corpus-based register profiling of texts from mechanical engineering

Sabine Bartsch, Richard Eckart**, Monica Holtz*, Elke Teich**
Darmstadt University of Technology,
Department of Linguistics and Literature,
Hochschulstrasse 1, 64289 Darmstadt, Germany,
eMail: {LASTNAME}@linglit.tu-darmstadt.de*,
eckart@rgb.tu-darmstadt.de**
http://www.linglit.tu-darmstadt.de

1. Introduction

Corpora are a helpful source of authentic data for the characterisation of domain specific language. The project described in this paper compiles and annotates a corpus of texts from the domain of data processing in construction, a sub-domain of mechanical engineering. The aim of this project is to develop profiles of the registers of texts from this target domain based on a multi-layer annotated corpus which comprises a variety of registers that are prevalent in that domain such as academic articles, text books, web-based teaching materials and teaching-induced texts (texts written by students). For this purpose, a corpus of English and German texts from the target domain is being compiled and enriched with multiple layers of linguistic annotation at the levels of lexis, grammar, and textual structure. The corpus is annotated in terms of formal and functional categories (part-of-speech, syntactic structure, transitivity, rhetorical structure and generic structure) (see Section 3 below). These annotations are the vantage point for the subsequent register analysis which serves as the basis for the development of profiles of the different text types in the corpus. The theoretical disposition and aims of the register profiling are described in Section 2 below. Section 3 describes the compilation of the corpus and the annotation.

The paper, furthermore, describes the development of an XML-based corpus repository and workbench that allows users to upload and process files in various original formats (pdf, html, doc, txt). The functionality of the PACE-Ling workbench is to include automatic pre-processing and integration of files from varied sources as well as integration and querying of strings and annotations or combinations thereof. The corpus repository and the workbench are described in Section 4 below.

2. Corpus-based register profiling

The theoretical foundations of the project are rooted in Systemic Functional Linguistics (SFL) (Halliday 1985/1994/2004) and register linguistics (Halliday, Mackintosh, Strevens, 1964; Quirk et al., 1985; Biber, 1988; Biber et al., 1998). This theoretical disposition suggests a primarily corpus linguistic methodology. Because the texts under study are instances of different domain-specific registers, they require a linguistic characterisation in terms of register specific features. These features are described according to the tenets laid out by SFL and include the parameters *field*, *tenor* and *mode*.

The parameter *field* characterises texts in terms of their domain-specificity (described in terms of lexis, specialised terminology and collocations, subcategorisation features etc.). The parameter *tenor* characterises texts in terms of the interaction between the participants involved in the interaction (e.g. *expert-to-expert* or *teacher-to-student*). The parameter *mode* refers to the realisation of the communication process in terms of *channel* and *medium*, in the texts under study this is for example indirect, non-face-to-face communication (*channel*) and the texts are written to be read (other possibilities are e.g. written to be spoken, written to be read) (*medium*). Whereas the texts under study are relatively uniform in terms of *mode* features, i.e. all of the texts are instances of indirect face-to-face communication and are written to be read, there is some variation in the corpus in terms of *tenor*, i.e. there are texts instantiating expert-to-expert, teacher-to-student and student-to-teacher interaction. The most interesting variation, however, is to be expected in terms of *field* reflecting e.g. variation in terms of the level of domain-specific terminology and collocations due to differing expert levels of the interactants. It can, for example be expected that variation is to be found in terms of level of domain-specificity in lexis and domain-specific collocations in academic articles, teaching materials and teaching induced text as produced by students in a teaching context.

One of the reflexes of the parameter *field* is found in the transitivity of linguistic structures which is one aspect of the ideational metafunction (Halliday 2004). Transitivity describes the fact that experience is construed as a set of different process types with different participants involved and which are associated with different types of circumstances. There are principally six process types to be distinguished in the transitivity analysis: *material processes* describing actual physical actions (1), *mental processes* describing the inner, mental experience (2), and *relational processes* which are processes of identification and classification (3). Material, mental and relational processes are the most central processes, according to Halliday (2004: 171), the other three types are on the borderlines between these three processes respectively. *Behavioural processes* represent "outer manifestations of inner workings, the acting out of processes of consciousness" (4) (ibid.) and are thus on the borderline between material and mental processes. *Verbal processes* are processes of saying and meaning and establish symbolic relations enacted linguistically (5); verbal processes are thus on the borderline between mental and relational processes. *Existential processes* are processes concerned with existence in which phenomena are recognized 'to be' (6). These are on the borderline between relational and material. Examples of these process types are given in the examples (1) – (6) below (processes in **bold**, process + participant underlined):

(1) material: (Technologies) that can automatically **construct** physical models from Computer-Aided Design (CAD) data.

