Towards a Quality Improvement in Machine Translation: Modelling Discourse Structure and Including Discourse Development in the Determination of Translation Equivalents

Karin Haenelt

Integrated Publication and Information Systems Institute (IPSI) of GMD
Dolivostraße 15, D 6100 Darmstadt, Germany
haenelt@darmstadt.gmd.de, phone ++497(0)6151/869-811, fax -818

1 Introduction

Recent results in the field of machine translation research and development have shown that there will be much progress in future. The results have even encouraged researchers to see indications "that the time may finally be coming to strive once again for the promise of true automated translation" [Carbonell/Tomita 87:69]. With respect to a more general applicability of the methods involved one of the most crucial research requirements is the ability to cope with the aspect of textuality in translation. The requirements stated include:

- a text theoretical foundation of textual operations (such as text analysis, text generation, text translation) (cf. e.g. [Tsujii 86,88], [Hauenschild 88], [Nagao 89]),
- text analysis/text understanding components (cf. e.g. [Bar-Hillel 60], [ALPAC 66] [Carbonell/Cullingford/Gershman 81], [Tsujii 86,88], [Carbonell/Tomita 87], [Nagao 87,89], [Nirenburg/Raskin/Tucker 87],
- text generation components (cf. e.g. [Tsujii 86], [Bateman 90]).

In this paper a contribution to dealing with the aspect of textuality is proposed. The proposal is based on the KONTEXT text model [Haenelt/Könyves-Tóth 91] which has been developed as a theoretical foundation of textual operations. It is shown how textual aspects of machine translation can be treated on this basis. The KONTEXT model explicitly describes mechanisms of text constitution and textual communication of knowledge. It defines a multi-layered multi-state text representation which explicitly describes textual content and its contextual organization (text structure). This representation can be regarded as a text interface structure which provides the basis for further text transformations which essentially involve operations on text structure and text content, like translation.
At first a survey of the KONTEXT model is given under the aspect of its contribution towards textually controlled processing of natural language texts. Then the multi-layered multi-state text representation is described in its function as a text interface structure. Machine translation is viewed on this basis. Finally examples of textually controlled translation operations are given.

2 The KONTEXT model as a basis of textually controlled processing of texts

The KONTEXT model [Haenelt/Könyves-Tóth 91] basically consists of two components, namely a multi-layered text representation and the notion of discourse.
In the text representation different kinds of information which constitute a text are distinguished and structured into five layers:

- sentence structure
- thematic structure
- referential structure
- view (on background knowledge)
- background knowledge

The two lower layers (view and background knowledge) model conceptual information, where the view can be regarded as representing the text's content. The three upper layers (sentence structure, thematic structure, referential structure) describe the contextual organization of these concepts.The layers are ordered according to their degree of abstraction and dynamism. The information of lower layer structures is clustered by upper layer structures, and lower layer structures con-

tain more static information which is independent of the actual sequence of the textual presentation. The notion of discourse is used in order to describe the sequential establishment and use of information in a text. A discourse is defined as sequences of transitions between discourse states, and discourse states are defined by the information represented in the layers. A transition of a discourse state is the effect of the textual interpretation of a linguistic expression, which is determined by the textual function of a linguistic expression. The textual function of a linguistic expression is its contribution to all the layers of the text representation. It is described in a lexicalized text grammar which is modeled in feature structures [Firzlaff/Haenelt 92a,b] [Könyves-Tóth 91] [Böttcher 91] (cf. [Kasper 89]).

It is important to note that the contribution of linguistic expressions towards the constitution of concepts and text structure depends on the actual context. Textual interpretations of linguistic expressions are operations on previous discourse states, and the result of every operation depends on the preceding textual development. If, however, during the processing of texts, linguistic expressions are divorced from their context - as they are in sentence oriented approaches -, they lose their text specific contribution. In order to achieve a linguistic control of the processing of texts it is necessary to base the processing of texts on the context dependent contribution of linguistic means towards text constitution.

