The experimental MT system of the project KIT-FAST

Wilhelm Weisweber

Technical University of Berlin, Germany
ww@cs.tu-berlin.de

Abstract

Within the project KIT-FAST an experimental machine translation (MT) system has been developed and implemented, which translates written German texts into English. For that reason a syntactic, semantic and aspects of a conceptual level of representation have been realised. In general each level has three dimensions, which are a sentence and a text representation, which are constructed during translation, as well as pre-defined background knowledge (domain and world knowledge, linguistic knowledge).

Our first step towards the translation of German texts instead of single sentences was to interpret anaphoric relations in the source language. For that reason a knowledge representation system has been integrated into the MT system in order to represent text and background knowledge in a terminological logic.

The syntactic and semantic sentence representations are structures which are generated by Generalised Phrase Structure Grammars (GPSG), and Functor-Argument Structures (FAS), respectively. The conceptual sentence representation is the ABox Tell Language, with the help of which a representation of the text content is constructed.

In the MT system two text representations are employed. One for representing the structural information of a text and another for representing the text content. The text representations are constructed incrementally from the sentential ones during translation. In principle a text representation is needed on every level, but this would lead to redundant representations on the syntactic and semantic level. For that reason we decided to take the more general semantic level for the representation of structural aspects of the text. The text content is represented as assertional knowledge in the ABox of the knowledge representation system.

The background knowledge has to be pre-defined. The objective is to represent all forms of it on each level as terminological knowledge in the TBox of the knowledge representation system. At the moment this is only realised on the conceptual level. On the syntactic and semantic level only aspects of the linguistic background knowledge are represented.

The MT system is implemented mainly with four algorithms: a component for morphological analysis and synthesis, a GPSG parser for syntactic analysis, an interpreter for non-confluent term- rewrite systems for semantic and conceptual analysis, transfer and generation, and a component for the evaluation of anaphoric relations after conceptual analysis and before transfer.

1 Models for MT

An MT model is based on abstraction and idealisation. It should describe the functional properties of human translation in order to derive the presuppositions for the development of an MT system.

Models for natural language processes of a human being do not completely describe the reality, but only cover certain cuts. Most of the MT models even do not consider the required linguistic and extra-linguistic background knowledge, which is necessary for high quality MT.

Additional limitations to the power of MT systems, which would simulate such reality models, result from the requirement that the algorithms employed should be computable and not too complex.

For that reason MT will approximate the power of human translation in the best case.

In this section a systematisation for MT models is introduced. After that the systematisation is refined in order to show that the MT model of the project KIT-FAST is an instance of it.

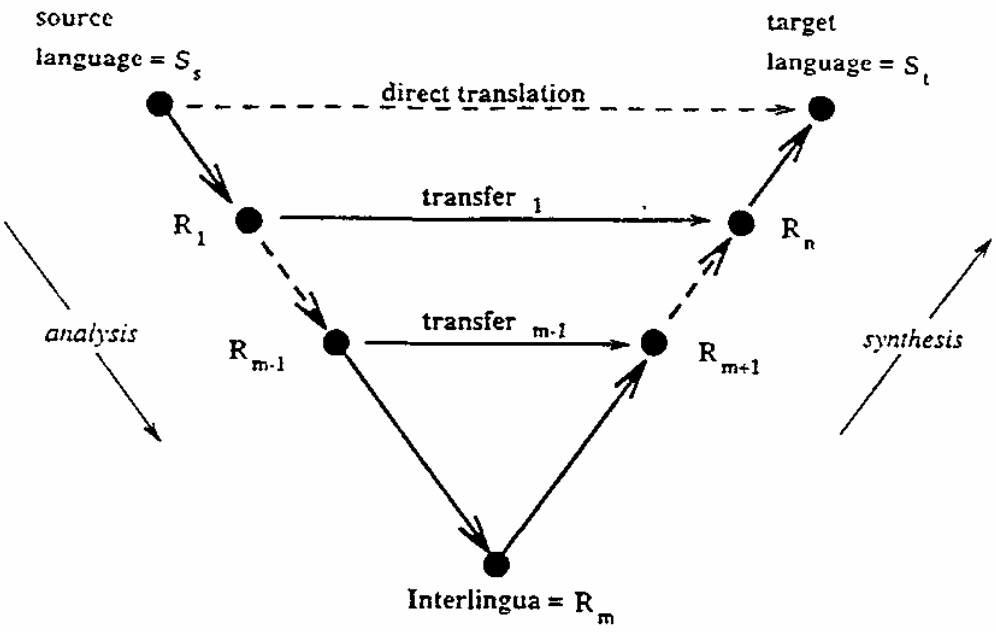

Figure 17 - 1: A systematisation for MT models

1.1 A systematisation for MT models

A systematisation for MT models in Figure 17 - 1, on which many existing MT systems are based, assumes that the representational levels are getting more abstract the deeper they occur.

In the analysis phase the representations

Ri (m >= i >= 1),

and in the synthesis phase the representations

Rj (n >= j >= m)

are getting more abstract with increasing i and with decreasing j respectively. The source and target language (SL and TL respectively) representations converge with increasing depth and are identical on the deepest level. Below, it will become clear that this can only be regarded as an idealisation in case of multi-lingual systems.

Following [Tsujii 86] the systematisation in Figure 17 - 1 can be interpreted in two ways. On the one hand one representation can be replaced by an adjacent one within analysis or synthesis. Consequently each representation has to contain more or less explicitly the complete information which is necessary for the translation of an SL sentence. This automatically leads to redundancies.

On the other hand different representations may contain different information and so they are without redundancies. This approach is assumed in [Boitet/Gerber 84].

The dashed arrows between the representations R1 and Rm-1 or Rm+1 and Rn in Figure 17 - 1 induce that the depth of the analysis is in principle not fixed. There are two criteria for the determination of the representations for analysis and synthesis. First, they should be linguistically and translation- theoretically motivated and second, each single analysis or synthesis step should not be too complex in order to ease the definition of the relations between them. The depth of analysis should depend on the needs of the transfer, i.e. the analysis can be finished when the transfer is definitely possible.

The horizontal arrows in Figure 17 - 1 (R1 --> Rn, ..., Rm-1 --> Rm+1) mean that the representations for synthesis are influenced by the corresponding representations for analysis. They also depend on the requirements of the TL, which result from its syntactic, semantic and pragmatic structure. For that reason the synthesis is oriented at the SL text if it is not completely determined by the TL.