(2) mental: … which is also **known** as the solider process.

(3) relational: The term rapid prototyping (RP) **refers to** a class of technologies ….

(4) behavioural: Prototypes are also useful for testing a design, to see if it **performs** as desired or needs improvement.

(5) verbal: … the same **could be said** for colour laser printing …

(6) existential: … several prototyping techniques **exist**.

By distinguishing the different process types and their distribution in the text as a whole as well as their sequence and clustering in different subparts of the text we find for example that the introductory section often displays a clustering of relational processes. This is due to the fact that

typically terminology and concepts are introduced and defined in this section. In contrast, the main expository sections of a text often display a clustering of material processes which e.g. describe procedures involved in the CAD construction process or the processes carried out by a particular type of system. The following short section shows a clustering of material processes in a text describing different types of rapid prototyping systems:

> Patented in 1986, stereolithography **started** the rapid prototyping revolution. The technique **builds** three-dimensional models from liquid photosensitive polymers that **solidify** when **exposed** to ultraviolet light. As shown in the figure below, the model is **built** upon a platform situated just below the surface in a vat of liquid epoxy or acrylate resin. A low-power highly focused UV laser **traces** out the first layer, **solidifying** the model's cross section while **leaving** excess areas liquid.
> (Source: Palm, W. 05.1998. Rapid Prototyping Primer. Penn State Learning Factory)

An analysis of the textual structure complements the analysis of ideational features in order to shed light on the organisation of information in a text. The analysis of textual structure according to the tenets of SFL covers thematic structure, cohesion and generic structure. Thematic structure, a reflex of *mode*, focuses on the way information is organised in the clause. According to Halliday (2004: 64) the thematic structure gives the clause its character as a message. *Theme* is signalled in a variety of ways in different languages. In English and German, *theme* is signalled by its position at the beginning of the clause; it combines with the *rheme* to constitute the message. The prominent function of the *theme* stems from its function as the point of departure of the message. Different types of experiential elements, i.e. the process itself, participants in the process, or circumstantial factors of time, manner and cause, can fulfil the function of *theme* as is illustrated in examples (7) and (8).

(7) Topical process NP: **Developed by Carl Deckard for his master's thesis at the University of Texas, selective laser sintering** was patented in 1989.

(8) Imperative topical process: "The steps are:
1. **Create** a CAD model of the design
2. **Convert** the CAD model to STL format
3. **Slice** the STL file into thin cross-sectional layers
4. **Construct** the model one layer atop another
5. **Clean and finish** the model"

The corpus is, furthermore, annotated for coherence features such as *anaphora*, *cohesion* (Halliday, Hasan 1976) and *collocations*. *Anaphoric relations*, i.e. coreference relations between an expression and its antecedents, contribute importantly to the way in which elements of a text "hang together", thereby ensuring that the text is perceived as a coherent whole. Anaphora therefore plays an important role in the characterisation of text. The annotation is carried out automatically and is described in Section 3.2 below.

The annotation also includes *spatial-temporal relations*. The text extract below shows an example of temporal conjunctive relations to illustrate their contribution to the overall coherence of the text:

> **First**, photosensitive resin is sprayed on the build platform. **Next**, the machine develops a photomask (like a stencil) of the layer to be built. […] The mask is **then** exposed to UV light, which only passes through the transparent portions of the mask to selectively harden the shape of the current layer.

In this example, temporal expressions mark the temporal sequence of a process thereby guiding the reader through the sequential steps of a procedure. This is of special relevance to the texts in an engineering domain which frequently describe procedures as part of e.g. construction processes.

