Preprint: to appear in: Dieter Metzing und Andreas Witt (eds.): Linguistic Modelling of Information and Markup Languages. Dordrecht: Springer.

Angelika Storrer

MARK-UP DRIVEN STRATEGIES FOR TEXT-TO-HYPERTEXT CONVERSION

1. INTRODUCTION

Hypertext technology is not only used to build hypertext applications from scratch. It may just as well be used to transform existing material into a format that can be processed by a hypertext system. In this context of text-to-hypertext conversion one is confronted with three types of conversion issues:

- *Segmentation issues*: What are the criteria for segmenting documents into hypertext units (henceforth called "hypertext nodes")?
- *Linking issues*: What are the guidelines and principles for reconnecting these nodes via hyperlinks?
- *Reorganization issues*: What kinds of transformations are necessary to unchain text segments from their linkage to the reading path of the sequential document, so that they may be integrated into different user-selected pathways?

Early conversion approaches concentrated on text types that naturally profit from the linking and searching capabilities of hypertext: dictionaries and other reference works. For the conversion of such text types, reorganization issues are less important. Documents of this type are commonly composed of text blocks, e.g. dictionary articles, which are designed as "stand-alone" units that may be consulted selectively and in arbitrary sequence. In contrast, sequential text types like text books, scientific papers or monographs are designed to be read completely and in the sequence presented by the author. When these documents are transformed into hypertext nodes, they may still contain explicit and implicit cohesive markers (anaphoric expressions, connectives, text-deictic expressions) related to units of the preceding or subsequent text. Conversion approaches for such text types are thus naturally confronted with reorganization issues. This paper discusses conversion strategies that use markup on several annotation layers for the segmentation, linking and reorganization of sequentially organized document types.

Conversion approaches usually transform the structure of the sequentially organized documents into a new hypertext structure. The approach described in this paper does not intend to irreversibly convert sequential documents into hypertext networks. Instead, it implements a flexible set of segmentation, linking and reorganization rules which automatically generate hypertext views as additional layers while preserving the original sequence and content of the sequential documents. These rules process information of mark-up at different annotation layers in order to segment the documents into hypertext nodes, achieve their cohesive closedness and establish hyperlinks. The approach was developed and evaluated using a corpus of German scientific texts coming from two domains, namely text technology and hypertext research. The semantics of technical terms in these domains were represented in a WordNet-style semantic network. This network is used as a basis for generating glossary views that are linked to the term occurrences in the corpus.

The approach has been developed in the framework of the project HyTex[1]. This paper describes the basic concepts, guidelines and strategies that are substantial for our segmentation, linking and reorganization rules. The article by Lenz (in this volume) discusses implementation issues and presents a specialized hypertext transformation language that she has developed in this framework.

[i] The acronym HyTex is spelled out "Hypertextualisierung auf textgrammatischer Basis" (= Hypertextualization on a text-grammatical basis; cf. http://www.hytex.info). The project is funded by the German Science Foundation DFG in the framework of the research group "Texttechnologische Informationsmodellierung" (= text-technological information modelling; cf. www.text-technology.de).

2. USER SCENARIO AND CONVERSION GUIDELINES

In order to simplify a later evaluation, our conversion approach is developed with the following usage scenario in mind: hypertext users are in search of information in a scientific domain in which they have previous but no expert knowledge. Their time is constrained, and they have to solve a specific type of problem. Such a scenario may occur in the course of an interdisciplinary research project, in scientific journalism and specialized lexicography. In these contexts users tend to read excursively and only perceive parts of longer documents. When these documents are sequentially organized, i.e. designed to be read from the beginning to the end, this selective reading may result in coherence problems. For example, a reader, jumping right in the middle of a sequential document, may not understand (or may misunderstand) a paragraph because he lacks the prerequisite knowledge given in the preceding text. The objective of our conversion approach is to avoid such coherence problems and make selective reading and browsing more efficient and more convenient than it would be possible with printmedia. To accomplish this objective our approach follows two guidelines, namely 1) recoverability and 2) coherence-based conversion.

ad 1) By *recoverability* we mean that we generate hypertext views as additional layers while preserving the original sequence and content of the sequential documents. In this way, the reader still has the option to perceive the text in its original sequential form, provided he has the time to do so. The hypertext views mark an offer for those readers who only have the time to scan the text. Our goal is to offer this sort of reader a better support in text understanding than it would be possible while reading printmedia excursively.

ad 2) *Coherence-based conversion* means that the way in which the documents are split up into nodes and linked to other nodes is governed by the concept of coherence. The guideline was introduced in Kuhlen (1991)[2] as an alternative to purely form-based conversion approaches. Below, I want to outline the main differences between form- and

[2] Cf. Kuhlen (1991, 163ff).

coherence-based approaches and explain how this guideline is implemented in our approach.

Form-based approaches segment a sequentially organized text according to its structure of chapters, sections, subsections and paragraphs. In many cases, paragraphs are regarded to be the smallest units, and the segmentation follows the principle “one paragraph is one node”. The nodes generated by such form-based principles are then reconnected via hyperlinks.

Common strategies for form-based linking may be explained on the basis of the hypertext structure visualized in figure 1:

- Form-based hierarchical linking: This principle creates hyperlinks that reconstruct the hierarchical relations between chapters, sections, subsections and paragraphs. In hypertext navigation bars, links of this type are usually represented by up(ward)- and down(ward)-arrows. In addition, most hyperdocuments provide links from all nodes to an index page – typically a “clickable” table of contents in which the headings are linked to the first nodes of the respective sections and subsections.
- Form-based sequential linking: This principle reconstructs the sequence of the original sequential text by creating a reading path leading in a depth-first-strategy through the hierarchy of nodes. In navigation bars these links are typically represented by left- and right-arrows. Users that follow this reading path will perceive the document in exactly the sequence that the author of the sequential document had in mind.

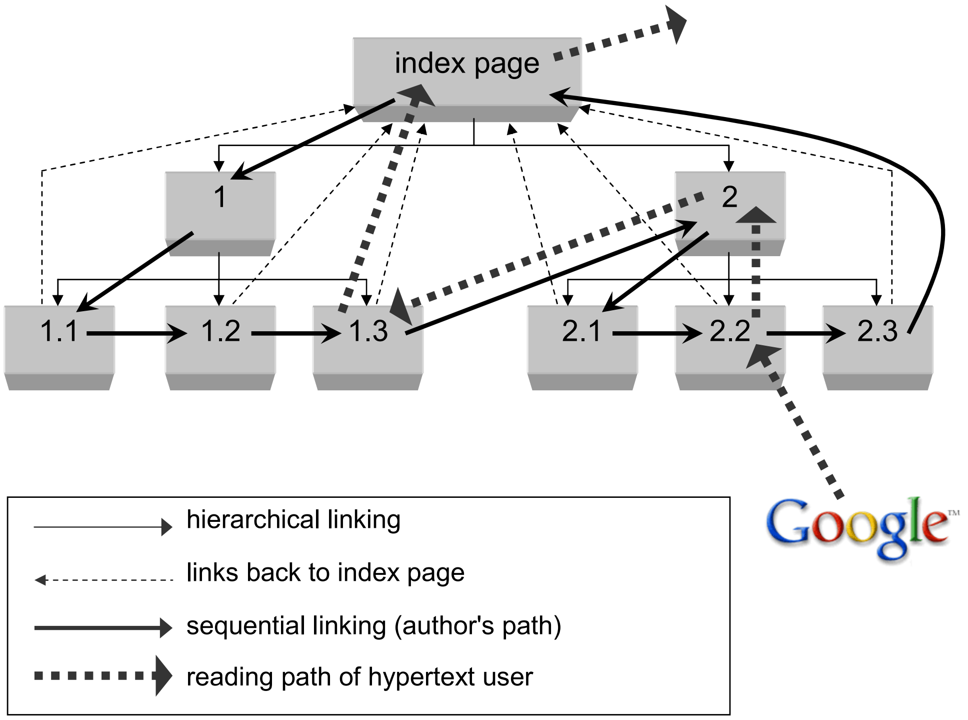


Figure 1. Resulting structure of a form-based conversion approach

Reading paths created by the sequential linking principle are only an option. Most hypertext users will select their own paths. Browsing the web with a search engine, a user may directly be ushered to a node. From there, he may click to a higher level, climb down again in order to pick an interesting detail, then jump to the homepage and afterwards surf to a different site. A user path of this type is illustrated in figure 1. The crucial point for our discussion is that users of form-based hypertexts which do not follow the author's path but search their own paths may be faced with two types of problems that are both related to the concept of coherence:

1) Some problems are located on the *micro-level* and are related to the concept of cohesive closedness[3]. These problems are caused by the fact that paragraphs in sequential documents may contain cohesion markers (anaphoric and textdeictic expressions; connectives) related to information that is located in the preceding or

[3] Cf. Kuhlen (1991, 33f and 87f).

in the subsequent text. Coherence-based conversion strategies that cope with this problem aim at liberating cohesive markers from their linkage to the reading path of the sequential document. These strategies will be described in section 4.

2) Other problems are located on the *macro-level*. They are caused by the fact that an author of sequential text, who verbalizes its content, presupposes that the reader is already acquainted with the content in the preceding text[4]. Hence, he may not repeat information that has been given in the preceding text. The selective reader, who is sent directly to a node, like in the example illustrated in figure 1, may, therefore, lack important knowledge prerequisites. Our solution to problems like these is linking according to knowledge prerequisites. That means that – by creating hyperlinks – we offer those knowledge units that a selective reader needs for properly understanding the current node. The strategies that we have developed in this context will be described in section 5.

3. PROJECT ARCHITECTURE: INFORMATION LEVELS AND ANNOTATION LAYERS

Our conversion approach processes information from two levels:

- On the *document level* we annotate the documents in our corpus on different linguistic and text-grammatical annotation layers which will be described below. This markup is then used for automatic segmentation, linking and reorganization.
- On the *domain knowledge level*, we represent the main concepts of our subject domains in a WordNet style semantic net, called TermNet. The technical basis for this representation is XML Topic Maps[5]. All technical terms are represented as word topics and related to their definitions and term occurrences in the documents.

A dynamic-adaptive component which processes logs of user paths is planned for the second phase of our project. This *hypertext usage level* would supply information about

[4] Cf. Foltz (1996), Fritz (1999), Storrer (2002).

[5] Cf. Pepper , Moore (2001), Lenz, Storrer (2002).

the hypertext nodes already visited by a user and, with this, about the knowledge prerequisites that he already has.

The following subsections give an outline of the annotation layers on the document level (section 3.1) and of the semantic net on the domain knowledge level (section 3.2). In section 4 and 5 we will explain how these levels and layers are used in conversion rules that we have implemented in the first phase of our project. Implementation issues are discussed in more detail in Lenz (this volume).

3.1 Annotation layers on the document level

In the first phase of our project, we gathered a corpus with documents from two domains: hypertext research and text technology. We developed XML document grammars to annotate this corpus on different linguistic and text-grammatical information layers: the document structure layer, the terms and definitions layer, the thematic structure layer and the cohesion layer. Additional linguistic information was provided by *morphosyntactic annotations* automatically assigned by the *KaRoPars* (v.0.36) technology developed at the University of Tübingen.[6] The *KaRoPars* output provides part-of-speech information[7], lemmatization and a “flat” syntactic analysis. This syntactic analysis includes the demarcation of “topological fields” (“Vorfeld”, “Mittelfeld”, “Nachfeld”) relevant for German word order regularities.

Below we illustrate the mark-up used in these annotation layers using the following text segment as an example:

[6] Cf. Müller (2004). We want to thank the Erhard Hinrichs research group for their cooperation.

[7] The part-of-speech categories used are those of the “Stuttgart-Tübingen-Tagset” (STTS, cf. http://www.sfs.nphil.uni-tuebingen.de/Elwis/stts/stts.html).

Example text 1:

Tochtermann (1995) spezifiziert einen Anker als eine eineindeutige Zuordnung zwischen einem Identifikator und einem Ankerobjekt, das sich durch fünf Felder charakterisieren lässt:

– das Hyperdokument,
– das betreffende Modul,
– die Komponente,
– der Ankerbereich,
– Attribute zum Anker (z.B. für Informationen zur Gewichtung oder zu Zugriffsrechten).

Engl.: *Tochtermann specifies an anchor as a reversibly unambiguous assignment between an identifier and an anchor object, which is characterized by five positions:*
– the hyperdocument
– the respective hypertext node
– the node component
– the position of the anchor
– further anchor attributes (e.g. information on relevance ranking, or on access rights)

On the *document structure layer* we annotate structural units (such as chapters, paragraphs, footnotes, enumerated and unordered lists) using an annotation scheme derivated from DocBook. On this layer our example text would be annotated in the following way:

```
<doc:para>
  Tochtermann (1995,76) spezifiziert einen Anker als eine eineindeutige Zuordnung zwischen einem Iden-
tifikator und einem Ankerobjekt, das sich durch fünf Felder charakterisieren lässt:
  <doc:itemizedlist>
    <doc:listitem>
      <doc:para>das Hyperdokument,</doc:para>
    </doc:listitem>
    <doc:listitem>
      <doc:para>das betreffende Modul,</doc:para>
    </doc:listitem>
    <doc:listitem>
      <doc:para>die Komponente,</doc:para>
    </doc:listitem>
    <doc:listitem>
      <doc:para>der Ankerbereich,</doc:para>
    </doc:listitem>
    <doc:listitem>
      <doc:para>Attribute zum Anker (z.B. für Informationen zur Gewichtung oder zu Zugriffsrech-
ten).</doc:para>
    </doc:listitem>
  </doc:itemizedlist>
</doc:para>
```

On the *terms and definitions layer* we annotate occurrences of technical terms as well as text segments in which these terms are explicitly defined. Definitions typically consist of three functional components: the *Definiendum* (the term to be defined), the *Definiens* (meaning postulates for the term) and the *Definitor* (the verb which relates the definiens component to the definiendum component). Our document grammar specifies mark-up for each of these components. In addition, we explicitly annotate the occurrences of all terms that are included in our semantic net described in section 3.2. . The definition in our example text is annotated in the following way:

```
<definitions>
 [...]
 <defSegment>
   <def type="Fremdzuschreibung">
     Tochtermann (1995,76)
     <dfnSegment> spezifiziert </dfnSegment>
     <definiendum> einen <term normalForm="Anker" baseForm="Anker">Anker</term> </definiendum>
     <dfnSegment> als </dfnSegment>
     <definiens> eine eineindeutige Zuordnung zwischen einem Identifikator und einem Ankerobjekt, das
sich durch fünf Felder charakterisieren lässt: das
     <term normalForm="Hyperdokument" baseForm="Hyperdokument"> Hyperdokument </term>, das
betreffende <term normalForm="Modul" baseForm="Modul">Modul</term>, die <term normal-
Form="Komponente" baseForm="Komponente"> Komponente</term>, der Ankerbereich, Attribute zum
<term normalForm="Anker" baseForm="Anker">Anker</term> (z.B. für Informationen zur Gewichtung o-
der zu Zugriffsrechten)
     </definiens>.
   </def>
 </defSegment></definitions>
```