In most cases MT systems are said to be Interlingua- or transfer-based. In Interlingua-based systems the analysis proceeds to a common semantic or deeper representation for SL and TL (Interlingua). The more languages to be considered in a multi-lingual system the more universal the Interlingua has to be. The level of universality depends also on the languages. The TL is generated directly from the Interlingua. The disadvantage of such an approach is that some ambiguities can arise in the SL or TL, which are caused by another language involved in the multi-lingual MT system.

This situation is illustrated by[1] Figure 17 - 2.

Let us assume, for example, that the translation of a concept *a* of a SL into a TL concept *e* is definitely possible, if only SL and TL are considered. An Interlingua of a multi-lingual MT system has to represent the semantic differences of all other languages involved in the system. In Figure 17 - 2 this leads to an artificial ambiguity of the concept *a* of the SL, because it has the readings *b''*, *c* and *d'* in another language, which have to be represented in the Interlingua. These three readings have to be related to the concepts *e'*, *e''* and *e'''*, respectively, which have to be collapsed to the concept *e* by the synthesis. This normally leads to a considerable overhead for analysis and synthesis, which can be eliminated by a direct transfer from SL to TL.

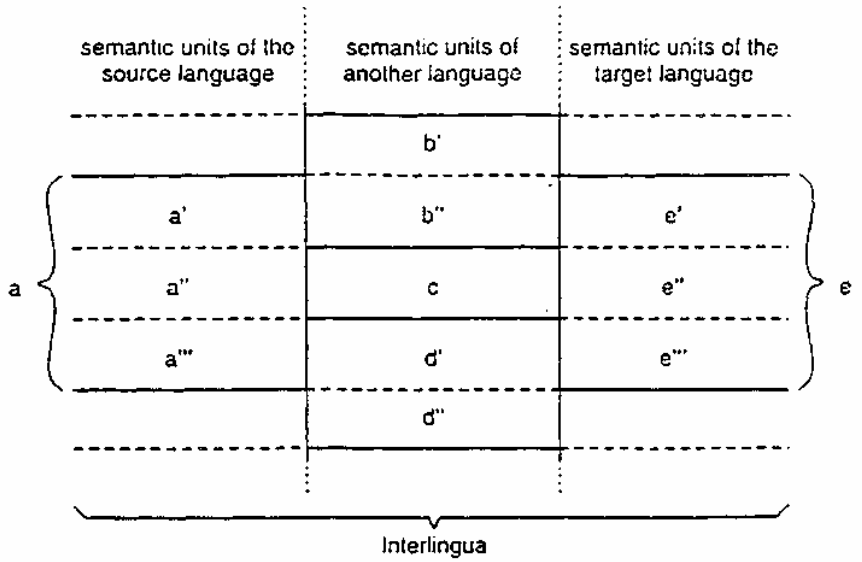

Figure 17 - 2: Ambiguities in an Interlingua-based multi-lingual MT system

1.2 Refinement of the systematisation for MT models

Many MT systems use only sentence representations, i.e. the representations of each level are two-dimensional (grammar and representation). The research in MT has provided evidence that at least another dimension has to be assumed on each level. This is a representation of the text. Apart from the grammar other kinds of background knowledge have to be represented (cf. [Preuß et al. 93]). All these dimensions are illustrated in Figure 17 - 3. The text representation includes the assertible knowledge which is conveyed by the text to be translated. The sentence representations can then be considered either as integrated parts of the text representation (this is indicated in Figure 17 - 3) or each can be represented separately.

The background knowledge consists of linguistic and extra-linguistic knowledge. The linguistic knowledge contains knowledge about the structure of sentences and texts. The extra-linguistic background knowledge includes encyclopaedic knowledge and knowledge about translational theory.

---

[1] Figure 17 - 2 is simplified because there are no such strict borderlines between the concepts in a natural language as indicated in the sketch.

The text and sentence representations and the background knowledge have different status. The former are constructed during the translation process and the latter has to be pre-defined, which is indicated by the fact that the representations of the background knowledge have no incoming arrows in Figure 17 - 3. The dashed arrows from the representations of background knowledge to the text representation indicate that the linguistic background knowledge determines how the text and sentence representations have to be constructed.

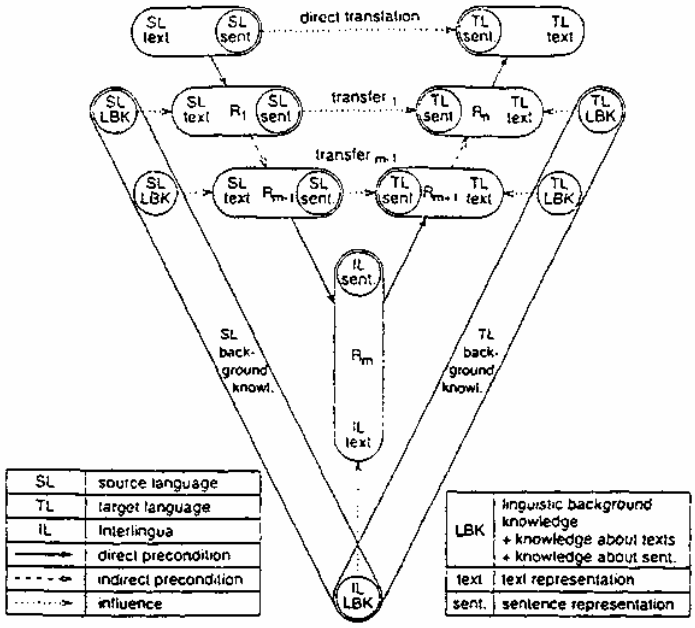

Figure 17 - 3: Refined systematisation for MT models

1.3 The KIT-FAST MT model

The model, on which the experimental MT system of the project KIT-FAST is based, is shown in Figure 17 - 4 (cf. [Hauenschild 88], [Weisweber/Hauenschild 90] and [Hauenschild 91]). The model is transfer-based and includes three representations for each of the SL and TL, and one language-independent representation.

The syntactic representations are generated by Generalised Phrase Structure Grammars (GPSG, cf. [Gazdar et al. 85], [Busemann/Hauenschild 88] and [Hauenschild/Busemann 88]). The semantic representations are called Functor-Argument Structures (FAS, cf. [KIT-FAST 93] and [Mahr/Umbach 90]). The FAS was developed by the project KIT-FAST especially for the needs of transfer and generation. It was designed in such a way that it supports a formal model-theoretic interpretation in the sense of a mapping onto expressions of intensional logic. The FAS allows for the representation of functor-argument relations, the thematic structure, semantic roles and features, and references to discourse objects of the conceptual level (see section 3.1).