Generic Structure Potential (GSP) (Hasan 1977) builds on the different genre specific configurations of *field*, *tenor* and *mode* features to result in a description of a generic type of structure that is characteristic of a particular register such as e.g. an academic article. The Generic Structure description identifies different process type configurations as well as *tenor* and *mode* features unfolding in the development of different sections of a text (e.g. the introductory section comprises predominantly *relational* processes, e.g. in definitions; the explanatory section explains typical actions and processes predominantly comprising *material* processes). The aim is to develop a register specific profile according to those features which allow it to be identified as a text as belonging to a generic type of genre such as academic article, textbook, news report etc.
Rhetorical Structure Theory (RST) (Mann, Thompson 1987) is applied in the corpus analysis in order to investigate hierarchical discourse structures which contribute to the coherence of a text. RST provides an inventory of relations which are applied to non-overlapping units of the text, typically units at clause or sentence level. These rhetorical relations describe the way in which a *nucleus*, the central proposition, and its *satellites*, those units providing information referring to the *nucleus*, e.g. elaboration, background, antithesis, concession, restatement etc., relate to one another thereby creating a coherent whole. A Rhetorical Structure analysis assumes that a text can be organised – ideally – into a complete hierarchical structure, based on a so-called schema. Figure 1 below shows an example of an RST analysis.

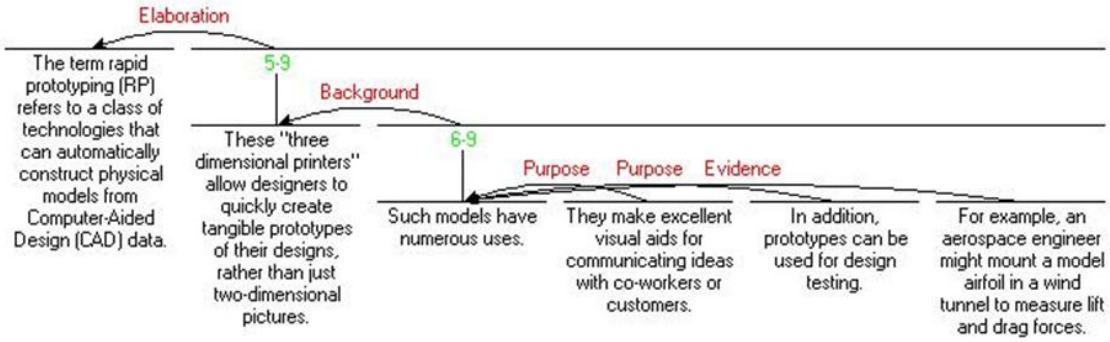


Figure 1: Example RST analysis

It can thus be shown how different elements of a text relate to one another in terms whether an element subsequent to the *nucleus* of a proposition is an elaboration of the information expressed by the *nucleus* or whether it states a purpose or gives evidence in order to corroborate a statement made by the proposition of the *nucleus*. This type of annotation allows for a fine-grained characterisation of the distribution and organisation of different types of information in a

text and allows the development of profiles of the rhetorical organisation of a text. Differences in rhetorical structure between texts with different tenor dispositions can thus be observed. Thus, texts written for *teacher-to-student* communication such as teaching materials or text books display a denser distribution of *purpose* and *evidence* as well as *restatement* relations which serve the purpose of exemplifying or restating the information presented in a teaching context. In contrast, *expert-to-expert* communications such as in academic articles, which presuppose expert knowledge on the part of the reader, may be expected to supply less explanatory information, but will e.g. introduce *antitheses* as part of the discussion of different academic stances in the discussion of the state of the art.

Based on this theoretical disposition, the corpus is annotated at multiple levels and with features based on the outlined parameters and the general theoretical disposition. The following section describes the multi-layer annotation of the corpus from a conceptual perspective, while Section 4 explains the technical issues concerning the data model and the possible queries which form the basis for register profiling.

3. Multi-level corpus annotation for register profiling

3.1 Building a corpus for register profiling

The text archive of the present project currently comprises 1.25 million running words of texts in English and German (500.000 running words of English, 750.000 of German) and is still in the process of being expanded. The text archive comprises the following text types:

- academic articles
- teaching materials
- text books
- texts produced by students

The texts are supplemented by XML header information according to the standards established by the TEI (Sperberg-McQueen, Burnard 2004). The headers contain information comprising bibliographic and source information as well as a characterisation by means of the above mentioned systemic-functional linguistic parameters *field*, *tenor* and *mode*. The parameter *field* in the header characterises texts in terms of their domain-specificity by means of keywords characterising a text, the parameter *tenor* characterises texts in terms of the participants involved in the interaction, in the case of the texts comprised in the corpus *expert-to-expert*, *teacher-to-student* and *student-to-teacher*. The parameter *mode* refers to the realisation of the communication process in terms of *channel* and *medium*. Figure 1 shows a section from a typical header with *field, tenor, mode* information:

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<teiHeader>
  <fileDesc>
       <titleStmt>
         <title>Direct Fabrication of Polymer Composite
                Structures with Curved LOM</title>
           <author>Donald Klosterman</author>
           <author>Richard Chartoff</author>
           <author>Mukesh Agarwala</author>
           <author>Ira Fiscus</author>
           <author>John Murphy</author>
           <author>Sean Cullen</author>
           <author>Mark Yeazell</author>
       </titleStmt>
       <editionStmt>
         <edition>
           <date>Friday, November 12, 1999 12:03:31 PM</date>
         </edition>
       </editionStmt>
       <publicationStmt>
         <publisher> University of Texas at Austin <publisher/>
           <pubPlace>Austin, Texas<pubPlace/>
           <availability>online<availability/>
       </publicationStmt>
  </fileDesc>
  <profileDesc>
    <textClass>
      <keywords scheme="PACE.field.experientialDomain">
        <list>
          <item>mechanical engineering</item>
          <item>rapid prototyping</item>
          <item>polymer matrix composites</item>
          <item>laminated object manufacturing process</item>
          <item>hardware</item>
          <item>software</item>
        </list>
      </keywords>
      <keywords scheme="PACE.field.goalOrientation">
        <list/>
      </keywords>
      <keywords scheme="PACE.field.socialActivity">
        <list/>
      </keywords>
      <keywords scheme="PACE.tenor.agentiveRoles">
        <list>
          <item>expert-expert</item>
        </list>
      </keywords>
      <keywords scheme="PACE.mode.channel">
        <list>
          <item>non-face-to-face interaction</item>
        </list>
      </keywords>
      <keywords scheme="PACE.mode.medium">
        <list>
          <item>written-to-be-read</item>
        </list>
      </keywords>
    </textClass>
  </profileDesc>
</teiHeader>
```

Figure 2: Header from PACE text archive

Header information of this kind is required in order to compile register specific corpora which are a subset of the texts comprised in the text archive according to the respective parameters in order to be able to draw register specific distinctions between the different text types.

3.2 Multi-layer corpus annotation

The corpus is annotated with multiple levels of annotation by means of formal and functional categories at word, clause and phrase level as well as larger scale textual features such as generic structure and rhetorical structure.

The first set of automatically generated annotations are annotations at the levels of part-of-speech and syntactic phrase structure. For this purpose, the texts have to be tokenized in the first step because the PoS tagger requires tokenized, one-word per line text as input. The PoS annotation is carried out by means of the automatic PoS tagger Trigrams 'n' Tags[1] (TnT) (Brants 2000). In the next step, the texts are parsed syntactically by means of an automatic parser such as the Charniak parser[2]. A variety of processing steps have to be carried out on the output of PoS tagger and parser in order to obtain a native XML format that can be integrated into the multi-layer annotation. This process and the associated issues are described in Section 4 below. Anaphora resolution is carried out by the automatic tool Guitar 1.1[3] (A General Tool for Anaphora Resolution), a Java-based tool that integrates scripts for the required text-to-XML conversion as well as a chunker, LT-Chunk[4]. LTChunk is part of LT-XML suite developed at the University of Edinburgh and has to be obtained separately, but integrates with Guitar to produce the chunked input format required by the actual anaphora resolution. The Guitar toolset also comprises a syntactic pre-processor as well as the actual anaphora resolution procedure.

All other annotation steps have to be carried out manually because there are, as yet, no tools that allow the automatic or even semi-automatic annotation of transitivity features or rhetorical structure annotations. These annotation levels, therefore, have to be built up manually.

