ELSEVIER

International Journal of Medical Informatics 53 (1999) 1–28

International Journal of Medical Informatics

Discourse structures in medical reports—Watch out! The generation of referentially coherent and valid text knowledge bases in the MEDSYNDIKATE system

Udo Hahn [a,*], Martin Romacker [a,b], Stefan Schulz [a,b]

[a] *Freiburg University, Computational Linguistics Lab, Werthmannplatz 1, D-79085 Freiburg, Germany*
[b] *Department of Medical Informatics, Freiburg University Hospital, Stefan-Meier-Str. 26, D-79104 Freiburg, Germany*

Received 15 February 1998; received in revised form 20 March 1998; accepted 25 March 1998

Abstract

The automatic analysis of medical narratives currently suffers from neglecting text structure phenomena such as referential relations between discourse units. This has unwarranted effects on the descriptional adequacy of medical knowledge bases automatically generated from texts. The resulting representation bias can be characterized in terms of incomplete, artificially fragmented and referentially invalid knowledge structures. We focus here on four basic types of textual reference relations, *viz*. pronominal and nominal anaphora, textual ellipsis and metonymy and show how to deal with them in an adequate text parsing device. Since the types of reference relations we discuss show an increasing dependence on conceptual background knowledge, we stress the need for formally grounded, expressive conceptual representation systems for medical knowledge. Our suggestions are based on experience with MEDSYNDIKATE, a medical text knowledge acquisition system designed to properly deal with various sorts of discourse structure phenomena.

Keywords: Natural language processing: text understanding; Knowledge acquisition from texts; Knowledge representation: description logics; Ontology and terminology: pathology domain

1. Introduction

With the overall diffusion of electronic text processing technology in clinical offices and at the physician's workplace and, more recently, the unlimited access to text resources in the Internet, a vast potential for medical information supply arises. The natural language processing community, therefore, faces the challenge to meet the requirements of cursory as well as in-depth analysis of large

* Corresponding author. Tel.: +49 761 2033255; fax: +49 761 2033251; e-mail: hahn@coling.uni-freiburg.de

1386-5056/99/$ - see front matter
PII S1386-5056(98)00091-4

corpora of texts. So far, its response to the urgent needs of real-world text processing in the medical domain has shaped in two text analysis paradigms, each one corresponding to different information needs.

One branch, mainly with an information retrieval background, aims at determining a subset of *relevant documents* from a large text collection by means of lexical co-occurrence statistics and probabilistic measures incorporating lexical distribution data (a survey with focus on the health care domain is given by [1]). Linguistic knowledge comes into play at a modest level only, *viz*. in terms of lists of stop words and simple morphological stemming procedures. Usually no grammar and no *a priori* domain knowledge is considered, except for medical classifications and thesauri (such as ICD, MeSH or SNOMED). The approach is appealing for the task of reference retrieval, while it seems hardly extendible to more sophisticated forms of content-oriented text analysis.

This is the major concern of methodologies which aim at the extraction of *relevant facts* from a large document collection by means of efficient grammar and parsing devices at the level of syntactic linguistic analysis. These mechanisms also make complementary use of large commonsense lexical repositories (such as WordNet [2]) as well as comprehensive domain-specific taxonomies or ontologies (such as UMLS [3], or GALEN [4] in the medical domain), which provide for semantic and conceptual knowledge. While proponents of this approach vary with respect to the emphasis they place on structural aspects of syntactic processing [5–8] or on additional inferential processing of domain knowledge [9,10], it is striking that they converge on neglecting the influence of discourse structure phenomena on text analysis, such as reference-establishing anaphoric relations between sentences (for a survey, cf. [11]). It seems then that they share the implicitly held assumption that medical texts (discharge summaries, findings reports, etc.) can be considered a sequence of phrases or sentences lacking any further *inter*dependencies.

In this article, we shall challenge this view. We claim that medical texts, as any other text genre, exhibit textual structures and that disregarding these structural relations will lead to artificially fragmented, incomplete or even invalid content representations. As a consequence, the results of sentence-centered medical text analysis would be of limited value only. Moreover, considering the regularities underlying the analysis of textual phenomena in more depth, we will provide evidence that in order to properly account for discourse structure phenomena quite sophisticated knowledge representation structures have to be supplied. Finally given the appropriateness of a knowledge-based approach to text analysis, we will formulate concept modeling requirements for knowledge representation platforms capable of adequately supporting the analysis of these text structures.

This article is organized as follows. In Section 2 we will introduce the architecture of MEDSYNDIKATE, a knowledge acquisition system for German pathology texts, by concentrating on the exposition of its major knowledge sources and underlying design decisions. In order to illustrate the basic principles of linguistic analysis, we will discuss in depth the processing steps a sentence undergoes from lexical scanning to conceptual interpretation in Section 3. As we proceed with the analysis of sentences of a selected text fragment in isolation, we are able to demonstrate the unwarranted implications of not textual structures accounting for and show how incomplete, artificially fragmented (referentially incoherent) and referentially invalid text knowledge representation structures are likely to emerge. Our approach to text analy-

sis is based on the centering model which is introduced in Section 4.1. We then turn to the consideration of the major reference phenomena one encounters in medical texts. We start with pronominal anaphora in Section 4.2, turn to nominal anaphora in Section 4.3, consider textual ellipsis in Section 4.4 and end up with the discussion of metonymies in Section 4.5. In all of these sections we will give a descriptive account of the phenomena involved and the regularities underlying proper text understanding. We also provide sample analyses, each showing how the recognition of the particular type of discourse structure improves the quality of the resulting target representation structures in the underlying text knowledge base. In order to lend support to our argument that the recognition of text coherence structures is at all relevant medical text analysis we report in Section 4.6 on an empirical investigation of findings reports from a large clinical text database. Since we regularly identify the prerequisites for proper text analysis at the grammar as well as the concept description level, we are able to derive basic requirements for adequate representation languages serving the needs of text structure understanding in (Section 5). Finally, in Section 6 we will place our approach in the context of related research efforts in the field of medical language processing.

2. MEDSYNDIKATE: requirements and system design

At our lab, we are currently developing a large-scale text knowledge acquisition system called SYNDIKATE (SYNthesis of DIstributed Knowledge Acquired from TExts). The SYNDIKATE core system is currently adapted to serve two application domains. The first prototype, ITSYNDIKATE, covers portions of the information technology (IT) domain and analyzes product reviews as well as test reports selected from various technology magazines. The second one, MEDSYNDIKATE, operates in the medical domain and treats gastro-intestinal pathology reports as selected from the clinical information system of the University Hospital in Freiburg. The domain-independent task of the SYNDIKATE system is to acquire from each text a maximum number of *simple facts* ('The size of the fragment is *4 mm*'.), *complex propositions* ('In *all* samples from the subcutaneous fatty tissue lymph node metastases of an undifferentiated carcinoma are found.') and *evaluative assertions* such as physicians' staging and grading assertions ('A chronical duodenitis of an *average* degree.'). The task of SYNDIKATE's text understanding kernel is then to map each incoming text into a corresponding *text knowledge base*, which contains a formal representation of the text's contents (facts, complex propositions and evaluative assertions). Given such a task-independent representation layer, we are then able to exploit this knowledge in MEDSYNDIKATE for various medical information services considered relevant from the viewpoint of hospital practice. For instance, the knowledge can be used for common information retrieval applications such as the automatic classification of findings according to ICD or other clinical coding systems, or it may be used for document or text passage retrieval as well as document filtering. It is additionally useful for inferentially supported fact retrieval and may even be envisaged as a source for knowledge-intensive applications such as text summarization [12]. One of the high priority future applications, however, aims at the combination of a large collection of different text knowledge bases into a single text knowledge pool, a procedure we refer to as *text knowledge base synthesis*. Given such a level

of knowledge aggregation, it would become feasible to provide medical records for each patient whose data was automatically collected and formatted.

At the level of system architecture (cf. Fig. 1), we therefore distinguish broadly between the language processing component, the so-called *text understanding kernel* system and the knowledge transformation component, a library of application-oriented procedures that operate on text knowledge bases. In this paper, however, we will concentrate on the text understanding kernel proper, in which we distinguish two major components: the parser carries out the language analysis (mapping text strings to text knowledge representation structures); while the *learner* is concerned with the ongoing augmentation of the domain knowledge base by new concepts

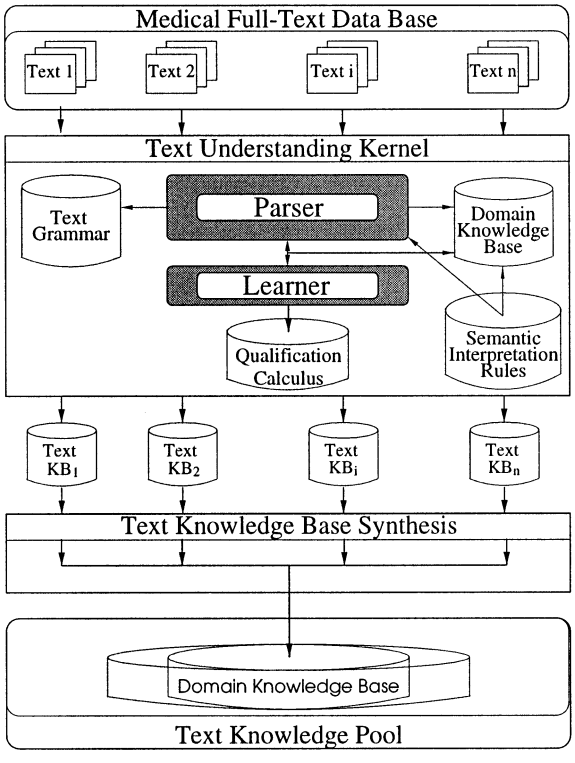


Fig. 1. Architecture for medical text understanding.

(for more details of the learning component and, in particular, the qualification calculus, a formal system for the assessment of different learning hypotheses, cf. [13]). The parsing component uses *grammar* knowledge that incorporates sentential as well as textual regularities of the underlying natural language and it uses knowledge about the underlying *domain*. At the intermediary semantic level, we supply *semantic interpretation* rules which provide transformations from language-oriented conceptual representations to the more canonical level of a text knowledge base. Note already that semantic representations and conceptual representations are based on a common representation format in terms of description logics.

2.1. Design principles

Since our focus in this article is on the language processing aspects of MEDSYNDIKATE, we shall briefly review the design principles it is built on. They reflect the exposure of our system to so-called *real-world* (or naturally-occurring) full-texts, i.e. reports, notes, memos, etc. taken from external sources (like a clinical text database) without any kind of pre-editing. Given such natural input data and provided the overall goal is to acquire a maximum degree of knowledge encoded in these texts, we set up the following design criteria (cf. also [14]):

2.1.1. Robustness

Robust language processors are able to deal with real-world textual input in a *fail-soft* manner—ideally, no matter how complicated or idiosyncratic its linguistic structure is, how deficient their knowledge sources are, nor how corrupted and fragmentary their input is. The degree of understanding they achieve will be dynamically limited to those portions they can deal with, while those out

of their scope are disregarded without causing the system to block (hence, *partial* or *limited-depth understanding*). Two major phenomena must be taken into consideration. *Ungrammaticality* refers to erroneous language input such as typos, agreement failures, or violations of case restrictions. Robust processing requires then the relaxation of some of the grammar constraints involved and the making of the right guesses in order to 'repair' the failures being recognized on the fly. *Extragrammaticality*, on the other hand, refers to perfectly well-formed language input for which, nonetheless, no appropriate lexical, grammatical, or conceptual specifications yet exist in the system. Robust processing then amounts to somehow geting around these specification gaps at the price of a limited depth of understanding[1]. Also, especially in the analysis of medical texts, one often encounters a densely written, almost *telegrammatic* style in physicians' narratives. This is a genre-specific phenomenon which requires grammars to either be specially tuned to this kind of jargon or particularly flexibilized to handle various sorts of exceptions to standard well-formedness conditions as quasi-norms.

2.1.2. Knowledge-based processing

Given that the target structure of text analysis is a representation of the text's contents which can be reasoned about, the task of the language processor consists not only in deriving the grammatical structure of each utterance but also covers its interpretation in terms of meaning and knowledge representation structures. Hence, two major knowledge sources come into play. Linguistic knowledge about the structural aspects of the underlying natural language, the *grammar* system, and conceptual knowledge about the universe of discourse, the *domain knowledge* base. The parser has to account for the fact that both knowledge sources interact heavily in the course of the language understanding process. The final result of text analysis consists of a so-called *text knowledge base*, which, at the beginning of the analysis, is just a copy of the domain knowledge base that then gets continuously augmented by new facts, propositions, evaluations and new concepts contained in the text under consideration.

Our concern about powerful reasoning capabilities associated with the representation of domain knowledge has led us to adopt a *description logics* framework for knowledge representation (for a survey, cf. [15]). Due to the fairly developed requirements these formalisms impose on the corresponding domain models, the majority of terminologies and classifications already available in the medical domain cannot be used directly for domain modeling and representation. We have, however, taken a pragmatic position during ontological engineering in order to incorporate available knowledge structures into the specifications of our knowledge base as much as possible.

2.1.3. Text structures

The influence of referential chaining between utterances by way of text coherence mechanisms has long been neglected in the natural language parsing community. Reference at the text level is centered around the broad notion of anaphora and includes diverse phenomena, e.g. the use of pronouns or definite noun phrases to refer to already introduced discourse units. The problems text

[1] Of course, another strategy could be to actively remedy existing specification gaps by learning mechanisms for grammar or domain knowledge and, thus, to continuously improve the coverage of the system's knowledge sources. But the learning task involved is currently too complex to achieve the desired results at both system levels. In SYNDIKATE we currently aim to improve the coverage of the domain knowledge base by concept learning mechanisms [13].

phenomena cause (unless they are properly accounted for) have a strong impact on the adequacy of the representation structures resulting from natural language processing and are centered around the notions of incomplete, invalid and incohesive knowledge bases. *Incomplete* knowledge bases emerge when references to already established discourse entities are simply not recognized, as in the case of *pronominal anaphora* (cf. Section 4.2). *Invalid* knowledge bases emerge when each entity which has a different denotation at the text surface is treated as a formally distinct item at the symbol level, although it refers literally to the same entity. These false referential descriptions will be illustrated in the discussion of *nominal anaphora* (Section 4.3) and *metonymies* (Section 4.5). Finally, *incohesive* or artificially fragmented knowledge bases emerge when entities which are linked by various conceptual relations at the knowledge level occur in a text such that an implicit reference to these relations is made, although this is not made explicit at the symbol level of the text knowledge base. These lacking referential relations will be illustrated in the discussion of textual ellipsis (Section 4.4) and metonymies (Section 4.5). It is interesting to note here that the arguments concerning the necessity of adequately recognizing text structures are valid primarily at the level of text knowledge representation structures. However, tracking the proper referents becomes also immediately relevant for the adequate semantic interpretation and even syntactic processing (selectional restrictions, theta patterns, etc.) of subsequent utterances.

2.2. The PARSETALK performance grammar framework

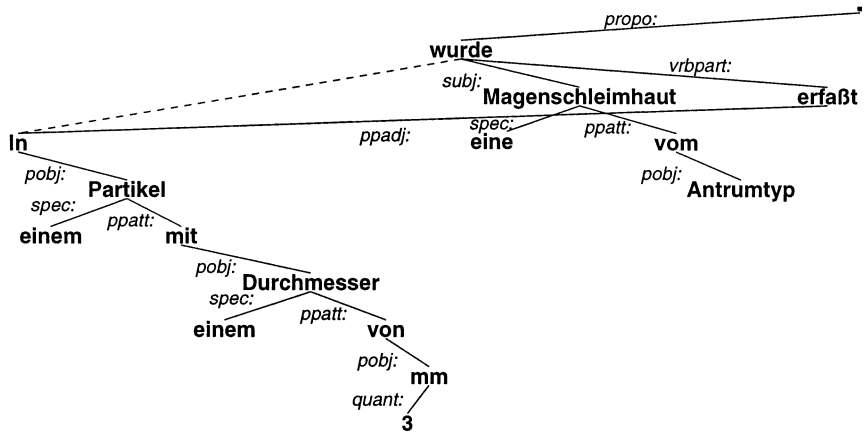
Trying to combine the above requirements—robustness, close interaction between grammar and domain knowledge, integration of text phenomena—into the currently dominating paradigms of language engineering (mainly, finite-state devices [16]) or language theory (mainly, unification-based formalisms [17]) turned out to be extremely difficult and cumbersome. While language engineering approaches are strong in terms of robustness, they are not concerned with integrating domain knowledge and accounting for text structure issues at all. Conversely, language theory approaches require perfect, i.e. complete and deep specifications (and are, therefore, unable to respond to the robustness requirement), incorporate a low-profile, functor-argument-based style of semantics without considering deeper inferencing issues involved in reasoning about domain knowledge, and tend to widely ignore the impact of discourse phenomena in terms of text understanding.

We, therefore, decided in favor of a radical re-design in terms of a so-called *performance grammar*. On the one hand, a performance grammar contains *declarative* knowledge like any other natural language grammar formalism, e.g. part-of-speech information, morphosyntactic features (for gender, number, tense, etc.). On the other hand, and this is quite uncommon in the entire natural language processing community, we consider the *procedural* aspects of how grammar specifications are used an integral part of *grammar knowledge*, too. Since we treat data (grammar) and procedures (parsing) on a par, this homogeneous view is best realized by an *object-oriented* specification of the grammar and a corresponding object-oriented implementation of the parser, an approach formally based on the actor computation model [18]. Considering the above requirements, the distribution of a complex computation (e.g. the determination of a complete parse tree for a sentence) to single objects (e.g. a set of

autonomous lexical processes) eases the generation of partial results in a very natural way. Partial parse trees, e.g. for noun phrases or main clauses omitting embedded clauses, can already be interpreted although a complete solution is still lacking, either because it is infeasible (computationally too complex) or because it is impossible to compute (due to lacking specifications). Together with the additional advantages of this paradigm, e.g. the inheritance in object hierarchies or the encapsulation of computation processes, this specification approach offers some inherent opportunities for dealing with specification gaps without system breakdowns. While this is a highly rewarding feature for any robust processing, further support for partial text analysis requires additional explicit descriptions at the level of a performance grammar.

The performance grammar we use in the SYNDIKATE system, the PARSETALK system [19,20], combines this object-oriented specification framework with a strictly lexical approach to grammar encoding. Lexicalization of knowledge about language constitutes a natural unit of decomposition and, as will become evident below, also provides a high potential for partial analyses. All grammar knowledge resides in lexicalized specifications, so-called *word actors*. Word actors contain valency descriptions which state for each lexical head (e.g. a noun) the kind of modifiers it may accept (e.g. determiners, quantifiers, adjectives, noun phrase attributes), the morphosyntactic and semantic requirements it imposes on each possible modifier and the word order constraints that must hold between different modifiers. A binary *dependency* relationship between a head and a modifier based on the fulfillment of these declarative constraints is established by local computations, only involving the head and the specific modifier. Unlike phrase structure grammars which require a complete coverage, dependency grammars inherently allow for incomplete analyses, since unspecified or underspecified modifiers may not succeed in finding their appropriate head. Hence, partial analysis does not affect the overall pursuit of the analysis.

The procedures involved in determining a concrete dependency relation are specified in terms of a particular message passing protocol we elaborate on in Section 4. In short, a word actor representing a concrete lexical item in the parser asks its left context whether another word actor is capable of accepting the requesting actor as a possible modifier. The query is concurrently passed in a linguistically legitimate way (i.e. searching the outer right 'rim' of the already built dependency graph) to any possible head. Each word actor being addressed carries out local constraint computations reflecting the dependency criteria from above. When a word actor arrives at a positive evaluation of the request, it sends an accept message to the querying word actor (if more than one message arrives at the querying actor, a structural ambiguity has been detected which results in multiple structural readings). Upon the reception of an acceptance message the modifier creates the corresponding dependency relation by linking itself with the determined head. Additional protocols for several processing strategies relate to, e.g. *partial* parsing (by which specification gaps are accounted for), *predictive* parsing (by which word classes are predicted to occur during the incremental parsing process in the still unprocessed right context, given in the already processed left context), *preferential* parsing (which makes linguistically plausible selections from sets of ambiguous readings), or *referential* parsing (which establishes reference relations between different utterances). A detailed description of these protocols is available in [21,22]. In Sections 4.2, 4.3, 4.4



In a [...] particle with a diameter of 3 mm -- was a gastric mucosa of the antrum type -- seized

Fig. 2. Dependency graph for sentence (1).

and 4.5 we will discuss in some detail the protocols involved in referential parsing, while in the following section we elaborate on the basic protocol for establishing dependency relations at the level of sentence analysis.

3. Sentence analysis

We now turn to the discussion of a concrete fragment of a pathology report. It contains all the types of text phenomena we will be dealing with in this article. The way we discuss this fragment, however, proceeds in two steps. In this section, we will consider the procedures underlying a sentence-level analysis only. First, we will introduce the basic dependency protocol, and some semantic interpretation rules we apply, thus, motivating the resulting conceptual representation structures for the first sentence in considerable depth. The remaining sentences are dealt with merely at the level of the (quite deficient) conceptual interpretation. Hence, we demonstrate how artificially fragmented, incomplete and invalid text knowledge representation structures emerge in the course of sentence-centered analysis. In Section 4 we will then turn to remedy these shortcomings and introduce the basic mechanisms which restore the referential linkage between the utterances of the text fragment by means of more adequate text parsing procedures.

1. In einem reiskorngroßen Partikel mit einem Durchmesser von 3 mm wurde eine Magenschleimhaut vom Antrumtyp erfaßt. (A gastric mucosa of the antrum type was seized in a rice-grain-sized particle with a diameter of 3 mm.)
2. Sie zeigt ein ödematöses Stroma. (It reveals an edematous stroma.)
3. Die Schleimhaut wird zudem dicht von Lymphozyten infiltriert. (The mucosa is, moreover, densely infiltrated by lymphocytes.)
4. Im Oberflächenschleim konnten Helicobacter-pylori nachgewiesen werden. (Helicobacter-pylori could be identified in the surface mucus.)
5. Der Patient muß weiterhin bioptisch kontrolliert werden. (The patient must still be checked with a biopsy).

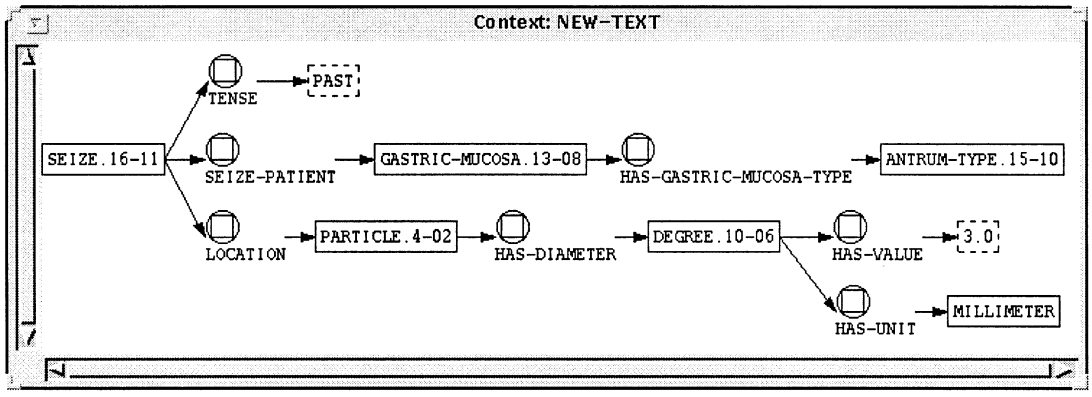


Fig. 3. Concept graph for sentence (1).

The result of the syntactic analysis for the first of these sentences is depicted in the dependency graph in Fig. 2. Labeled solid lines indicate a dependency relation between two words, with the type of the dependency relation being indicated by the label. As the German language has a relatively free word order, the position of a modifier may be fixed by a syntactic head not immediately preceding this modifier. This kind of positional dependency is represented by a dashed line. Both kinds of edges occur in Fig. 2. The finite verb form `wurde` (‘*was*’) governs the participle `erfaßt` (‘*seized*’) via the syntactic relation **vrbpart**[2]. Additionally, `wurde` fixes the position of the prepositional phrase (henceforth, PP) ‘*In einem Partikel…*’, as indicated by the dashed line between `wurde` and `In`, the latter being the syntactic head of the PP. The PP itself, however, is governed by `erfaßt` via the dependency relation **ppadj** as the participle’s prepositional adjunct.

Fig. 3 depicts the concept graph representation for sentence (1) in the common graphical format of description logics. Instantiations of concepts are visualized by rectangles, dashed rectangles contain atomic symbols, whereas the labeled and directed arrows represent instance roles. The numbers attached to each instance are composed of two kinds of information, namely, first, the sentence position of the word denoting that particular instance and, second, the unique identifier of that instance in the knowledge base.

The concept graph provides sort of a ‘simplified’ representation structure compared with the ‘deeper’ dependency graph. The mapping from the syntactic representation level of dependency graphs to the conceptual one of concept graphs is carried out incrementally by semantic interpretation rules. We will demonstrate the interaction between the syntactic analysis and the semantic interpretation with respect to the verb `erfaßt` in sentence (1). Each word actor not only contains grammatical feature information relevant to the corresponding lexeme (e.g. part of speech, morphosyntactic features, valency frame for modifiers, restrictions on the modifiers’ word order, etc.) but also makes available a repertoire of protocols especially adjusted to the corresponding part of speech. Based on these protocols each word actor communicates

[2] Courier fonts indicate the lexical form of a word or its corresponding word actor, **bold** fonts stand for dependency relations and sentence fragments we refer to are quoted in ‘*italics*’. Knowledge base objects will be referred to by SMALL-CAPS.

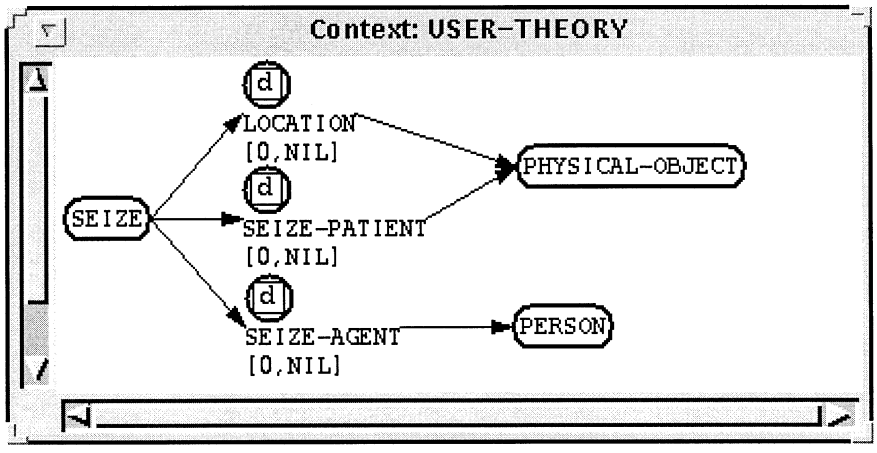


Fig. 4. Concept graph for the SEIZE activity.

with other (word) actors by message passing in order to determine dependency, referential and other kinds of relations. Note also that most of the word actors (all those representing an open-class lexical item such as nouns, verbs, or adjectives) have a conceptual correlate assigned to them, which gets instantiated upon their creation in the parsing process. As an example, consider the conceptual correlate SEIZE.16-11 of the word actor for `erfaßt`. The conceptual representation of the concept SEIZE classifying SEIZE.16-11 is depicted in Fig. 4. Rounded rectangles denote concepts and labeled arrows denote roles, the latter carry information about possible number restrictions attached to that role, as well as information whether a role is definitory (necessary and sufficient, indicated by D) or just implied (necessary). The terminological definition of the surgical procedure SEIZE can be rephrased in the following way: the filler of the LOCATION and the SEIZE-PATIENT slot must have the conceptual type PHYSICAL-OBJECT, whereas the SEIZE-AGENT must belong to the conceptual type PERSON[3]. In principle, an unlimited number of locations, patients and agents are allowed. These sortal restrictions will later be used to reduce the number of structural ambiguities during the parsing process.

In order to determine its syntactic dependencies the word actor of `erfaßt` initiates a *local* search for its potential head. The goal actor must hold a valency in its grammatical specification that can be occupied by `erfaßt`. The protocol applied to check for a dependency relation corresponds to a multi-step message exchange between the word actors involved (for a more technical description of this protocol, cf. [21]):

1. `erfaßt` sends a query to the immediate left context asking whether a participle valency of a preceding word can be occupied. This query is simultaneously passed along the right 'rim' of the dependency

[3] With AGENTS and PATIENTS we refer to representational constructs which have their own status as common linguistic thematic role names, each one adapted to the particular activity to be described, e.g. SEIZE-AGENT or SEIZE-PATIENT. While an AGENT denotes an acting entity (e.g. one carrying out the SEIZE action), a PATIENT stands for an entity which is affected by the action (e.g. the object being seized in the SEIZE action). In the medical domain, unfortunately, some confusion may arise, since the mention of *patient* usually implies reference to a person receiving medical treatment. Unless otherwise indicated, we always refer to the thematic role meaning described above when using the term PATIENT.

graph already established by the word actors `Antrumtyp, vom, Magenschleimhaut` and `wurde`. No word actor other than those mentioned can be reached by `erfaßt`.

2. Each of the enumerated word actors checks locally and concurrently whether its individual valency restrictions (in our case, those for the word classes NOUN, PREPOSITION and AUXILIARY) allow for the acceptance of the participle as a modifier. During this check grammatical and semantic/conceptual constraints imposed by the underlying dependency predicate must be simultaneously satisfied.
3. PREPOSITIONS have no participle valency, although NOUNS have such a valency, in this case, word order constraints are violated (no participle may follow a noun in German). So, dependency relations are precluded in both cases. Hence, in the given example, only the auxiliary `wurde` can positively evaluate the dependency predicate and, therefore, passes an acceptance message to `erfaßt`.
4. Finally, `erfaßt` establishes the dependency relation **vrbpart** to `wurde`.

This basic protocol roughly described above can easily be extended to realize structural ambiguity and partial analysis (by means of the so-called skipping behavior [22]). Partial understanding based on the latter protocol mode is evident from the fact that the lexeme `reiskorngroß` (‘*rice-grain-sized*’) for which neither lexical nor conceptual specifications yet exist, as a consequence, does not show up in the dependency graph of Fig. 2. It also implies that we do not have a conceptual interpretation for this item in the text knowledge base. Further protocol extensions concern predicative, preferential and referential analyses which are elaborated in more depth in [22].

Let us now turn to the mediating role of semantic interpretation rules. As already mentioned in step (2) above, a successful check for the participle valency between word actors incorporates a check for conceptual integrity of the entities involved in the text knowledge base. The word actor wurde (‘*was*’) governing `erfaßt` (‘*seized*’) is marked with a PASSIVE voice feature therefore an interpretation of the syntactic **subject** (`Magenschleimhaut`) (‘*gastric mucosa*’) of the auxiliary verb `wurde` (‘*was*’) as the **direct-object** of the dependent full verb `erfaßt` attempted by a semantic rule turning passive readings into active ones. After this normalization, as a consequence of directly linking syntactic structures to conceptual ones, the syntactic role **direct-object** is projected on the corresponding thematic role (SEIZE-PATIENT) of the verbal concept (SEIZE) (in Fig. 3, the instance GASTRIC-MUCOSA.13-08 of the text knowledge base is the conceptual correlate of the word actor `Magenschleimhaut`). The establishment of dependency relations at the syntactic level requires the sortal constraints associated with thematic roles not to be violated (in our example, no sortal conflicts arise, because PHYSICAL-OBJECT subsumes GASTRIC-MUCOSA which is the concept type of the instance GASTRIC-MUCOSA.13-08).

Another prepositional phrase contained in sentence (1), ‘*vom Antrumtyp*’ (‘*of the antrum type*’), illustrates how the immediate coupling of syntactic analysis and semantic interpretation helps to constrain potential ambiguities—provided a fine-grained domain model is available. In German a prepositional phrase introduced by ‘*von*’ (‘*of*’, ‘*by*’) might thematically contain an AGENT which would allow a PP attachment to the verb (i.e. the SEIZE-AGENT of the SEIZE activity). As the concept type of the instance ANTRUM-TYPE.15-10 does not fulfill the conceptual constraints imposed on the filler of SEIZE-AGENT, i.e. it is not subsumed by PERSON, a syntactically possible dependency structure is rejected on the basis of a conceptual constraint.

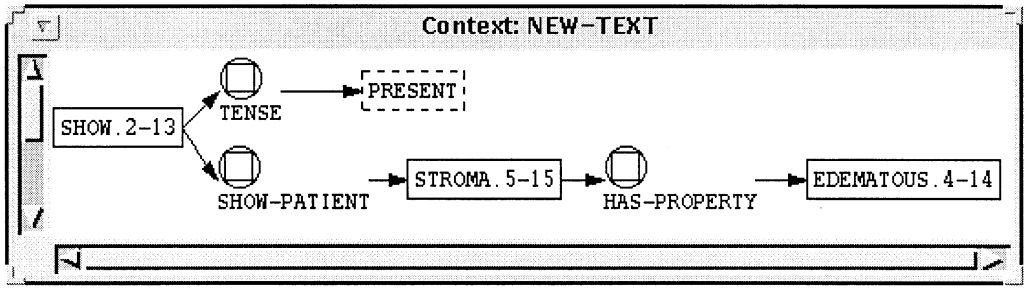


Fig. 5. Concept graph for sentence (2).

In order to obtain a valid conceptual representation various semantic interpretations have to be carried out. For sentence (1), we may distinguish the following ones: (a) as just discussed, the projection of the syntactic subject `Magenschleimhaut` ('*gastric mucosa*') within the passive onto the thematic role SEIZE-PATIENT of the concept SEIZE (representing the corresponding surgical procedure); (b) the determination of the conceptual relation(s) between the head and the modifiers in all PPs considering syntactic regularities and conceptual constraints as defined by the preposition itself (IN-LOCAL, for example, allows LOCATION, FUTURE-TIME-POINT, etc.). This amounts to the following triples: (`erfaßt, in, Partikel`) ('*seized-in-particle*'), (`Partikel, mit, Durchmesser`) ('*particle-with-diameter*'), (`Magenschleimhaut, vom, Antrumtyp`) ('*gastric mucosa-of-antrum type*') and (`Durchmesser, von, mm`) ('*diameter-of-mm*'); and (c) the creation and conceptual integration of an instance of the special concept type DEGREE incorporating the unit of measurement and its value (cf. [23] for more details).

Abstracting away from the details of syntactic analysis, let us now briefly summarize the conceptual representations that can be achieved by grammatical analysis at the sentence level only. For sentence (2), an incomplete conceptual representation is created due

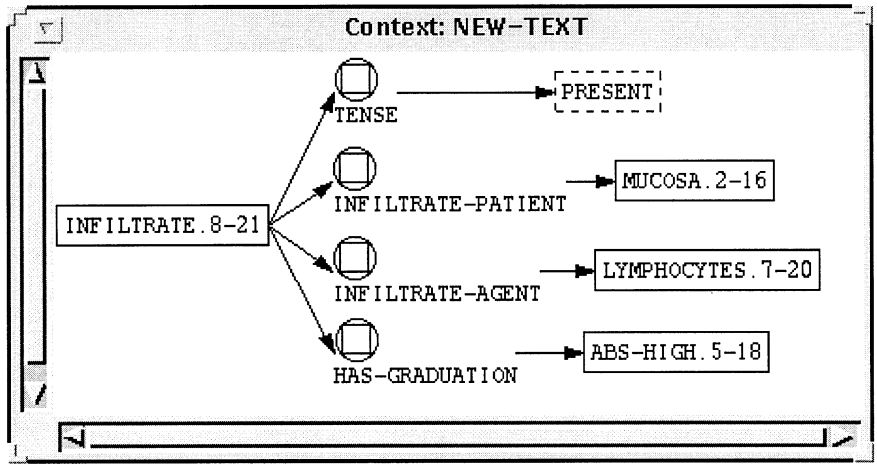


Fig. 6. Concept graph for sentence (3).

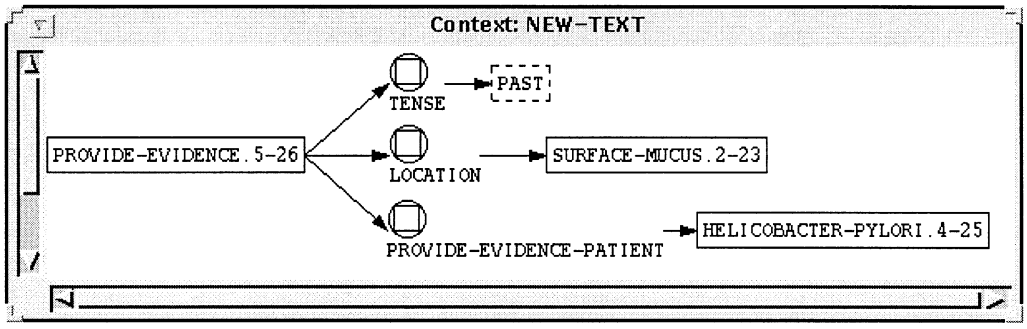


Fig. 7. Concept graph for sentence (4).

to the unknown referent of the personal pronoun `Sie` (‘*if*’) (cf. Fig. 5). The only information left is that some ‘*edematous stroma*’ has been detected but its relation to ‘*gastric mucosa*’ is lost entirely. The conceptual representation for sentence (3) is depicted in Fig. 6. By analogy with sentence (1), the passive voice necessitates a semantic interpretation of the text knowledge base instance MUCOSA.2-16 created for the word actor `Schleimhaut`. Hence, MUCOSA.2-16 is mapped to the thematic role INFILTRATE-PATIENT. In contradistinction to sentence (1), the PP introduced by the preposition `von` is conceptually interpreted in terms of the INFILTRATE-AGENT. The intensity of the infiltration is indicated by ABS-HIGH.5-18, saying that on an absolute scale the intensity of infiltration is high (cf. [23]). A new knowledge base item, MUCOSA.2-16, is introduced, because the coreference relation with GASTRIC-MUCOSA.13-08 is not recognized, thus generating an invalid conceptual entity. Sentence (4) shows no intrasentential phenomena that have not already been mentioned for the previous sentences (cf. Fig. 7). Only for the PP with the head `im` can a location-directed semantic interpretation rule be applied. All other mappings proceed directly from the syntax to the conceptual representation if the

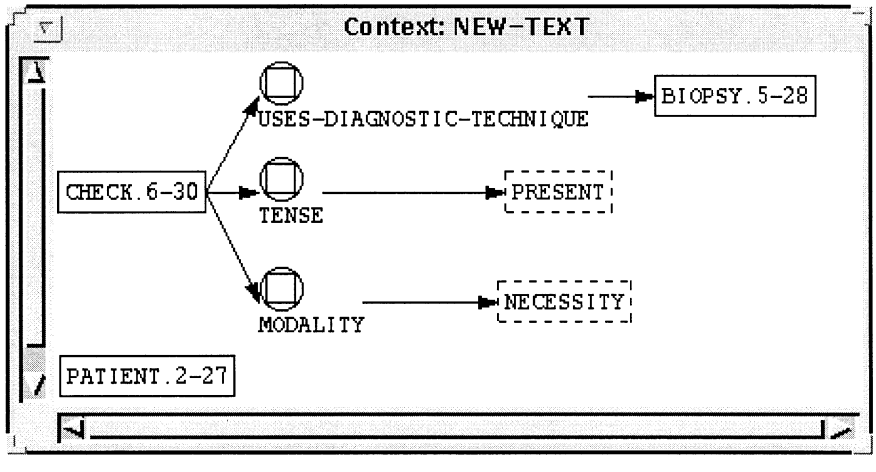


Fig. 8. Concept graph for sentence (5).

corresponding entities are specified. Note, however, that in this representation structure it remains entirely undetermined as to which entity THE SURFACE MUCUS.2-23 belongs. Finally, in sentence (5) a modal verb and an auxiliary have to be interpreted correctly, thus shifting the tense and the modality feature to the conceptual correlate of `kontrolliert` in the text knowledge base, *viz*. CHECK.6-30. Due to sortal constraints arising from the domain knowledge, PATIENT.2-27[4] is not a legal filler for the CHECK-PATIENT slot which would be the appropriate interpretation of the **subject** dependency relation possible between `kontrolliert` and `Patient`. Also, for sentence (5) we have an incohesive and invalid conceptual representation. It is invalid, as a referentially implausible knowledge base entity is created in the course of processing—PATIENT.2-27 is initialized, though its organ structure, GASTRIC-MUCOSA.13-08, is actually referred to. It is incohesive, as a consequence, because the referential relationship indicating that the GASTRIC-MUCOSA.13-08 is the proper entity that is subject to further checks is missed (`Patient` was only introduced as a phenomenon of figurative speech).

Summarizing the shortcomings of these analyses, the conceptual representations being generated for each sentence tend to be *incomplete* (as in cases of pronominal anaphora, because the reference to an already introduced discourse entity is missed). They also introduce different conceptual items for the same entity (as, e.g. with MUCOSA.2-16 and GASTRIC-MUCOSA.13-08 in sentence (3) and (1), respectively, or with PATIENT.2-27 for GASTRIC-MUCOSA.13-08 in sentence (5) and (3), respectively), thus resulting in *invalid* representation structures. At the same time, the representation structures appear artificially *fragmented*, as it is not made explicit that the SURFACE-MUCUS.2-23 can conceptually be related to GASTRIC-MUCOSA.13-08, thus relating sentence (4) with sentence (3)—assuming the nominal anaphor `Die Schleimhaut` ('the mucosa') has been properly resolved to GASTRIC-MUCOSA.13-08. Similarly, the CHECKing relation between GASTRIC-MUCOSA.13-08 and BIOPSY.5-28 is left unconsidered (relating sentence (5) with sentence (4)—assuming the textual ellipsis between SURFACE-MUCUS.2-23 and GASTRIC-MUCOSA.13-08 in sentence (4) and (3), respectively, has been properly resolved). Hence, we have gathered sufficient evidence for the claim that the representation structures resulting from sentence analysis only are likely to become deficient.

[4] PATIENT.2-27 is an instance of the concept type PATIENT which represents any person receiving medical treatment.

4. Text analysis

In the following section, we shall discuss four different types of text phenomena which are commonly used to establish referential links between the utterances constituting a lengthy text. After the introduction of the centering model underlying the tracking of local coherence in discourse in Section 4.1, we start in Section 4.2 with the consideration of pronominal anaphora and then turn in Section 4.3 to nominal anaphora. While pronominal anaphora still heavily depend on grammatical conditions—the agreement of the antecedent and the pronoun in gender and number,—the influence of grammatical criteria gradually diminishes for all other types of text phenomena. For nominal anaphora, number constraints are still valid, while a generalization relation IS-A between the anaphoric noun and its proper antecedent must hold, in addition. In the case of textual ellipsis, no grammar

constraint at all applies, while quite sophisticated role path conditions come into play (cf. Section 4.4). Finally, for metonymies (cf. Section 4.5) concept type coercion mechanisms apply that require an even more elaborated style of conceptual reasoning and so further extend the criteria underlying the resolution of text ellipsis. These observations are summarized in Table 1. They demonstrate how local text coherence increasingly is built on more and more sophisticated conceptual conditions that are rooted in the conceptual domain representation.

4.1. Brief survey of the centering model

In this section, we will briefly introduce our approach to dealing with reference relations between different utterances. The framework of this model is provided by the well-known *centering* mechanism [24]. The theory of centering is intended to model the local coherence of discourse, i.e. coherence among the utterances U_i in a particular discourse segment (say, a paragraph of a text). Local coherence is opposed to global coherence, i.e. coherence with other segments in the discourse (for a proposal extending the centering model to global reference relations in discourse, cf. [25]). Discourse entities serving to link one utterance to other utterances in a particular discourse segment are organized in terms of centers. Each utterance U_i in a discourse segment is assigned a set of *forward-looking centers*, $C_f(U_i)$, and a unique *backward-looking center*, $C_b(U_i)$. The forward looking centers of U_i depend only on the expressions which constitute that utterance, previous utterances provide no constraints on $C_f(U_i)$. The elements of $C_f(U_i)$ are partially ordered to reflect relative prominence in U_i. The most highly ranked element of $C_f(U_i)$ that is *realized* in U_{i+1} (i.e. is associated with an expression that has a valid interpretation in the underlying semantic/conceptual representation language) is the $C_b(U_{i+1})$. The ranking imposed on the elements of the C_f reflects the assumption that the most highly ranked element of $C_f(U_i)$ is the most preferred antecedent of an anaphoric expression in U_{i+1}, while the remaining elements are (partially) ordered according to decreasing preference for establishing referential links.

The main difference between Grosz et al.'s work [24] and our proposal [26] concerns the criteria for ranking the forward-looking centers. While Grosz et al. assume (for the English language) that *grammatical* roles are the major determinant for the ranking on the C_f, we claim that for German—a language with relatively free word order—it is the *functional* information structure of the sentence in terms of topic/comment or theme/rheme patterns.

Very briefly, the constraints on the ordering of entries in $C_f(U_i)$ that can be derived from these considerations say that, first of all, *context-bound* elements in an utterance (i.e. those that are already related to previously introduced discourse elements) are preferred over non-bound discourse elements for anaphora resolution, second, if several bound elements have to be considered, then resolved anaphors are given preference over textelliptical ant-

Table 1
Sources of well-formedness criteria for different textual phenomena

	Pronominal anaphora	Nominal anaphora	Textual ellipsis	Metonymy
Grammatical constraints	Number Gender	Number	— —	— —
Conceptual constraints	—	IS-A	Role path patterns	Type coercion

Table 2
Centering data for sentence (1)

(1)	**Cb:**	—
	Cf:	[PARTICLE.4-02: Partikel, DEGREE.10-06: Durchmesser, GASTRIC-MUCOSA.13-08: Magenschleimhaut, ANTRUM-TYPE.15-10: Antrumtyp]

ecedents, while textelliptical antecedents are given preference over textelliptical expressions for reference resolution purposes and, last but not the least, if multiple occurrences of the same type of anaphoric construction occur (e.g. two anaphora or two unbound elements), then preference is defined in terms of linear precedence in the source text.

When we apply these criteria to sentence (1) of the text fragment already introduced in Section 3, Table 2 depicts the order of forward-looking centers in $C_f(U_1)$ (note that no $C_b(U_1)$ can be determined, as its computation requires the consideration of the immediately preceding sentence which is not available at the beginning of a text). Since we have no bound elements in the first sentence, only textual precedence applies to the ordering of the center list items. Grammatically, only nouns and their conceptual correlates are taken into consideration. The tuple notation takes the conceptual correlate of the lexical item in the text knowledge base in the first place, while the lexical surface form appears in the second place.

4.2. Pronominal anaphora

Unlike any of the other coherence phenomena we discuss in this article, the use of pronouns marks referential relationships explicitly by a specific part-of-speech category. As soon as a dependency relationship between the word actor for a pronoun and its governing head can be established, the anaphora resolution protocol is triggered. Since pronouns may have sentence-internal as well as sentence-external referents, first, a reference check within the sentence is made (this requires certain governing relations between phrases to be considered in the dependency graph, ones we refer to as D-binding constraints; cf. [27] for technical details). Only when a sentence-external resolution of the pronoun is tried are the data structures of the centering model for the previous utterance taken into consideration. In the ranked order they appear, the entries in the center-forward list of the previous sentence are checked as to whether the lexical items agree in number and gender with the pronoun in the currently considered utterance. Only in the case that no agreement conditions are violated are the conceptual constraints for establishing a dependency relation checked, too. This is done by determining whether the conceptual correlate of the considered lexical item (the potential antecedent) can be taken as a role filler for the conceptual correlate of the grammatical head the pronoun may be bound to by a dependency relation. If no sortal conflicts arise, the dependency relation between the pronoun's word actor and the grammatical head is, finally, established. Whenever sortal conflicts occur in the course of these checks, the next item of the list of forward-looking centers will be tried.

Consider the word actor for the pronoun `Sie` (*'it'*) in sentence (2). Its morphological features consist of the disjunctive feature lists: (a) (gender: feminine; number: singular); or (b) (gender: feminine, masculine, neutrum; number: plural). Feature unification with the finite verb form `zeigt` (*'shows'*) reduces the ambiguity to case (a). Therefore, the list of forward-looking centers of the previous sentence is scanned for entries that meet the associated morphological conditions. Since the only entry marked by the morphological feature (gender: feminine) is `Magenschleimhaut` (*'gastric mucosa'*) a suitable antecedent has been determined. Consequently, it is attempted to compute a conceptual rela-tion between the antecedent's instance GASTRIC-MUCOSA.13-08 and the instance for the

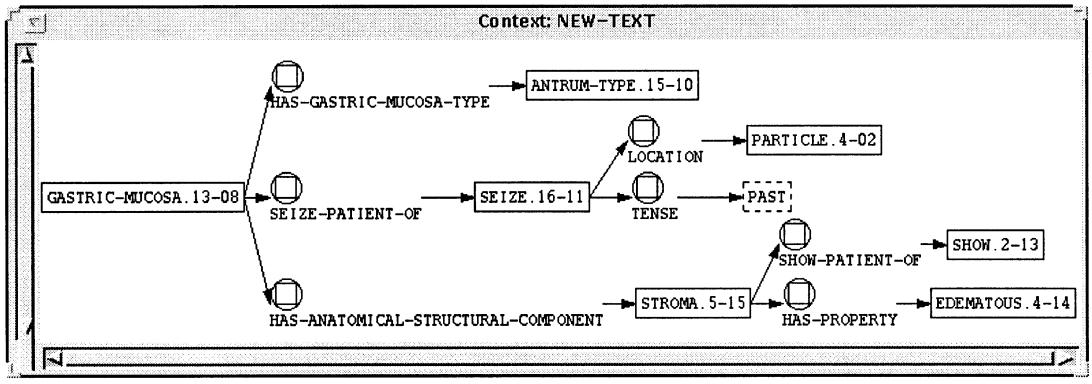


Fig. 9. Concept graph for the accumulation of sentences 1 and 2.

syntactic head SHOW.2-13 respecting the conceptual restrictions imposed by the syntactic relation **subject**. When this integrity check succeeds, `Magenschleimhaut` is accepted as antecedent and deleted from the centering list. Because SHOW.2-13 is an instance of a conceptual type that belongs to a class indicating a close conceptual relationship between its AGENT and its PATIENT, a production rule is triggered that tries to conceptually relate GASTRIC-MUCOSA.13-08 and STROMA.5-15. The only way that this can be acheived is by linking them via the relation HAS-ANATOMICAL-STRUCTURAL-COMPONENT. The final result of the conceptual interpretation after resolution of the pronominal anaphora is shown in Fig. 9[5]. GASTRIC-MUCOSA.13-08 and STROMA.5-15 are linked by the relation HAS-ANATOMICAL-STRUCTURAL-COMPONENT. At the end of the sentence analysis the centering list is constructed incorporating the results of the anaphora resolution process (cf. Table 3).

Table 3
Centering data for sentence (2)

(2)	**Cb:**	GASTRIC-MUCOSA.13-08: Sie
	Cf:	[GASTRIC-MUCOSA.13-08: Sie, STROMA.5-15: Stroma]

4.3. Nominal anaphora

Compared with pronominal anaphora nominal anaphora reveal simpler agreement constraints, as only number agreement between the antecedent and the anaphoric noun phrase is required. On the other hand, a new constraint is introduced at the conceptual level, *viz*. the conceptual type of the anaphoric expression must subsume the conceptual type of the antecedent (i.e. the anaphoric expression is conceptually more general than the antecedent). The triggering condition for the resolution of a nominal anaphora is fulfilled when the word actor of the syntactic head of a definite noun phrase (henceforth NP) carrying an anaphoric expression finds its governing syntactic head[6]. The search in the center-for-

[5] A serious problem for our incremental parsing approach arises from the fact that the time point of anaphora resolution is crucial. For instance, in case of a sortal conflict between GASTRIC-MUCOSA.13-08 and STROMA.5-15 the establishment of a syntactic dependency relation should be rejected as a consequence of failing anaphora resolution. But at this time point the information about a possible sortal conflict is not available since `Stroma` has not yet been parsed. Hence, backtracking over centering structures has to be carried out.

[6] This strategy was already successfully applied to technical domain texts in ITSYNDIKATE. Unfortunately, texts from the medical domain exhibit the already mentioned telegram-like style in which the (definite) article is often omitted. Hence, checks have to be made for any occurrence of a head noun of an NP lacking a proper determiner.

ward list of the previous utterance proceeds in the already described way, by checking for each discourse unit in the order of appearance in this list as to whether the lexical items agree in number and the conceptual correlate of the anaphoric expression subsumes the one of the antecedent. Note that it is not necessary to check whether the conceptual correlate of the antecedent violates sortal restrictions as a slot value in the conceptual structure the anaphoric expression is inserted in, since the antecedent's correlate is conceptually more specific than that of the anaphoric expression. When actually processing the anaphor resolution the two instances involved must be merged in the text knowledge base. The process of instance merging relates two instances by asserting a referential identity between them. As a consequence, the role filler representing the conceptually more general instance is identified with the more special one (i.e. that of the antecedent). Note that this exchange of referents is also mirrored at the level of the centering lists, as is illustrated by Table 4, in which the lexical item `Schleimhaut` (*'mucosa'*) is made referentially identical to GASTRIC-MUCOSA.13-08 after successful anaphora resolution.

Taking sentence (3) more deeply into consideration shows that the triggering condition for anaphora resolution is fulfilled when `Schleimhaut` has bound its specifier `die` (*'the'*) and successfully found its syntactic head `wird` (*'is'*). Processing of the centering list from utterance (2) (cf. Table 3) results in a query as to whether GASTRIC-MUCOSA is subsumed by MUCOSA. As this relationship obviously holds, MUCOSA.2-16, the literal instance identifier, is replaced by GASTRIC-MUCOSA.13-08, the referentially valid identifier in the conceptual representation structure of sentence (3) (cf. Fig. 6). Instead of having unlinked sentence graphs (cf. Figs. 3, 5 and 6), the resolution of reference for (pro)nominal anaphora leads to a joining of them in a coherent and valid text knowledge graph. Fig. 10 depicts the final result of the conceptual interpretation after resolution of the nominal anaphora. In particular, it shows how the instance GASTRIC-MUCOSA.13-08 has been inserted into the relation INFILTRATE-PATIENT-OF formerly occupied by MUCOSA.2-16. The corresponding construction of the centering list at the end of the analysis of sentence (3) is illustrated in Table 4.

4.4. Textual ellipsis

In contradistinction to the first two anaphorical phenomena introduced before, textual ellipses exhibit no grammatical constraints at all. Instead, at the conceptual level, a textual ellipsis relates a quasi-anaphoric expression to its extrasentential antecedent by conceptual attributes (or roles) associated with that antecedent. These are encoded at the symbol level of knowledge representation, but not made explicit at the literal text level. Thus, it can be viewed as a complementary phenomenon to nominal anaphora, where the anaphoric expression is related to its antecedent in terms of conceptual generalization. In the case of textual ellipsis, the missing conceptual link between two discourse elements of the text knowledge base occurring in adjacent sentences must be

Table 4
Centering data for sentence (3)

(3)	**Cb:**	GASTRIC-MUCOSA.11-08: Schleimhaut
	Cf:	[GASTRIC-MUCOSA.11-08: Schleimhaut, LYMPHOCYTES.6-21: Lymphozyten]

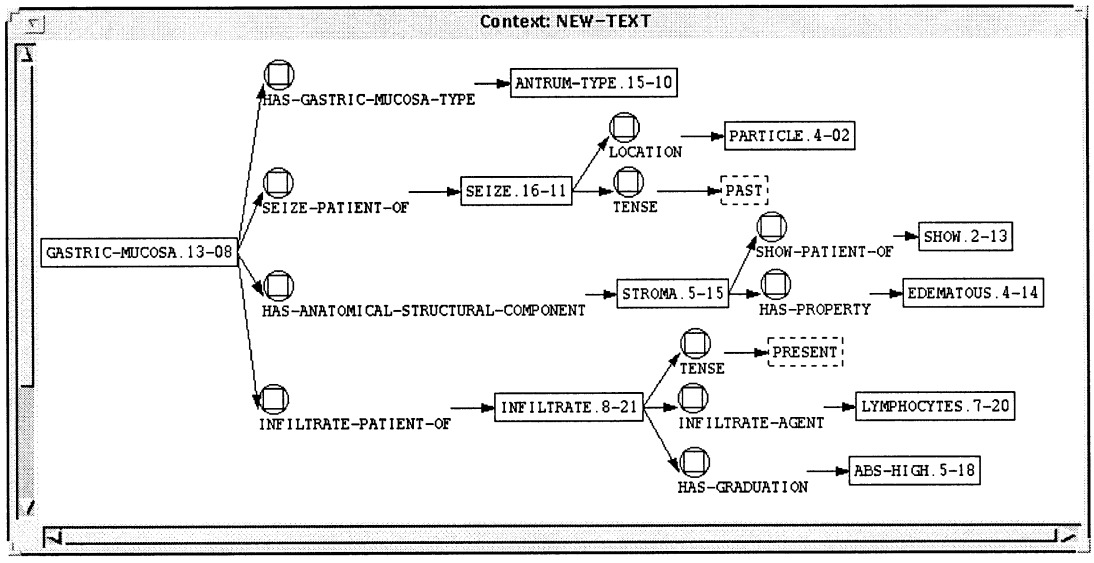


Fig. 10. Concept graph for the accumulation of sentences 1–3.

inferred in order to establish the local coherence of discourse. This inference crucially contributes to referentially cohesive text knowledge bases, as it makes explicit a referent only implicitly given by the elliptical expression. The deduction of the linking path from the concept type of the textelliptical expression to the concept type of the antecedent necessarily requires a fine-grained domain knowledge base with tied up restrictions of role fillers and relation ranges. Otherwise, no reasonable results could be computed, as the search space might explode. In the process of *path finding* in the directed acyclic graph of the domain knowledge base, an extensive unidirectional search is carried out. Furthermore, formal *well-formedness* conditions must hold for the paths between the two concepts considered, *viz*. complete connectivity (compatibility of domains and ranges of the included relations), non-cyclicity (exclusion of inverses of relations) and non-redundancy (exclusion of including paths). The search results are then evaluated according to empirically validated criteria of plausibility (for technical definitions of these terms, cf. [28]).

As with nominal anaphora, the triggering condition for textual ellipsis resolution is the occurrence of a definite NP carrying the textelliptical expression. Diverging, however, from the conditions for the analysis of nominal anaphora, the resolution of textual ellipsis is processed at the end of the sentence analysis as part of the sentence's 'wrap up'. In order to determine the textelliptical antecedent, all forward-looking centers of the previous utterance are examined in the order of their appearance and evaluated as to whether a discourse unit (the possible antecedent) relates to the textelliptical expression according to three different types of conceptual strengths (in decreasing order, plausible, metonymic, implausible). The one with the highest strength is chosen (even if it occurs in a position in the center-forward list that follows another, though weaker candidate). Candidates of equal strength at the

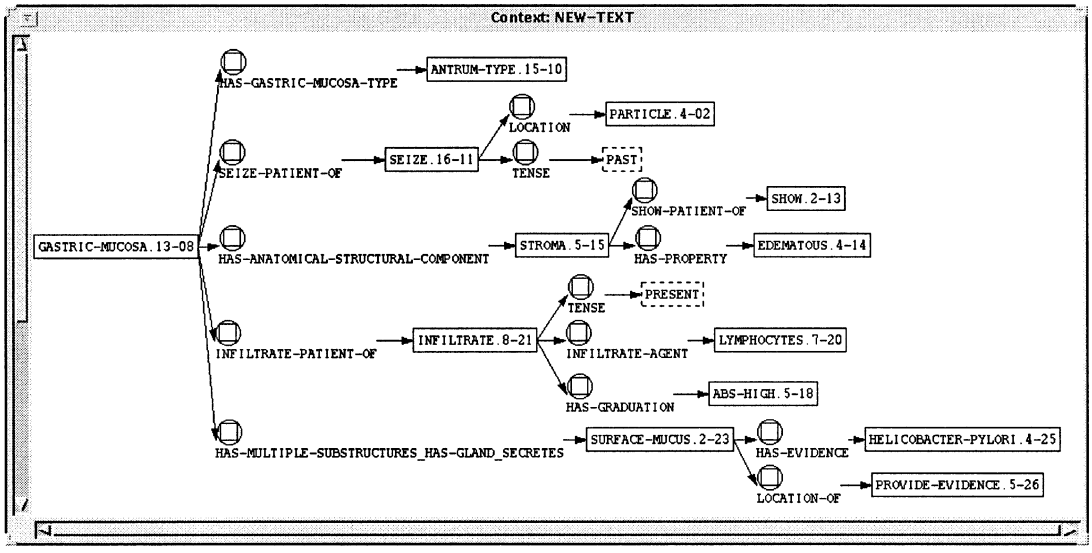


Fig. 11. Concept graph for the accumulation of sentences 1–4.

highest level possible, however, are chosen according to the priority expressed in the center-forward list.

In our text fragment the fourth sentence contains a definite NP (the definite article is somewhat hidden in the elided concatenation of 'in' and 'dem' to `Im`). Since the process to determine a textelliptical referent is initiated at the end of the sentence, the forward-looking centers for utterance (3) (cf. Table 4) are checked. For GASTRIC-MUCOSA.13-08, its first element, the path finder finds a plausible path between the concept types GASTRIC-MUCOSA and SURFACE-MUCUS. The path-finding algorithm immediately terminates, because a stronger path cannot be found (it is a plausible one, i.e. of highest conceptual strength and it is the first element of the center-forward list, hence, of highest priority for reference resolution). The linking between these concepts is made via the relations HAS-MULTIPLE-SUBSTRUCTURES, HAS-GLAND and SECRETES. Accordingly single new relation, HAS-MULTIPLE-SUBSTRUCTURES:HAS-GLAND:SECRETES, is generated from these basic relations and the instances in the text knowledge base corresponding to the two concepts GASTRIC-MUCOSA.13-08 and SURFACE-MUCUS.2-23 are related by this new relation[7]. Fig. 11 depicts the final result of the conceptual interpretation after resolution of the textual ellipsis. Note how the incoherence evident from the Figs. 6 and 7 is remedied in Fig. 11, now linking the SURFACE-MUCUS.2-23 to GASTRIC-MUCOSA.13-08, as intended by the author of the text. The corresponding construction of the centering list at the end of the analysis of sentence (4) is illustrated in Table 5. Note that the textelliptic antecedent, GASTRIC-MUCOSA.13-08, is assigned utmost priority as it appears on top of the centering list though it has not been mentioned explicitly.

[7] It is still an unsolved issue whether to create a new composed relation out of the search result or just to instantiate the 'missing linking' concepts. However, the latter solution might result in a much more complicated strategy for computation of the centering list. But we have not gathered sufficient empirical evidence so far to finally decide on this issue.

Table 5
Centering data for sentence (4)

(4)	**Cb:**	GASTRIC-MUCOSA.13-08: —
	Cf:	[GASTRIC-MUCOSA.13-08: —, SURFACE-MUCUS.2-23: Oberflächenschleim, HELICOBACTER-PYLORI.4-25: Helicobacter-pylori]

4.5. Metonymies

Metonymies further extend the patterns characteristic of textual ellipsis in terms of figurative speech. Like textual ellipsis metonymies impose no agreement constraints on a possible antecedent, instead even more complex conceptual conditions are required. An expression *A* is considered a metonymy, if *A* deviates from its 'standard denotation' (often causing a so-called sortal conflict which gives rise to some kind of type coercion) in that it stands for an entity *B* which is not expressed explicitly but is conceptually related to *A* via a (usually conventionalized) conceptual relation *r*. The metonymic expression (a noun or a NP) is related to an extrasentential antecedent in the preceding text by placing a corresponding conceptual role constraint upon both. As with text ellipsis, the missing metonymic conceptual path between those two discourse elements must be inferred via an extended conceptual graph search in the domain knowledge base in order to establish the local coherence of the discourse. Also metonymies are required to occur in a definite NP. Their resolution, however, is carried out as soon as a definite NP has found its syntactic head (in analogy to nominal anaphors). We do not give a preferential treatment to a 'literal-meaning-first' interpretation, as the path finding algorithm searches all possible conceptual relations in parallel and selects the most reasonable one according to a plausibility ranking (for a detailed discussion cf. [29]).

In sentence (5) of our text fragment a check for conceptual constraints is carried out when an attempt is made to bind the definite NP '*der Patient*' ('*the patient*') to `kontrolliert` ('*checked*'). The type restriction in the concept definition for CHECK, however, does not allow for a filler with the concept type PATIENT, thus producing a sortal conflict, which prohibits the establishment of a dependency relation given the literal interpretation of PATIENT[8].

However, the results of metonymy resolution permit a syntactic binding, since they provide a valid conceptual interpretation. For its computation the centering data for sentence (4) (cf. Table 5) is searched for an antecedent whose type is not subsumed by PATIENT. If this condition is met, which is obviously the case for the first entry GASTRIC-MUCOSA.13-08, it is checked whether a metonymic path can be computed linking CHECK.6-30 (cf. Fig. 8) via GASTRIC-MUCOSA.13-08 to PATIENT.2-27. Since in our example an occurrence of a (highly conventional) *whole-for-part* metonymy is encountered, a connecting path is retrieved for PATIENT.2-27 and GASTRIC-MUCOSA.13-08 by the transitive HAS-ANATOMICAL-PART relation. Because GASTRIC-MUCOSA is subsumed by ORGANISM-SUBSTRUCTURE it is a legal filler for the CHECK-PATIENT slot. Hence, a connecting conceptual path also

[8] In the discussion of sentence (5), the concept PATIENT refers to the person receiving medical treatment. Furthermore, the concept CHECK refers to actions on substructures (e.g. ORGANISM-SUBSTRUCTURES like the gastrointestinal tract) of a whole (e.g. an organism), thus precluding instances of PATIENT as valid role fillers. In a non-metonymic usage, the substructures are literally available in the immediate textual context.

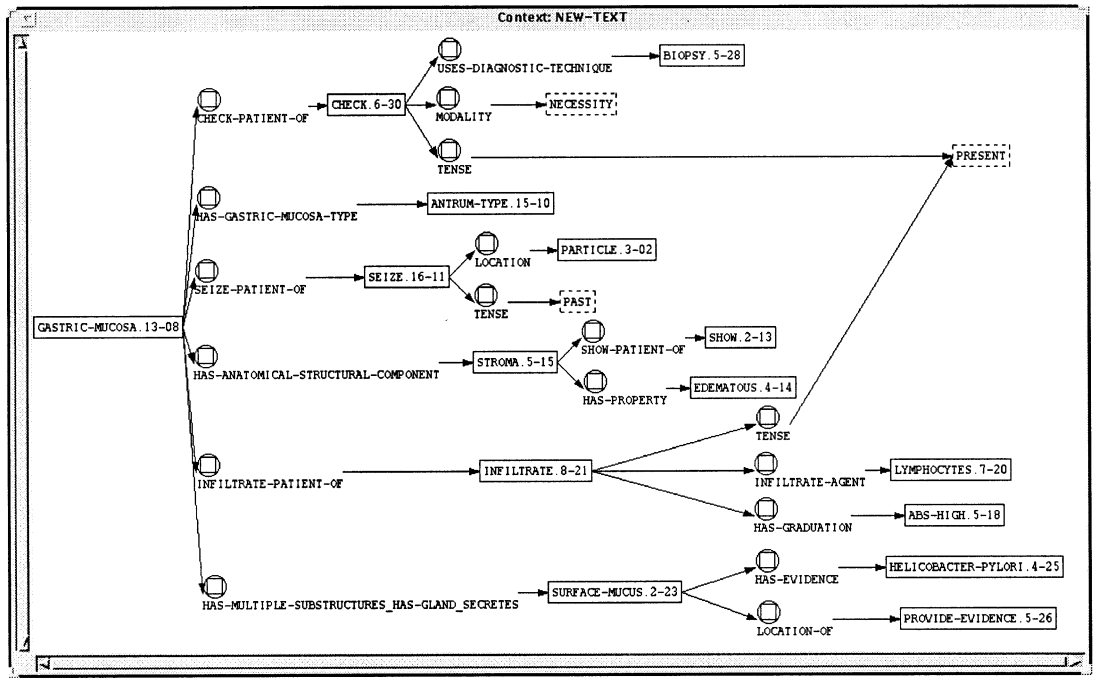


Fig. 12. Concept graph for the accumulation of sentences 1–5.

exists between CHECK.6-30 and GASTRIC-MUCOSA.13-08. In Fig. 12 the resulting conceptual representation is depicted showing that the metonymic expression PATIENT.2-27 ($=A$) has the intended denotation GASTRIC-MUCOSA.13-08 ($=B$) which is connected to CHECK.6-30. Again, comprehensive analysis of all preceding text phenomena is assumed, i.e. the proper elliptic antecedent GASTRIC-MUCOSA.13-08 to which the elliptic expression SURFACE-MUCUS.2-23 refers has already been resolved in sentences (4) and (3), respectively, via the conceptual role $r =$ HAS-MULTIPLE-SUBSTRUCTURES:HAS-GLAND:SECRETES.

4.6. Empirical study on the distribution of text phenomena in medical texts

In order to land substance to our claim that accounting for text structures is vital for medical text processing, we analyzed a randomly chosen sample of 100 reports on histological findings taken from the clinical information system of the University Hospital at Freiburg. These (German language) texts deal with biopsy material from a great variety of locations. The total number of words amount to approximately 14000 giving an average of 140 terms per document. Single texts range from a minimum of 23 up to a maximum of 925 words depending on the complexity of histological analyses and the severity of the findings.

Since the SYNDIKATE system, prior to porting it to the medical domain, was originally developed for the analysis of IT test reports in the ITSYNDIKATE system, we have already gathered empirical data on the occurrence of textual phenomena in the IT test domain (details of these results are discussed in [26]). In IT texts, anaphors and textual ellipses occur at an almost balanced rate (we also gathered quantitative evidence that

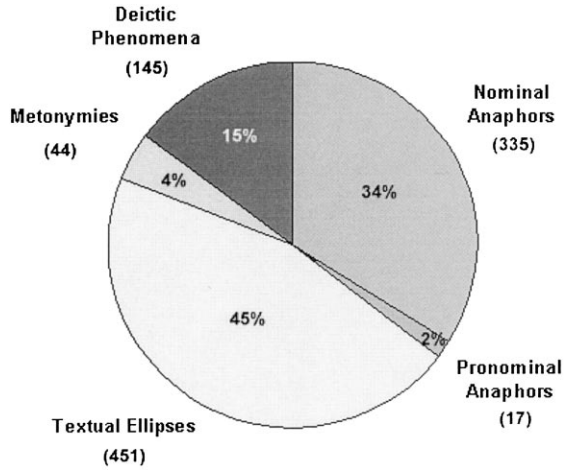


Fig. 13. Empirical distribution of textual phenomena.

anaphora are the dominating textual phenomenon in newspaper and, in particular, in literary texts). The results of the empirical study of medical texts are summarized in Fig. 13.

The quantitative distribution of textual phenomena in the medical texts we investigated exhibits a surprising result compared with the IT domain.. The data show that *textual ellipses* are the major glue for establishing local coherence in medical texts (almost half of all textual phenomena), while anaphora, pronominal anaphora in particular, play a far less important role than in any other text genre. This is interesting insofar as the phenomenon of textual ellipsis, unlike the broad coverage of anaphoric phenomena, has received only marginal attention in the field of natural language processing so far (cf. [28] for a fully worked out algorithmic proposal).

The immense importance of textual ellipsis (45%) [42–48%][9] and the remarkable ratio of nominal (34%) [31–37%] compared with extrasentential pronominal anaphora (2%) [1–3%] is clearly an indication of the primary orientation in medical texts of conveying facts in a very compact manner presupposing a considerable degree of medical background knowledge. Stylistic criteria, mainly the source of using pronominal anaphora, have far less impact. Also, diverging from the IT study, metonymies play a less important role as they account for only 4% [3–6%] of the data. In the sample, two basic metonymic patterns could be identified, The use of *hematopoesis* to refer to determinate cell populations in the bone marrow and the metonymic use of *biopsy* to characterize the material yielded. Deictic expressions which account for 15% [12–17%] of all phenomena mainly occur in the introduction of the final interpretation of the findings '*Die Befunde entsprechen…*' ('*the findings correspond to…*'), thus referring in a quite unspecific way to the whole of the text.

Summing up, local text coherence structures are frequent phenomena in medical texts. In particular, the high impact text ellipses have on the quantitative distribution of text coherence patterns provides a great challenge for medical concept languages, since these address the kinds of knowledge structures usually not made available in comprehensive medical terminologies.

5. Requirements for conceptual representation languages

In the previous subsections, we outlined the basic mechanisms for the interaction of the grammatical processes and the domain knowledge base. Most of the inferences undertaken in the course of the text understanding process require a comprehensive and fine-grained domain ontology. This relates to nominal anaphora, where generalization relations play a significant role, and even more so to textual ellipses and metonymies which

[9] For all percentage numbers 95% confidence intervals are supplied in square brackets.

make heavy use of the linkage induced by conceptual roles.

In the medical domain, a number of knowledge repositories are already available, but they are only useful to a limited extent given the requirements of discourse understanding. The ICD (International Classification of Diseases) system [30] covers only diseases and disorders, lacks systematicity and has only a coarse granularity. The combinatorial nomenclature SNOMED [31] (Systematized Nomenclature of Human and Veterinary Medicine) is a multi-axial coding system which provides a fine-grained coverage of the whole domain of medical sciences and health care. Unfortunately, partitive and generic hierarchies are continuously mixed and virtually no constraints are available so that any combination of axes is allowed. Hence, inconsistent or redundant coding is likely to result from its application. More than 330000 concepts from over 30 vocabularies and classifications (including ICD and SNOMED) are contained in the UMLS (Unified Medical Language System) Metathesaurus [32,33]. The consistency problem is reduced by a thorough typing of all concepts using a semantic network which contains 135 semantical types and 51 relations. On the basis of this semantic network, UMLS is able to carry out some simple semantic type checking. Reasoning that goes beyond generic hierarchies (Is-A relations) is beyond the scope of UMLS.

It should have become evident that the requirements of adequate text coherence analysis are more far-reaching than the capabilities of the above mentioned terminologies. A suitable framework for our work consists of knowledge representation languages with a solid formal semantics and a high degree of expressiveness. Most likely candidates are terminological knowledge representation languages in the tradition of the KL-ONE family (e.g. GRAIL [34], which allows classification along part-whole hierarchies as a special feature, or LOOM [35], which we use in the SYNDIKATE system). Also, conceptual graphs [36] are widely used in the field of medical natural language processing.

Considering the variety of medical knowledge sources, a trade-off between expressiveness and coverage (in the numbers of concepts) can be observed. Highly expressive and fine-grained domain models contain only a small number of concepts, whereas the broad coverage systems clearly lack expressiveness as well as consistency of expressions that can be derived from them. They usually provide, however, an enormous coverage of the medical domain. From a natural language text understanding point of view joining both worlds would be of real benefit in building massively scaleable systems.

6. Related work

After the discussion of medical terminologies and their usability in our framework, in this section we will briefly relate our text analysis approach to other work in the medical language processing field. Given a recent survey of medical language processing [37], one may conclude that the treatment of text phenomena so far has not been recognized as a pressing research problem.

By far the most prominent project ever run on medical texts using a natural language processing methodology is the Linguistic String Project (LSP) [5]. Only little attention, however, has been paid to reference relations holding between sentences. The procedures available consist of crude regularization and normalization heuristics which operate on the content of information formats—static, table-like representation units which are filled

during the parsing process [38]. There is no systematic account either of sentence-external anaphora or of textual ellipsis incorporating additional knowledge sources other than sublanguage-specific information formats. In particular, LSP supplies no conceptual reasoning facilities comparable to knowledge representation systems. The most far-reaching attempt in LSP which accounts for coherence phenomena treats temporal relations in texts and makes explicit partial time orderings of the medical events reported [39].

Baud et al. [9] derive a semantic representation of medical narratives by exploiting the conceptual relations of proximate words in a sentence using a simple pattern matching method for syntactic analysis. Based on the clustering of words in sentences, conceptual graphs are used as knowledge representation target formalism [36] to determine the relevant semantic relations. Although very effective for the task the system is designed for, this approach is likely to run into problems when it has to deal with text phenomena. This is due to the fact that grammar-dependent government relations and appropriate grouping at the phrase level are relevant both for sentence-internal as well as sentence-external anaphora (cf. [27]). Also, the lack of any sort of discourse memory (as supplied, e.g. by centering lists) may severely limit the extension of the systems under development by this group to deal with text structure phenomena.

In contradistinction, the approach by Zweigenbaum et al. [10] is based on a linguistically principled attitude to natural language parsing and a strong deductive reasoning component using the conceptual graph representation approach. It lacks, however, an explicit account of text structure phenomena. More recently, Bouaud et al. [40] have dealt with the problem of resolving metonymies based on a graph traversal approach similar in spirit to our work on metonymies [41].

Lots of projects focus on information extraction from medical texts (cf. e.g. [42,8]). The corresponding systems often employ a phrase-oriented style of linguistic analysis, acquire simple facts only, focus on prefixed information templates and do not account for in a systemic way textual phenomena.

What is lacking in all these studies is a unified methodology for accounting for a broad spectrum of referential phenomena. This is where we see the major contribution of our work.

7. Conclusions

The task of analyzing the contents of real-world medical texts consists of deriving a valid and coherent representation of the knowledge they contain. Hence, text understanding must be a knowledge-based process. In order to make this claim concrete we have introduced the basic architecture and design decisions underlying the development of MEDSYNDIKATE, a text knowledge acquisition system for German pathology reports. We propose a close interaction of the knowledge about language, encoded in a performance grammar, and the domain knowledge, encoded in a description logics style, in the course of sentence and text analysis. The constraints provided by the domain knowledge are intended to significantly reduce the number of ambiguous readings at the sentence level. The mapping of the syntactic structures to a normalized conceptual representation is mediated by a semantic level that supports the incremental and asynchronous parsing process.

The major hypothesis underlying our approach is that considering the sentences of a text in isolation leads to referentially invalid, incomplete and artificially fragmented text

knowledge bases. These shortcomings result from systematically ignoring the referential relations that hold between subsequent utterances in terms of text coherence. In order to determine plausible discourse units for reference resolution, we complement linguistic and domain knowledge by a model of discourse memory and associated management principles in terms of the centering model. This allows us to deal more adequately with various forms of textual phenomena in medical texts, *viz*. pronominal anaphora, nominal anaphora and textual ellipses. Nevertheless, this model provides only the organizational platform for dealing with text phenomena—it makes accessible possible referents for anaphoric expressions according to the current discourse context. In order to achieve local coherence in texts these discourse entities have to be conceptually linked. In this sense, text structures only reflect conceptual structures—the concept system which encodes the background knowledge of a particular domain provides the real foundation on which text coherence is actually built.

Considering the representational requirements underlying an adequate treatment of these reference relations we claim that only sophisticated knowledge representation languages with powerful terminological reasoning capabilities, such as those from the KL-ONE family, are able to deal with the full range of challenges of referential text understanding.

Acknowledgements

We would like to thank our colleagues in the CLIF group and of the Department of Medical Informatics for fruitful discussions. Martin Romacker and Stefan Schulz are supported by a grant from DFG (Ha 2097/5-1).

References

[1] W. Hersh, Information Retrieval. A Health Care Perspective, Springer, New York, 1996.
[2] C. Fellbaum, (Ed.). WordNet. An Electronic Lexical Database, Cambridge, MA, MIT Press, 1998.
[3] D. Lindberg, B. Humphreys, A. McCray, The unified medical language system, Methods Inf. Med. 32 (4) (1993) 281–291.
[4] A. Rector, W. Solomon, W. Nowlan, T. Rush, A terminology server for medical language and medical information systems, Methods Inf. Med. 34 (2) (1995) 147–157.
[5] N. Sager, C. Friedman, M. Lyman, Medical Language Processing. Computer Management of Narrative Text, Addison-Wesley, Reading, MA, 1987.
[6] K. Spackman, W. Hersh, Recognizing noun phrases in medical discharge summaries: An evaluation of two natural language parsers, in: J.J. Cimino, (Ed.), Proceedings of the 1996 AMIA Annual Fall Symposium (formerly SCAMC). Beyond the Superhighway: Exploiting the Internet with Medical Informatics, pp. 155–158. Washington, DC, October 26–30, 1996, Hanley and Belfus, Philadelphia, PA, 1996.
[7] C. Sneiderman, T. Rindflesch, A. Aronson, Finding the findings: Identification of findings in medical literature using restricted natural language processing, in: J.J. Cimino, (Ed.), Proceedings of the 1996 AMIA Annual Fall Symposium (formerly SCAMC). Beyond the Superhighway: Exploiting the Internet with Medical Informatics, pp. 239–243. Washington, DC, October 26–30, 1996, Hanley and Belfus, Philadelphia, PA, 1996.
[8] D. Evans, N. Brownlow, W. Hersh, E. Campbell, Automatic concept identification in the electronic medical record: An experiment in extracting dosage information, in: J.J. Cimino, (Ed.), Proceedings of the 1996 AMIA Annual Fall Symposium (formerly SCAMC). Beyond the Superhighway: Exploiting the Internet with Medical Informatics, pp. 388–392. Washington, DC, October 26–30, 1996. Hanley and Belfus, Philadelphia, PA, 1996.
[9] R. Baud, A.-M. Rassinoux, J.-R. Scherrer, Natural language processing and semantical representation of medical texts, Methods Inf. Med. 31 (2) (1992) 117–125.
[10] P. Zweigenbaum, Consortium MENELAS, MENELAS: An access system for medical records using natural language, Comput. Methods Programs Biomed. 45 (1–2) (1994) 117–120.

[11] B. Grosz, M. Pollack, C. Sidner, Discourse, in: M.I. Posner (Ed.), Foundations of Cognitive Science, MIT Press, Cambridge, MA, 1989, pp. 437–468.

[12] U. Reimer, U. Hahn, Text condensation as knowledge base abstraction, in: CAIA'88—Proceedings of the Fourth IEEE/AAAI Conference on Artificial Intelligence Applications, pp. 338–344. San Diego, CA, March 14–18, 1988. Washington, DC, Computer Society Press of the IEEE, 1988.

[13] U. Hahn, M. Klenner, K. Schnattinger, Learning from texts: A terminological metareasoning perspective, in: S. Wermter, E. Riloff, G. Scheler (Eds.), Connectionist, Statistical and Symbolic Approaches to Learning for Natural Language Processing, Springer, Berlin, 1996, pp. 453–468 LNAI, 1040.

[14] U. Hahn, Making understanders out of parsers: Semantically driven parsing as a key concept for realistic text understanding applications, Int. J. Intell. Syst. 4 (3) (1989) 345–393. Special Issue on 'Knowledge-Based Information Retrieval Systems'.

[15] W. Woods, J. Schmolze, The KL-ONE family, Comput. Math. Appl. 23 (2–5) (1992) 133–177.

[16] E. Roche, Y. Schabes (Eds.), Finite-State Natural Language Processing, MIT Press, Cambridge, MA, 1997.

[17] B. Carpenter, The Logic of Typed Feature Structures. With Applications to Unification Grammars, Logic Programs and Constraint Resolution, Cambridge University Press, Cambridge, 1992.

[18] G. Agha, I.A. Mason, S.F. Smith, C.L. Talcott, A foundation for actor computation, J. Function. Program. 7 (1) (1997) 1–72.

[19] U. Hahn, S. Schacht, N. Bröker, Concurrent object-oriented natural language parsing: The PARSETALK model, Int. J. Human-Comput. Stud. 41 (1–2) (1994) 179–222. Special Issue on 'Object-oriented Approaches in Artificial Intelligence and Human-Computer Interaction'.

[20] N. Bröker, M. Strube, S. Schacht, U. Hahn, Coarse-grained parallelism in natural language understanding: Parsing as message passing, in: D.B. Jones, H.L. Somers (Eds.), New Methods in Language Processing, UCL Press, London, 1997, pp. 301–317.

[21] P. Neuhaus, U. Hahn, Restricted parallelism in object-oriented lexical parsing, in: COLING'96—Proceedings of the Sixteenth International Conference on Computational Linguistics, Copenhagen, Denmark, August 5–9, 1996, vol. 1, 1996, pp. 502–507.

[22] U. Hahn, P. Neuhaus, N. Bröker, Message-passing protocols for real-world parsing: An object-oriented model and its preliminary evaluation, in: Proceedings of the Fifth International Workshop on Parsing Technologies—IWPT'97, Massachusetts Institute of Technology (M.I.T.), Boston, MA, September 17–20, 1997, 1997, pp. 101–112.

[23] S. Staab, U. Hahn, 'Tall', 'good', 'high'- compared to what? In: IJCAI'97—Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence, Nagoya, Japan, August 23–29, 1997, vol. 2, Morgan Kaufmann, San Francisco, CA, 1997, pp. 996–1001.

[24] B. Grosz, A. Joshi, S. Weinstein, Centering: A framework for modeling the local coherence of discourse, Comput. Linguistics 21 (2) (1995) 203–225.

[25] U. Hahn, M. Strube, Centering in-the-large: Computing referential discourse segments, in: Proceedings of the Thirty Fifth Annual Meeting of the Association for Computational Linguistics and the Eighth Conference of the European Chapter of the Association for Computational Linguistics, Madrid, Spain, July 7–12, 1997, Morgan Kaufmann, San Francisco, CA, 1997, pp. 104–111.

[26] M. Strube, U. Hahn, Functional centering, in: ACL'96—Proceedings of the Thirty Fourth Annual Meeting of the Association for Computational Linguistics, Santa Cruz, CA, June 23–28, 1996, Morgan Kaufmann, San Francisco, CA, 1996, pp. 270–277.

[27] M. Strube, U. Hahn, PARSETALK about sentence- and text-level anaphora, in: EACL'95—Proceedings of the Seventh Conference of the European Chapter of the Association for Computational Linguistics, Dublin, Ireland, March 27–31, 1995, Assoc. Comput. Linguist. 1995, pp. 237–244.

[28] U. Hahn, K. Markert, M. Strube, A conceptual reasoning approach to textual ellipsis, in: W. Wahlster (Ed.), ECAI'96—Proceedings of the Twelfth European Conference on Artificial Intelligence, Budapest, Hungary, August 11–16, 1996, Wiley, Chichester, 1996, pp. 572–576.

[29] U. Hahn, K. Markert, In support of the equal rights movement for literal and figurative language: A parallel search and preferential choice model, in: M. Shafto, P. Langley (Eds.), CogSci'97—Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society, Stanford, CA, USA, August 7–10, 1997, Lawrence Erlbaum, Mahwah, NJ, 1997, pp. 289–294.

[30] WHO, International Statistical Classification of Diseases and Health Related Problems, Tenth Revision, The World Health Organization, Geneva, 1992.
[31] R. Cote, SNOMED International, Coll. Am. Pathol. 1993.
[32] A. McCray, A. Razi, The UMLS knowledge source server, in: R.H. Greenes, H.E. Peterson, D.J. Protti (Eds.), MEDINFO'95—Proceedings of the Eighth World Congress on Medical Informatics, North-Holland, Amsterdam, 1995, pp. 144–147.
[33] A. McCray, S. Nelson, The representation of meaning in the UMLS, Methods Inf. Med. 34 (1–2) (1995) 193–201.
[34] A. Rector, S. Bechhofer, C. Goble, I. Horrocks, W. Nowlan, W. Solomon, The GRAIL concept modeling language for medical terminology, Artif. Intell. Med. 9 (1995) 139–171.
[35] R. MacGregor, A description classifier for the predicate calculus, in: AAAI'94—Proceedings of the Twelfth National Conference on Artificial Intelligence, Seattle, WA, July 31–August 4, 1994. AAAI Press/MIT Press, Menlo Park, CA, 1994, pp. 213–223.
[36] J. Sowa, Conceptual Structures. Information Processing in Mind and Machine, Addison-Wesley, Reading, MA, 1984.
[37] P. Spyns, Natural language processing in medicine, Methods Inf. Med. 35 (4–5) (1996) 285–301.
[38] G. Story, L. Hirschman, Data base design for natural language medical data, J. Med. Syst. 6 (1) (1982) 77–88.
[39] L. Hirschman, Retrieving time information from natural-language texts, in: R.N. Oddy, S.E. Robertson, C.J. van Rijsbergen, P.W. Williams (Eds.), Information Retrieval Research, Butterworth, London, 1981, pp. 154–171.
[40] J. Bouaud, B. Bachimont, P. Zweigenbaum, Processing metonymy: A domain-model heuristic graph traversal approach, in: COLING'96—Proceedings of the Sixteenth International Conference on Computational Linguistics, Copenhagen, Denmark, August 5–9, 1996, pp. 137–142.
[41] K. Markert, U. Hahn, On the interaction of metonymies and anaphora, in: IJCAI'97—Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence, Nagoya, Japan, August 23–29, 1997, vol. 2, Morgan Kaufmann, San Francisco, CA, 1997, pp. 1010–1015.
[42] G. Hripcsak, C. Friedman, P. Alderson, W. DuMouchel, S. Johnson, P. Clayton, Unlocking clinical data from narrative reports: A study of natural language processing, Ann. Intern. Med. 122 (9) (1995) 681–688.