The representation of the text content (see section 3.1) is a collection of all predications about the discourse objects asserted in the text. Different expressions of the text which refer to the same discourse object are realised as one formal object.

The three levels of transfer reflect three different aspects of translational equivalence, namely aspects of form, of meaning and of communicative function. This model is, of course, not meant to be a final solution, but rather a working hypothesis the project KIT-FAST started with.

The synthesis is viewed as a kind of decision process. In those cases where choices are left open from the point of view of the message to be conveyed and of the TL, the result ought to be as close as possible to the SL. This implies the well-known conflicts between the faithfulness to the original and the constraints of the TL. These conflicts are indicated by the two arrows pointing to each TL representation.

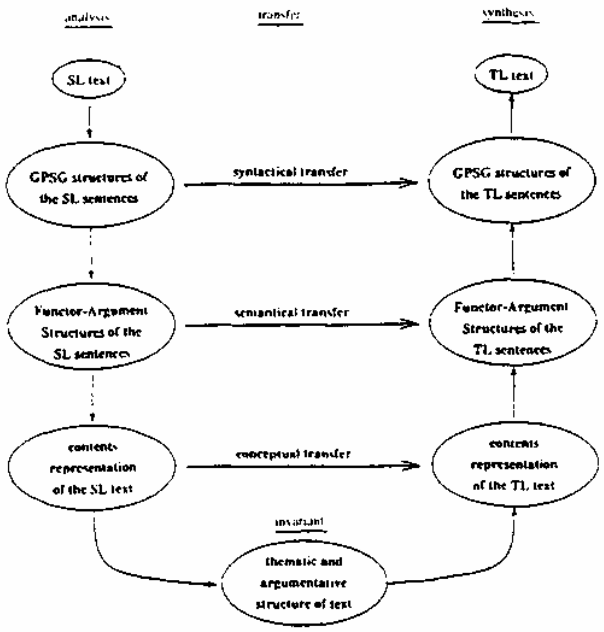

Figure 17 - 4: The MT model of the project KIT-FAST

The model in Figure 17 - 4 assumes the use of informative texts. The thematic and argumentative text structure is supposed to be an invariant of these kinds of text, i.e. it must not be changed by the transfer and represents a partial Interlingua. The thematic text structure includes hierarchical and conceptual relations of the text. The argumentative text structure consists of the logical and rhetorical relations between the assertions of the text.

2 MT system architectures

The systematisation for MT models presented in Figure 17 - 3 of the preceding section represents conceptual relations between the representations of an MT system and is not a data flow chart. For that reason no corresponding architecture can be derived directly from it. In the following two sections the architecture of the KIT-FAST MT system is developed with the help of the systematisation for MT models in Figure 17 - 3 and the KIT-FAST MT model in Figure 17 - 4.

2.1 A scheme for an architecture

In order to develop an architecture for an MT system from the KIT-FAST MT model and from the systematisation for MT models, some things have to be said about the interface between sentence and text representations and about the construction of a representation from an adjacent one. Figure 17 - 5 illustrates the interface between sentence and text representations. There are two possibilities to connect them. The possibility in (1) indicates the integration of the sentence representations into the text representation. In this case it is important that the borderlines of the sentences remain transparent. This can be achieved for example by realising the sentence

representation as one constituent of the text representation. (Typed) feature structures as in HPSG (cf. [Pollard/Sag 87]) offer a possibility for the integration of sentence representations into a textual one. An example for the integration of the sentence representations in the text representation is the KBMT system, which is an Interlingua-based MT system (cf. [Nirenburg et al. 92]). On the interlingual level the sentence representations are completely integrated into the text representation. The authors of the KBMT system do not use the notion transfer, but all facts support the assumption that the augmentation step, which disambiguates multiple readings relative to the TL after the analysis phase with the help of terminological knowledge stored in the concept lexicon, is something like a transfer step.

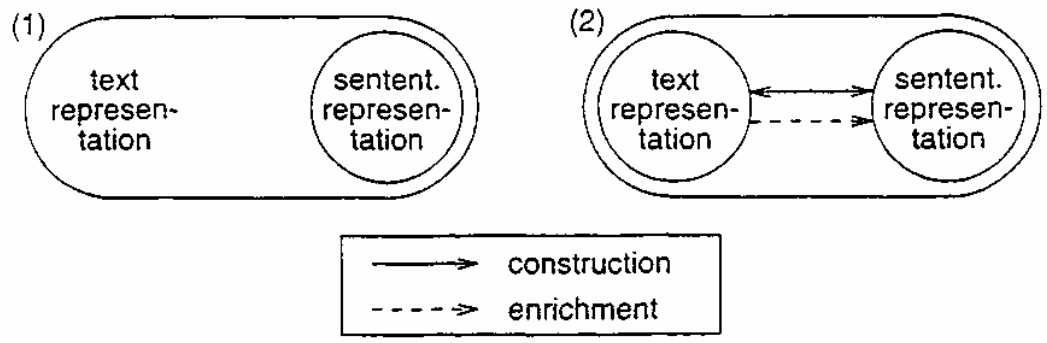

Figure 17 - 5: *Interface between sentence and text representation*

After this little excursion to the KBMT system I return to the interface between sentence and text representations. Point (2) of Figure 17 - 5 illustrates the separate realisation of the two representations. In this case some information has to be exchanged between them. The exchange of information depends on whether the text is translated sentence by sentence or as a whole. In both cases the text information has to be available for the sentence representations. This can be done by enrichment of the sentence representations with text information.

I think that the integration of the sentence representations into the text representation is preferable because no additional exchange of information between them is necessary. But it is also possible to combine the possibilities (1) and (2) in Figure 17 - 5 in one MT system by selecting the first for one level of representation and the second for another as is done in the KIT-FAST MT system.

If the text is translated as a whole then all sentence representations have to be constructed from the corresponding textual ones. The enrichment of the sentence representation with the text information can be realised by this construction by transmitting the text information to the sentence representations with the help of the mapping rules.

The translation of a text as a whole is only considered as a theoretical possibility because of its complexity. But an advantage would be more flexibility with respect to text planning compared to the translation sentence by sentence, i.e. restructuring of the text by collapsing several sentences of the SL to one of the TL, by expanding one sentence of the SL to several ones of the TL or by permuting sentences.