The annotation of transitivity features comprises the identification and annotation of the process types introduced in Section 2 above, which, in the case of the registers under study, are predominantly *material* and *relational* as well as – to a considerably lesser degree – *existential*, *verbal*, and *mental* processes, *behavioural* processes very are rare in the register under study. The different process types are further annotated for their respective subtypes, e.g. *material* processes may be *creative* or *transformative* in terms of the *Type of doing* and *transitive* or *intransitive* in terms of *Impact*. *Relational* processes are further subdivided according to the so-called *Mode of Relation* (*attributive* or *identifying*) and *intensive*, *possessive* or *circumstantial* in terms of the *Type of relation*.

Instantiations of these process types in the corpus are identified and annotated manually. The annotation can be aided by the tool Systemic Coder (O'Donnell 1995) which allows the specification of annotation schemes and provides XML output, or coded directly in an XML-editor such as Altova's XMLSpy©. The latter procedure was actually adopted in preference to Systemic Coder in the course of the annotation process as it proved to be more practical in view of the required XML output format. The encoding in XML is crucial for the implementation of this annotation layer in the overall multi-layer annotation as all text annotation storage and organisation as well as corpus-processing is XML-based as described in Section 4 below. A

[1] TnT – Statistical Part-of-Speech Tagging http://www.coli.uni-saarland.de/~thorsten/tnt/
[2] Eugene Charniak's homepage http://www.cs.brown.edu/people/ec/
[3] Guitar - A General Tool for Anaphora Resolution http://privatewww.essex.ac.uk/~malexa/GuiTAR/
[4] LTG Software http://www.ltg.ed.ac.uk/software/chunk/index.html

progressively assembled verb lexicon, which identifies each verb with the processes in which it commonly occurs as well as the typical participants and circumstantial elements accompanying them, is currently under development. This lexicon is being assembled to aid annotators in the decision process and ensure better inter-annotator agreement. The annotation of process types is based on an SFL style annotation scheme as exemplified in Figure 3:

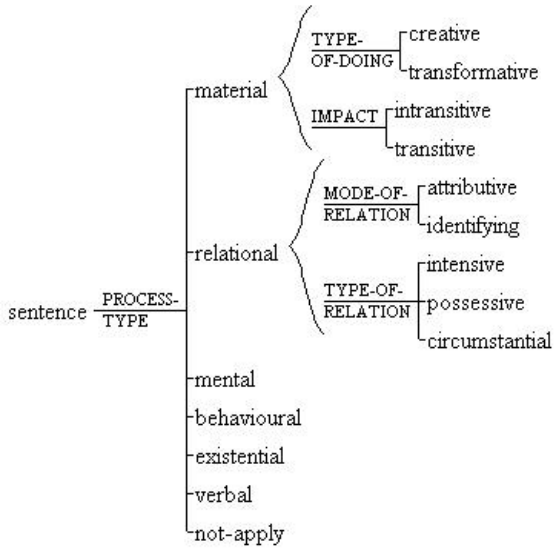


Figure 3: Annotation scheme for process type annotation

Examples (9) – (16) below show some segments of the process type annotation:

Examples of annotated relational processes:

(9) **The term rapid prototyping (RP) refers to a class of technologies**

(10) **the techniques are often collectively referred to as solid free-form fabrication, computer automated manufacturing, or layered manufacturing.**

(11) **Rapid prototyping is an "additive" process,**

(12) **In contrast, most machining processes (milling, drilling, grinding, etc.) are "subtractive" processes**

Examples of material processes:

(13) The term rapid prototyping (RP) refers to a class of technologies **that can automatically construct physical models from Computer-Aided Design (CAD) data.**

(14) **In addition to prototypes, RP techniques can also be used to make tooling (referred to as rapid tooling) and even production-quality parts (rapid manufacturing).**

(15) **A software package "slices" the CAD model into a number of thin (~0.1 mm) layers,**

(16) **which are then built up one atop another.**

The annotation for transitivity features furthermore includes the annotation of circumstantial elements where appropriate.

In order to enable an analysis of larger scale textual structures, the corpus is annotated for Rhetorical Structure. The rhetorical structure annotation is carried out manually by means of RSTTool[5] (O'Donnell 1997; 2000), a tool that allows the implementation of coding schemes for rhetorical structure analysis as well as the actual text annotation. Rhetorical structure may be annotated at both clause and sentence level. Examples (17) – (23) show a section of the rhetorical structure annotation.