On the *thematic structure layer* we want to capture the way in which topics are introduced, elaborated in the subsequent text and related to subordinate topics (subtopics) or more general topics (macro-topics). The annotation schema is based on the typology of thematic progression proposed by Ludger Hoffmann[8]. This typology presents five basic patterns of thematic progression: topic continuation, topic splitting, topic composition, topic subsumption and topic association. These basic patterns can be combined into more complex clusters representing the thematic structure of paragraphs. The basic idea of our schema is to segment each paragraph in a top-down-fashion into thematic clusters and basic patterns. According to this document grammar, the first part of our example text is annotated in the following way:

[8] Cf. Zifonun, Hoffmann et al.. (1997, chapter C6, 535-591) and Hoffmann (2000).

```
<tCluster type="associate">
(...)
<tCluster role="associatedTopic" type="compose">
  <tsegment role="compoundTopic">
    (Tochtermann 1995,76) spezifiziert einen <topic type="word" topicConceptName="Anker"> Anker
</topic> als eine eineindeutige Zuordnung zwischen einem Identifikator und einem Ankerobjekt, das sich
durch fünf Felder charakterisieren lässt:
  </tsegment>
  <tsegment role="componentTopic">
    das <topic type="concept"> Hyperdokument </topic>,
  </tsegment>
  <tsegment role="componentTopic">
    das betreffende
    <topic type="concept" topicConceptName="Modul"> Modul </topic>,
  </tsegment>
  <tsegment role="componentTopic">
    die <topic type="concept" topicConceptName="Komponente"> Komponente </topic>,
  </tsegment>
  <tsegment role="componentTopic">
      der <topic type="concept"> Ankerbereich </topic>,
  </tsegment>
  <tsegment role="componentTopic">
    <topic type="concept">Attribute zum Anker</topic>
      (z.B. für Informationen zur Gewichtung oder zu Zugriffsrechten).
  </tsegment>
</tCluster>
</tCluster>
```

This annotation implies that the compound topic “anchor” is composed of five subordinate component topics. When topic words are included in our semantic net, we specifiy their word forms as values of the optional attribute topicConceptName (e.g. the topic words “Anker”, “Modul” and “Komponente” in our example). Accordingly, the thematic structure on the document level is linked to the topic map representation on the domain knowledge level.

On the *cohesion layer* we annotate text-grammatical information of various types, e.g. co-reference, connectives, text-deictic expressions. This layer is crucial for our reorganization strategies, i.e. for generating cohesively closed hypertext nodes. Therefore, we will describe the mark-up on this annotation level in section 4.

Following the approach developed by Witt et al. (2005), we store our annotation layers in separate files. Thus, each layer can be annotated and maintained separately and can be validated against its corresponding document grammar (DTD or schema file). In a subsequent unification step, the different annotation layers of our corpus documents are merged. The resulting unified representation is the basis for an XSLT transformation, which automatically generates the hypertext views along the guidelines of our linking and segmentation strategies.

3.2 Structure of the terminological wordnet on the domain knowledge level

Two-level architectures of hypertext supplement the hypertext documents with a formalized knowledge representation (e.g. Mayfield, 1997; Carr et al., 2001). Following this idea, our architecture connects the annotated documents on the document level with a semantic net on the domain knowledge level. This semantic net, called TermNet, represents the semantics of the technical terms that are relevant for the subject domains in our documents in a WordNet-style representation. In our approach, we use information from this domain knowledge level to automatically generate glossary views, which show how a technical term is linked to other terms and concepts of the domain. These glossary views also contain hyperlinks to text segments, in which the respective terms are explicitly defined. The glossary views are connected to all term occurrences in the documents; but the glossary can also be used as an additional stand-alone component. The interplay between the two architectural levels is illustrated in figure 2. Using an example, we will explain in section 5 how information from the document and the domain knowledge level is used for our coherence-based linking strategies. In the following section we will concentrate on the main structural features of our semantic net and outline some implementation issues. More detailed descriptions (in German) are given in Beißwenger et al. (2003) and Lenz et al. (2003).

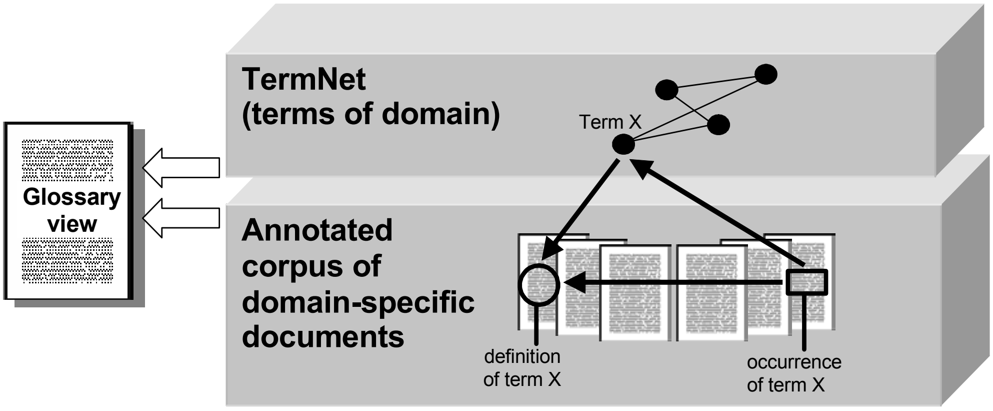


Figure 2. Interplay between the two architectural levels

Fundamental for the structure of TermNet are the entities and relations introduced for the Princeton WordNet (Fellbaum, 1998) and the German word net GermaNet (Kunze/Wagner, 2001). The two main entity types in the WordNet representation model are “words” (lexical units) and “synsets”, i.e. sets of synonymous word senses. Synonymy in the strong sense of interchangeability in all contexts is rare in natural language. Therefore, WordNet uses a “smoother” criterion: two word senses belong to the same synset when they may be interchanged in some context (Miller, 1998, 23f). The two main entities – words and synsets – are related by lexical relations between words and conceptual relations between synsets. The number and the definition of these relations are slightly different in the Princeton WordNet and in GermaNet. In our approach we concentrated on a subset of conceptual relations used in both approaches. Furthermore, we introduced some additional lexical relations that we found useful for our application context.

In TermNet the two basic entities are *terms* (the equivalent to “word/lexical unit” in the WordNet model) and *termsets* (the equivalent to “synsets” in the WordNet model). Terms in TermNet are linguistic expressions, the technical meaning of which is explicitly defined in our corpus, i.e. a term in our TermNet is related to one or more definitions in the corpus. As described above, these definitions are explicitly annotated in our "terms and definitions layer". The version of TermNet that we developed in the first phase of the project comprises mainly nouns, many of them multiword units composed of a noun and an adjective modifier such as "bidirektionaler Link" (engl. bidirectional link). We treat these multiword units as "words-with-blanks" and provide information about the inflected forms of the nouns and adjectives in a separate list. This list is used for the automated annotation of the terms on the "terms and definition layer" described in the previous section.