If the text is translated sentence by sentence then the text representations on each level have to be constructed incrementally from the corresponding sentence representations. The enrichment of the sentence representations with text information can be realised by (inter-sentential) bindings of variables. In order to allow the same flexibility with respect to text planning (see above) a specific module has to be developed.

In the following it is assumed that a text is translated sentence by sentence. There are two possibilities for this with respect to the data flow. They depend on whether the representations for analysis and synthesis are free of redundancies or not (see section 1.1).

A transfer step with representations which are free of redundancies is illustrated in Figure 17 - 6. This figure comes from [Tsujii 86]. There the representations are called factors. They determine the (syntactic) surface of the SL and TL sentences. The factors of all levels allow for a systematic comparison of two languages.

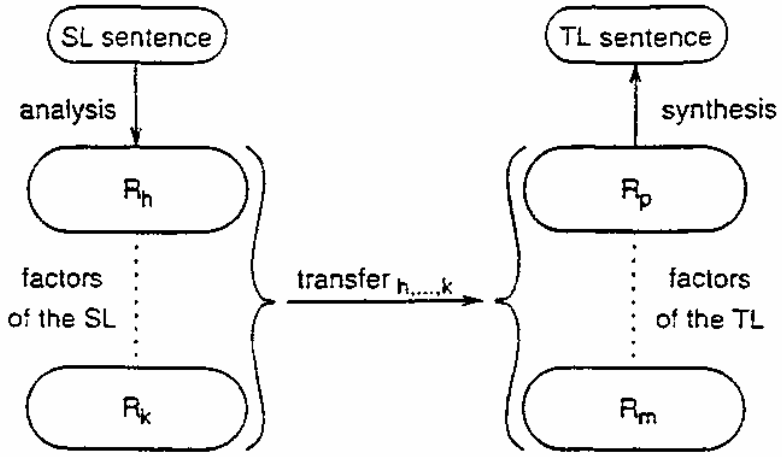

Figure 17 - 6: Transfer with representations which are free of redundancies

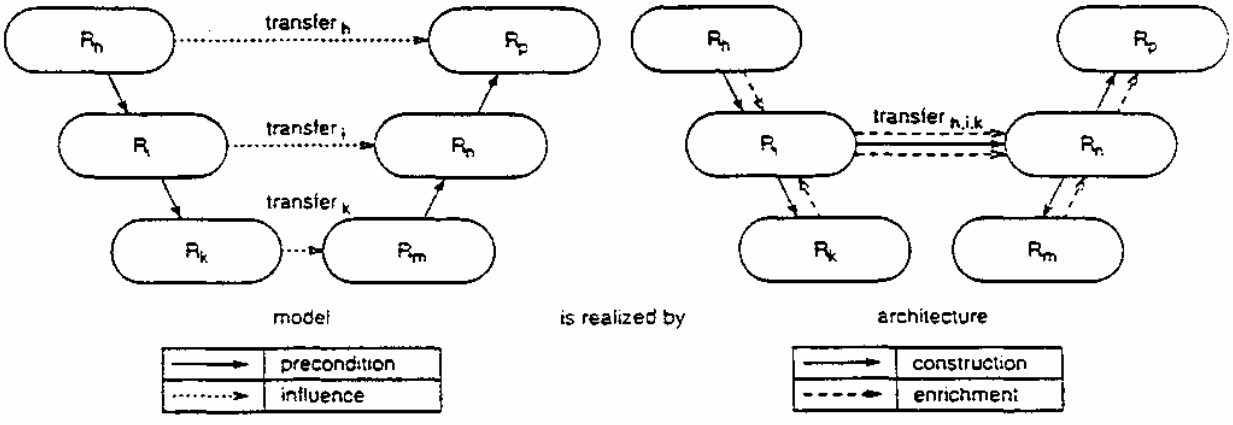

Figure 17 - 7: Enrichment of the transfer level with information which is relevant for transfer

The factors are regarded as constraints which have to be fulfilled by the transfer and synthesis steps in order to generate the corresponding TL factors and sentences, respectively. The transfer has access to background knowledge which is needed for the comparison of two languages (not displayed in Figure 17 - 6).

There are two reasons why I do not pursue Tsujii's proposal. On the one hand a uniform architecture[2] (see [Zajac 91] and [Weisweber 92]) of the MT system becomes impossible because transfer and generation have to be realised by other algorithms than analysis. On the other hand the quasi parallel mappings needed for transfer and synthesis are probably too complex.

An alternative conception has been developed in the project KIT-FAST. In order to have a transfer step only on one level of representation and to have access to the information from all other levels at the transfer level, representations are used which are partially redundant. Figure 17 - 7 sketches how the transfer representations can be enriched with the information from other levels which are relevant for transfer. If the arrows for construction and enrichment point in the same direction then both actions can be performed in one step. The information for the enrichment can be transmitted with the help of the mapping rules needed for the construction. If both arrows point in different directions then the enrichment can be realised for example by (inter-sentential) bindings of variables or by a specific module which has access to the representations involved.

If two-dimensional representations, e.g. sentential ones, are assumed then the architecture of an

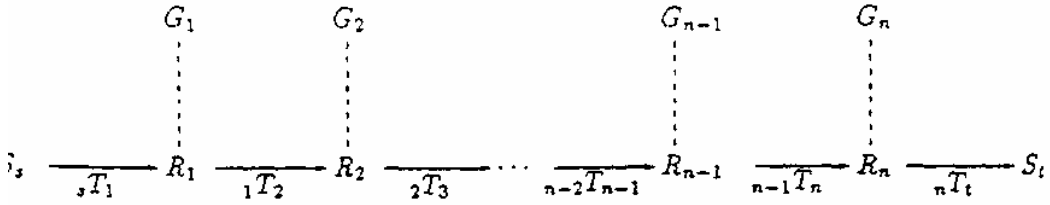

Figure 17 - 8: MT system architecture without text representation and background knowledge

MT system looks like the one in [Arnold et al. 86], which is given in Figure 17 - 8. The disambiguation steps which are theoretically necessary after each mapping $_iT_{i+1}$ are neglected here. They would need the text representations and the background knowledge.