(17) **The term rapid prototyping (RP) refers to a class of technologies**

(18) **that can automatically construct physical models from Computer-Aided Design (CAD) data.**

(19) **These "three dimensional printers" allow designers to quickly create tangible prototypes of their designs, rather than just two-dimensional pictures.**

(20) **Such models have numerous uses.**

(21) **They make excellent visual aids for communicating ideas with co-workers or customers.**

(22) **In addition, prototypes can be used for design testing.**

(23) **For example, an aerospace engineer might mount a model airfoil in a wind tunnel to measure lift and drag forces.**

Rhetorical structure annotation allows statements about the hierarchical organisation of a text in terms of the rhetorical relations between text segments at clause and sentence level. It also enables quantitative statements concerning the frequency and clustering of different rhetorical relations in the text as a whole, or different sections of a text, thereby allowing for a fine-grained

[5] RST Homepage http://www.sfu.ca/rst/

characterisation of the discourse structure. RSTTool allows the storage of the annotated output in XML format which is important for the later integration of this annotation into the multi-layer annotation.

4. The PACE-ling suite of corpus tools

The text archive is enhanced for the purpose of linguistic analysis with the annotations described above. This process entails the conversion of the diverse output file formats to native XML as well as the integration of the corpus with *stand-off* annotations on different linguistic levels. These annotations will later be exploited to compile corpora and pose complex linguistic queries over strings and multiple annotation layers which will allow more global statements about recurrent characteristic features of the texts in the corpus.

The integration of multiple layers of annotation of a corpus with a diverse set of input formats and annotations produced by different tools with diverse output formats raises a number of issues that need to be addressed in turn.

The first issue arises from the fact that the text archive is comprised of texts in a variety of original file formats such as html, pdf, word processing software formats and plain text. While the original formats need to be retained for later reference concerning questions of layout etc., the actual linguistic processing cannot be carried out directly on those file formats as linguistic annotation software requires plain text as input. The first requirement is thus a process whereby new texts can be converted from their respective formats into plain text. A second set of issues arises from the fact that a number of different linguistic pre-processing steps need to be carried out as different annotation tools require their own specific input formats (e.g. tokenised input streams). The third set of issues concerns the diversity of the annotation tools employed in the project. Ideally, all tools should be useable within a single integrated processing framework in order to allow users to invoke or carry out manually all annotation steps required in the annotation procedure and that the workbench may be used on all platforms. The problem with this last requirement is the diversity of the tools: some, like the PoS tagger and the Charniak parser, are command line tools under Windows and UNIX, others, like RSTTool, are GUI based Windows programs, while the anaphora resolution suite Guitar is a platform-independent Java-based command-line tool. The tools are thus not integrateable in a straightforward way. A fourth issue are the diverse output formats produced by the tools. Some produce plain text output, but in different formats (e.g. line-by-line double column plain text as in the case of the PoS tagger), while others produce their own XML output format. These output formats have to be transformed into a uniform XML model by means of standard XML-processing such as XSLT in order to allow integration into a multi-layer stand-off annotation which can be processed and queried by a uniform set of tools. For these reasons we have adopted XML as the standard text format for the corpus repository. This, of course, means that all output eventually has to be converted into a dedicated native XML format.