3 The text representation as a multi-layered multi-state text Interface structure

The text representation defined by the KONTEXT text model can be regarded as a multi-layered multi-state text interface (as has been explicitly or implicitly asked for by e.g. [Nagao 87], [Tsujii 88], [Hauenschild 88]). The layers provide an explicit description of different kinds of textual information and their relationships. The states describe the sequence of the actual textual presentation and use of these kinds of information. The view grows incrementally from state to state. The multi-layered multi-state text representation is a structure of linguistic functions, and no claims towards interlinguality or even universality are made.
On the basis of this representation the relationship of different texts can be described and different text-to-text-operations can be defined and performed, where one text forms the source of the creation of a new one. Such operations are e.g. paraphrasing, condensation, and also translation. The complete information of all the five layers and of the whole sequence of discourse states exactly describes one particular text. Selection or variation of certain features, however, may lead to a set of related texts. The more the information of the upper layers and of the sequence of discourse states is dropped the more the restrictions on realization decrease and the variety of possible linguistic realizations increases, i.e. the information of the deeper layers and of isolated discourse states describes less and less a particular text.

4 Operations on the interface structure: machine translation

Different text-to-text-operations differ with respect to the degree they adhere to the features of the original text and with respect to their task specific operation rules. The most characteristic quality requirement for translation seems to be to preserve features of the source text as much as possible. Range and degree of preservation, however, are notoriously difficult to determine. The requirements vary depending on text type and translation task (cf. e.g. [Hauenschild 88]).
In machine translation we especially want to deal with texts which are produced for information purposes. For these texts especially the preservation of the 'text's content' is required, but even this is a very complex requirement. Although the complexity cannot be fully described yet, the KONTEXT model may help to explain, what the implications of this requirement are, and thus, why it is a fairly strong one. A necessary (though possibly not sufficient) condition for fulfilling this requirement can be made explicit in the following way: In terms of the KONTEXT model it is the final state of the view, that can be regarded as the text's content. This state of the view, however, is the result of an incremental construction. There does not necessarily exist a direct correlation between any part of that view and particular literal passages of the text that might have caused its generation. Thus, translating a text under the condition of 'preserving its content' means: It is the final state of the view, which is to be preserved; but in order to make the generation of the final state possible, the construction instructions must be translated, and this must be done in a way which preserves the final view.

Machine translation, however, can either start from the multi-layered multi-state interface structure or from the final state of the view (or from those states which are final states with respect to a certain thematic passage). In the latter case a multi-layered multi-state text structure must be generated, which leads from the concepts assumed as background knowledge to the view. In this case the pragmatic solution found by the target language generator, namely the chain of operations on the background knowledge which lead to the view, may be different from the solution provided by the source text, and it may not meet the pragmatic requirements of the specific communication situation.

5 Example: Text model based determination of translation equivalents

Two examples may serve to illustrate how the KONTEXT model supports the determination of translation equivalents depending on the textual development and with respect to view construction capacities of linguistic means.
The following sample text is used: "*The electronic dictionaries that are the goal of EDR will be dictionaries of computers, by computers, and for computers. Of computers means that they can be processed and recompiled with computers into various forms. These dictionaries are stored in storage media such as CD-ROM. [...] By computers means that these dictionaries are being developed, or more precisely, can only be developed by using the current computer and natural language processing technology. [...] For computers means [...]"*[EDR 88]

At first a graphical overview of the development of the text representation from state to state (fig. 1.1 and 1.2) is presented, and a selection of feature structures (fig.2) is also shown. On this basis two examples of text model based determination of translation equivalents are discussed (5.1 and 5.2).

