Terminological Knowledge in Multilingual Language Processing

Jörg Schütz

IAI, Saarbrücken, Germany
joerg@iai.uni-sb.de

Abstract

Over the last decade (1984-1994) machine translation (MT) systems, commercial systems as well as research prototypes, have steadily improved regarding their linguistic capabilities. However, they still lack the command of language a human translator possesses, in particular with regard to the interpretation of the textual units to be translated in their contextual, situational and cultural background.

In this decade, the most innovative approaches are those that are based on *unification grammars* (UG). UGs were first introduced by Martin Kay as Functional Unification Grammar (FUG) and suggested for MT ([Kay, 1984]). In the field of MT, the unification paradigm was first adopted by the EUROTRA project of the European Commission ([Copeland et al. (Eds.), 1991]), which aimed at developing an MT system for all nine languages of the European Union. However, this very ambitious goal was not achieved during the life cycle of EUROTRA (1985-1992). To some extent, it was the basis for more successful MT research prototypes based on unification, such as the CAT2 system of IAI, Saarbrücken (an official EUROTRA sideline), and the LFG-based CHARON system of IMS, Stuttgart. On the international MT scene, unification was applied in particular by the AI inspired interlingua- and knowledge-based systems, e.g. [Nirenburg et al., 1992].

Today, there exist several unification-based systems which are not entirely dedicated to MT but to natural language processing (NLP) in general. What these systems lack is efficient processing[1], but they are very well suited for language engineering (LE) because unification grammars are the first grammar models that are shared by theoretical and applied computational linguists. Recently, the research work in this field has concentrated on better general algorithms and the restriction of the inventory of applied logic.

1 Introduction

LE in general is faced with the following serious problem areas:

1. The cost factor, because currently LE is very costly in terms of the resources and manpower needed.

[1] The Trace Unification Grammar (TUG) system of Siemens AG, Munich, might be an exception in this respect.

2. The time factor; at the moment the development of grammars designed with recent unification formalisms and with a large coverage takes about eight to ten years.
3. The lack of concepts for real distributed grammar development.
4. The unavailability of optimal tools for LE.
5. The non-existence of reusable grammar and lexicon resources.

However, recently, larger *competence grammars* with a broad coverage are under development. What they lack is extensibility, transparency, algorithmic independence and very good runtimes. For the resolution of these problems we need *performance models* that integrate knowledge processing into the linguistic processes, filter out improbable readings and allow for dealing with uncovered input. Applying such performance models to existing competence grammars will lead to real usable grammars which can be characterised by their:

- application-oriented grammatical coverage,
- efficient runtime behaviour, and
- robust processing.

For the specification of a performance model we have several possibilities, which are mainly based on the improvement of the underlying computational interpretation strategies. We distinguish between[2]:

1. Algorithmic control which is concerned with the underlying processes and data structures;
2. Compilation techniques;
3. Coverage modification.

Algorithmic control includes the ordering of the linguistic tasks involved during processing, the suppressing of solutions (*clipping*), the suppressing of failures (*grafting*) and the collapsing or expanding of formalism specific expressions. Compilation techniques are mainly concerned with the reduction of formal language expressions into efficient code, e.g. *bit vector* representations and vector operations for efficient unification. Coverage modification can be based on specific reading distinctions according to the subject field of an application or an explicit reduction of the grammatical constructions (sublanguage grammars, restricted language).

In this paper we will argue for terminology-based MT as one particular instance of performance control, and discuss the various knowledge sources involved. The performance control that we have implemented is an instance of the coverage modification which is achieved by filtering language construction according to domain specific information, i.e. the terminological/conceptual knowledge of the subject field. In this sense it is a pure technical application which does not take into account any psycholinguistic dimension, such as the performance control of humans which certainly would lead into a philosophical discussion. We have restricted our work to the domain of telecommunications, in particular that of satellite communication.

[2] This classification is due to Hans Uszkoreit. I am grateful to him for sharing the ideas about performance control with me.

Our aim is to provide for comprehensive knowledge resources which are organized so that sublanguage information and extra-linguistic information about a specific domain is accessible in a concise, efficiently machine-tractable form, and which are formalized so as to ensure consistency across organizations of related grammatical and lexical strata. However, the depth of the natural language (NL) analysis is restricted to the needs of the translation task, i.e. the access to the different information sources is defined by the application.

To achieve this, different sorts of knowledge are used to build a *sublanguage information repository* which can be applied within a unification-based natural language processing (NLP) framework. The approach has been implemented and tested in the Advanced Language Engineering Platform (ALEP) environment ([ALEP, 1993]), a general purpose NLP development platform, based on an object-centered architecture and a typed feature logic based linguistic formalism, promoted by the European Commission (EC) for the Linguistic Research and Engineering (LRE) action line and the forthcoming Fourth Framework Programme.

The primary driving force of our MT approach is the conceptual organization of the domain of telecommunications. Its purpose is to provide domain-specific constraints that ensure the control of the analysis, translation and generation process of sublanguage expressions. On the one hand, this is done by providing a model of the domain - the ontology - which represents the concepts of the domain and their generic and partitive relationships. On the other hand, knowledge about the terminology of the domain in terms of conceptual roles and conceptual modifiers defines the multi-dimensional relationships of the concepts of the domain. When linked together these sorts of knowledge correspond to the *intensional meaning* of a sublanguage expression (proposition).

The overall leading idea for the integration of the conceptual (terminological) knowledge into the linguistic processes is to control a competence grammar for general language by means of conceptual (terminological) constraints. The actual engineering is carried out entirely on a lexical basis by so-called *terminological anchors*, which provide the links from the different conceptual dimensions to the general semantic relations of the competence grammar. What is new in our approach is that no specialized interface between the different sorts of knowledge has to be designed because they are modelled using the same formal device, the ALEP formalism. The advantage we gain from this approach is that we have the *full grammar* as a *fall-back* in cases were the conceptual knowledge cannot contribute to the disambiguation process, due to ambiguities inherited from the domain itself.

2 Sublanguage Information and NLP

2.1 NL Analysis controlled by Conceptual Information

The development of the ontology has proceeded from two global research strands: the acquisition, organization and representation of knowledge in lexical and terminological resources, and their conceptual modelling. Since the overall purpose of the conceptual structure is to be maximally supportive for the computational processing of sublanguage expressions in an NLP environment, there is a third direction from which research on this topic has proceeded: the investigation of the linguistic realization of terminological expressions. in their sentential and textual context, of a corpus dealing with the domain, and the investigation of the question of what the sentential and textual context may contribute to the (human and computational) interpretation of terms. Here, the main focus is on an extended conceptual and linguistic analysis of the corpora, which in particular takes into account the role of

domain dependent and general language verbs within the specific subject field, and on how this analysis may support the entire conceptual analysis of the domain.

The major concepts of the domain, i.e. those concepts which are realized by nouns, nominalized verbs and verbs (in terms of processes), are represented in the domain's ontology; they are characterized by *descriptors* which list the properties of the real world thing the concept denotes. For example, the concept TELECOMMUNICATION_EQUIPMENT can be defined as being a subconcept of EQUIPMENT (the generic relation among concepts of the ontology) with multidimensional relations, such as *input*, *output*, *location*, *channel_capacity*, *frequency_range*, *linearity* and *digital_rate* which must have values of type SIGNAL, WAVE, EARTH_STATION, CAPACITY_VAL, FREQUENCY_VAL, LINEARITY_VAL and DIGITAL_RATE_VAL respectively.

These relations are specified in so-called *term definition forms* according to ISO and DIN specifications, as well as in existing de facto standards in the subject field, developed within the EC-sponsored ET-10 project on *'Terminology and Extra-linguistic Knowledge'* ([Ripplinger et al., 1994]). Part of the information, i.e. a general classification schema, was derived from an existing multilingual termbank of the domain of telecommunications (EIRETERM) which was developed in the context of the MT project EUROTRA.

Since this information is not sufficient for the envisaged NLP task, we have further enhanced these descriptions by a thorough textual and conceptual analysis of a domain corpus, in particular by analyzing the verbs of the domain, which enabled us to define so-called *conceptual templates* ([Schütz, 1994]). Such a template consists of a number of properties that characterize a general concept as either a type, i.e. a thing that can have instances, or a class which governs types that specialize the class. The common classes are ENTITIES, SITUATIONS and PROPERTIES. ENTITIES are those types that can have real world instances and which are realized linguistically as nouns and nominalized verbs, i.e. a subset of the elements of the ontology. SITUATIONS are facilitated by types that express *time* and *place relations*, and that identify *participant*, *agent* and *result roles* (STATES and EVENTS), i.e. the PROCESS subset of the ontology. PROPERTIES are types that denote modifiers (adjectives and adverbs) that describe details of a thing (e.g. MEASURE_VAL), relationships (RELATION) that identify relational properties to other things (RELATED_THING), types that denote attributes that (partially) describe a thing (PARTS), and types that denote constraints which are logical assertions that impose some restrictions on one or more properties of a thing (CONSTRAINTS). This classification schema is derived from known classifications in *compositional semantics* (cf. e.g. [Jackendoff, 1990] and [Pustejovsky, 1991]) and knowledge-based NLP (e.g. the Text And Meaning Representation Language - TAMERLAN - of [Nirenburg et al., 1992]).

The term definition forms and the conceptual templates can be automatically transformed into the typed feature structure representation of the ALEP formalism ([Ripplinger et al., 1994]). This TERM_FS structure contains general terminological information, i.e. the classification schema as provided by the EIRETERM termbank and the concept definition, the concept feature which identifies the CONCEPT and thus provides the link to the ontology of the domain, the CONCEPT_ROLES_FS structure which specifies the role slots of the concept, derived from the SITUATION class and parts of the PROPERTY class, and the conceptual modifiers which are listed in the CONCEPT_MODIFY_FS structure, also derived from the PROPERTY class.

For the semantic descriptions of general language we have used the *semantic relations* (SRs) approach developed in the MT project EUROTRA for German. The SRs define the SEM_FS feature structure of an HPSG inspired competence grammar for German, which specifies a *functor-argument-modifier* structure. The domain-specific conceptual information is associated with these relations: the concept type (CONCEPT) is associated with the semantic functor, the conceptual frame elements (CONCEPT_ROLES) with the semantic arguments and

the conceptual modifiers (CONCEPT_MODIFIY) with the semantic modifiers. The TERM_FS structure is embedded in the SEM_FS structure to permit the testing and evaluation of different sublanguage templates in an appropriate way (modularization).

The semantic and conceptual information structure (SEM_FS with TERM_FS) is embedded together with the syntactic (SYN) and phonological information (PHON) in the overall SIGN feature structure. This organization establishes, on the one hand, the global structure of a KB entity and, on the other hand, the complete lexical information for the implementation. With this organization of information the NL analysis is controlled either by unification or by inferences on the sentence level (for which the competence grammar is designed) in order to check, for example, *selectional restrictions* and *subcategorization frames* based on general semantic and domain-specific information, *type coercion* for the identification of metaphorical senses, or *conceptual classification* information (generic and partitive relations).

The application of these information structures to the analysis process results in a language-independent representation of the intension expressed in a sentence by means of a conceptual organization. We call this the *micro-structure* of the sentence; this term is adopted from evaluation strategies applied to human translations, where similar representations are employed (cf. [Gerzymisch-Arbogast, 1994]). This micro-structure can then be used as input for multilingual language processing such as translation (cf. below).

2.2 Conceptual Information and Translation

In general terminography, such as the EIRETERM database, the focus is on concepts and their linguistic form expressed in terms which are extracted from texts (term identification). In translation the focus is on *production*, i.e. a dynamic process, concerned with the movement from the textual substance in one language to the textual substance in another language. Inside this process there is a procedure in which *units of meaning* of one culture are matched with those of another before finding their textually and situationally appropriate linguistic realization. In view of terminology these units are not of interest because they are temporary and casual collocations of concepts brought into a particular relationship by an author. Translation has to work with concepts and terms in context, whereas terminology isolates terms from their context (decontextualization) and then associates them with concepts, i.e. matching between term and concept vs. matching between textual units through concepts.

Concept correspondence is discovered when comparing the terminologies of different languages, subject fields and cultural systems. Based on this assumption there are thus four possibilities for the process of translation based on the *intension* of a conceptual representation. By intension we mean the set of characteristics, i.e. the formal representation of the properties of an object serving to form and delimit its concept, which constitutes the concept. We distinguish:

1. *Complete co-incidence* of intensions, i.e. the conceptual meaning can be expressed in the languages under consideration in terms of a linguistically realized proposition.
2. *Inclusion* of one intension in the other, i.e. there are conceptual meanings of a concept which do not exist in another language, for example, the concept PALACE has one specific meaning which is only valid in a monarchy. Another example of this kind is the process DIE in its metaphorical meaning in telecommunications and computer science: in English we may have the realization with an active verb *'The signal died.'* but in German this has to be realized by the ergative verb 'abbrechen' (break down), i.e. *'Das Signal brach ab.'*. This is in contrast to *'Hans brach das Signal ab.'* (∗*'Hans broke down the signal.'*) vs. *'Hans killed the signal.'*

3. *Overlapping* of intensions, i.e. there are in either language conceptual meanings of a concept which do not have a corresponding value. For example, the concept PICTURE, which in English is a superconcept of PAINTING, DRAWING and PHOTOGRAPH, has no direct correspondence in Japanese. Only the subconcepts have such a correspondence.
4. *No co-incidence* of intensions, i.e. either the concept does not exist in another language, or the conceptual meanings are different. For example, the term *zapping* with its meaning of the frequent switching between TV channels didn't exist in German a few years ago. In the field of virtual reality (VR), we can find many of these examples.

Cases 2, 3 and 4 above are called *conceptual* or *intensional mismatches*. Mismatches are mostly caused on social, political and cultural grounds, although the conceptual structures are not bound to particular languages.
Case 1 needs no specific translation rule. Cases 2 and 3 need inferencing capabilities over the concept system for the identification of *common* superconcepts, which, however, will cause a degradation of the granularity of the concept's intensional description. In order to keep the granularity of the source and target language as close as possible, as well as to save costly inferences during generation, it might be worth considering the application of explicit translation rules, as is done for case 4. In the actual implementation (cf. below), we have applied the latter approach, due to the missing inference capabilities in the current ALEP system.

3 Demonstrator Implementation

In the previous sections we have briefly outlined the theoretical framework for the integration of different sorts of knowledge into the analysis and translation process of an NLP system: in this section we describe the actual implementation in the ALEP framework.

3.1 Implementation Overview

The general architecture of our analysis module is based on *staged processing*, which was selected for reasons of efficiency (runtime behaviour). In our approach, analysis is therefore composed of two steps: 1. *shallow syntactic analysis* for efficient parsing with a competence grammar for German, and 2. *conceptual refinement* of the parsing result as performance control.

With the second step we achieve a sublanguage-specific filtering of the parsing results. For parsing we have used the grammar and the parts of the lexical entries which specify the syntactic and phonological information, including the terms of the domain, but without any particular domain information. For the refinement process (filtering) we have used those parts of the lexical entries which specify the general semantics and the domain-specific information. In this step the grammar rules function as the navigator through the parsing structures; the actual filtering process is done by unification (cf. below).

For the translation module which has been designed for mapping German analysis output (so-called *linguistic structures*) to English synthesis input, we have adopted an approach which calls translation on a specific type contained in the top-most feature structure of the input linguistic structure, i.e. the conceptual (sub-) feature structure. At the moment, compared to the German analysis module, the transfer module as well as the English synthesis modules have a limited coverage. This is mainly due to the fact that the focus of our work was on the conceptual organization of the domain and the performance control of the analysis process through conceptual knowledge.

3.2 Knowledge Sorts

The formal specifications for the conceptual and sortal (semantics) organization can be directly expressed in terms of the *type system* facility of the ALEP formalism (cf. above).

In the parsing grammar we have specified the information distribution of the semantic feature structure (SEM_FS) which includes as a substructure the conceptual knowledge organization (TERM_FS) about the domain. During parsing these information slots are opened, and during refinement they are filled in. by the appropriate information by unification. Unification failure then triggers the disambiguation process in the refinement phase and thus the performance control in analysis.

In the refinement part of the lexicon we have stated the selectional restrictions for different semantic and conceptual reading distinctions, as well as the appropriate subcategorization frames and type coercion information. This information is used during the refinement process to identify valid parsing results by unification. The result of the refinement process is a fully specified intensional representation according to the selected semantic and conceptual information.
Consider, for instance, the lexicon entry for *adaptieren (adapt)*; in the entry the semantic subject *agent* is linked to the conceptual role *agent* which is of type EQUIPMENT, which is a type of the domain's ontology, and the semantic object *affected* is linked to the conceptual role *result*, which is of type SIGNAL.
Selectional restrictions based on specific domain information for nouns are linked to the noun's subcategorisation frame and which can be associated with the appropriate prepositions, such as *von (of)* and *mit (with)* which have a specific interpretation in the domain, e.g. *'... die Abstimmung von Hochfrequenzträgern mit Niedrigfrequenzsignalen ...'* (... the modulation of very high-frequency carriers with low-frequency signals ...).

Similar to these selectional restrictions, domain dependent restrictions, for example for the subject/object identification, can be formulated, e.g. *'Fernübertragungsausrüstungen umfassen auch Modulationsgeräte.'* (Telecommunication equipment also comprises modulating equipment.). In this example, the concept associated with the object must be more specific than the concept assigned to the subject (generic relationship).

According to the domain-specific information, the sentence *'Diese Geräte überlagern die Audiofrequenzsignale auf der IF-Trägerwelle.'* (This equipment superimposes the audio-frequency signals on the IF-carrier.) is well-formed, as opposed to the sentence *'Diese Geräte überlagern die Audiofrequenzsignale auf der Erde.'* (This equipment superimposes the audio-frequency signals on the earth.) which is not well-formed, although grammatically correct.

3.3 Translation Relations

Within the translation module there is one rule for initializing the translation process. Once translation is called on the conceptual (sub-) feature structure specified as the value of the linguistic structure's top-most SEM_FS structure. translation is called recursively on type SEM_FS and all subordinate types respectively.
In cases of complete co-incidence of source and target structures, no specific translation rule is applied; only in cases of mismatches are explicit translation rules applied. In this case, when translation is called on type SEM_FS, the predicate string specified by the *pred*-attribute of the functor feature structure is translated from one language into the other. For the translation of the appropriate conceptual information, rules for the different conceptually dependent arities are then used. This approach also allows for

a straightforward account of instances of complex transfer where changes have to be performed according to the argument structure of the predicate that has to be translated.
A domain-specific role structure of a concept, identified by the TERM_FS attribute *concept_roles*, is translated by a rule dedicated to the relevant subtypes of type CONCEPT_ROLES_FS. For instance, the role structure assigned to the predicate is translated by a rule operating on the conceptual role subtypes and calling recursively for translation on type CONCEPTUAL_FS which is the type assigned to the roles of a concept. Type CONCEPTUAL_FS will, then, be translated by a rule which, in turn, calls for translation on type TERM_FS again.
The translation of the modifier-list of a concept in TERM_FS, finally, is performed by distinct rules with each of them accounting for a specific number of elements specified in the modifier list (including the empty modifier list).

In each case, the result of the translation is a fully specified conceptual representation of the intension of the analyzed sentence. In cases of mismatches, the representation is augmented by an appropriate semantic description for ease of generation.

3.4 Synthesis

Ideally, the basic SIGN feature structure and, more specifically, the conceptual feature structures should be the same for all languages. With this assumption, it should only be the syntactic feature structure which has to be revised in designing the type and feature specification for an English synthesis grammar.

Since no refinement can be applied in synthesis (in the current ALEP release), the English synthesis grammar operates in one step. Here, the conceptual descriptions (in some cases augmented by general semantics) trigger the access to the (generation) lexicon.

4 Extended Example

We illustrate the actual processing of our approach by an explanation of the different internal representation levels for the sentence *'Diese Geräte überlagern die Audiofrequenzsignale auf der IF-Trägerwelle.'*.

Due to the fact that we have not applied a morphological analysis we have used a full-form lexicon for parsing and a separate lexicon for refinement. This design decision was made, since the current ALEP release does not provide a *lexicon specifier* with which different sorts of lexicon accesses can be specified (cf. above). Thus, we are not able to control the lexicon lookup for the morphemes in the case of a morphological analysis, and the appropriate pattern (substructures) of the semantic feature structure in the case of parsing. This would result in a dramatic increase of time and (temporarily dynamic) space requirements at runtime. However, our current design decision accounts for a well balanced trade-off between the overall time and space requirements: time and dynamic space is drastically reduced, and static space is slightly increased because of the separate refinement lexicon[3].

[3] The second ALEP release contains the missing *lexicon specifier*. Thus, the lexicons can be merged and the appropriate access is controlled by the *lexicon specifier* for parsing and refinement.

4.1 Shallow Parsing

According to the conceptual analysis of the domain, the different terms and non-terms of the sentence have to be linked to the concepts EQUIPMENT (*Gerät*), AUDIO_SIGNAL (*Audiofrequenzsignal*) and IF_CARRIER (*IF-Trägerwelle*) which are organised in the monotonic inheritance lattice of the ontology of the domain. Special attention has to be taken for the verb *überlagern* which has to be linked to the process concept SUPERIMPOSE with its conceptual environment and the preposition *auf* which, in the domain, accounts for a nominal phrase of a specific type, at least the type has to be PRODUCT. The information about the process concept and its environment, as well as the concept types that can be assigned to the conceptual location relation, are specified in the refinement lexicon as *terminological anchors*.

The processing starts with a *segmentation* of the input sentence and a *lift operation* which assigns an underspecified dominance relation to the input sentence, which permits the identification of the parsing *axiom* (transformation into a feature structure representation). Operating on this input structure the *head-driven* parser produces, according to the competence grammar for general language, different parsing results either in separate structures (basic version) or in a packed representation (record version). In the record version of the parser structure packing is controlled by the subsumption relation that holds between feature structures.

4.2 Conceptual Refinement

The refinement process is started as soon as the first parsing result is available (basic version) or after parsing is completed (record version). The refinement process then operates on the internal parsing representations and is controlled by either a specific *refinement grammar* or the parsing grammar, and a *refinement* lexicon. For our implementation we have chosen the latter approach. Thus the parsing results are further constrained only by lexical information. In this case the parsing grammar is used for controlling the traversal of the parsing results. When having defined an additional refinement grammar it is also possible to constrain the refinement process with specific structural conditions. For our purpose, i.e. constraining the language analysis by terminological information, the lexicon-based approach has proven as being sufficient, also with respect to the time behaviour of ALEP.

The parser produces four results which are ambiguous in two respects: one ambiguity on the structural level which is concerned with the attachment of the prepositional phrase, and one on the functional level which is concerned with the assignment of the subject and object relations. Figure 14 - 1 shows the ambiguous structures which are the input to the refinement process.

The subject/object ambiguity is resolved through the conceptual classification of the agent and patient types of the concept SUPERIMPOSE which have to be EQUIPMENT and SIGNAL respectively. The refinement lexicon entry that resolves this ambiguity is listed[4] in Figure 14 - 2. The entries for *Gerät* and *Audiofrequenzsignal* are shown in Figure 14 - 3 and 14 - 4.

In order to account for the specific location relation of the preposition *auf* in the domain, the type of the argument of the preposition is being restricted to CARRIER as can be seen in Figure 14 - 6. Thus, a sentence like *'Diese Geräte überlagern die Audiofrequenzsignale auf der Erde.'* would be rejected by the refinement process. Figure 14 - 5 lists the entry for *IF-Trägerwelle*.

[4] For reasons of readability we show the representation in the ALEP macro notation.

```
S [ NP [ diese Gerate ] - SUBJECT / OBJECT
    VP [ VP [ Vfin [ uberlagern ]
              NP   [ die Audiofrequenzsignale ] - SUBJECT / OBJECT
         PP [ auf der IF-Tragerwelle ]
       ]
  ]


S [ NP [ diese Gerate ] - SUBJECT / OBJECT
    VP [ Vfin [ uberlagern ]
         NP   [ Det  [ die ]                          - SUBJECT / OBJECT
                NP [ N [ Audiofrequenzsignale ]
                     PP [ auf der IF-Tragerwelle ]
                   ]
              ]
       ]
  ]
```

Figure 14 - 1 *Abbreviated parsing results*

```
uberlagern ~
mLEXde_SIGN_refine[
        mLEXde_SYN_MAJOR[ _,
                mLEXde_HEAD_V[],
                m_SUBJ[ sign:{m_COMPL_N[nom,ARG1]})],
                m_SUBCAT_1[ sign:{m_COMPL_N[acc,ARG2]})]],
        m_SEM_term_yes[ m_GOV_V[uberlagern,action],
                m_ARGS_BI[ m_ARGselec[agent,ARG1,sem_fs:{term=>term_yes:{
                                                        concept=>equip:{}})],
                           m_ARGselec[affected,ARG2,sem_fs:{term=>term_yes:{
                                                        concept=>signal:{}})]],
                term_yes:{
                  term_info     => term_info_fs:{
                                        class         => class_fs:{
                                                           c1_type => tcomm,
                                                           c2_type => tranm,
                                                           c3_type => process},
                                        definition    => _,
                                        form          => no_mwt},
                  concept       => superimpose:{},
                  concept_roles => bi_c_role:{
                                     concept_role1 => conceptual_fs:{
                                                        concept_role  => agnt:{},
                                                        concept_type  => equip:{},
                                                        concept_descr => term_fs:{}
                                                                         },
                                     concept_role2 => conceptual_fs:{
                                                        concept_role  => ptnt:{},
                                                        concept_type  => signal:{},
                                                        concept_descr => term_fs:{}
                                                                         }
                                                   },
                  concept_modify => _ })].
```

Figure 14 - 2 *Refinement entry for 'überlagern'*

```
gerat ~
mLEXde_SIGN_refine[
        mLEXde_SYN_MAJOR[ _,
                mLEXde_HEAD_N[_], [], []],
        m_SEM_term_yes[
                m_GOV_N[gerat,no,_,common,_,
                        m_N_PROPS[ concrete:{}, artificial:{},
                                m_STRUCT_PROP[ count, inhomogeneous, nil_compl, nil_gran],
                                nil_temp:{},
                                m_SPAT_PROP[ nil_shape:{}, nil_norm, nil_intr]]],
                m_ARGS_ZERO[],
                term_yes:{
                  term_info     => term_info_fs:{
                                        class         => class_fs:{
                                                              c1_type => tcomm,
                                                              c2_type => genter,
                                                              c3_type => equipment},
                                        definition    => 'General communication system supply',
                                        form          => no_mwt},
                  concept       => equip:{},
                  concept_roles => _,
                  concept_modify => _ }]].
```

Figure 14 - 3 *Refinement entry for 'Gerät'*

```
audiofrequenzsignal ~
mLEXde_SIGN_refine[
        mLEXde_SYN_MAJOR[ _,
                mLEXde_HEAD_N[_], [], []],
        m_SEM_term_yes[
                m_GOV_N[audiofrequenzsignal,no,_,common,_,
                        m_N_PROPS[ semiotic:{}, nil_animacy:{},
                                m_STRUCT_PROP[ count, inhomogeneous, nil_compl, nil_gran],
                                nil_temp:{},
                                m_SPAT_PROP[ nil_shape:{}, nil_norm, nil_intr]]],
                m_ARGS_ZERO[],
                term_yes:{
                  term_info      => term_info_fs:{
                                        class         => class_fs:{
                                                              c1_type => tcomm,
                                                              c2_type => genter,
                                                              c3_type => product},
                                        definition    => 'An analogue electrical signal',
                                        form          => mwt},
                  concept       => audio_signal:{},
                  concept_roles  => _,
                  concept_modify => _ }]].
```

Figure 14 - 4 *Refinement entry for 'Audiofrequenzsignal'*

```
'IF-Tragerwelle' -
mLEXde_SIGN_refine[
        mLEXde_SYN_MAJOR[ _,
                mLEXde_HEAD_N[_], [], []],
        m_SEM_term_yes[
                m_GOV_N['IF-tragerwelle',no,_,common,_,
                        m_N_PROPS[ concrete:{}, artificial:{},
                                m_STRUCT_PROP[ count, inhomogeneous, nil_compl, nil_gran],
                                nil_temp:{},
                                m_SPAT_PROP[ surface:{}, nil_norm, nil_intr]]],
                m_ARGS_ZERO[],
                term_yes:{
                  term_info      => term_info_fs:{
                                        class         => class_fs:{
                                                           c1_type => tcomm,
                                                           c2_type => genter,
                                                           c3_type => product},
                                        definition    => 'A wave in an intermediate
                                                          processing stage',
                                        form          => mwt},
                  concept        => intermediate_frequency_carrier:{},
                  concept_roles  => _,
                  concept_modify => _ }]].
```

Figure 14 - 5 *Refinement entry for 'IF-Trägerwelle'*

```
auf_PLACE_POS -
mLEXde_SIGN_refine[
        mLEXde_SYN_MAJOR[ _,
                mLEXde_HEAD_P[_], [],
                m_SUBCAT_1[ sign:{m_COMPL_N[dat,ARG ]}]],
        m_SEM_term_yes[
                m_GOV_P[auf,place_pos,qual_place_pos],
                m_ARGS_MONO[
                        m_ARGselec[ prep_arg, ARG,
                                m_ARGselec_N_Term[ _,
                                        m_N_PROPS[ _, _, _, _,
                                                m_SPAT_PROP[ surface:{}, _, _]],
                                        carrier:{}]]],_]].
```

Figure 14 - 6 *Refinement entry for preposition 'auf'*

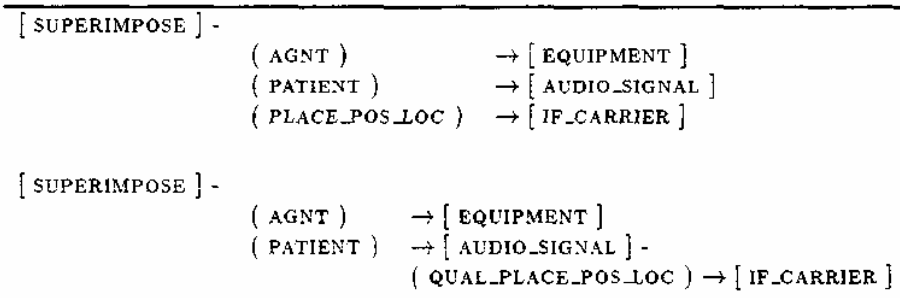

Figure 14 - 7 *Simplified conceptual refinement results*

Since in the domain, according to the information we have got so far, both locational reading distinctions, i.e. a place position of type PLACE_POS_LOC that is associated to the location role of the concept SUPERIMPOSE, and a quality place position of type QUAL_PLACE_POS that modifies the patient role of the concept, are possible, the refinement process has two output structures which are shown in Figure 14 - 7 in an abbreviated simple conceptual graph ([Sowa, 1991]) notation. This is the information which is represented in the TERM_FS feature structure associated with the axiom of the parsing grammar.

The representation is language independent and forms the conceptual part of the complete analysis result. In order to generate from such a representation a sentence in another natural language, we have to enhance the conceptual representation with additional semantic properties which are represented in the subfeature structures of type SEM_FS of the axiom, i.e. the feature structures of the attributes *n_props*, *n_temp_prop* and *n_spat_prop* of the agent and patient roles, and the location relation. To select such a feature structure, i.e. parts of SEM_FS and the complete TERM_FS feature structures, the ALEP system does not provide a direct mechanism. Therefore, we have decided to use the ALEP translation formalism to filter out the appropriate information structures. This operation is briefly described above.

Since we have used a separate feature structure for the conceptual information, i.e. the feature structure of type TERM_FS, instead of integrating it entirely into the SEM_FS feature structure, an exchange of the information structures between several domains is easy to perform; this being an additional advantage of our approach.

The following table is the listing ALEP produces for the sample sentence. In order to show the different parsing results we have used the basic version of the parser which results in higher runtime figures. In cases where the refinement process filters a reading (`refine_failure`) the respective non-matching structure is given in its internal representation (`No lexical entry for ld( ...)`); we have omitted this for readability reasons. The last time figure gives the total runtime for the analysis of the sentence.

```
text analysis succeeded
CPU time of basic_lift is 0.020000.
CPU time of basic_analyse is 11.470000.
No lexical entry for ld(sign(spec_fs(de,proc_spec(_2621,y,_2623),data_type( ...
CPU time of refine_failure is 0.700000.
refinement_failed
CPU time of basic_analyse is 1.450000.
No lexical entry for ld(sign(spec_fs(de,proc_spec(_3291,y,_3293),data_type( ...
CPU time of refine_failure is 0.650000.
refinement_failed
CPU time of basic_analyse is 1.730000.
CPU time of refine is 1.090000.
```

```
Object (lg_LS,lsTerm118,js,[iai_exTerm,de],base) is asserted.
The elapsed time to get the solution was 73 sec.
The total elapsed time to get the solutions was 73 sec.
CPU time of refine_failure is 6.660000.
CPU time of basic_analyse is 8.400000.
CPU time of refine is 1.110000.
Object (lg_LS,lsTerm119,js,[iai_exTerm,de],base) is asserted.
The elapsed time to get the solution was 42 sec.
The total elapsed time to get the solutions was 115 sec.
CPU time of refine_failure is 6.570000.
CPU time of basic_analyse_failure is 25.900000.
CPU time of basic_lift_failure is 49.000000.
The elapsed time to get the solution was 47 sec.
The total elapsed time to get the solutions was 162 sec.
```

5 Concluding Remarks

In this paper we have briefly described a new approach which certainly deserves the attention of the machine translation community and further exploration in additional subject domains. Currently, a similar approach is being investigated into the translation of spontaneous speech (dialogues for appointment scheduling) in the German VERBMOBIL project (cf. [Ripplinger, 1994]).

One important advantage of the suggested approach is its modularity. The basic linguistic knowledge is represented in a competence grammar. The terminological knowledge for the subject domain is encoded in a hierarchy of typed feature terms. This specialized knowledge for a sublanguage constrains grammatical analysis and transfer. The depth of semantic/conceptual analysis is restricted to the needs of the translation task. The separation of the knowledge sources facilitates extensibility and portability to other sublanguages as well as to the macro-structural handling of texts (cf. below).

Our approach, as one instance of performance control, is innovative because we do not need a specially designed interface between the different knowledge sorts, because each sort is realized by means of the ALEP formalism. The use of ALEP seems to be a good choice, since ALEP builds on de-facto standards for notation and will be freely available for European research and development.

In addition, in accordance with the work of [Gerzymisch-Arbogast, 1994] it is possible to extend our approach to the text level by introducing further relations which address the conceptual macro-structure of a given text by so-called *discourse grammar rules*. A *discourse grammar* is then the sentence grammar augmented by a set of discourse rules. At present, we have applied this only to the resolution of anaphora across sentence boundaries as, for example, in: *The following section describes modulating equipment. It superimposes the audio-frequency signals on the IF-carrier. This equipment extracts them from the IF-carrier.* Research in this direction addresses in particular the application of modified unification processes, such as higher-order unification (e.g. [Dalrymple et al., 1991]), which, however, is beyond the scope of this paper.

References

[ALEP, 1993] ALEP Documentation Package, Vol. I and II. CEC and PE International, