ON TEXT COHERENCE PARSING

Udo Hahn

Albert-Ludwigs-Universität Freiburg
Linguistische Informatik / Computerlinguistik
Friedrichstr. 50
D-W-7800 Freiburg i. Brsg.
Germany

email: hahn@supreme.coling.uni-freiburg.de

ABSTRACT

In this paper global patterns of thematic text organization are considered within the framework of a distributed model of text understanding. Based on the parsing results of prior text cohesion analysis, specialized text grammar modules determine whether some well-defined text macro-organization pattern is computable from the available text representation structures. The model underlying text coherence parsing formalizes hitherto entirely intuitive textlinguistic notions whose origin can be traced back to Danes's work on thematic progression patterns.

1 INTRODUCTION

During the last years it has become increasingly apparent that dialog and text understanding systems must account for connectivity relations that extend over sentence boundaries. This has led to a bulk of work dealing with various forms of cohesion-preserving language mechanisms, mainly in the field of anaphora, which contribute to connectivity among sentences. From the focus on these linguistic phenomena one might obtain a misleading picture of textual connectivity, viz. one that considers it basically as a 'flat', continuous stream of formally connected utterances lacking additional structure. Far less research has been devoted to the internal organization of cohesive utterances by mechanisms at a more global level of dialog/text architecture, the level of *text coherence*.

Major computational approaches related to coherence aspects within a dialog processing framework are due to Reichman's [1978], McKeown's [1985] and Scha & Polanyi's [1988] formalizations of dialog grammars. Coherence criteria of written texts have been investigated in the context of 'Rhetorical Structure Theory' [Mann & Thompson 1988] and related extensions [e.g., Alterman 1982, Tucker, Nirenburg & Raskin 1986] of the original theory of coherence relations in discourse [Hobbs 1982]. A second major methodology which deals with the global structuring of written texts is the model of text macro propositions and superstructures [Kintsch & van Dijk 1978, van Dijk 1980], the latter sharing all relevant properties one generally attributes to story grammars [Rumelhart 1975]. The problem with this kind of methodology is that, unlike the coherence relation approach, the grammars which have been proposed so far are fairly idiosyncratic for each application domain (narratives, weather reports, etc.). Common to all these approaches is the requirement of a deep, propositionally guided understanding of the underlying discourse; in particular, a complete theory of its domain and an exhaustive specification of a natural language grammar must be supplied in order to guarantee proper operation of implemented systems. This might explain why, with only few exceptions, these models of text coherence have resisted further computational treatment as evidenced by operational systems.

We here make an alternative and computationally more tractable proposal on how to deal with global text structures at the text coherence level. Its roots can be traced back to the seminal work of F. Danes [1974], in which he informally developed the notion of *thematic progression patterns*, distinguishing between three prototypical patterns, viz. constant theme, continuous thematization of rhemes, and derived theme (see section 3). The model outlined in this paper starts from a thorough *formalization* of (one of) these notions and places it into the environment of a *fully operational text parsing system* whose design is mainly oriented towards the proper recognition of text cohesion and coherence phenomena. Pertinent reasons for our choice of a Danes-type model of text coherence are:

(1) The text parser forms part of the text understanding system TOPIC. It operates in a real-world domain [Reimer & Hahn 1988], i.e. textual input is taken from a permanent stream of test reports in major German information technology magazines. As it seems that it will remain infeasible for a long time to come to provide exhaustive domain and grammar specifications for routinely operating text understanders, a particularly robust *partial parsing* approach capable of handling potential specification gaps has been adopted. These conditions obviously preclude the consideration of RST-style coherence relation computing as a text coherence analysis strategy, since relevant knowledge portions might be lacking for determining specific instances of coherence relations. Conversely, the coherence relation approach seems currently infeasible for the routine processing of large-scale text collections in real domains.

(2) The description of coherence structures in terms of coherence relations or text macro propositions requires the availability of deep *assertional* knowledge from their application domain (A-box level specifications in Krypton terminology; cf. Brachman et al. [1985]). The TOPIC system, however, emphasizes the role of *terminological* knowledge of its domain, i.e. the description of prototypical properties and inference rules related to basic conceptual units of the domain (Krypton's T-box level knowledge). As TOPIC is rather weak with respect to full-blown assertional knowledge, coherence relation computing, however valuable it might be, is currently out of reach for this system. Fortunately, Danes-type coherence patterns primarily refer to the level of terminological knowledge.

(3) Prototypical patterns of thematic progression are fairly *general* and *independent of particular domains* that expository texts deal with. Linguistic studies have

collected empirical evidence for this claim through investigations of texts from diverse domains [Giora 1983a, Kurzon 1984]. This coincides with the generality of use of most coherence relations, but is in sharp contrast to the highly constrained and domain-dependent model of superstructures and story grammars.

(4) Major thematic progression patterns are correlated with particular search styles and retrieval modes in full-text information systems. Hence, providing typed coherence operators inherently supports graphics-based user interactions with the TOPIC system in terms of advanced *conceptual orientation and navigation* tools for *semantically* guided text graph tours (see section 5.3).