Ss and St are the SL and TL sentences, respectively. The Ri are different sentence representations, which are generated by certain formal rule systems G_i. The G_i can be regarded as parts of the linguistic background knowledge, but a text representation and extra-linguistic background knowledge are not considered here. The mapping systems $_iT_{i+1}$ map one

[2]An MT system has a uniform architecture if there is one single algorithm which is able to perform all mappings from one representation onto an adjacent one.

representation R_i onto a representation R_{i+1}. If the MT system is Interlingua-based then one of the sentence representations is interlingual or at least bilingual[3]. If the MT system is transfer-

based then one of the mappings ${}_iT_{i+1}$ is the transfer system.[4]

At the end of this section the design criteria for the architecture of an MT system are summarised. The following parameters have to be fixed (the decisions for the KIT-FAST MT system are in italics):

- *transfer-* vs. Interlingua-based
- in case of transfer-based: level of transfer
- number and contents of the representations
- technical realisation of the representations and transitions between them
- the representations are free of redundancies or *(partially) redundant*
- *two- vs. more-dimensional representations*
- in case that sentence and text representation are realised on a level: integration of the sentence
- representations into the text representation vs. separation of sentence and text representations
- translation of a text, sentence by sentence or as a whole
- language-specific vs. interlingual background knowledge
- integration of disambiguation and structural mappings vs. disambiguation as filters for representations

The work presented in [Hauenschild 88] and [Preuß et al. 93] indicates that several arguments are against Interlingua-based MT systems. This is valid for interlingual sentence and text representation as well as for interlingual background knowledge.

---

[3]From the technical point of view I think that it is possible to develop a bilingual representation for one pair of languages. This "bilingua" has to be fine-grained enough to represent all properties of SL and TL which are relevant for the translation. But this would lead to the necessity to disambiguate multiple TL readings during analysis and to take SL ambiguities into account during synthesis.

[4]Transfer-based MT systems are best fitted to fulfil the requirements for translation because the contrastive aspect plays a crucial role. Transfer rules allow explicit description of what has to remain equivalent and what has to be changed in a translation (cf. [Hauenschild 88]).

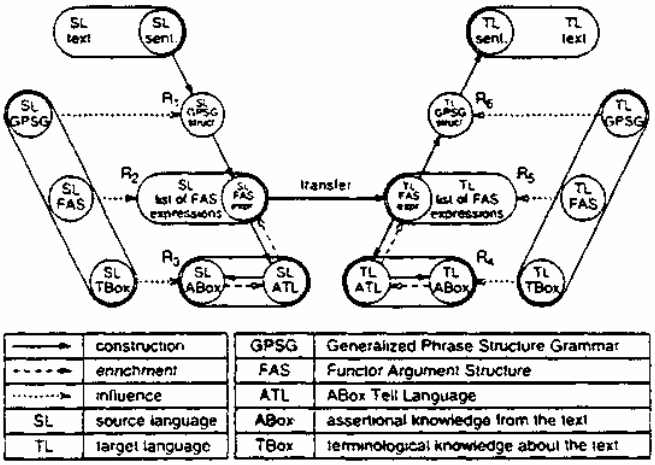

Figure 17 - 9: Architecture of the KIT-FAST MT system

2.2 The architecture of the KIT-FAST MT system

In the project KIT-FAST an experimental transfer-based MT system has been developed. The architecture is shown in Figure 17 - 9, i.e. the representations and transitions between them are given. The representations are partially redundant. The transfer is done on the semantic level (FAS) which is enriched with syntactic and conceptual information. The text is translated sentence by sentence, i.e. for every sentence the corresponding sentence representations are constructed along the direction of the arrows. These representations are integrated into the corresponding text representation (semantic level, FAS) or the text representations are constructed from them (conceptual level, ATL). The transitions are realised with the help of an algorithm on the basis of term-rewriting (cf. [Weisweber 92] and [Weisweber 94]).

The syntactic representations R1 and R6 are two-dimensional. They consist of a GPSG and the corresponding structures generated from it. On the syntactic level there is no text representation. The sentential GPSG structures are used quasi as intermediate representations in order to generate the semantic representation of the TL and to ensure the grammaticality of the TL sentences. The semantic representations R2 and R5 are three-dimensional. FAS expressions which are generated by corresponding context-free grammars. On this level the sentence representations are integrated into the textual one (see section 3.1).

The conceptual representations R3 and R4 are three-dimensional. They are realised with the help of the knowledge representation system BACK (see section 3.1).

3 Anaphora resolution in the KIT-FAST MT system

Anaphora resolution is regarded as the starting point for the investigation of text phenomena in MT. The aim is to disambiguate and translate texts instead of single sentences. In order to achieve this the evaluation of anaphoric relations has to be combined with structural and lexical disambiguation. In general there are different factors which determine the antecedent of an anaphor. Some factors refer to structural properties of an antecedent and others to conceptual properties which have to be represented in the MT system. This is done by the structural and referential text representation sketched in the first section below. The factors realised in the KIT-FAST MT system are outlined in the second section. The algorithm for anaphora resolution and the mutual dependencies are described in the third section. This section is based on the work by [Hauenschild 91], [Schmitz et al. 92], [Preuβ et al. 92], [Preuβ et al. 93], [Dunker/Umbach 93] and [KIT-FAST 93].

3.1 Two levels of text representation

A text can at least be viewed from two different perspectives. On the one hand the text is structured at least by the sequence of the sentences, in which it is uttered or read. But in general more complex structures have to be adopted. On the other hand the text makes predications about some referents occurring in it. For that reason two levels of text representation are realised in the KIT-FAST MT system which reflect these two aspects. They are called structural and referential text representation respectively. The factors for the resolution of anaphoric relations rely on them in order to determine the structural prominence and the conceptual consistency of a possible antecedent.

3.1.1 The structural text representation

The structural text representation includes information about:

- functor-argument relations, e.g. between nouns, verbs and adjectives and their complements,
- semantic roles of arguments, e.g. agent, affected, attribuand, associated, location, aim, etc. (cf. [Steiner et al. 88b]),
- the thematic structure of a sentence (cf. [Sgall et al. 73] and [Firbas 74]),
- semantic features that express local or temporal conceptualisation as known from cognitive sources,
- grammar (cf. [Zellinsky-Wibbelt 88]) and
- anaphoric relations represented by coindexation.

This information is represented by the FAS (see section 1.3). The FAS can be regarded as an abstract syntax with additional semantic features. It represents the functional structure of sentences and does not contain redundant information for checking the wellformedness conditions of the surface syntax as for example syntactic agreement.

Actually the structural text representation consists of a root node with the FAS expressions of the sentences as daughter nodes, such that intra- and extra-sentential constituents, which are co-referent, can be coindexed. For adequate disambiguation, a more hierarchical structure of the text has to be constructed which represents more complex relations between the sentences. A proposal in this direction has been made by [Grosz/Sidner 86].