Termsets contain technical terms that denote the same or a quite similar concept in different approaches to a given scientific domain. For instance, the books by Kuhlen (1991) und Tochtermann (1995) both introduced a terminology for hypertext concepts that influenced the technical terms used in German papers on hypertext research. Both authors provide definitions for the concept of a "hyperlink" and specify a taxonomy of subclasses (1:1-link, bidirectional link etc.). But Kuhlen uses the loan word "Verknüpfung" in his taxonomy (1:1-Verknüpfung, bidirektionale Verknüpfung) while Tochtermann's taxonomy uses the loan word "Verweis" (with subconcepts like 1:1-Verweis, bidirektionaler Verweis). In addition, the definitions of the concepts and subconcepts given by these authors are slightly different, and the two taxonomies are not isomorphic. As a consequence, in a scientific document on the subject domain, a term of the Kuhlen taxonomy can not be replaced by the corresponding term of the Tochtermann taxonomy. After all, the purpose of defining terms is exactly to bind their wordforms to the semantics provided in the definition. The usage of these terms in documents may serve, in contrast, as an indicator to which theoretical framework or "scientific school" the paper belongs. "Verknüpfung" and "Verweis" are, thus, not synonyms in the sense of being interchangeable in at least some contexts. For this reason, we do not use the term "synset" but introduced the term "termset": the members of termsets are terms that denote the same or very similar categories in competing taxonomies on the same scientific domain. This relationship of "Kategorienähnlichkeit" (category correspondence) is not determined by their interchangeability in corpus documents, but by their extension: two terms A and B are categorially correspondent if the set of objects in the research domain that are instances of term A has a high intersection with the set of those objects that are instances of term B.

But not all terms in a termset belong to different taxonomies. We recall two cases, in which the same technical term has alternative wordforms: (1) a multiword term has an equivalent abbreviated form (e.g. “Hyperlink” and “Link” or “hypertext markup language” and “HTML”); (2) a term has two orthographic variants (e.g. “Hyper-Link” and “Hyperlink”). In these cases, the respective term forms are actually synonyms in the strong sense that they denote equivalent classes of instances and may be interchanged in all contexts. In TermNet we represent this strong equivalence by means of lexical relations between terms of the same termset: the relation “is-abbreviation-of” and its inverse relation “is-expansion-of”, and the symmetric relation “is-orthographic-variant-of”. In order to support multilingual linking in a later stage of our project, we link German technical terms to their English equivalent using the additional lexical relations “is-loanword-of” (“Link” is loanword of “link”) and “is-loan-translation-of” (“Verknüpfung” is loan translation of “link”).

Many concept-based terminology representations label one of the terms as the preferred term. However, when terms belong to competing approaches and schools – as it is frequently the case in scientific domains – this decision may be hard to make because all approaches have their benefits and complement each other. For this reason, we do not use the preferred term label in our representation. The objective of our representation is to connect competing technologies with each other because, in our usage scenario, it is often quite useful to know that the term A used in document x denotes merely the same category that term B used in document y. If the user is interested, he may easily reconstruct in which semantic aspects they differ because all terms of a given termset are linked to their definitions in the documents.

In addition to the lexical relations described above, TermNet represents conceptual relations between termsets: the taxonomic relation “is-hyponym-of” and its inverse relation “is-hypernym-of”, the part-of relation “is-meronym-of” and its inverse relation “is-holonym-of”. In addition, we relate termsets that denote opposite categories by the relation “is-antonym-of”. Here we deliberately deviate from the standard WordNet model that represents antonymy as a lexical relation because we feel that, for our usage scenario, it may be important to know that the terms “monodirektionaler Verweis” and “monodirekionale Verknüpfung” both denote a category that is complementary to the category denoted by the terms “bidirektionaler Verweis” and “bidirekionale Verknüpfung”. In our subject domain we found that termsets on the same hierarchical level often form groups of mutually disjoint concepts. For example, one may use multiple classification features to subdivide the general concept of a hyperlink. The class of links may be subclassified into monodirectional and bidirectional links depending on whether their underlying relation is asymmetric or symmetric. According to the position of their target, anchor links may be further subclassified into internal and external links. These subclasses are, in most cases, the same in competing taxonomies. We, thus, find a bunch of termsets with similar terms for the same specific concept, i.e. “monodirektionaler Link”, “monodirektionaler Verweis”, “monodirektionale Verknüpfung”, that are all hyponyms to the superordinated termset for the concept “Link”. If only this hyponymy relation is encoded, an aspect that is vital for inferences is concealed: an individual link in a document may be simultaneously monodirectional and external. But it cannot be simultaneously monodirectional and bidirectional, since these subclasses are defined to be mutually disjoint. In order to account for this fact, we enhanced the standard WordNet model by (optional) attributes that specify classification features for subordinate termsets. Termsets that have the same hypernym and the same classification feature are defined as denoting disjoint classes of instances.

In the first stage of our project, TermNet was represented as an XML Topic Maps (Pepper and Moore, 2001) application. In order to facilitate the construction and the maintenance of TermNet, we used K-Infinity[9], a tool for building and maintaining knowledge networks with a comfortable graphical editor. K-Infinity has an internal representation that already performs consistency checks (e.g. it prevents cycles in hyponymy relations) and is enhanced by export facitities, e.g. an XSLT stylesheet that transform the internal K-Infinity representation into an XML Topic Map representation. We conduct some additional consistency checks on this XTM representation and enrich it by relations that are not explicitly encoded but can be automatically inferred, e.g. the disjointness of subclasses with the same classification feature that we explained above (cf. Lenz et al., 2003). The resulting XTM representation forms the basis for our hypertextualisation strategies described in chapter 5.

4. COHERENCE-BASED STRATEGIES ON THE MICRO-LEVEL: COHESIVE CLOSEDNESS IN HYPERTEXT NODES

Form-based conversion approaches segment larger documents according to structural units, i.e. sections, subsections and paragraphs. In our approach we aim at a very "granular" segmentation that is based on the general principle that one paragraph becomes one hypertext node. The respective segmentation rules process mark-up from the document structure layer, especially mark-up indicating section, subsection and paragraph boundaries; subrules handle special cases like unordered and ordered lists, tables, figures and their respective captions. These rules construct the basic units of our hypertext view: the hypertext nodes[10]. However, these nodes quite often contain cohesion markers related to information that is located in the preceding or in the subsequent text, e.g. anaphoric pronouns or anaphoric noun phrases, textdeictic expressions like "siehe oben" (E: *see above*) and various types of connectives. This is due to the fact that sequential documents are generally designed to be read completely and in the sequence prepared by the author. A subtask in the conversion of sequential documents into hyperdocuments is to liberate cohesive markers in hypertext nodes from their linkage to a specific reading path, i.e. to achieve "cohesive closedness" in hypertext nodes[11].

We transform paragraphs in cohesively closed hypertext nodes by rules that use annotations from the cohesion layer. This layer provides mark-up for anaphoric pronouns and noun phrases, text-deictic expressions and connectives. On this basis we implemented

[9] K-Infinity is a commercial knowledge engineering software developed and distributed by Intelligent Views: www.i-views.de/. We thank Intelligent Views for their valuable and kind support.

[10] Cf. Lenz (in this volume) for implementation details.

