# Exploiting Coreference Annotations for Text-to-Hypertext Conversion

**Anke Holler[1], Jan Frederik Maas[2], Angelika Storrer[2]**

[1] Ruprecht-Karls University of Heidelberg
Department of Computational Linguistics
Karlstr. 2, D - 69117 Heidelberg, Germany
holler@cl.uni-heidelberg.de

[2] University of Dortmund
Department of Cultural Studies, HyTex Project
D – 44221 Dortmund, Germany
angelika.storrer@uni-dortmund.de
jfmaas@gmx.de

**Abstract**

The paper describes an annotation scheme for coreference developed within the application context of text-to-hypertext conversion. In this context coference is used (1) for generating document-internal and cross-document hyperlinks, and (2) for resolving anaphoric expressions in order to achieve cohesive closedness in hypertext nodes. We will argue that for the purpose of cross-document linking it is necessary to separate the annotation of coreference relations from the annotation of anaphoric relations. To account for this requirement, we developed a knowledge-based annotation scheme that relates referential expressions in the text to entities in a knowledge representation, which is modeled using XML Topic Maps.

## 1. Project Framework

Converting linear text documents into documents that can be published in a hypertext environment is a complex task requiring conversion software on the technical side as well as conversion strategies and methods on the conceptual side. In the project HyTex[1], which is the framework of the approach discussed in this paper, we concentrate on principles and strategies for handling conceptual problems of text-to-hypertext conversion such as:

- *Segmentation*: What are the criteria for segmenting documents into text segments to be used as hypertext nodes?
- *Reorganization*: What are the guidelines for generating "cohesive closedness" in hypertext nodes, i.e. 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?
- *Linking*: What are the guidelines and principles for reconnecting the nodes via hyperlinks?

Using XML as the technical basis for hypertext modeling and viewing, the project develops strategies and methods which (semi)-automatically create hypertext layers and views based on text-grammatical annotations. By storing the hypertext as additional document layers, our approach preserves structure and content of the original text documents, and thus provides the reader with the choice between sequential and selective reading modes. The general aim of the project is to support selective hypertext readers in finding coherent pathways through the document network and thus make selective reading and browsing more efficient and more convenient than it would be possible with printmedia. Feasibility and performance of the methodology is tested and evaluated using a German text corpus, which comprises documents that deal with two subject domains, namely "text technology" and "hypertext research" (Lenz & Storrer, 2002).

The central idea of the conversion approach in HyTex is to base strategies for segmentation, reorganization and linking on information coming from two levels:

- On the **document level**, we explicitly markup the text-grammatical structures and relations between text segments, e.g. coreference relations, semantics of connectives, text-deictic expressions, and expressions indicating topic handling.
- On the **domain knowledge level**, we represent the main concepts of this subject domain and their interrelations, using the WordNet model (Fellbaum, 1998) as the conceptual and XML Topic Maps (XTM, 2001) as the technical basis (Beißwenger & Storrer & Runte, 2004; Lenz & Birkenhage & Maas, 2004).

A dynamic-adaptive component that processes logs of usage has been considered but not been put into practice during the current phase of the project. In a later stage, this document usage level would supply information about

[1] The acronym „HyTex" is spelt out as *Hypertextualisierung auf textgrammatischer Grundlage* ('Hypertext conversion on a textgrammatical basis'). The project was launched in November 2001 as part of the research group *Texttechnologische Informationsmodellierung* ('Text-technological information modelling'), cf. http://www.text-technology.de. For more information on the HyTex project see http://www.hytex.info.

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

## 2. Annotation of Coreference Phenomena

The focus of this paper is on an annotation scheme for coreference phenomena. This scheme serves two purposes in our approach: firstly, generating document-internal and cross-document hyperlinks, cf. 2.3., and secondly, resolving anaphoric expressions in order to achieve cohesive closedness in hypertext nodes, cf. 2.4. Focusing on these two tasks, we will now discuss how a proper annotation of the relations of coreference and anaphora can be exploited for text-to-hypertext conversion in the above-mentioned framework. We argue that existing annotation schemes need to be extended in order to meet this task. As a result, a new annotation scheme is proposed that encodes coreference as a relation between the document level and the domain knowledge level. Thereby, it is possible to strictly separate the annotation of the relation of coreference from the annotation of anaphoric relations. Furthermore, the paper describes how the presented scheme can be employed to annotate the sequentially organized documents enclosed in the HyTex text corpus.