3.1.2 The referential text representation

The referential text representation contains aspects of the text content, namely the discourse referents and the conceptual relations between them. Co-referent expressions are represented by one discourse referent. Every relation that holds for an antecedent is also valid for an anaphor that refers to it. Not only nouns but also verbs and adjectives are represented as discourse referents, because anaphors can refer to events (the denotation of a verb) and properties (denotation of an adjective) respectively.

The referential information is represented in a terminological logic with the help of the knowledge representation system BACK (cf. [Peltason et al. 89], [Quantz/Kindermann 90] and [Hoppe et al. 93]). The BACK system, which has been developed independently, has been integrated into the MT system. In general such systems distinguish between descriptions and

definitions. Definitions introduce concepts and roles and represent terminological knowledge in the so called TBox. A description describes an object, which is either an instance of a concept or is related to another object by a role, and represents assertional knowledge in the so called ABox. In the TBox, aspects of the background knowledge are represented (cf. [Preuβ et al. 93]). Actually the concept definitions provide information about the semantic type of a lexeme, its semantic roles and the semantic types of the role fillers (selectional restrictions).

The discourse referents are represented by descriptions in the ABox of the BACK system. They are connected by (semantic) roles. The ABox is incrementally constructed with the help of the ABox Tell Language (ATL) which is generated from the sentential FAS expressions (see section 2.2). The references for the discourse referents in the ABox are returned to the FAS level via variable binding.

At the moment the ABox and TBox are only used for the interpretation of anaphoric relations in texts (cf. [Hauenschild 91] and [Preuβ et al. 93]), but in principle they can also be used for other disambiguation purposes.

As already mentioned above the dual text representation allows for distinguishing three aspects of anaphoric expressions (cf. [LuperFoy/Rich 90] for a similar proposal):

1. their position in the linguistic structure,

2. their relation to other linguistic expressions and

3. the type of relation between the discourse referents, which is introduced by the anaphor

(e.g. identity of sense, identity of reference or part-whole relation, cf. [Quantz 92]).

The first two aspects are expressed by coindexation of the corresponding expressions in the structural text representation. With respect to the third aspect the most common type of relation for personal and possessive pronouns is the identity of reference (coreference). This is expressed in the referential text representation by using the same discourse referent for an anaphor and its antecedent.

Our definition of anaphor and antecedent is based on the assumption of a dual text representation. Both are complex objects which consist of the discourse referent and their structural position. It is not sufficient to define antecedent candidates only on the basis of their structural position, because in some cases this leads to spurious ambiguities, which are caused by regarding different structural occurrences of the same referent as different antecedent candidates.

Another reason for the twofold definition of antecedent comes from the binding principle which is described in section 3.2.1. This principle depends on structural information, but has also an influence on the referential structure because if it excludes an antecedent candidate for an anaphor then all coreferential candidates are also excluded.

3.2 Factors for anaphora resolution

In the experimental KIT-FAST MT system every antecedent candidate is evaluated by a set of factors.

Some of them refer to the structural text representation and the others to the referential text representation. The factors can be classified to the groups proximity and binding, themehood, parallelism, and conceptual consistency. According to this classification each factor is outlined.

3.2.1 Proximity and binding

The relative distance between an anaphor and its antecedent is a factor that determines the structural prominence of an antecedent candidate. It is covered by the co-operation of the proximity and binding principle, which reduce the search space for antecedent candidates in a complementary way. The search space can be divided in a local and an outer part on which the binding and proximity principle is applied respectively.

The proximity principle

The proximity principle accounts for the fact that personal pronouns are most likely to have their antecedents in the superordinate or preceding sentence, while there is preference for each possessive pronoun to refer to a noun occurring in the same sentence.

The proximity principle counts the number of superordinated nodes which represent a verbal or nominal predicate or a co-ordinated structure in the structural text representation, ignoring the parts which are excluded by the binding principle. The antecedent candidate with the lowest distance is preferred by the proximity principle.

The binding principle

The binding principle excludes all sisters of a pronominal argument in the structural text representation as antecedent. This definition corresponds to condition B of the HPSG approach (cf. [Pollard/Sag 87]). Additionally the functor of a pronominal argument and functors which c-command the anaphor are excluded, where a constituent of the structural text representation X c-commands a constituent Y if and only if the mother of X dominates Y and Y does not c-command X.

Co-ordinated structures are treated in the way, that generally all constituents with number plural may be possible antecedents, but in cases where the anaphor is part of one of the conjuncts, the corresponding co-ordinated constituent is excluded.

3.2.2 Themehood

This class of factors define structurally prominent constituents as for example the subject or the topic of a sentence. They also contribute to the determination of the text's theme.

Preference for the semantic subject

Since the factors refer to the structural and referential text representations, they have no access to purely syntactic information like subject. For that reason a notion of semantic subject is defined on the basis of the structural text representation.

For every functor a list of arguments is defined in canonical order. The first argument is regarded as semantic subject. In most cases this definition yields the same results as the traditional syntactic definition. Passive verbs, which change the syntactic realisation of the most prominent semantic roles, are an essential exception. If the optional agent role is not filled then the second argument of the list automatically becomes the semantic subject.

Actually we have no adequate solution for the situation in which the agent role of a passive verb is filled. In this case there is a conflict between the semantic and syntactic subject.

Topic preference

This factor refers to the thematic structure of a sentence which is represented by a scalar order from the most to the least topical constituent of a sentence. This definition adequately accounts for the gradability of themehood in languages with free word order. The most topical antecedent candidate is preferred.

Negative preference for free adjuncts

Free adjuncts are represented as fillers of certain semantic roles in the structural text representation and are bad antecedent candidates for personal and possessive pronouns. Free adjuncts seem to be good antecedents for anaphors of their ontological type, but this has not been elaborated up to now.

3.2.3 Parallelism

The fact that anaphors and antecedents prefer to be parallel with respect to some syntactic properties or to their semantic role is captured by factors about agreement or role identity, respectively.

Agreement

The structural text representation includes information about number and gender of a discourse referent. There is preference for an anaphor to refer to an antecedent which agrees with it in these features.

Up to now only syntactic number and gender are considered and their semantic counterparts have not been elaborated, but it would be no problem to add this information to the structural text representation.

Identity of roles

The semantic roles which are filled by the discourse objects are represented in the structural text representation. The antecedent candidate which fills the same role as the anaphor is preferred.

3.2.4 Conceptual consistency

Although even inconsistent propositions can lead to a more or less coherent text we assume that the texts to be translated are consistent. In order to check the conceptual consistency of an antecedent candidate the predications about the anaphor and the antecedent stored in the referential text representation are checked for compatibility. For this a semantic representation of the lexemes and phrases together with background knowledge is necessary, which includes encyclopaedic knowledge and knowledge about translational theory. The representations should support inferences.