[11] Cf. Kuhlen (1991, 33f and 87f).

four basic operations that transform the paragraphs of sequential documents into "stand-alone" hypertext nodes that may be integrated into various reading paths:

1. *Anaphora resolution*: some paragraphs contain anaphoric pronouns or noun phrases, the antecedents of which are found in the previous paragraph. In these cases a pop up element with the antecedent is displayed above the pronoun.
2. *Node expansion*: some connectives indicate that the content of the paragraph is strongly related to the previous (or the subsequent) text, e.g. "außerdem" (*in addition*), "allerdings" (*though*), "darüber hinaus" (*furthermore*). In these cases, we provide the option to expand the current node and display the preceding or subsequent paragraph. With this option the user may accumulate as much context as he desires for properly understanding the content of the node.
3. *Linking*: in many cases we find expressions pointing to other text segments in the document. These expressions are transformed into hyperlinks that are related to their target segments. These target segments may be identified quite precisely, e.g. in expressions like "siehe Kapitel 3.4.2" (see chapter 3.4.2.). Other text-deictic expressions, e.g. "siehe oben" (see above) or "wie bereits erwähnt" (as mentioned already), are bound to the position of the current node in the author's reader path. In some of these cases, it is not easy to locate and to delimit the text segment to which the deictic expression is pointing.
4. *Deletion*: some occurrences of connective particles like "noch" or "also" seem to be stylistically motivated, i.e. they serve first and foremost the creation of a fluent text. Although they indicate how the current node is related to the previous paragraph, the content of the previous paragraph is not a prerequisite for the correct interpretation of the current node. In these cases, we decided to delete the connective particles in order to obtain a more "stand-alone" text version.

We will illustrate below how the mark-up of the cohesion layer is used to automatically obtain cohesive closedness. Example text 2 is a paragraph of a text book on hypertext.[12] According to our segmentation rules, this paragraph would constitute a hypertext node.

[12] We did not find a paragraph in which all rules and procedures could be demonstrated. Therefore, the example is slightly modified - the original paragraph does not contain an anaphoric pronoun. However, our corpus contains several examples with ana-

Example text 2:

Weiterhin unterscheidet **er noch** nach der Anzahl der in einen Link involvierten Anker in 1:1-Links, in denen ein Ausgangsanker mit genau einem Zielanker verknüpft ist, 1:n-Links, in denen ein Ausgangsanker mit mehreren Zielankern verbunden ist, und n:m-Links, in denen mehrere Anker unabhängig von der Traversierungsrichtung miteinander zu einem Linking-Muster kombiniert sind. Im Linking-Element von HTML sind nur 1:1-Links vorgesehen; **die obige Spezifikation** und das Konzept des “Extended Link” (im Sinne der Xlink-Spezifikation) sehen auch Links mit mehreren Ankern vor.

English: *According to the number of anchors that are involved in a link,* ***he further*** *differentiates between one-to-one-links, which connect a source anchor to exactly one target anchor, one-to-many links which connect a source anchor to several target anchors, and many-to-many-links in which several anchors are combined into a linking pattern that is independent from the direction of traversal. The link element in HTML only provides 1:1 links; the* ***above-mentioned specification*** *and the concept of an “extended link” (as defined in the XLINK specification) also provide links with multiple target anchors.*

This paragraph contains four cohesive markers related to elements of the preceeding text: (1) the connectives “weiterhin” (further) and “noch” (in addition), (2) the anaphoric pronoun “er”, (3) the textdeictic expression “die obige Spezifikation”. These markers would be annotated in the cohesion layer as follows:

phoric pronouns, the antecedents of which are placed in the previous segment. We handle these cases in the way that is described in our example.

<connective connectedTo="backward"> Weiterhin </connective> unterscheidet **<discourseEntity deID="de_n_30" deType="nom"> er </discourseEntity> <connective pragType="stylistic" connectedTo="unspecified"> noch </connective>** nach der Anzahl der in einen Link involvierten Anker in 1:1-Links, in denen ein Ausgangsanker mit genau einem Zielanker verknüpft ist, 1:n-Links, in denen ein Ausgangsanker mit mehreren Zielankern verbunden ist, und n:m-Links, in denen mehrere Anker unabhängig von der Traversierungsrichtung miteinander zu einem Linking-Muster kombiniert sind.

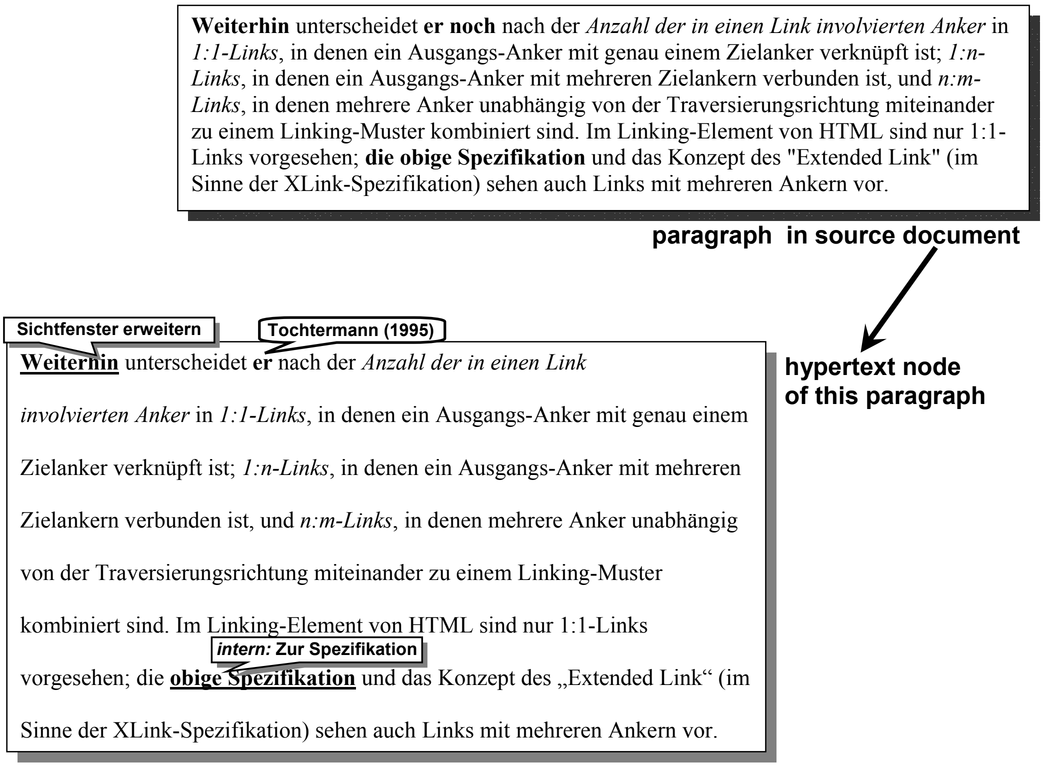
<semRel>
  <cospecLink relType="**propName**" phorIDRef="de_n_30"
  antecedentIDRefs="de_n_27"/>
</semRel>

Im Linking-Element von HTML sind nur 1:1-Links vorgesehen;
<connective connectedTo="specifiedByID" connectedToID="caID_52"> die obige Spezifikation </connective> und das Konzept des “Extended Link” (im Sinne der Xlink-Spezifikation) sehen auch Links mit mehreren Ankern vor.

Our reorganization rules process these annotations to generate the hypertext node illustrated in figure 3. In this reorganization process, all of the above-mentioned operations are applied:

1. *Anaphora resolution*: the antecedent of the anaphoric pronoun “er” is displayed in a pop up element. This operation uses the antecedentIDRefs attribute of the cospecLink element and identifies the antecedent by its value. Our antecedent assignment was annotated manually[13]. But in principle, this operation could also be applied to documents with anaphora that were resolved automatically. Since automated anaphora resolution is not correct in all cases, we display antecedents as pop up elements (instead of replacing the pronouns by their antecedents).
2. *Node expansion (Sichtfelderweiterung)*: many connectives are directly related to the previous or the subsequent node; an example of this type is “weiterhin” (furthermore). In our annotation, we specify this relatedness by means of the values backward or forward assigned to the attribute connectedTo. When a connective has one of these values in its connectedTo attribute, it will be transformed into a

link that displays the previous node (if the value is backward) or the subsequent node (if the value is forward).