(5) The investigation of thematic progression patterns is of value in its own methodological right. They constitute a basic *structural* model of text macro organization as opposed to model-theoretic and plan/goal-based approaches (a distinction made by Pustejovsky [1987]). As such they might complement current text understanding methodologies whose emphasis, so far, has been on fairly knowledge-expensive assertional models (such as coherence relations and text macro propositions) or stereotyped text-semantical models (such as superstructures and story grammars).

2 MOTIVATING THE NEED FOR TEXT COHERENCE PARSING

The model of text structure parsing we propose draws a careful distinction between text cohesion and text coherence phenomena. As to the illustration of text cohesion mechanisms in natural language texts, consider the following text passage:

[1] The ***Delta-X*** from *ZetaMachines Inc.* is a computer system that runs *Unix V.3*.

[2] The system is based on a *68020 processor*.

[3] It has a *12-inch monochrome display* and an integrated *telephone handset* and built-in *modem*.

[4] Internally, there's a *40-megabyte hard disk*, a *1.2-megabyte 51/4-inch floppy disk drive*, *4.5 megabytes of RAM*, *three RS-232C ports*, and an *ST-506 port*.

Repeated occurrences of various text cohesion phenomena are illustrated by nominal anaphora (*'The system'* in [2]), pronominal anaphora (*'It'* in [3]), both referring to the unique antecedent *Delta-X* (in [1]), while *'Internally, there's a ... hard disk'* (in [4]) is linked to *Delta-X* via textual ellipsis. The basic cohesion among these sentences yields the common thematic background for constantly elaborating on a single topic (*Delta-X*). An appropriate *text* parser should, first of all, recognize these multiple cohesion phenomena and produce something like the following representation structures (indicated by $[\ldots]_R$):

$[1]_R$ Delta-X < manufacturer: { ZetaMachines Inc. } >
Delta-X < operating system: { Unix V.3 } >

$[2]_R$ Delta-X < CPU: { 68020 } >

$[3]_R$ Delta-X < peripheral devices: { 12-inch monochrome display } >
Delta-X < peripheral devices: { telephone handset } >
Delta-X < communication devices: { modem } >

$[4]_R$ Delta-X < external storage devices: { 40-megabyte hard disk } >
Delta-X < external storage devices:
{ 1.2-megabyte 51/4-inch floppy disk drive } >
Delta-X < main memory: { 4.5 megabytes of RAM } >
Delta-X < ports: { 3 RS-232C ports } >
Delta-X < ports: { ST-506 port } >

What is still lacking is a representation facility which characterizes this sequence of single assertions *constantly* referring to a single topic (*Delta-X*) as constituting a coherent whole. Recognizing linguistic forms of text coherency and providing appropriate thematic grouping operators for text knowledge bases is what *text coherence parsing* mainly is about. Even if parsers would perfectly recognize and normalize all occurrences of text cohesion phenomena in texts, missing recognition capabilities for text coherence phenomena would nevertheless produce under-structured, incoherent text knowledge bases in the sense that global pragmatic indicators of discourse bracketing would be lacking.

3 BASIC TEXT COHERENCE PATTERNS

In this section, we informally describe the basic patterns of text coherence focused on in this paper. According to Danes [1974] three categories of thematic developments can be distinguished:

❒ **Constant Theme.** This pattern is characterized by the *constant* elaboration of *one* specific topic within a text (passage) by considering several of its conceptual facets. The following two paragraphs serve to illustrate this major pattern of thematic progression (the reference points to the constant theme (***Delta-X***) are indicated by italics):

[T1.1]. The ***Delta-X*** from *ZetaMachines Inc.* is a *multiuser*, *multitasking* computer system that runs *Unix V.3* and comes complete with most of the *software needed for business applications*. The combination host computer/workstation is based on a *68020 processor*, with *dual 68000 processors providing peripheral processing*. It has a *12-inch monochrome display* and an integrated *telephone handset* and built-in *modem*.

Internally, there's *a 40-megabyte hard disk, a 1.2-megabyte 51/4-inch floppy disk drive, 4.5 megabytes of RAM, a network controller, three RS-232C ports*, and *an ST-506 port*.

❒ **Continuous Thematization of Rhemes.** In contrast to constant themes, this pattern realizes a *continuous shift* of topics (visualized by bold italics). The process starts with a theme and some comment on that theme which we shall call rheme (actually, an elaboration on one of its conceptual facets). Now this rheme is focused on as the next theme that is elaborated by a corresponding rheme, etc.:

[T1.2]. The $12,000 ***Delta-X*** host/workstation can be supplied from ***ZetaMachines Inc.***, 2999 State St., Santa Barbara, CA 93105. Zeta-Machines' sales manager, Brian Wilson, says that they also plan to market the ***Gamma-Z***, a CAD/CAM workstation based on a ***Connection Machine architecture***. The underlying theoretical foundations are due to ***D. Hillis***, a former M.I.T. student who first developed an experimental prototype based on connectionist principles.