The project KIT-FAST decided to begin with modelling selectional restrictions with the help of the TBox of the knowledge representation system BACK (see section 3.1). It is stressed here that this is only a first step towards adequate knowledge representation for MT. The objective is to represent all kinds of background knowledge in the TBox of the BACK system.

3.3 The algorithm for anaphora resolution

The algorithm for anaphora resolution, which is described in [Dunker/Umbach 93], determines the antecedent for an anaphoric expression occurring in the SL text. Actually only the references of personal and possessive pronouns are resolved. The most significant achievement is the

uniform treatment of both types of pronouns. This is possible because we do not treat possessive pronouns as determiners but as arguments of nouns.

All nominal phrases including other pronouns and co-ordinated phrases, that occur in the same or preceding sentences, are taken to be possible antecedents. The number of preceding sentences to be considered is a parameter of the algorithm. In this way intra-sentential cataphoric relations are also covered.

The factors mentioned above are used to evaluate anaphoric relations. They are treated as preference rules and refer to structural properties of an antecedent candidate, e.g. to be the subject of a sentence or the distance (proximity) between an anaphor and its antecedent, as well as to referential (conceptual) properties, which are equivalent for all coreferential objects. Problems arise if an anaphor has a possible antecedent from which it is known that another anaphor refers to it. In this case it is unclear whether the factors should refer to the structural or referential properties or both. In order to remedy this situation the algorithm uses the two levels of text representation outlined in section 3.1.

In order to find the best antecedent for an anaphor the algorithm evaluates every possible antecedent. For that reason each factor adds a positive or negative score to every antecedent candidate of an anaphor. The scores depend on the type of text and have to be found out empirically. Factors can reinforce each other or are in competition. All reinforcing factors assign a score with the same sign. Competitive factors have the same score with different signs. After the application of all factors the antecedent candidate with the highest score and the anaphor become the same object in the referential text representation, i.e. the same ABox object.

Factors which impose excluding constraints on anaphoric relations like agreement and binding distribute very high negative scores. But this would not lead to an exclusion of an antecedent candidate if it violates for example agreement. For that reason the algorithm has two additional parameters: the absolute and relative minimum score. If a candidate is worse than the absolute minimum then it is excluded. If the difference of the scores of a candidate and the best antecedent exceeds the relative minimum then the candidate is also excluded.

For efficiency reasons factors which assign high negative scores should be applied before others with lower scores. For that reason the order for the application is another parameter of the algorithm.

If there is more than one solution the anaphoric relation may be ambiguous even for a human reader because she or he lacks information or the text is not homogeneous. But it is more likely that the factors developed so far are not sufficient. In this case the user has to decide interactively which antecedent is the best one.

It may be the case that there is no antecedent because there is no antecedent candidate in the same or preceding sentences, or all candidates have been excluded for exceeding the absolute or relative minimum. Such anaphors are interpreted as deictic expressions, i.e. they constitute an autonomous object in the referential text representation.

4 Conclusion

This paper describes the systematic development of the architecture of the KIT-FAST MT system from a given MT model. A list of parameters for implementing MT systems has been elaborated, which seems to be sound but not complete. A particularity of the MT system is that it essentially consists of two algorithms:

- an algorithm on the basis of term-rewriting, which realises semantic and conceptual analysis, transfer and generation (cf. [Weisweber 92] and [Weisweber 94]), and
- an algorithm for the evaluation of anaphoric relations (see section 3).

The syntactic analysis is realised by a GPSG parser (cf. [Weisweber 87] and [Weisweber/Preuß 92]). But [Weisweber 94] shows that parsing can also be realised by the term-rewriting algorithm and that it makes sense to do syntactic and semantic analysis in one step.

The main achievements of the algorithm for anaphora resolution are the uniform treatment of personal and possessive pronouns and the introduction of a structural and referential text representation. An evaluation of anaphoric relations is only realised for the SL, but that is not sufficient. If constituents of the SL have to be translated into fixed syntactic constructions or idioms of the TL then it is possible that new discourse referents are introduced, which can be referred to by an anaphor. For that reason the evaluation of anaphoric relations has to be realised for the TL as well. Anaphora resolution is of course only a first step towards the translation of texts with a computer, but it opens many promising perspectives.

The (referential) text representation in the ABox of the BACK system crucially depends on the pre-defined background knowledge in the TBox. It can be expanded in two ways. On the one hand the linguistic knowledge of all levels of representation can be added. On the other hand, forms of background knowledge other than selectional restrictions, as for example encyclopaedic knowledge and knowledge about translational theory, are necessary. Terminological knowledge expanded in the way described can be used for disambiguation purposes other than anaphora resolution.

The structural text representation, which actually consists of a root node and the FAS expressions of the sentences as daughters, can be further elaborated. In order to use it for an adequate disambiguation a more hierarchical text structure has to be constructed. A proposal in this direction has been made by [Grosz/Sidner 86].

The MT system described in this paper has been implemented on a UNIX workstation. It is available in Quintus and SWI Prolog. Linguistic data for a fragment from German to English has been developed and tested. Analysis grammars without anaphora resolution for small fragments of French and Russian have also been implemented.

References

[Arnold et al. 86] J. Arnold, S. Krauwer, M. Rosner, L. de Tombe, G.B. Varile, The <C,A>,T Framework in EUROTRA: A Theoretically Committed Notation for MT, in: Proc. 11th COLING-86, Bonn 1986, pp. 297-303

[Boitet/Gerber 84] C. Boitet, R. Gerber, Expert Systems and other Techniques in MT Systems, in: Proc. 10th COLING-84, Stanford 1984

[Busemann/Hauenschild 88] S. Busemann, C. Hauenschild, A Constructive View of GPSG or How to Make it Work, in: Proc. 12th COLING-88, Budapest 1988, pp. 77-82

[Dunker/Umbach 93] G. Dunker, C. Umbach, Verfahren zur Anaphernresolution in KIT-FAST, KIT Internal Working Paper 28, Technical University of Berlin 1993

[Firbas 74] J. Firbas, Some Aspects from the Czechoslovak Approach to Problems in Functional Sentence Perspective, in: F. Daneš (ed.), Papers in Functional Sentence Perspective, Mouton, Den Haag/Paris 1974, pp. 11-37