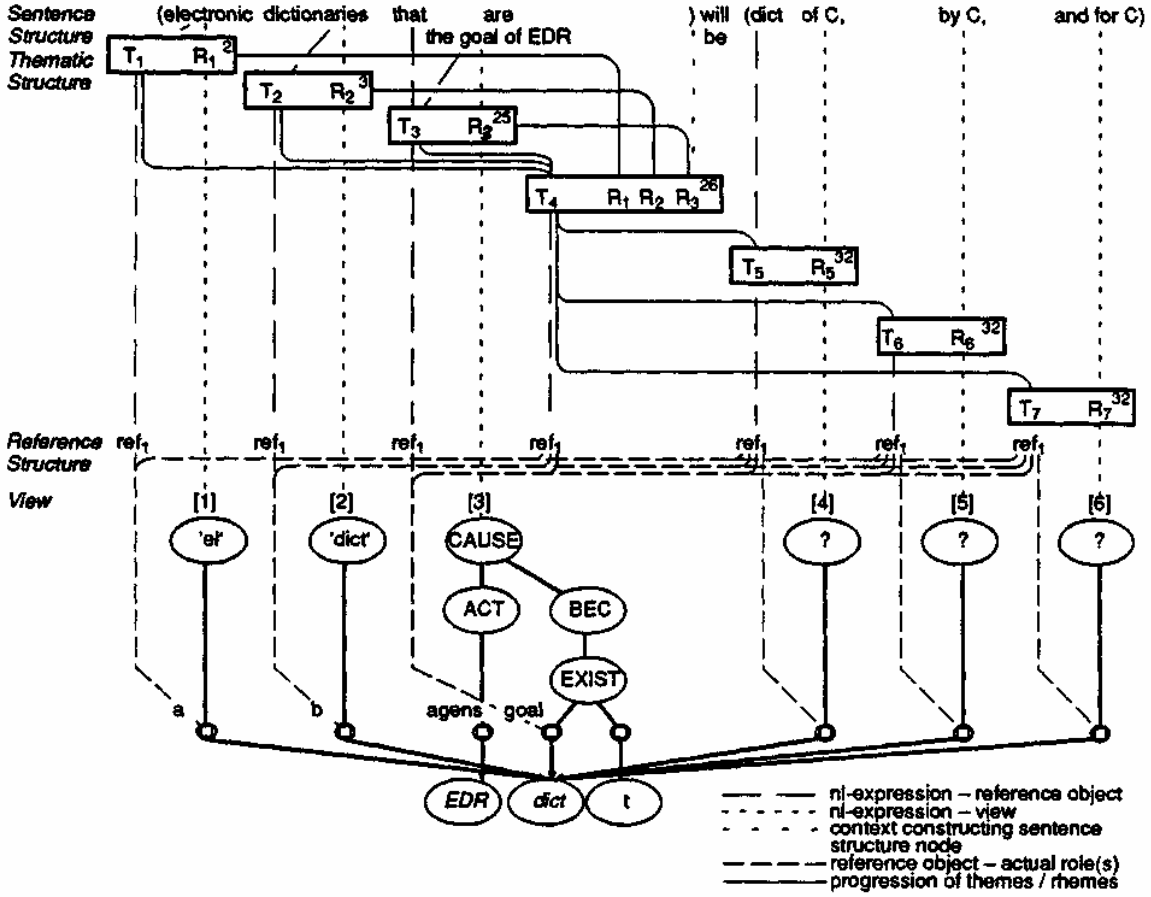

Figure 1.1: Textual modelling of "*The electronic dictionaries that are the goal of EDR will be dictionaries of computers, by computers, and for computers."*