❒ **Derived Theme.** Global text structure can also be introduced by a variety of *topics* which *share conceptual commonalities* (facets) at the knowledge representation level (not necessarily need this be paralleled with properties actually mentioned in the text!) without the general concept being explicitly stated in the text. Technically this is realized by a set of sub-

ordinates or instances of a common (only implicit) superordinate/prototype. Suppose the illustrative text [T1] composed of its two constituent parts from above, [T1.1] and [T1.2], is augmented by several paragraphs dealing with ***Gamma-Z*** and ***Sigma-P*** machines on a similar level of detail as those passages which consider the ***Delta-X*** in [T1]:

[T2]. The ***Delta-X*** from *ZetaMachines*... [T1.1⊕T1.2]

The ***Gamma-Z*** is a *MS-DOS* machine. *Peripheral devices* include an *8- inch color display*, a *matrix printer*, and a *keyboard*. ...

The ***Sigma-P*** system makes available a lot of desirable *application software* such as a *database system*, *word processing*, and a variety of *games*. ...

This text implicitly has ***workstation*** as a derived theme, since that is the immediate prototype concept of those three instances (***Delta-X***, ***Gamma-Z***, ***Sigma-P***) explicitly mentioned in [T2].

4 THE KNOWLEDGE SOURCES INVOLVED IN TEXT PARSING

This section deals with the knowledge sources involved in actually parsing a text. Basically (see Figure 1), these are constituted by the PARSE BULLETIN, a blackboard-type memory which records the single events of the parsing process, the DOMAIN KNOWLEDGE BASE, which contains the domain-specific background knowledge needed for the parse, and various EXPERTs for actually driving the parse through the text grammar specifications they incorporate (cf. Hahn [1990] for a more comprehensive presentation).

The **PARSE BULLETIN** has a flat list structure. It records the sequence of text tokens as they appear in the text and, if relevant (see below), notes their *class identifiers* (FRAME item, ADJective, etc.). More important, constructive parsing activities based on operations of the knowledge base and the parser are indicated at several positions (so-called *parse points*) in the PARSE BULLETIN. The type of operation being performed is indicated by a particular *parse descriptor*. Some are internal to the management of the knowledge base, e.g., DEFACT (default concept activation), while others indicate grammatical relations recognized by the parser, such as NounATT (conceptual attribution relations between nouns), AdjATT (conceptual attribution relations between adjectives and nouns). The items affected by an operation form a so-called *parse tuple*.

The parser does not consider every token it receives from the input text at the same level of detail. Instead, it distinguishes between words which are significant to its performance (conceptually relevant ones, such as nouns or adjectives which denote concepts in the domain knowledge base, or linguistically relevant ones, such as negation particles, certain conjunctions, quantifiers, etc.), and those that are not (among them a wide variety of semantically indifferent nouns, verbs, particles, etc., each of which is assigned the class identifier NIL). The latter are simply discarded from further analysis, while the former are assigned lexicalized grammar specifications. The parser has thus been tuned towards *partial parsing* in a spirit similar to that advocated by Schank et al. [1980] and achieves text understanding primarily on a terminological level of knowledge representation.

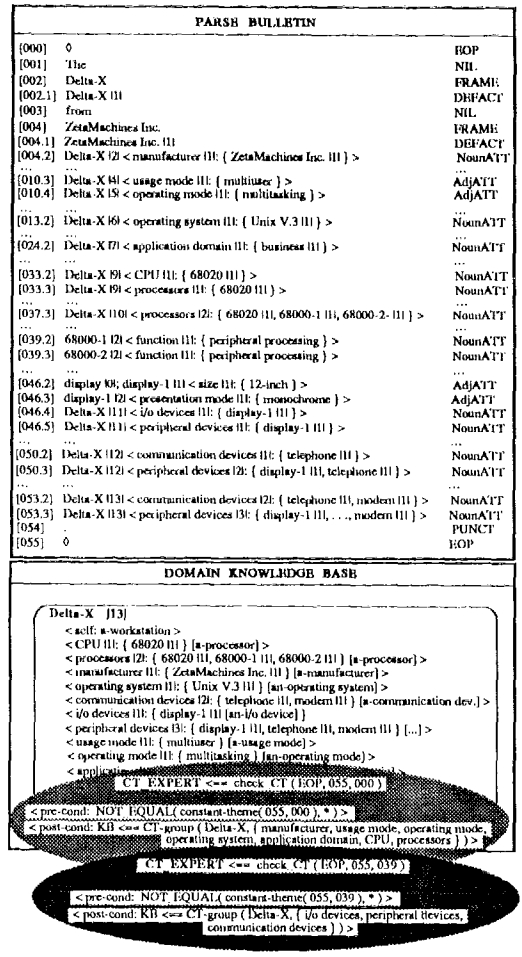

Figure 1 A Snapshot of the Parser (also Pre-Conditions Holding with respect to a Constant Theme Pattern)

The **DOMAIN KNOWLEDGE BASE** (KB for short) contains frame representation structures. Each *frame* identifier (in bold face) is assigned a list of *slots* (enclosed by angular brackets). These slots are associated with two different kinds of slot fillers. *Permitted slot fillers* are enclosed in square brackets, [a-frame name], which characterizes the range of possible slot fillers by all those frames which are a subordinate or an instance of frame name. *Actual slot fillers* are enclosed in curly braces and can be taken as facts either known *a priori* to the system or acquired continuously from the text as its understanding proceeds during the parse.