[Gazdar et al. 85] G. Gazdar, E. Klein, G. Pullum und I. Sag, Generalised Phrase Structure Grammar, Blackwell, Oxford 1985

[Grosz/Sidner 86] B. Grosz, C. Sidner, Attention, Intentions, and the Structure of Discourse, in: Computational Linguistics Vol. 12 No. 3 (1986), pp. 175-204

[Hauenschild 88] C. Hauenschild, Discourse Structure - Some Implications for Machine Translation, in: D.Maxwell, K. Schubert, A.P.M. Witkam (eds.), New Directions in Machine Translation, Proc. of the Conference, Foris, Dordrecht 1988, pp. 145-156

[Hauenschild 91] C. Hauenschild, Anaphern-Interpretation in der maschinellen Übersetzung, Zeitschrift für Literaturwissenschaft und Linguistik 84 (1991), Vandenhoeck & Ruprecht, pp. 50-66

[Hauenschild/Busemann 88] Ch. Hauenschild, S. Busemann, A constructive version of GPSG for machine translation, in: [Steiner et al. 88a], pp. 216-238

[Hoppe et al. 93] T. Hoppe, C. Kindermann, J. Quantz, A. Schmiedel, M. Fischer, BACK V5 - Tutorial & Manual, KIT-Report 100, Technical University of Berlin 1993

[KIT-FAST 93] Project KIT-FAST: Ch. Hauenschild, B. Mahr, S. Preuß, B. Schmitz, C. Umbach, W. Weisweber, L. Beheshty, G. Dunker, M. Rickard, Ch. Werner-Meier, E. Ziegler, Schlussbericht des Berliner Projekts der EUROTRA-D-Begleitforschung "Anaphorinterpretation in der maschinellen Übersetzung", KIT-Report 108, Technical University of Berlin 1993

[LuperFoy/Rich 90] S. LuperFoy, E. Rich, A Computational Model for the Resolution of Context Dependent References, in: MCC Technical Report, Austin 1990

[Mahr/Umbach 90] B. Mahr, C. Umbach, Functor-Argument Structures for the Meaning of Natural Language Sentences and Their Formal Interpretation, in: K.H. Bläsius, U. Hedtstück, C. Rollinger (eds.), Sorts and Types in Artificial Intelligence, Lecture Notes in AI, Springer, Berlin 1990, pp. 286-304

[Nirenburg et al. 92] S. Nirenburg, Ja. Carbonell, Ma. Tomita, Ke. Goodman (eds.), Machine Translation: A Knowledge-Based Approach, Morgan Kaufmann, San Mateo/CA 1992

[Peltason et al. 89] C. Peltason, A. Schmiedel, C. Kindermann, J. Quantz, The BACK System Revisited, KIT-Report 75, Technical University of Berlin 1989

[Pollard/Sag 87] C. Pollard, I. Sag, Information-Based Syntax and Semantics, Volume 1: Fundamentals, CSLI Lecture Notes No. 13, Stanford 1987

[Preuß et al. 92] S. Preuß, B. Schmitz, C. Hauenschild, Anaphora Resolution Based on Semantic and Conceptual Knowledge, in: S. Preuß, B. Schmitz (Hrsg.), Text Representation and Domain

Modelling - Ideas From Linguistics and AI, Proceedings des Workshops. KIT-Report 97, Technical University of Berlin 1992, pp. 1-13

[Preuβ et al. 93] S. Preuβ, B. Schmitz, C. Hauenschild, C. Umbach, Anaphora Resolution in Machine Translation, to appear in: W. Ramm, P. Schmidt, J. Schütz (eds.), Studies in Machine Translation and Natural Language Processing, Volume on "Discourse in Machine Translation"

[Quantz 92] J. Quantz, Semantische Repräsentation anaphorischer Bezüge in terminologischen Logiken, KIT-Report 96, Technical University of Berlin 1992

[Quantz/Kindermann 90] J. Quantz, C. Kindermann, Implementation of the BACK System Version 4, KIT-Report 78, Technical University of Berlin 1990

[Schmitz et al. 92] B. Schmitz, S. Preuβ, C. Hauenschild, Textrepräsentation und Hintergrundwissen für die Anaphernresolution im maschinellen Übersetzungssystem KIT-FAST, KIT-Report 93, Technical University of Berlin 1992

[Sgall et al. 73] P. Sgall, E. Hajičová, E. Benešová, Topic, Focus and Generative Semantics, Scriptor, Kronberg 1973

[Steiner et al. 88a] E. Steiner, P. Schmidt, C. Zellinsky-Wibbelt, From Syntax to Semantics. Insights from Machine Translation, Frances Pinter, London 1988

[Steiner et al. 88b] E. Steiner, U. Eckert, B. Roth, J. Winter-Thielen, The Development of the EUROTRA-D System of Semantic Relations, in: [Steiner et al. 88a] pp. 40-104

[Tsujii 86] J.-I. Tsujii, Future Directions of Machine Translation, in: Proc. 11th COLING-86, Bonn 1986, pp. 655-668

[Weisweber 87] W. Weisweber, Ein Dominanz-Chart-Parser für generalisierte Phrasenstrukturgrammatiken, KIT-Report 45, Technical University of Berlin 1987

[Weisweber 92] W. Weisweber, Term-Rewriting as a Basis for a Uniform Architecture in Machine Translation, in: Proc. 14th COLING-92, Nantes 1992, pp. 777-783

[Weisweber 94] W. Weisweber, Termersetzung als Basis für eine einheitliche Architektur in der maschinellen Sprachübersetzung, Sprache und Information 28, Niemeyer Verlag, Tübingen 1994

[Weisweber/Hauenschild 90] W. Weisweber, C. Hauenschild, A Model of Multi-Level Transfer for Machine Translation and Its Partial Realization, KIT-Report 77, Technical University of Berlin 1990 and in: Proc. Seminar Computers & Translation '89, Tbilisi 1989

[Weisweber/Preuβ 92] W. Weisweber, S. Preuβ, Direct Parsing with Metarules in: Proc. 14th COLING-92, Nantes 1992, pp. 1111-1115

[Zajac 91] R. Zajac, A Uniform Architecture for Parsing, Generation and Transfer, in: T. Strzalkowski (ed.), Proc. Workshop on Reversible Grammar in NLP, Berkeley 1991, pp. 71-80

[Zellinsky-Wibbelt 88] C. Zellinsky-Wibbelt, From Cognitive Grammar to the Generation of Semantic Interpretation in Machine Translation, in: [Steiner et al. 88a], pp. 105-132