Weiterhin unterscheidet **er noch** nach der *Anzahl der in einen Link involvierten Anker* in *1:1-Links*, in denen ein Ausgangs-Anker mit genau einem Zielanker verknüpft ist; *1:n-Links*, in denen ein Ausgangs-Anker mit mehreren Zielankern verbunden ist, und *n:m-Links*, in denen mehrere Anker unabhängig von der Traversierungsrichtung miteinander zu einem Linking-Muster kombiniert sind. Im Linking-Element von HTML sind nur 1:1-Links vorgesehen; **die obige Spezifikation** und das Konzept des "Extended Link" (im Sinne der XLink-Spezifikation) sehen auch Links mit mehreren Ankern vor.

Weiterhin unterscheidet **er** nach der *Anzahl der in einen Link involvierten Anker* in *1:1-Links*, in denen ein Ausgangs-Anker mit genau einem Zielanker verknüpft ist; *1:n-Links*, in denen ein Ausgangs-Anker mit mehreren Zielankern verbunden ist, und *n:m-Links*, in denen mehrere Anker unabhängig von der Traversierungsrichtung miteinander zu einem Linking-Muster kombiniert sind. Im Linking-Element von HTML sind nur 1:1-Links vorgesehen; die **obige Spezifikation** und das Konzept des „Extended Link“ (im Sinne der XLink-Spezifikation) sehen auch Links mit mehreren Ankern vor.

Figure 3. Cohesive closure in hypertext nodes

3. *Linking*: we annotate textdeictic expressions like “die obige Spezifikation” (the above-mentioned specification) as connectives which have the value specifiedByID assigned to the connectedTo attribute. The value of the additional connectedToID attribute identifies a text segment in the previous or subsequent text; in our example, this text segment is a specification that is annotated in the following way on our cohesion layer:

[13] The annotation scheme was described in Holler et al. (2004) and Holler (2003).

(...)**<connectiveAnchor ID="caID_52">** Tochtermann (1995, 68) spezifiziert einen Link als eine eineindeutige Zuordnung zwischen einem Identifikator und einem Linkobjekt, das durch fünf Felder charakterisiert wird: (1) einen oder mehrere Ursprungs-Anker, (2) einen oder mehrere Ziel-Anker bzw. Berechnungsvorschriften für Ziel-Anker, (3) die Richtungsinformation zur Spezifikation der Traversierungsrichtung, (4) Attribute für zusätzliche Informationen zum Link-Typ, zur Gewichtung oder zu den Zugriffsrechten, (5) Operationen, die bei der Aktivierung des Verweises ausgeführt werden (optional). **</connectiveAnchor>**(...)

Engl: *Tochtermann (1995, 68) specifies a link as a reversibly unambiguous assignment between an identifier and a link object, which is characterized by five positions: (1) one or several source anchors, (2) one or several target anchors or an algorithm for the computation of target anchors, (3) information about the direction of traversal, (4) attributes for additional information about the link type, about the weighting or about access rights, (5) operations which are executed when the link is activated (optional).*

When a connective has the value specifiedByID in its connectedTo attribute, it will be transformed into a link that displays the node containing the connectiveAnchor element. It should be noted that, from a linguistic viewpoint, the expression “die obige Spezifikation” could just as well be treated as an anaphoric expression. The decision to treat it as a textdeictic connective is, in this case, in the first place motivated by the size of the text segment which would not fit nicely in a pop up, and only in the second place by the adjective “obig” and its deictic function. But in many other cases (e.g. “siehe oben”) the difference is clear, although it is not always easy to identify the boundaries of the connective anchors to which these expressions refer. References to text segments that can automatically be identified by the document structure (e.g. “siehe Kapitel 3.2.4”) are easier to handle. In all of these cases, the basic operation is to transform the connective into a hyperlink that is related to the node containing the anchor element.

4. *Deletion*: some connectives and particles first and foremost serve the creation of a fluent text, like the connective “noch” in our example paragraph. These connectives

have the value stylistic in the pragType attribute, which describes the pragmatic functions of connectives. Connectives with this value are deleted.

As can be seen by our example, our annotation of cohesion phenomena is quite selective, i.e we annotate only those markers that are relevant for transforming text segments in cohesively closed hypertext nodes. A full annotation of all cohesion phenomena would imply a complete reconstruction of anaphoric and co-reference relations between text segments and an elaborate set of different types of connectives. In the framework of our project, we did not have the means to annotate our corpus documents in such a fine-grained manner, and German corpora with cohesion annotations are not available yet. Our selection was made intellectually, and the annotation was done manually; the resulting mark-up forms, thus, the basis for the automated generation of cohesive closure in hypertext nodes.

5. COHERENCE-BASED STRATEGIES ON THE MACRO-LEVEL: LINKING ACCORDING TO TERMINOLOGICAL KNOWLEDGE PREREQUISITES

The conversion strategies described in the previous section are concerned with cohesion markers, i.e. with phenonema that are related to verbal units on the surface structure of the text segment. The goal of these strategies was to revise those cohesion markers that point to segments in the previous or subsequent text in a way that fits the resulting hypertext nodes into multiple reading paths. However, the revision of cohesive markers on the micro level, i.e. inside the current node, does not solve another problem that has been the focus of research on hypertext coherence[14]: the author of a sequential text assumes that the user is acquainted with the discourse referents and information which he introduced in the preceding text. Hence, he does not need to mention them again explicitly. For this very reason, the hypertext reader, who does not follow the author's reading path, may lack essential knowledge prerequisites.

The problem may be explained using the example hypertext structure that was illustrated in figure 1. We can imagine that the sequential text that formed the basis of this hyperdocument was a hypertext textbook. The author of the sequential document supplied a definition for a technical term in section 1.2., e.g. he defined the term “link”. He may then presuppose that the reader in the subsequent paragraphs understands this term according to this definition. But if the sequential document is converted into a hyperdocument, it will typically be read selectively and in a non-predictible sequence. Our hypothetical hypertext user in figure 1, for instance, has not “visited” node 1.2. When reading node 1.3, he may come across an occurrence of the term “link”, but may not be familiar with its technical meaning — as we explained in section 2, our usage scenario focuses on users with previous but no expert knowledge in the domain. In the best case, he will notice this knowledge gap and search for the definition. In the worse case, he will interpret the term in a non-technical sense or according to its technical meaning in another scientific domain (e.g. “link” as used in Artificial Intelligence). In this case, he risks missing important knowledge prerequisites and misunderstanding the content of the node.

The conversion strategies that will be discussed below aim to compensate for coherence problems of this type by generating additional links to text segments that may be prerequisite for the correct understanding of the current node. In contrast to the linking rules described in section 4, these coherence phenomena are not explicitly indicated by cohesive markers (e.g. explicit references, textdeictic or anaphoric expressions), but are implicitly presupposed by the author, who verbalized his content with a fixed and predefined reading path in mind. In the first stage of our project, we concentrated on knowledge related to the meaning of technical terms because, for our user scenario, technical terms play a central role. Whoever wants to become acquainted with a particular knowledge domain has to understand the concepts denoted by the technical terms in this domain, i.e. has to be informed as to how these terms are defined.