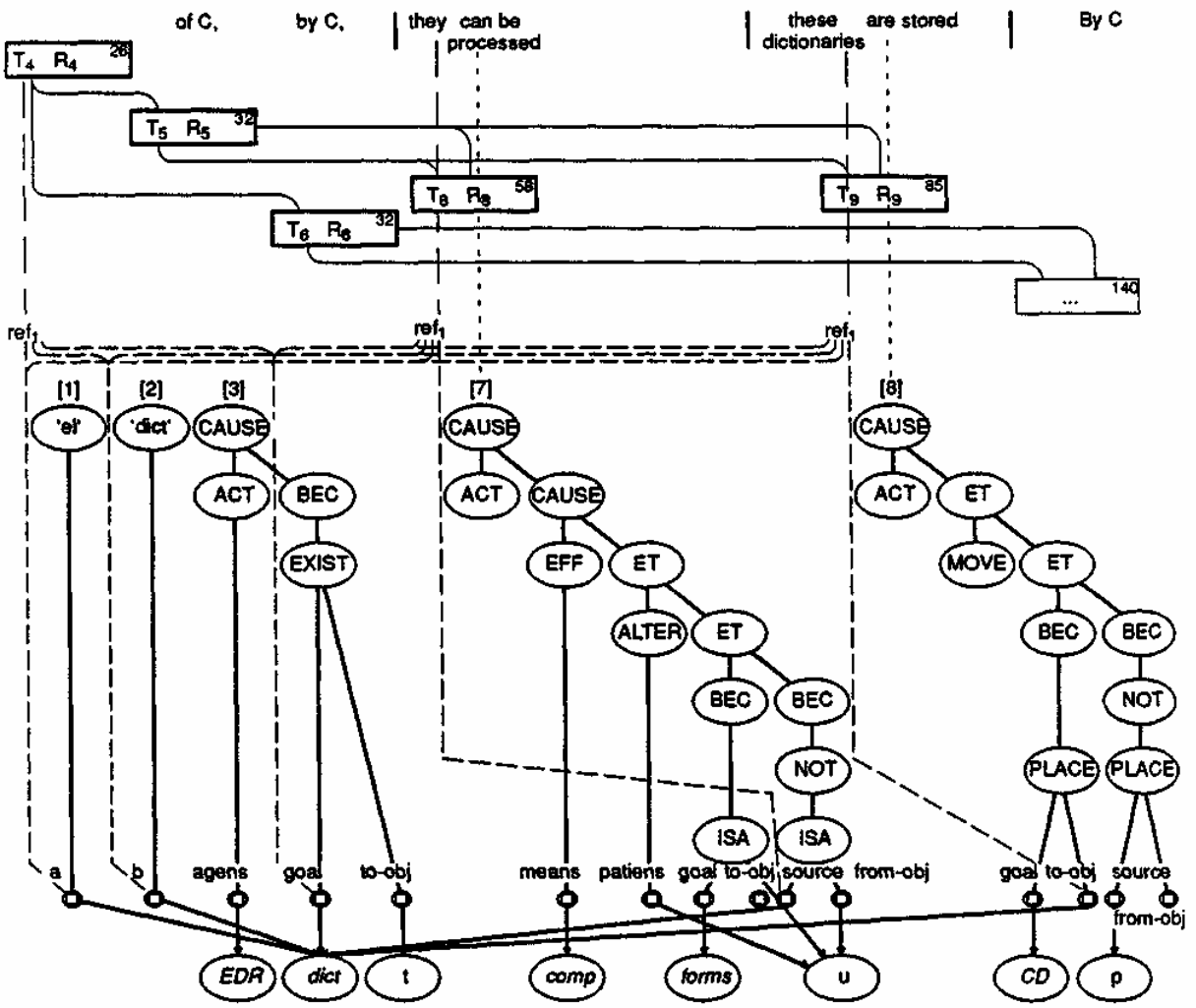

Figure 1.2: Textual modelling of the sample text

Figures 1.1 and 1.2 show a text representation of different text passages. Figure 1.1 shows the text representation which is achieved after the interpretation of "*The electronic dictionaries that are the goal of EDR will be dictionaries of computers, by computers, and for computers",* fig. 1.2 shows the continuation.

The figures show a text representation with selected discourse states. Discourse states can be traced in the layer of 'thematic structure'. The boxes in the figure represent contexts, and the numbers (2-140) in the boxes refer to the discourse states in which they have been created. The lines between the boxes represent the thematic progression. They correspond to discourse state transitions. Each discourse state is defined by the information represented in the layers of the text representation.