In order to enable an integrated workflow in the processing of the text archive, a data base application is being developed that allows users to add new texts to the repository and automatically includes all processing steps necessary to convert files to any format that any of the linguistic processors require as input and to convert any type of output into XML for storage

and querying in the repository and workbench. The following automatic processing tools are being integrated into the application:

- Tokenizer: dedicated
- Part-of-Speech tagger: TnT
- Chunker: LTChunk
- Syntactic parser: Charniak parser
- Anaphora resolution including tokenisation, chunking and syntactic pre-processing necessary for anaphora resolution: Guitar 1.1

The manual annotations concerning the transitivity analyses and the Rhetorical Structure annotation have to be carried out with the help of external tools such as Mick O'Donnell's RSTTool and by means of an XML editor. The corpus repository enables the integration of these layers of annotation into a multi-layer stand-off annotation (the data model and procedures are described in more detail in Teich et al. 2001; Teich et al. 2005).

As a last and crucial step, the application is to be completed by a workbench that allows queries over strings and annotations and displays the results in appropriate formats. Thus, the user can e.g. create word lists or pose simple and complex queries over strings and annotations. The results can be displayed in standard linguistic formats such as e.g. KWIC (key-word in context) concordances (see Figure 4) below.

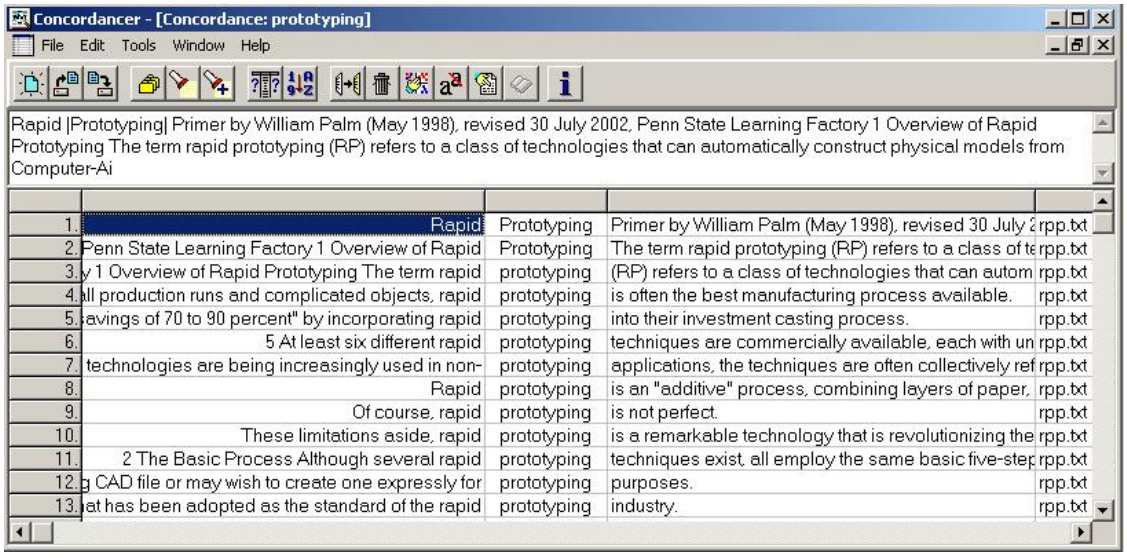
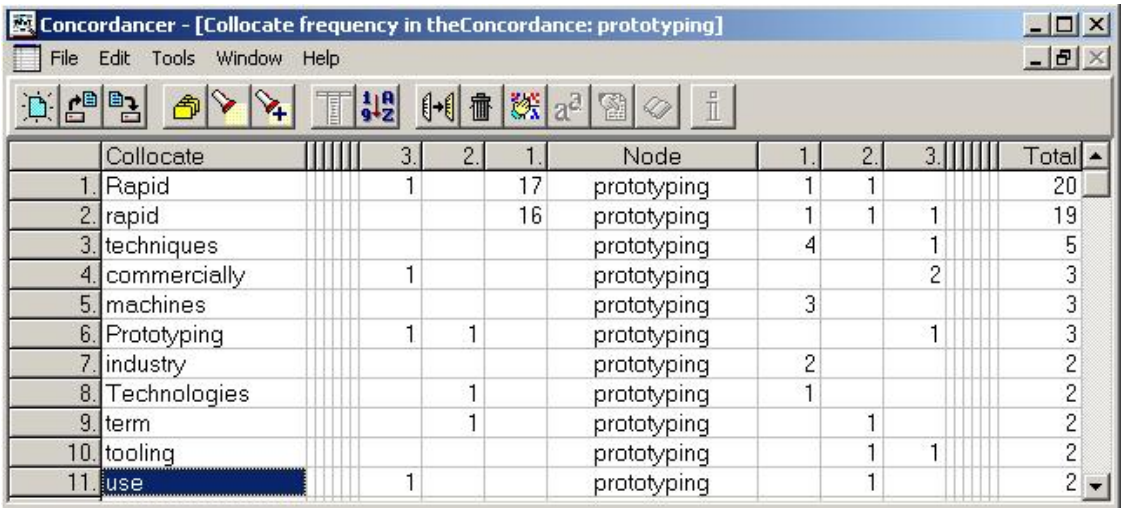


Figure 4: KWIC Concordance for the node „prototyping"

The functionality also includes collocation frequency analyses which allow the identification and extraction of collocations based on statistical measures (Bartsch 2003). An example of the format is shown in Figure 5 below:



	Collocate	3.	2.	1.	Node	1.	2.	3.	Total
1.	Rapid	1		17	prototyping	1	1		20
2.	rapid			16	prototyping	1	1	1	19
3.	techniques				prototyping	4		1	5
4.	commercially	1			prototyping			2	3
5.	machines				prototyping	3			3
6.	Prototyping	1	1		prototyping			1	3
7.	industry				prototyping	2			2
8.	Technologies		1		prototyping	1			2
9.	term		1		prototyping		1		2
10.	tooling				prototyping		1	1	2
11.	use	1			prototyping		1		2

Figure 5: Collocation table for the node „prototyping“

The workbench also allows the extraction of statistical profiles on the distribution of different combinations of grammatical and textual features from the stand-off annotation. It enables, e.g. the extraction of statistical information on the relative frequencies of the different process types in combination with the phrase types instantiating them for each individual text as well as for a register specific corpus.

The workbench will thus aid the linguist in all pre-processing steps required for the import of files in diverse file formats and prepare the plain text input for processing by the different annotation tools. The intention underlying the development of this workbench is to provide the user with an integrated set of tools for the processing of newly imported files and to ensure that all tools receive the required input with minimal manual intervention by the user. It will also ensure that all output from the linguistic annotation tools is converted to an XML output that can be automatically integrated into a multi-layer stand-off annotation.