In addition, each concept has attached to it an *activation weight counter*. The values of the weight factors are enclosed by vertical bars attached to each item; if no bars explicitly occur, a zero weight is assumed. Activation weights are incremented (starting from zero-level activation) whenever a noun denoting its associated concept occurs in the text, and whenever structure-building operations in KB affect that concept. The ma-

nipulation of activation weights serves several purposes, the major one being their use as an indicator of salience of concepts during the text condensation phase, during which text summaries are generated from the text representation structures resulting from the text parse [Reimer & Hahn 1988].

The text grammar is composed of a set of distributed grammar experts, each one responsible for some specific linguistic function (e.g., concept attribution via nominal, adjectival or prepositional phrases, anaphora). Each expert is characterized by a unique **EXPERT_NAME** and is activated by a message event, i.e., by receiving a *message text* which may contain some parameters. In order to check its competence in contributing to the parse, **pre-conditions** composed of complex *test predicates* are evaluated. If these pre-conditions hold for that expert, the **post-conditions** immediately apply, i.e. messages are sent to qualified actors (to other grammar experts, to the domain KB or to the bulletin).

5 A DISTRIBUTED MODEL OF TEXT COHERENCE PARSING

In this paper, we shall not go into the details of phrasal, clausal, and text cohesion parsing (cf. Hahn [1989] for an in-depth consideration of related technical issues). Instead, we assume that these preliminary activities have already been carried out properly and that some initial structural representation is already available from the bulletin. These requirements are fulfilled in the snapshot of the PARSE BULLETIN in Figure 1, taken after all local parsing events have terminated; this characterizes a state ready to turn to the activation of global text structure computing experts.

We here consider the end of the paragraph (denoted by the symbol ◊ and the class identifier EOP) as an anchoring point for coherence computation. It is motivated by the observation that -- at least in the sublanguage domain we are currently working in -- major topic movements occur predominantly at paragraph boundaries. This coincides with linguistic evidence for the (text)grammatical status of paragraphs [Hinds 1979, Giora 1983b, and Zadrozny & Jensen 1991]. Therefore, the proper recognition of textual macro structures is always initialized at the end of a paragraph.

5.1 Considering Constant Theme

Constant theme is a coherence pattern which is characterized by multiple occurrences of a single ***frame*** in the PARSE BULLETIN within one paragraph. Most of its occurrences, in turn, are accompanied by a slot and/or slot filler indicating that some knowledge base operation with respect to ***frame*** has been carried out in KB (e.g., slot filling as indicated by NounATT or AdjATT for which we shall introduce the LC* descriptor as a convenient shorthand notation). It is the continuous elaboration of that particular concept that makes the corresponding text passage coherent. While the bulletin maintains the *sequential* order of these operations, KB provides the *conceptual* background for continuous references to the same frame object.

Figure 2 visualizes the description for **constant theme**; the DOMAIN KNOWLEDGE BASE window displays all properties of ***frame*** dealt with in a text (passage) in the shadowed area of the frame box, while those not mentioned in the text are in the remaining white part. Consequently, it is neither necessary that all slots of a frame available in the knowledge base be referred to in the text (as with $slot_{n+1}$, ..., $slot_m$), nor that there be any ordering constraint relating single slots of a frame in KB to the sequence of slot filling operations in the PARSE BULLETIN.

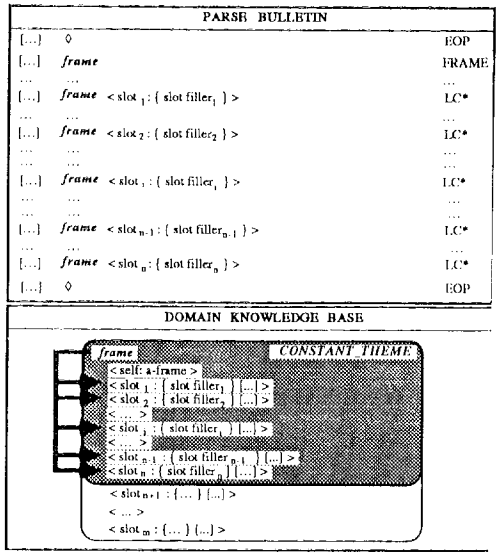

Figure_2 The General Constant Theme Configuration Pattern

The general pattern from Figure 2 is already present in Figure 1. This contains a description of the parsing results of the first paragraph of text [**T1.1**]. The entries in the PARSE BULLETIN have been worked out by experts for linguistic phenomena on the local level of phrasal, sentence and text cohesion analysis. For the purpose of constant theme computation, we need only consider those entries whose parse descriptor designates manipulations of slots or slot values of some frame (LC*-type descriptors, such as NounATT or AdjATT). Other descriptors are irrelevant here and have been left out on purpose in Figure 1. From this we construct the set THEMES. It consists of triples (*frame, slot, bullpos*) where *frame* is the name of a frame, and *slot* is the name of a slot of that frame, both co-occurring as lexical parameters of some parse tuple in the PARSE BULLETIN with a LC*-type parse descriptor; *bullpos* gives the parse point in the PARSE BULLETIN where *frame* and *slot* occur instantaneously. With respect to Figure 1 THEMES is given by:

THEMES = { (*Delta-X*, manufacturer, 004),
(*Delta-X*, usage mode, 010),
(*Delta-X*, operating mode, 010),
(*Delta-X*, operating system, 013),
(*Delta-X*, application domain, 024),
(*Delta-X*, CPU, 033),
(*Delta-X*, processors, 033),
(*Delta-X*, processors, 037),
(68000-1, function, 039),
(68000-2, function, 039),
(display-1, size, 046),
(display-1, presentation mode, 046),
(*Delta-X*, i/o devices, 046),

(*Delta-X*, peripheral devices, 046),
(*Delta-X*, communication devices, 050),
(*Delta-X*, peripheral devices, 050),
(*Delta-X*, communication devices, 053),
(*Delta-X*, peripheral devices, 053) }

When considering THEMES, we want the criterion for constant theme to be specified in a way that accounts for the fact that up to parse point '037' each slot (value) manipulation refers to one particular theme (*Delta-X*). Between parse point '039' and '046' there is a minor thematical distortion in that there is no proper reference to that theme, although slots are mentioned which are associated with other concepts. However, from parse point '046' onward the already established theme is taken up again till the end of the paragraph. In conclusion, *Delta-X* seems to be a proper candidate for consideration as a constant theme of that paragraph.[1]

Figure 1 provides a snapshot of the pre-conditions that are encountered by the **CT_EXPERT**, the coherence expert for ConstantTheme. Running twice, supplied with different parameters, it works out the results alluded to above. The grammatical knowledge needed for the determination of a constant theme is incorporated in its pre-condition part. This expression is evaluated TRUE *iff* ***constant-theme*** produces some *theme* and an associated non-empty set *RHEMES* related to *theme*, otherwise it is FALSE. The conditions for a **constant theme** can now be stated more precisely:

***constant-theme*(textpos, testpos)**
= (theme, RHEMES, newpos) *iff*

(a) *testpos* < *textpos* &
(b) (*textpos*, ◊, EOP) is in the PARSE BULLETIN[2] &
(c) (*prepos*, ◊, EOP) is also in the PARSE BULLETIN such that *prepos* < *textpos* and such that no other triple with '◊' as text item intervenes between *prepos* and *textpos* in the PARSE BULLETIN &
(d) *newpos* ∈ [***max***(prepos, *testpos*)+1, *textpos*-1] &
(e) *theme* is a frame in the DOMAIN KNOWLEDGE BASE &
(f) ∀ k_i ∈ [***max***(prepos, *testpos*)+1, *newpos*-1]:
(*theme*, slot, k_i) ∈ THEMES
==> slot ⊂ *RHEMES* &
(g) ¬∃ k´ ⊂ [***max***(prepos, *testpos*)+1, *newpos*-1]:
(α) *alt_theme* (distinct from *theme*) is a frame in the DOMAIN KNOWLEDGE BASE &
(β) (*alt_theme*, slot´, k´) ∈ THEMES &
(χ) ¬∃ tsk´ ∈ THEMES:
tsk´ = (*theme*, slot, k´) &
(h) |*RHEMES*| > 2 &
(i) *newpos* is maximal in the sense that
¬∃ Apos ∈ [***max***(prepos, *testpos*)+1, *textpos*-1]:
Apos > *newpos* &
conditions (c) - (g) apply, too.

Otherwise, ***constant-theme*(textpos, testpos) = ***

Some comments related to this specification:

(a) The parameters supplied to ***constant-theme*** span the spatial extension in PARSE BULLETIN which is searched for a constant theme; *textpos* always denotes the end of the current paragraph, i.e. the upper bound of the search area, while *testpos* delimits its lower bound.
(b) The parse point characterized by *textpos* must contain the end-of-paragraph symbol ◊.
(c) Since *testpos* may be any arbitrary parse point preceding *textpos*, *prepos* denotes the parse point in PARSE BULLETIN that contains the end-of-paragraph symbol occurring right before the one on parse point *textpos*.
(d) After fixing the search interval in the bulletin for which a constant theme is going to be computed, *newpos* allows for various choices as to how far a constant theme may actually extend in that interval.
(e) *theme* may be any frame from KB.
(f) A *theme* is related to its various rhemes according to the following condition: at each bulletin position (k_i) where *theme* occurs in THEMES within the interval delimited by *newpos*, its associated *slot* (single rheme) is assigned to the set *RHEMES*.
(g) To guarantee that *theme* is the only topic dealt with in the text, we also require that no *alt theme* different from *theme* occur in the chosen interval such that it also forms part of THEMES -- (χ) accounts for more complicated cases where both, *alt theme* and *theme*, may occur at the same parse point.
(h) To rule out insignificant occurrences of *theme* the cardinality of *RHEMES* must exceed a certain level.
(i) The maximality criterion for *newpos* rules out choosing too small values of *newpos*.

Let us now consider an example of the computation processes involved in actual coherence parsing (see Figure 1). Various coherence experts start execution upon consumption of the ◊ symbol (indicating the end of a paragraph) by the administration expert of the parser, but we shall limit our attention to **CT_EXPERT** (since the others will eventually starve). After receiving `check CT( EOP, 055, 000 )` as its starting message, ***constant-theme*** is supplied with initial parameters: *textpos* = 055, *testpos* = 000. Obviously, *prepos* = 000, since the analysis starts for the first paragraph of the text. *newpos* may now range from '001' to '054'. Let us consider *Delta-X* as *theme*. (This is a proper choice. If improper choices were made, ***constant-theme*** would not produce a significant result.). The choice for *newpos* must accommodate the temporary breakdown of the selected *theme* beginning from position '039', since we have k' = 039 ∈ [001, 054] with *alt theme* = *68000-1* (or *68000-2*) in THEMES and no proper triple (*Delta-X*, slot, 039) as required by condition g(χ) above. So *newpos* has to be adjusted properly to the parse point '039', at which point the constant theme pattern for *Delta-X* eventually terminates for the first time. This produces:

***constant-theme*(055, 000) = (*Delta-X*,**
{*manufacturer, usage mode, operating mode, operating system, application domain, CPU, processors*}, 039)

and **CT_EXPERT** issues a CT-group reading to KB incorporating the constant theme together with its associated rhemes.

Since the PARSE BULLETIN has not exhaustively been investigated with respect to its coherence data

[1] Clearly, this discussion should not be taken such that the formal characterization given below only holds for the specific sample text referred to throughout this paper. Instead, it should indicate that, although the basic idea of thematic progression patterns is overwhelmingly simple, real-life texts tend to be less homogeneous with respect to these patterns than one may consider under clean laboratory conditions. Thus, formal descriptions have to be inherently robust towards such local forms of digressions.

[2] References to entries in the PARSE BULLETIN have the format (ParsePoint, ParseTuple, ParseDescriptor).

(*newpos*+1 < *textpos*), **CT_EXPERT** resumes execution, now starting with a second set of parameters: *textpos* = 055, *testpos* = 039 (see the second expert placed into the foreground in Figure 1). Again, *prepos* = 000, but due to the new *testpos* parameter *newpos* is now in the interval [40, 54]. The evaluation of ***constant-theme***(055, 039) starts with a proper choice of *newpos* = 054. *testpos*+1 excludes *68000-1* (*68000-2*) from further consideration. Finally, we obtain

constant-theme(055, 039) = (*Delta-X*, {*i/o devices, peripheral devices, communication devices*}, 054)

Note that the occurrence of *display-1* at parse point '046' does not conflict with criterion (g), since we also have *Delta-X* (thematically related to *i/o devices* and *peripheral devices*) at that parse point (cf criterion g(χ)). Since the end of the paragraph has been reached, the coherence computation process halts.

Figure 3 represents the effects of grouping a constant theme and the rhemes referred to in the text passage (cf. [055.1] and [055.2]) by the shadowed area of the (frame) box. This indicates that the grouped items are treated coherently in a text passage.

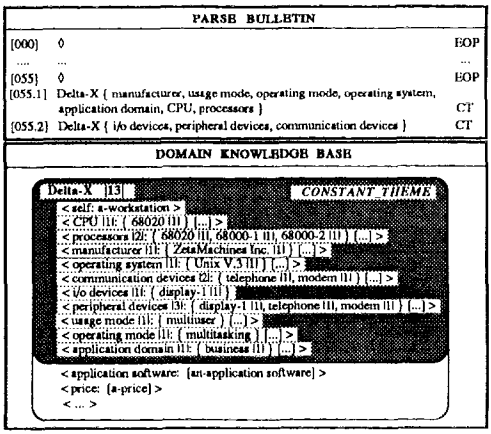

Figure_3 *Post-Conditions Holding with respect to a Constant Theme Pattern*

5.2 Remarks on Continuous Thematization of Rhemes and Derived Theme

Similarly, formal descriptions have been worked out for the other two basic text coherence patterns mentioned above. Instead of a full treatment, we give two rather informal sketches of the underlying regularities as they have been incorporated into our framework. *Continuous thematization of rhemes* most significantly departs from the constant theme schema just outlined (in fact, both are mutually exclusive) in that the former incorporates a continuous *shift* of the topics being considered. Figure 4 illustrates this permanent change of issues in a text. The PARSE BULLETIN contains a sequence of local theme-rheme pairs with $\boldsymbol{frame}_{T\,i}$ being the current local theme and $slot\,filler_{Ti}$ being its associated local rheme. Text coherence is due to the fact that the current local **rheme** ($slot\,filler_{Ti}$) becomes the next local **theme** ($\boldsymbol{frame}_{T\,i+1}$). This rheme-specific connectivity criterion is stressed by the double-sided black arrows in the DOMAIN KNOWLEDGE BASE which link the immediately preceding rheme to its identical theme successor, while local theme-rheme connections are indicated by the one-sided grey arrows which go from the local theme to its associated local rheme. A sequence of local theme-rheme pairs fulfilling the rheme-specific connectivity criterion in terms of overlapping parameters (current rheme becomes next theme) constitutes what is here called ***continuous* thematization of rhemes**, i.e. a *global* theme-rheme cluster.

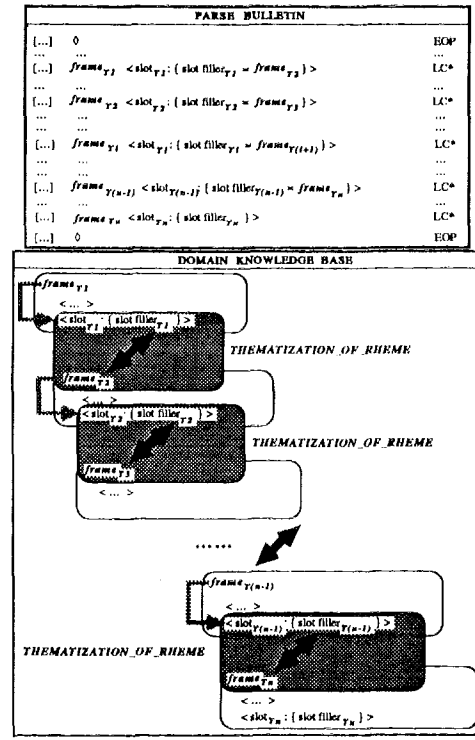

Figure_4 *The General Continuous Thematization of Rhemes Configuration Pattern*

An illustration is given by text fragment [**T1.2**] in section 3 where bold italics stress the emerging global theme-rheme cluster constituted by the following sequence of overlapping local theme-rheme pairs:

Delta-X - manufacturer - ***ZetaMachines Inc.***, ***ZetaMachines Inc.*** - product - ***Gamma-Z***, ***Gamma-Z*** - architecture - ***Conn. Machine architecture***, ***Conn. Machine architecture*** - developer - ***D. Hillis***

The third pattern further generalizes the results of the afore-going coherence computations on the paragraph level and extends them over various (adjacent) paragraphs and possibly over the whole text. Consider a series of paragraphs, each one dealing exclusively with one special topic (see Figure 5 below). The first paragraph deals with $\boldsymbol{frame}_{T\,1}$, the second one elaborates on $\boldsymbol{frame}_{T\,2}$, etc. A **derived theme** can be computed when all these different (sub)topics can be linked to the most specific general (super)topic ($\boldsymbol{frame}_{T}$). In technical terms, these subtopics are all instances of that

supertopic.Text [T2] illustrates this phenomenon: there are three paragraphs whose major topics are *Delta-X*, *Gamma-Z*, and *Sigma-P*; a conceptual generalization step links them to the derived theme *workstation*. In Figure 5 this relationship is indicated by the arrows pointing from each subtopic (of a single paragraph) to its supertopic, thematically characterizing these paragraphs on a more general level of conceptualization.

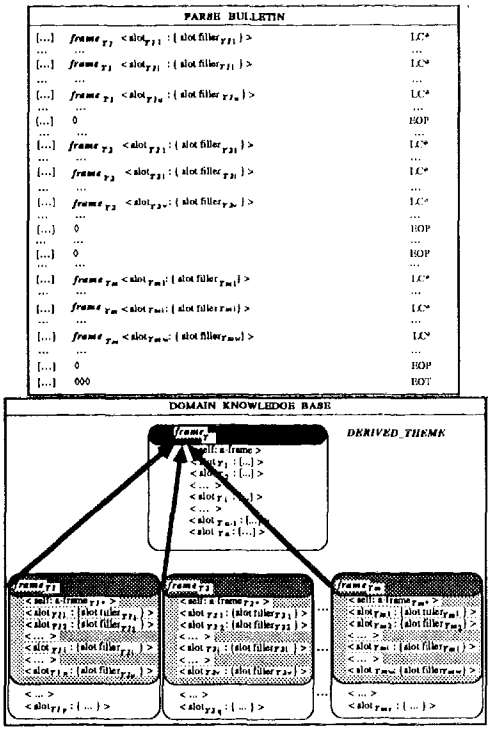

Figure_5 The General Derived Theme Configuration Pattern

5.3 The Merits of Text Coherence Parsing

Among the many advantages to having text coherence phenomena under computational control we here emphasize their potential for *information retrieval dialogs*. Evidence for this comes from our experiments with TOPOGRAPHIC, an interactive graphical interface to TOPIC's text knowledge bases [Thiel & Hammwöhner 1987]. In particular, we observed a close functional relationship between the selection of particular coherence patterns and particular search states during the retrieval process which is performed on network representations of text summaries, so-called text graphs:

1) ***Constant Theme*** coherently characterizes a variety of facts related to one particular topic. A CT-based search operation *enhances the user's knowledge of that topic* by presenting facets (or data related to those facets) the user is probably not aware of, although they may be relevant to the solution of his or her problem.

2) ***Continuous Thematization of Rhemes*** links a set of formerly unrelated topics by a coherent line of conceptual dependencies (current rheme becomes next theme). A CTR-based search operation therefore provides the basis for *thematical associations* and stimulates previously unconsidered lines of reasoning by *thematically constrained browsing*.

3) ***Derived Theme*** groups hierarchically related topics and thus may *enhance the knowledge of alternatives of the particular topic* (and facts related to it) under focused attention of the user (by way of stimulating comparisons, recognizing information gaps, etc.).

6 FINAL REMARKS

In this paper, a structural model of text coherence computation has been proposed that strongly exploits the knowledge chunking inherent to frame representations. These precompiled knowledge structures are instantiated by the topical evolution of a text as represented in the parser's bulletin. Thus, various coherence phenomena can be distinguished by particular instantiation patterns:

- ***constant theme*** is defined by multiple instantiations of aggregation (or conceptual association) relations for *one* particular frame item in KB;
- ***continuous thematization of rhemes*** is defined by multiple instantiations of aggregation relations for *continuously changing, though locally overlapping* frame items in KB;
- ***derived theme*** is defined by multiple instantiations of generalization/classification relations holding between *subparts of a frame hierarchy* in KB.

A more elaborated formal description of this model - including those parts which could only be treated rather sketchily in this contribution - is given in Hahn [1991]. The parser is currently running on SUN SPARCStations under Unix (SUNOS V4.1.1). The functionality described in this paper is fully operational and part of the TOPIC text understanding system.

REFERENCES

Alterman, R. [1982]. *A system of seven coherence relations for hierarchically organizing event concepts in text.* Univ. of Texas at Austin (TR-188).

Brachman, R.J.; V.P. Gilbert; H.J. Levesque [1985]. An essential hybrid reasoning system - knowledge and symbol level accounts of Krypton. *Proc. IJCAI 85*, pp.532-539.

Danes, F. [1974]. Functional sentence perspective and the organization of the text. In F. Danes, ed. *Papers on functional sentence perspective*. Academia, 106-128.

van Dijk, T.A. [1980]. *Macrostructures*. Hillsdale/NJ: L. Erlbaum.

Glora, R. [1983a]. Segmentation and segment cohesion: on the thematic organization of the text. *Text*, 3(2): 155-181.

Glora, R. [1983b]. Functional paragraph perspective. In J. Petöfi & E. Sözer, eds. *Micro and macro connexity of texts*. Hamburg: H. Buske, pp.153-182.

Hahn, U. [1989]. Making understanders out of parsers. *International Journal of Intelligent Systems*, 4(3): 345-393.

Hahn, U. [1990]. *Lexikalisch verteiltes Text-Parsing*. Berlin: Springer.

Hahn, U. [1991]. *Distributed text structure parsing*. Linguistische Informatik/Computerlinguistik, Univ. Freiburg, CLIF-Report 4/91.

Hinds, J. [1979]. Organizational patterns in discourse. In T. Givón, ed. *Syntax and semantics, Vol. 12*. New York/NY: Academic Pr., pp.135-157.

Hobbs, J.R. [1982]. Towards an understanding of coherence in discourse. In W.G. Lehner & M. Ringle, eds. *Strategies for natural language processing*. Hillsdale/NJ: L. Erlbaum, pp.223-243.

Kintsch, W.; T.A. van Dijk [1978]. Toward a model of text comprehension and production. *Psychological Review*, 85(5): 363-394.

Kurzon, D. [1984]. Themes, hyperthemes and the discourse structure of British legal texts. *Text*, 4(1-3): 31-55.

Mann, W.C.; S.A. Thompson [1988]. Rhetorical structure theory: towards a functional theory of text organization. *Text*, 8(3): 243-287.

McKeown, K. [1985]. Discourse strategies for generating natural-language text. *Artificial Intelligence*, 27(1):1-41.

Pustejovsky, J. [1987]. An integrated theory of discourse analysis. In S. Nirenburg, ed. *Machine translation*. Cambridge: Cambridge U.P, pp.168-191.

Reichman, R. [1978]. Conversational coherency. *Cognitive Science*, 2(4): 283-327.

Reimer, U.; U. Hahn [1988]. Text condensation as knowledge base abstraction. *Proc. 4th conf. on artificial intelligence applications (CAIA-88)*, pp.338-344.

Rumelhart, D.E. [1975]. Notes on a schema for stories. In D. Bobrow & A. Collins, eds. *Representation and understanding*. New York: Academic P., 211- 236.

Scha, R.; L. Polanyi [1988]. An augmented context free grammar for discourse. *Proc. COLING88*, pp.573-577

Schank, R.C.; M. Lebowitz; L. Birnbaum [1980]. An integrated understander. *American Journal of Computational Linguistics*, 6(1): 13-30.

Thiel, U.; R. Hammwöhner [1987]. Informational zooming: an interaction model for the graphical access to text knowledge bases. *Proc. 10th ACM SIGIR conf. on research & development in information retrieval*, pp.45-56

Tucker, A. B.; Nirenburg, S.; Raskin, V. [1986]. Discourse and cohesion in expository text. *Proc. COLING 86*, pp.181-183.

Zadrozny, W.; Jensen, K. [1991]. Semantics of paragraphs. *Computational Linguistics*, 17(2): 171-209.