The layers basically contain the following information: Conceptual information (view and background knowledge -the latter is not shown in the graphics) is modelled in terms of situations which are described by the model of Semantic Emphasis [Kunze 91] and expressed as predicate-argument-structures. The background of e.g. "*to process"* is the proposition: CAUSE (ACT (x), ET (ALTER (u), ET (BEC (ISA (q,u)), NOT (ISA (p,u))))) . This can be paraphrased as: an action of 'x' causes a 'u' to alter, where 'u' becomes a 'q' and ceases to be a 'p'. In the example the basic form is extended in accordance with the rule for integrating a means-relation (cf. view [7]). Similarly to Conceptual Dependency approaches [Schank 75] in our approach verbs are modelled as denoting situations (events or states), and concrete nouns as denoting participants of situations. Adjectives are modelled as denoting situations and focusing one participant of the respective situation, preposi-

tions as denoting those situations which include the preposition's meaning and which have an actant which in the surface realization of the situation can be realized with the respective preposition. Defining situations of word classes other than verbs and deverbative abstracts are determined and composed in accordance with the meaning paraphrases of [COBUILD 87] (cf. [Firzlaff/Haenelt 92a,b]). The example e.g. contains situations introduced by *"electronic"* (view [1]), *"dictionary"* [2], *"goal"* [3], *"of"* [4], *"by"* [5], *"for"* [6], *"process"* [7], and "*store*"[8].
Reference objects are participants of the situations (cf. e.g. state 25 where ref_1 is viewed as the goal of EXIST in view [3]). The 'thematic structure' traces the discourse development ([Danes 70], [Firbas 71], [Hajicová/Sgall 88], [Hajicová/Vrbova 82]). The development basically follows the left to right textual sequence. In our modelling initially reference objects can be themes and situations can be rhemes. Theme candidates are indicated by emphatically realized actants (basically non-prepositional noun phrases -- cf. [Kunze 91]) and by anaphorical resumptions. During discourse, however, rhemes can also become themes. The 'sentence structure' contains the lexemes, their dependency structure (and further syntactic descriptions not shown here, cf. [Hellwig 80]).

5.1 Context dependent determination of translation equivalents

In the first sentence of the sample text it is not clear, how "*of*" in *"of computers"* should be translated. If it can be paraphrased as *"the dictionary is possessed by the computer",* the German translation equivalent is *"von Computern".* If its meaning can be paraphrased as *"the dictionary is somehow part of the computer"* its translation depends on the specification of this *"somehow".* These ambiguities are resolved in a later text passage *("These dictionaries are stored in storage media such as CD-ROM").* Thus the development of discourse must be involved in determining the translation equivalent.

In figures 1.1, 1.2 and 2 this development can be traced as follows: In state 32 (cf. fig. 1.1) *"of"* brings in a further situation which involves *"electronic dictionaries"* as participants, and thus contributes to the definition of *"electronic dictionaries".* The situation, however, is not clearly denoted by the prepositional phrase (cf. fig.2, state 32, view:[4:{..}], which is a disjunction of possible situations which can be denoted by *"of").* Thus state 32 does not give enough information for the determination of a translation equivalent. The ambiguity is resolved in state 58 (cf. <theme 8>) and 85 (cf. <theme 9>) which refer to the object in question (<ref 1>) within the same thematic passage and give a specification of the aspects (<source isa_2> of "process", and <to-obj $place\text{-}on_2$> of "store") under which the object (<ref 1>) is viewed. The German translation equivalent of "*of*" must be compatible with these aspects. So a possible translation would be *"auf Computern".*

5.2 Text structure dependent determination of meaning units

The text representation of the sample text also shows an example of the determination of meaningful segments of the text. Such segments are e.g. the passage about *"dictionaries of computers",* the passage about *"dictionaries by computers",* and the passage about *"dictionaries for computers".* These passages are identified as thematic units in the layer of 'thematic structure'. In the graphical overview (fig. 1.2) e.g. the first context (T_5 R_5) in state 32 is resumed in state 58, and the sequence of resuming ends in state 85. In this state the respective reference objects are conceptually defined. In feature structures this information is represented as follows: In state 32 'rhema:2' has been introduced into the thematic structure (<theme 5>). But in this state it does not get assigned a useful value. Its value is defined via the path <view 4> which leads to a disjunction of hypotheses. In state 58 (cf. <theme 8>) reference object <ref 1> becomes further defined by <view 7>, and in state85 (cf. <theme 9>) by <view 8>. The development of a subtheme can be regarded as temporarily terminated if the discourse proceeds with shifting contexts (cf. state 140: <theme 10> does not carry on <theme 9>, it rather resumes <theme 4> and starts a new grouping of situations.

The determination of meaningful segments of a text is a very basic requirement of translation. Different languages access knowledge in different ways. Therefore different discourse states may be required for the stepwise creation of the same view. In one language it may be necessary to tell a whole story as construction instructions, while in the other language there may be a term which denotes the view immediately and at once. Different linguistic means with different textual functions will lead to the construction of different discourse states.

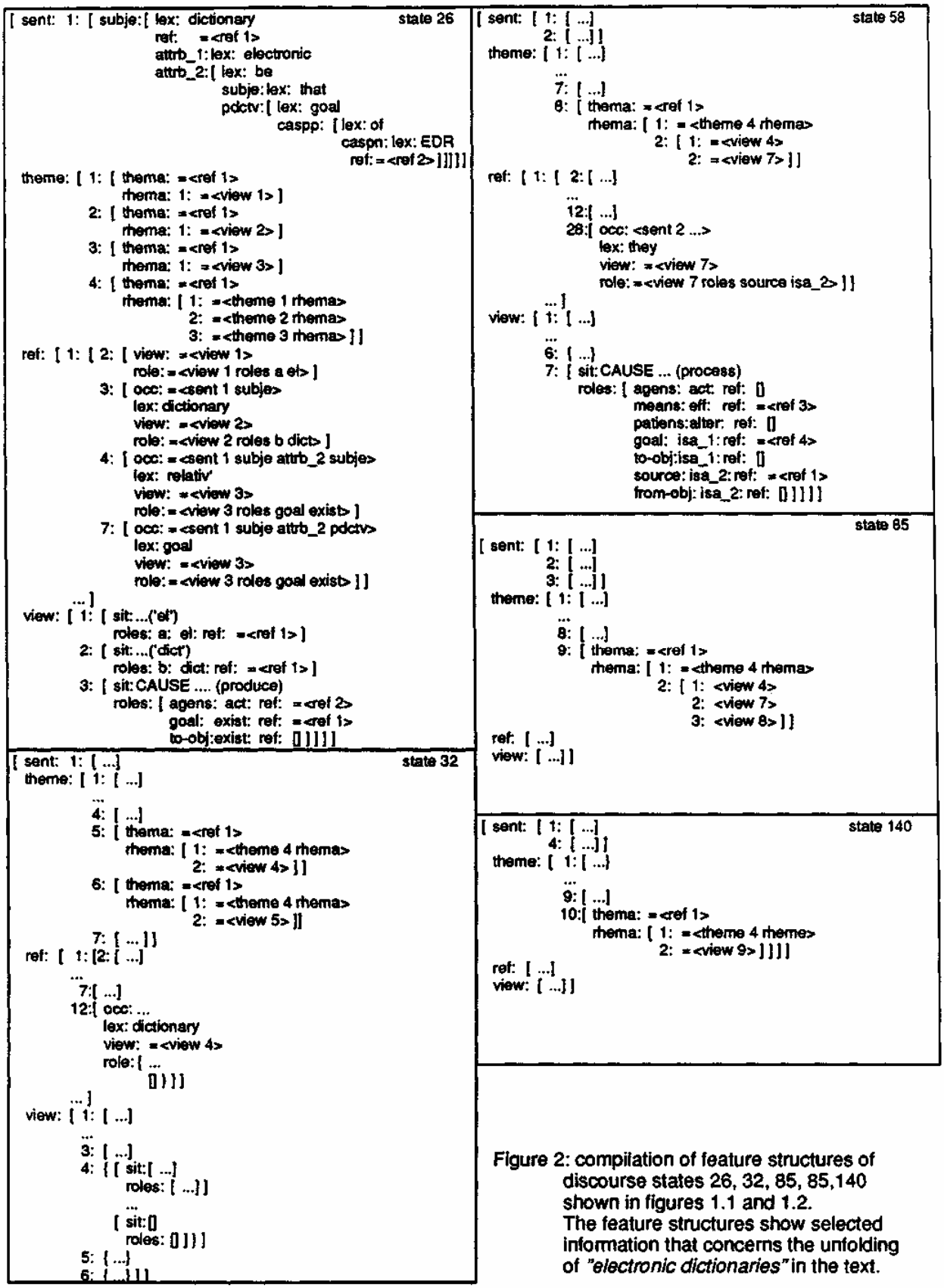

Figure 2: compilation of feature structures of discourse states 26, 32, 85, 85,140 shown in figures 1.1 and 1.2. The feature structures show selected information that concerns the unfolding of *"electronic dictionaries"* in the text.

Many translation decisions depend on the intermediate thematic final views that can be produced by the linguistic means available. One example has been shown in the previous section (5.1). Another example are task specific translation requirements. If not only the preservation of the source text's

content (final view) is required, but furthermore the preservation of as many features of the source text as possible, it is helpful to further adhere at least to the intermediate thematic final views. This contributes to preserving the text structure. This functionality is not only required for the translation of written texts, but furthermore for the interpretation of spoken dialogues where the interpreter warts for meaningful segments and reformulates the content in the target language.

6 Conclusion: Future Research

In the previous sections it was illustrated how further content and text structure dependent phenomena of translation can be made explicit and thus formalized on the basis of a text model. The text model presented is applied to modelling German and English descriptive texts. Further research work includes a detailed study of the textual usage of linguistic means in order to describe text organization mechanisms explicitly. It also includes a comparison of German and English text constitution mechanisms. Experiments have been designed which test this approach in a transfer and in an interlingua mode. These experiments will be carried out in the near future.

7 References

[ALPAC 66] *Language and Machines: Computers in Translation and Linguistics.* A report by the Automatic Language Processing Advisory Committee (ALPAC). Washington: NAS/NRC, 1966

[Bar-Hillel 60] Bar-Hillel, Y.: *The Present Status of Automatic Translation of Languages.* In: Alt, F.L. (ed.): Advances in Computers, vol. 1. New York, 1960. pp. 91-163

[Bateman 90] Bateman, John A.: *Finding Translation Equivalents: An Application of Grammatical Metaphor.* In: Proceedings of the 13th International Conference on Computational Linguistics, Helsinki, Finland, August 1990. pp. 13-18

[Böttcher 91] Böttcher, Martin: *The CFS System User Manual.* Technical report. IPS11991

[Carbonell/Cullingford/Gershman 81] Carbonell, Jaime G.; Cullingford, Richard E.; Gershman, A.V.: *Steps towards knowledge-based machine translation.* In: IEEE Translations on Pattern Analysis and Machine Intelligence. 1981,3, pp. 376-392

[Carbonell/Tomita 87] Carbonell, Jaime G.; Tomita, Masaru: *Knowledge-based machine translation, the CMU approach.* In: [Nirenburg 87] pp. 68-89

[COBUILD 87] Sinclair, John (ed. in chief): *Collins COBUILD English Language Dictionary.* London, 1987.

[Danes 70] Danes, Frantisek: *Zur linguistischen Analyse der Textstruktur.* In: Folia Linguistica 4, 1970, pp. 72-78

[EDR 88] *Japan Electronic Dictionary Research Institute.* Tokyo, 1988

[Fillmore 68] Fillmore, Ch. J.: *The Case for Case.* In: Bach, E.; Harms, R.T. (Eds.): Universals in Linguistic Theory. New York 1968.

[Firbas 71] Firbas, Jan: *On the Concept of Communicative Dynamism in the Theory of Functional Sentence Perspective.* In: Sbornik Praci Filosoficke Faculty Bmske University A 19,1971. pp. 135-144

[Firzlaff/Haenelt 92] Firzlaff, Beate; Haenelt, Karin: *On the Acquisition of Conceptual Definitions via Textual Modelling of Meaning Paraphrases.* In: Proceedings of the 14th International Conference on Computational Linguistics, Nantes, France, July 1992.

[Firzlaff/Haenelt 92] Firzlaff, Beate; Haenelt, Karin: *Applying Text Linguistic Principles to Modelling Meaning Paraphrases.* In: Proceedings of the Fifth EURALEX International Congress, Tampere, Finland, 1992

[Haenelt/Könyves-Tóth 91] Haenelt, Karin; Könyves-Tóth, Michael: *The Textual Development of Non-Stereotypic Concepts.* In: Proceedings of the 5th Conference of the European Chapter of the Association for Computational Linguistics, Berlin, Germany, 1991, pp. 263-268

[Haenelt/Könyves-Tóth 89] Haenelt, Karin; Könyves-Tóth, Michael: *A Creative Text Model. Text Understanding as Creative Operations on Knowledge Bases.* (Arbeitspapiere der GMD, Nr. 380) Birlinghoven: GMD, 1989

[Hajicová/Sgall 88] Hajicová, Eva; Sgall, Petr: *Topic and Focus of a Sentence and the Patterning of a Text.* In: Petöfi, Janos S. (ed.): Text and Discourse Constitution. Berlin: 1988. pp. 70-96

[Hajicová/Vrbová 82] Hajicová, Eva: Vrbová, Jarka: *On the Role of the Hierarchy of Activation in the Process of Natural Language Understanding.* In: Horecky, J. (ed.): Proc. COLING 1982, pp. 107-113

[Hauenschild 88] Hauenschild, Christa: *Discourse Structure - Some Implications for Machine Translation.* In: [Maxwell/Schubert/Witkam 88] pp. 145-156

[Hellwig 80] Hellwig, Peter: *Bausteine des Deutschen.* Germanistisches Seminar, Universität Heidelberg 1980

[Kasper89] Kasper, Robert T.: *Unification and Classification: An experiment in Information-Based Parsing.* In: International Workshop on Parsing Technologies. Carnegie Mellon University, August 1989.

[Könyves-Tóth 91] Könyves-Tóth, Michael: *Incremental Evaluation of Disjunctive Feature Terms.* Arbeitspapiere der GMD. Sankt Augustin: GMD, November 1991.

[Kunze 91] Kunze, Jürgen: *Kasusrelationen und Semantische Emphase.* (Studia grammatica XXXII) Berlin, 1991.

[Maxwell/Schubert/Witkam 88] Maxwell, Dan; Schubert, Klaus; Witkam, Toon (eds): *New Directions in Machine Translation.* Conference Proceedings, Budapest 18-19 August, 1988. Dordrecht, 1988

[Nagao 89] Nagao, Makoto: *A Japanese View of Machine Translation in Light of the Considerations and Recommendations* reported by ALPAC, U.S. Machine Translation System Research Committee. Tokyo: Japan Electronic Industry Development Association, 1989

[Nagao 87] Nagao, Makoto: *Role of Structural Tranformation in a Machine Translation System.* In: [Nirenburg 87] pp. 262-277

[Nagao/Tsujii/Nakamura 86] Nagao, Makoto; Tsujii, Jun-ichi; Nakamura, Jun-ichi: *Machine Translation from Japanese into English.* In: Proceedings of IEEE, vol. 74, No.7, July 1986. pp. 993-1012

[Nirenburg 87] Nirenburg, Sergei (ed.): *Machine Translation. Theoretical and Methodological Issues.* Cambridge / London/ New York, 1987.

[Nirenburg/ Raskin/ Tucker 87] Nirenburg, Sergei; Raskin, Victor; Tucker, Allen B.: *The Structure of Interlingua in TRANSLATOR.* In: [Nirenburg 87] pp. 90-103

[Tomita/Carbonell 86] Tomita, Masaru; Carbonell, Jaime: *Another Stride Towards Knowledge-Based Machine Translation.* In: Proceedings of the 11th International Conference on Computational Linguistics, Bonn, Germany, 1986. pp. 633-638

[Tsujii 88] Tsujii, Jun-ichi: *What is a Cross-Lingustically Valid Interpretation of Discourse?* In: [Maxwell/Schubert/Witkam88] pp. 157-166

[Tsujii 86] Tsujii, Jun-ichi: *Future Directions of Machine Translation.* In: Proceedings of the 11th International Conference on Computational Linguistics, Bonn, Germany, 1986. pp. 655-668