6. Conclusions

The combination of SFL and register theory with a corpus-based methodology offers a promising approach to register profiling. What is lacking so far, however, are corpora annotated for the relevant linguistic features in order to enable register analysis at multiple integrated levels. The project within which the work reported is situated attempts to compile and annotate such corpora in order to enable register profiling of texts in the domain of data processing in construction as well as other domains in future projects.

References

Bartsch, S. (2003) *Structural and Functional Properties of Collocations in English. A corpus study of lexical and pragmatic constraints on lexical co-occurrence*. (Tübingen: Narr).

Biber, D. (1988) *Variation across speech and writing.* (Cambridge: Cambridge University Press).

Biber, D. (1995) *Dimensions of register variation: A cross-linguistic comparison.* (Cambridge: Cambridge University Press).

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

Brants, T. (2000) "TnT - A Statistical Part-of-Speech Tagger." In: *Proceedings of the Sixth Applied Natural Language Processing Conference ANLP-2000*, Seattle, WA.

Charniak, Eugene. 2000. “A Maximum-Entropy-Inspired Parser”. In: Proceedings of NAACL-2000.

Christ, O. (1994) "A modular and flexible architecture for an integrated corpus query system". *COMPLEX'94*, Budapest, 1994.

Guitar 1.1 (2004) Homepage: http://privatewww.essex.ac.uk/~malexa/GuiTAR/

Halliday, M.A.K., McIntosh, A. and Strevens, P. (1964) *The Linguistic Sciences and Language Teaching. London*. (Longmans, Green and Co.)

Halliday, MAK, R. Hasan (1976) *Cohesion in English*. (London: Longman)

Halliday, MAK (2004) *An introduction to functional grammar*. (Third edition revised by C. Matthiessen) (London: Arnold).

Hasan, R. (1977) “Text in the Systemic-Functional Model”. In: Dressler, W. ed. *Current Trends in Textlinguistics*. (Berlin: Walter de Gruyter) 228 – 246.

Mann, W.C., S.A. Thompson (1987) ‘Rhetorical structure theory of text organisation.’ From: Polanyi, Livia. ed. *The Structure of Discourse.* (Norwood, N.J.: Ablex Publishing Corporation)

Mann, William C. (2005) *RST Homepage*. http://www.sfu.ca/rst/

Matthiessen, C. (1999) The system of TRANSITIVITY: an exploratory study of text-based profiles. *Functions of Language 6.1*.

O'Donnell, M. (1995) "From Corpus to Codings: Semi-Automating the Acquisition of Linguistic Features". In: Proceedings of the AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation, Stanford University, California, March 27 – 29.

O'Donnell, M. (1997) „RST-Tool: An RST Analysis Tool.“ In: Proceedings of the 6th European Workshop on Natural Language Generation March 24 – 26, 1997 Gerhard-Mercator University, Duisburg, Germany.

O'Donnell, M. (2000) "RSTTool 2.4 –A Markup Tool for Rhetorical Structure Theory". In: Proceedings of the International Natural Language Generation Conference (INLG'2000), 13–16 June 2000, Mitzpe Ramon, Israel. 253 –256.

O'Donnell, M. (2002) *Systemic Coder - a Text Markup Tool. Version 4.5*. Available online from: http://www.wagsoft.com/Coder/

Palm, W. 1998. Rapid Prototyping Primer. Penn State Learning Factory.

Quirk, Randolph, S. Greenbaum, G. Leech, and J. Svartvik (1985) *A comprehensive grammar of the English language*. (London: Longman)

Teich, E., S. Hansen, and P. Fankhauser (2001) “Representing and querying multi-layer corpora”. In *Proceedings of the IRCS Workshop on Linguistic Databases*, pages 228–237, University of Pennsylvania, Philadelphia, 11-13 December.

Teich, E., P. Fankhauser, R. Eckart, S. Bartsch, M. Holtz (2005) “Representing SFL-annotated corpus resources”. Paper presented at *Computational Systemic Functional Workshop*, University of Sydney, Australia, July 2005.

Sperberg-McQueen, C.M., L. Burnard (2004) *TEI P4, Guidelines for Electronic Text Encoding and Interchange. XML-compatible edition*. Text Encoding Initiative (TEI), URL: www.tei-c.org/.