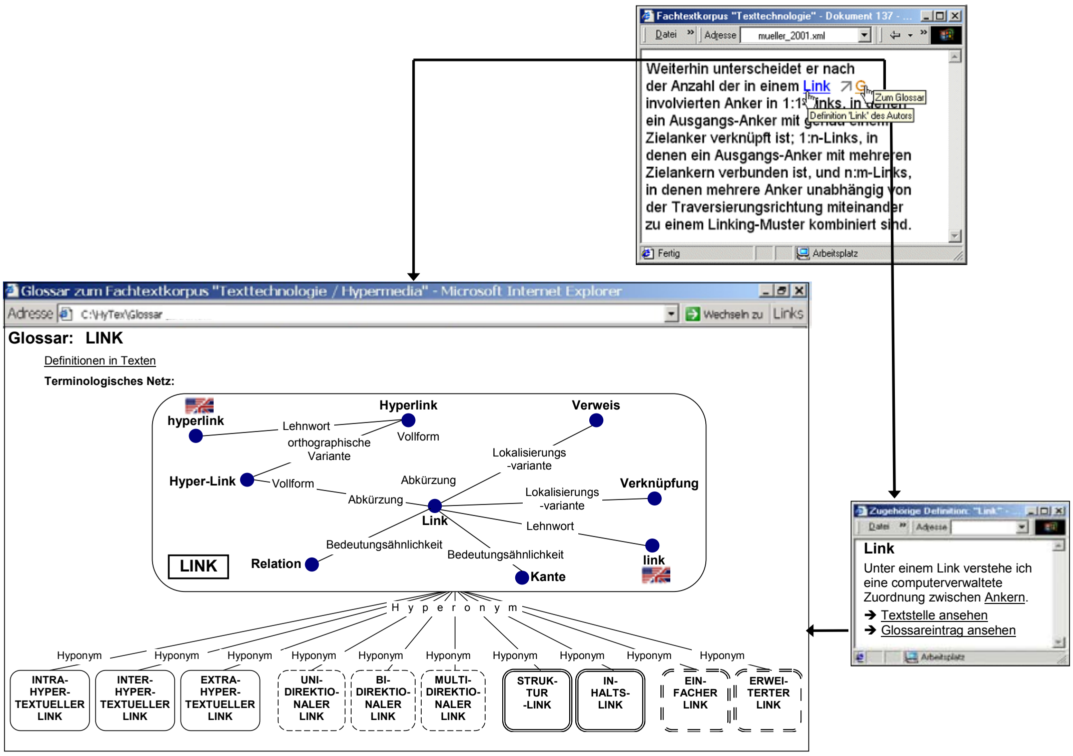
In our hypertext views we offer two options to assist selective readers in better understanding the terms and their underlying concepts:

[14] Cf. Hammwöhner (1990), Foltz (1996), Hammwöhner (1997), Fritz (1999), Storrer (2002).

- Term-to-definition links: if a technical term is defined in the document, all occurrences of this term are linked to the definition segment.
- Glossary views: all technical terms are linked to glossary views, which show how a given technical term is related to other terms and concepts of the domain. The glossary view for a term also provides links to all text segments in which the term is explicitly defined. Thus, the user gets a quick survey on how the term is used and defined in the respective domain, whether all authors agree on a definition, or whether various term variations compete.

These two strategies may be illustrated by the example in figure 4.

In this example, the term “Link” is marked as an occurrence of a technical term in the hypertext node. If the user does not know the technical meaning of this term, he may activate a link button which displays its definition in a pop-up window. To get more context, the definition pop up is linked to the node containing the definition. In addition, the user may activate the glossary window that visualizes the lexical and conceptual relation between the term and similar terms and concepts. Any of these terms are linked to their respective glossary entries. Each glossary entry is linked to all nodes that contain a definition for the respective term. With these linking structures, the user can, step by step, become familiar with the interrelations and differences between terms and concepts in the respective domain.



Fachtextkorpus "Texttechnologie" - Dokument 137 - ...
Datei
Adresse
mueller_2001.xml
Zum Glossar
Definition 'Link' des Autors
Fertig
Arbeitsplatz

Weiterhin unterscheidet er nach der Anzahl der in einem Link involvierten Anker in 1:1-Links, in denen ein Ausgangs-Anker mit genau einem Zielanker verknüpft ist; 1:n-Links, in denen ein Ausgangs-Anker mit mehreren Zielankern verbunden ist, und n:m-Links, in denen mehrere Anker unabhängig von der Traversierungsrichtung miteinander zu einem Linking-Muster kombiniert sind.

Link

Unter einem Link verstehe ich eine computerverwaltete Zuordnung zwischen Ankern.

→ Textstelle ansehen
→ Glossareintrag ansehen

Figure 4. Hypertext node enhanced by link to terminological knowledge prerequisites

The rules that generate these linking structures process information from the “terms and definitions layer” on the document level on the one hand, and from the Topic Map Representation of our semantic net on the other hand. The XML Topic Map representation of our semantic net forms the basis to generate our glossary views (cf. section 3.2). The “terms and definitions layer” (cf. section 3.1) is used to explicitly mark lexical units that are used as technical terms in our domains. In order to prevent “overlinking”, we only mark the first occurrence of a term in each node and filter out the other occurrences with the help of specialized rules. In addition, we rule out those special cases in which a technical term occurs in exactly the hypertext node in which this term is defined – in these cases, of course, we do not want to generate links. The “terms and definition layer” is also used to cut out the definition text segment and to display it in the definition window. This operation is quite simple when the document contains exactly one definition for the respective term. But in some cases, authors of scientific articles and textbooks discuss several definitions for the same term, e.g. definitions to be found by other authors or scientific schools, before they provide their own definition. In order to cope with this problem, we provide rules for the ranking of several definitions for the same term. This ranking is mainly based on the values of the type attribute of the def element, which classifies definitions according to their pragmatic function. One basic ranking rule is that terms that are explicitly defined by the author (the type value is "Selbstzuschreibung" = *self assignment*) are ranked higher than definitions that are assigned to other authors (when the type value is "Fremdzuschreibung" = *external assignment*). This basic ranking rule is complemented by other factors like the position of the definition in the document (cf. Beißwenger et al., 2002, Beißwenger, 2004). Since our ranking results are not always adequate, we display the texts of all definitions, ordered by the results of the ranking process.

6. CONCLUSION AND OUTLOOK

The conversion strategies discussed in this paper were implemented and tested using a corpus with 20 technical documents from two technical domains, namely hypertext research and text technology. On the basis of this corpus, we want to evaluate the effectiveness of these strategies with respect to the user scenario described in section 2. For this purpose, we generate two versions of our corpus:

(1) The hypertext version HyTex.1 offers hypertext views according to the rules described above: we offer cohesively closed hypertext nodes with links to related text passages, to definitions and to the glossary views.

(2) The sequential text version HyTex.0 displays the corpus documents in their original sequence and content. It offers no glossary views and no links – except for the possibility to click on a digital table of contents of the respective sections in the documents.

We plan to develop specific tasks that match our scenario, e.g. answering questions related to domain specific concepts. We want to compare the time needed to solve these tasks and the quality of the solutions with students of computer science that have no expert knowledge in hypertext research and text technology. Since we can conveniently create different versions of our corpus (cf. section 2 and Lenz (in this volume)), we may further experiment with additional versions. Ro study the effects of the glossary views, for example, we plan to create a sequential version Hytex.0+ with the glossary as an additional stand-alone component.

In the second phase of our project, we want to extend our approach in three ways:

(1) We want to extend our two-level approach by a third information level containing logs of individual user paths. This information will be used to adopt our linking strategies to the knowledge that the user already has acquired at the current point of his hypertext usage.

(2) We want to experiment with additional topic-based strategies that profit from the thematic annotation layer. These strategies include the automatic determination of a node's macro topic, the generation of clickable topic views for a corpus, and the refinement of our segmentation strategies.

(3) Although all our conversion strategies are automated, their information basis – the annotations on the document layers and the semantic net – is predominantly hand-coded. If we want to apply our approach to arbitrary technical domains, it will be important to automate the necessary preprocessing steps. As a first step, we currently experiment with methods that automatically detect definitions of technical terms in documents and annotate their components according to the annotation scheme described in section 3.1 (cf. Storrer, Wellinghoff 2006) . These automatically annotated definitions may not only be used for our term-to-definition linking (cf. section 5). We also want to use them to extract WordNet style semantic relations that will enrich and expand our semantic net.

7. REFERENCES

Beißwenger, M., Storrer, A., and Runte, M., 2003, Modellierung eines Terminologienetzes für das automatische Linking auf der Grundlage von WordNet, in: *Anwendungen des deutschen Wortnetzes in Theorie und Praxis*, C. Kunze, L. Lemnitzer, and A. Wagner LDV-Forum, 19 (1/2), pp. 113-125.

Beißwenger, M., Lenz, E. A., and Storrer, A., 2002, Generierung von Linkangeboten zur Rekonstruktion terminologiebedingter Wissensvoraussetzungen, in: *KONVENS 2002. 6. Konferenz zur Verarbeitung natürlicher Sprache. Proceedings, Saarbrücken, 30.09.-02.10.2002,* S. Busemann, Saarbrücken, DFKI Document D-02-01, pp. 187-191.

Beißwenger, M., 2004, Arbeitsbericht: Annotation definitorischer Textsegmente und "terminologiesensitives Linking"**.** Technical Report; http://www.hytex.uni-dortmund.de/hytex/publikationen.html.

Carr, L., Hall, W., Bechhofer, S. and Goble, C., 2001, Conceptual Linking: Ontology-based Open Hypermedia, in: *Proceedings of the Tenth International World Wide Web Conference*, Hong Kong, pp. 334–342.

Fellbaum, C., 1998, *WORDNET: An electronic lexical database*, MIT Press, Cambridge, MA.

Foltz, P. W., 1996, Comprehension, Coherence, and Strategies in Hypertext and Linear Text, in: *Hypertext and Cognition,* J.F. Rouet, J.J. Levonen and J. Jarmo et al., ed., Lawrence Erlbaum Associates Publishers, Mahwah/New Jersey, pp. 109-136.

Fritz, G., 1999, Coherence in Hypertext, in: *Coherence in Spoken and Written Discourse. Pragmatics and Beyond New Series*, W. Bublitz, U. Lenk, Uta et al., ed., John Benjamins, Amsterdam/Philadelphia, pp. 221-232.

Hammwöhner, R., 1990, Macro-Operations for Hypertext Construction, in: *Designing Hypermedia for Learning,* D.H. Jonassen, H. Mandl, Heinz, Springer, Berlin et al., ed., pp. 71-96.

Hammwöhner, R., 1997, *Offene Hypertextsysteme. Das Konstanzer Hypertextsystem (KHS) im wissenschaftlichen und technischen Kontext*, Konstanzer Universitätsverlag, Konstanz.

Hoffmann, L., 2000, Thema, Themenentfaltung, Makrostruktur, in: *Text- und Gesprächslinguistik -- ein internationales Handbuch zeitgenössischer Forschung. 1.Halbband* (Handbücher zur Sprach- und Kommunikationswissenschaft 16), K. Brinker, G. Antos et al., ed., de Gruyter, Berlin/ New York, pp. 344-356.

Holler, A., 2003, Spezifikation für ein Annotationsschema für Koreferenzphänomene im Hinblick auf Hypertextualisierungsstrategien. Technical Report; http://www.hytex.uni-dortmund.de/hytex/publikationen.html.

Holler, A., Maas, J.-F., and Storrer, A., 2004, Exploiting coreference annotations for text-to-hypertext conversion, *i*n: *Proceedings of the Third International Conference on Language Resources and Evaluation LREC 2004*, Lisboa, pp. 655-658.

Kuhlen, R., 1991, *Hypertext. Ein nicht-lineares Medium zwischen Buch und Wissensbank*, Springer, Berlin et al.

Mayfield, J., 1997, Two-level Models of Hypertext, in: *Intelligent Hypertext,* N. Charles and J. Mayfield, ed., Springer LNCS 1326, pp. 90–108.

Miller, G.A., 1998, Nouns in WordNet, in: *WORDNET: An electronic lexical database,* C. Fellbaum, ed., MA, Cambridge, pp. 23-46.

Kunze, C., Wagner, A., 2001, Anwendungsperspektiven des GermaNet, eines lexikalisch-semantischen Netzes für das Deutsche, in: Chancen und Perspektiven computergestützter Lexikographie, I. Lemberg, ed., Niemeyer, Tübingen, pp. 229-246.

Lenz, E. A., in this volume, HTTL – HYPERTEXT TRANSFORMATION LANGUAGE. A Framework for the Generation of Hypertext Views on XML Annotated Documents.

Lenz, E. A., Birkenhake, B., and Maas, J. F., 2003, Von der Erstellung bis zur Nutzung: Wortnetze als XML Topic Maps, in: *Anwendungen des deutschen Wortnetzes in Theorie und Praxis*, C. Kunze, L. Lemnitzer and A. Wagner, ed., LDV-Forum, 19 (1/2), pp. 113-125.

Lenz, E. A., Storrer, A., 2002, Converting a corpus into a hypertext: An approach using XML topic maps and XSLT, in: Proceedings of LREC 2002: Third International Conference on Language Resources and Evaluation M. Gonzàles Rodríguez, C. Paz Suarez Araujo, ed., pp. 432-436.

Lenz, E. A., Lüngen, H., 2004, Dokumentation: Annotationsschicht: Logische Dokumentstruktur. Technical Report; http://www.hytex.uni-dortmund.de/hytex/publikationen.html

Müller, H., 2004, Stylebook for the Tübingen Partially Parsed Corpus of Written German (TüPP-D/Z); http://www.sfs.uni-tuebingen.de/~fhm/Biblio/stylebook-04.pdf

Pepper, S., Moore, G., 2001, XML Topic Maps (XTM) 1.0. Topic-Maps.Org specification, (March 2001), http://www.topicmaps.org/xtm/1.0/.

Storrer,A.,Wellinghoff, S. (in press): Automated detection and annotation of term definitions in German text corpora. In: Proceedings of LREC 2006, Genoa 275-295, May 2006.

Storrer, A., 2004, Text und Hypertext, in: *Texttechnologie*, L. Lemnitzer, H. Lobin, ed., Stauffenburg, Tübingen, pp. 13-50.

Storrer, A., 2002, Coherence in text and hypertext, in: *Document Design* 3 (2), pp. 156-168.

Witt, A., Goecke, D., Sasaki, F., and Lüngen, H., 2005, Unification of XML Documents with Concurrent Markup. Lit Linguist Computing, 20(1), pp. 103–116.

Zifonun, G., Hoffmann, L. et al., 1997, *Grammatik der deutschen Sprache,* de Gruyter, Berlin/New Y-ork. 2 Bände.