### 2.1. Existing Annotation Schemes

Existing annotation formats such as the proposal of the *Text Encoding Initiative* (TEI), the task definition of the *Message Understanding Conferences* (MUC) and the annotation guidelines published by the project *Multilevel Annotation Tools Engineering* (MATE) treat coreference as one specific type of a generalized anaphoric relation. Arguing from a semantics viewpoint, (van Deemter & Kibble, 2001) point at some fundamental problems with this general practice by means of the MUC annotation exercise. They argue that in fact anaphora and coreference are two different things. Coreference constitutes an equivalence relation; anaphoric relations, by contrast, are irreflexive, nonsymmetrical, and nontransitive. Although anaphoric and coreferential relations can coincide, it is not generally the case that all coreferential relations are anaphoric, nor are all anaphoric relations coreferential. For instance, nonreferring NPs can enter anaphoric relations and thus should not be marked as coreferential, cf. (1a). Moreover, the notion of coreference may not be applied to bound anaphora, cf. (1b), and to intensional contexts, cf. (1c).[2]

(1) a. Whenever a solution emerged, we embraced it.
b. Every TV network reported its profits.
c. Henry Higgins, who was formerly sales director of Sudsy Soaps, became president of Dreamy Detergents.

In addition to these linguistic arguments, there are practical reasons that militate in favor of a separated annotation of coreference and anaphora. Investigating a hypertext base, one will encounter cases where two items are coreferential without being anaphoric. For example, two mentions of the same entity, e.g. a person, in two different hypertext documents are coreferential, but will not stand in anaphoric relation. This type of coreference is also termed 'cross-document coreference'; cf. (Baldwin & Bagga, 1998; Mitkov, 2003; Holler-Feldhaus, in press). Thus, marking coreference on the document level only leaves out of account that coreference is established with regard to entities in a real or mental world.
To account for the coreference phenomena observable in a hypertext base, a coreference scheme is needed which does not presuppose that coreferential items are in any case anaphoric as well.

### 2.2. Knowledge-based Annotation of Coreference

As mentioned before, HyTex implements a two-level architecture. The first level comprises the manually annotated documents of the hyperbase, whereas the second level represents the domain knowledge, which is modelled as an XML Topic Map, the so-called TermNet. Exploiting this architecture, we suggest regarding coreference as a relation between expressions occurring in a document and entries of the TermNet. Two expressions occurring in (maybe different) documents of the hyperbase are identified as coreferential if they point to the same term of the domain knowledge model. Since coreference is analyzed as a relation between items of the document level and items of the domain knowledge level, we do not presume any more that two expressions stand in anaphoric relation if they are marked as coreferential.
We will show next how this basic idea is realized by the definition of our annotation scheme.
First of all, relevant terms are marked as a nominal discourse entity by adding the tag <discourseEntity>. For this element two attributes are defined: *deID*, whose value enumerates all discourse entities, and *deType*, which specifies the semantic type of the respective entity.[3]
For annotating coreference, a link element <corefLink> as given in (2) is introduced. <corefLink> describes the relation between a text item given by its *deIDref* value and a referential anchor in the topic-map based TermNet represented by the value of the *tmIDRef* attribute.

(2) <corefLink deIDref = VALUE tmIDRef = VALUE />

The term 'link' in sentence (3a) taken from our text corpus is for example annotated as shown in (3b).

(3) a. Link L verknüpft A mit B im Hinblick auf C.
'Link L connects A with B with regard to C.'

b. <discourseEntity deID="deID_1" deType="nom">
Link </discourseEntity> verknüpft A mit B im Hinblick auf C.
<semRel><corefLink deIDRef="deID_1"
tmIDRef="TermNet-inferiert.xtm#Link"/>
</semRel>

In addition to the <corefLink>-element, a <cospecLink>-element is introduced into the annotation scheme, cf. (4). This element is used to annotate document-internal anaphoric relations. It bears three attributes: *relType*, *phorIDRef* and *antecedentIDRef*. As you can see from

[2] The examples are taken from (van Deemter & Kibble, 2001).

[3] In principle, it is possible to mark entities different from nominals such as abstract objects by using the developed annotation scheme.

example (5), the attribute *antecedentIDRef* describes the antecedent of the anaphor given by the value of the attribute *phorIDRef*, and the attribute *relType* describes the semantic relation (such as substitution, synonymy, hyperonymie etc.) established between the anaphor and its antecedent.

(4) <cospecLink relType = VALUE phorIDRef = VALUE > antecedentIDRef = VALUE />

(5) a. Link L verknüpft A mit B im Hinblick auf C. Verbalsprachliche Symbole, typographische Symbole und Ikonen können als Link fungieren.
'Link L connects A with B with regard to C. Linguistic symbols, typographic symbols and icons can pose as link.'

b. <discourseEntity deID="deID_1" deType="nom"> Link </discourseEntity> L verknüpft A mit B im Hinblick auf C. Verbalsprachliche Symbole, typographische Symbole und Ikonen können als <discourseEntity deID="deID_2" deType="nom"> Link </discourseEntity> fungieren.
<semRel><cospecLink relType="substitution" phorIDRef="deID_2 antecedentIDRef="deID_1"/> </semRel>

By the proposed method, it is easily possible to distinguish between coreferential and cospecified items (Sidner, 1979) in the annotations and to mark coreference independently of anaphoric relations.

### 2.3. Applying the Approach to the Linking Task

In this section, we will illustrate by two examples taken from our corpus how the presented approach has been applied to account for cross-document phenomena during the generation of hypertext layers and views in the HyTex framework.
Generally, two cases of cross-document phenomena have to be distinguished. First, two lexically identical expressions occuring in two different documents really corefer and hence should be marked as coreferential, and second, two lexically identical expressions occurring in two different documents do not corefer, and thus must not be marked as coreferential. The first case is exemplified by (6). Both sentences (6a) and (6c) contain the nominal expression 'Hypertextsystem' ('hypertext system'). By adding a <corefLink> tag whose *tmIDRef* value refers in both cases to the same entry in the TermNet, expressed by the topic map ID TermNet-inferiert.xtm#Hypertextsystem, the observed cross-document coreference is correctly expressed. This is depicted by the example annotations in (6b) and (6d), resp.

(6) a. Das von Kuhlen 1991 skizzierte Grundmodell eines Hypertextsystems orientiert sich am Vorbild von Datenbankmanagementsystemen.
'The basic model drafted by Kuhlen 1991 is geared to data management systems'

b. Das von Kuhlen 1991 skizzierte Grundmodell eines <discourseEntity deID="deID_3" deType="nom"> Hypertextsystems </discourseEntity> orientiert sich am Vorbild von Datenbankmanagementsystemen.
<semRel><corefLink deIDRef="deID_3 tmIDRef= "TermNet-inferiert.xtm#Hypertextsystem"/> </semRel>

c. Ein erstes Hypertextsystem, welches die Grundlage für das World Wide Web bildete, wurde 1989 von Tim Berners-Lee am CERN entwickelt.
'A first hypertext system that established the basis for the World Wide Web was developed by Tim Berners-Lee at CERN.'

d. Ein erstes <discourseEntity deID="deID_4" deType="nom"> Hypertextsystem </discourseEntity>, welches die Grundlage für das World Wide Web bildete, wurde 1989 von Tim Berners-Lee am CERN entwickelt.
<semRel><corefLink deIDRef="deID_4" tmIDRef= "TermNet-inferiert.xtm#Hypertextsystem"/> </semRel>

The sentences in (7) exemplify where one and the same term is used in two different ways and therefore a coreference relation shall not be established between the two terms. In the domain of hypertext research, the term 'annotation' means 'gloss'; whereas in the domain of text technology 'annotation' is a synonym to 'markup'. Consequently, both terms do not stand in a coreference relation. This falls out from our approach since the term 'annotation' expresses two different concepts and thus refers to two different entries in the Topic Map.

(7) a. Unter "Annotationen" werden in der Hypertext-literatur Anmerkungen und Notizen verstanden, die ein Hypertextnutzer während des Rezeptionsvorgangs zu den Inhalten eines Moduls anbringt.
'In the hypertext literature notes taken to the content of the modul by the hypertext user during the reception are understood as "annotations".'

b. Unter <discourseEntity deID="deID_5" deType="nom"> "Annotationen" </discourseEntity> werden in der Hypertextliteratur Anmerkungen und Notizen verstanden, die ein Hypertextnutzer während des Rezeptionsvorgangs zu den Inhalten eines Moduls anbringt.
<semRel><corefLink deIDRef="deID_5" tmIDRef="TermNet-inferiert.xtm#Annotation1"/> </semRel>

c. In der SGML/XML-Terminologie wird der Ausdruck "Annotation" allerdings meist in einem anderen Sinne verwendet, nämlich als Bezeichnung für die Auszeichnung von Dokumenten mittels Markup.
'Following the SGML/XML-terminology, the expression "annotation" is mostly used in a different way, viz. as description for the mark up of documents'

d. In der SGML/XML-Terminologie wird der Ausdruck <discourseEntity deID="deID_6" deType="nom"> "Annotation" </discourseEntity> allerdings meist in einem anderen Sinne verwendet, nämlich als Bezeichnung für die Auszeichnung von Dokumenten mittels Markup.

<semRel><corefLink deIDRef="deID_6"
tmIDRef="TermNet-inferiert.xtm#Annotation2"/>
</semRel>

### 2.4. Generating Cohesive Closedness

As mentioned above, segmentation strategies are applied to text documents in order to get easy access to comprehensible parts of the text. Thus a recipient can navigate through the text without reading from beginning to end but by selectively reading the relevant parts. Since cohesive markers – such as anaphors – may point to information which is located outside a text segment, it is necessary to unlink these markers from the original reading path to allow non-linear reading of the text nodes, and thereby gaining cohesively closed sections, cf. (Kuhlen, 1991). Consider the following example:

(8) Weiterhin unterscheidet er [...]
'Furthermore he differentiates between [...]'

A reader of (8) needs to know to whom the pronoun 'er' refers. This information can be provided exploiting the proposed annotation scheme. The resulting annotation is given in (9). (The entity "deID_7" which is not shown occurs in the preceding text.)

(9) Weiterhin unterscheidet <discourseEntity
deID="deID_8" deType="nom">er
</discourseEntity>
<semRel><cospecLink relType="ident"
phorIDRef="deID_8 antecedentIDRef="deID_7"/>
</semRel> [...]

In the above described scenario, the entity marked as "deID_7" would be inserted as antecedent of the pronoun if requested by the reader.
Depending on the relation type one of the following strategies is chosen:

a. Insertion
b. Linking
c. Expansion of the field of view (showing the next or preceding section)
d. Deletion

The choice of the appropriate strategy assures that the presentation of the information needed to generate cohesive closedness is clear and comprehensible.

## 3. Summary

In this paper, a knowledge-based annotation scheme has been developed that allows annotating coreference phenomena independently of anaphoric relations. Benefiting from a two-level architecture realized in the HyTex framework, we have argued for a coreference annotation that relates expressions of the documents to a WordNet-like model which represents terminological knowledge of the domain investigated. Further, it has been shown how the proposed method can be exploited for the creation of document-internal and cross-document hyperlinks and for generating cohesive closedness during text-to-hypertext conversion.

## References


Baldwin, B. & A. Bagga (1998). Coreference as the Foundations for Link Analysis over Free Text Databases. In Proceedings of the COLING-ACL'98 Content Visualization and Intermedia Represenation Workshop. CVIR'98. August 1998 (pp 19—24).

Beißwenger, M., A. Storrer & M. Runte (2004). Modellierung eines Terminologienetzes für das automatische Linking auf der Grundlage von WordNet. In Kunze, C. et al. (Eds.): GermaNet: Anwendungen des deutschen Wortnetzes in Theorie und Praxis. In LDV-Forum 19, 1/2.

Grishman, R. (1995). MUC-6 Coreference Task Definition. URL: http://www.cs.nyu.edu/cs/faculty/grishman/COtask21.book_1.htm.

Hirschman, L. & N. Chinchor (1997). MUC-7 Coreference Task Definition. URL: http://www-nlpir.nist.gov/related_projects/muc/proceedings/co_task.html.

Fellbaum, C. (Ed.) (1998). WordNet: An Electronic Lexical Database. Cambridge: MIT Press.

Holler-Feldhaus, A. (in press). Koreferenz in Hypertexten: Anforderungen an die Annotation. In Osnabrücker Beiträge zur Sprachtheorie (OBST).

Kuhlen, R. (1991). Hypertext : Ein nicht-lineares Medium zwischen Buch und Wissensbank. Berlin: Springer.

Lenz, E. A., B. Birkenhake & J. F. Maas (2004). Von der Erstellung bis zur Nutzung: Wortnetze als XML Topic Maps. In Kunze, C. et al. (Eds.): GermaNet: Anwendungen des deutschen Wortnetzes in Theorie und Praxis. In LDV-Forum 19, 1/2.

Lenz, E. A. & A. Storrer (2002). Converting a Corpus into a Hypertext: An Approach Using XML Topic Maps and XSLT. In M. Gonzàles Rodríguez & C. Paz Suarez Araujo (Eds.): LREC 2002: Third International Conference on Language Resources and Evaluation, Vol. II (pp. 432–-436).

Mitkov, R. (2002). Anaphora Resolution. London: Longman, Pearson Education.

Poesio, M. (2000). Annotating a corpus to develop and evaluate discourse entity realization algorithms: Issues and preliminary results. In M. Gavrilidou et al. (Eds.): Proceedings of the 2nd International Conference on Language Resources and Evaluation (pp. 211—218).

Sidner, C. (1979). Towards a computational theory of definite anaphora comprehension in English discourse. Ph.D. thesis. MIT.

Sperberg-McQueen, C.M. & L. Burnard (Eds.) (2001). TEI Guidelines for Electronic Text Encoding and Interchange. XML-compatible edition. URL: http://www.tei-c.org/P4X/

XTM: XML Topic Maps (2001). Specification. URL: http://www.topicmaps.org/xtm/1.0/

van Deemter, K. & R. Kibble (2001). On Coreferring: Coreference in MUC and related annotation schemes. In Journal of Computational Linguistics 26, 4 (pp. 629—637).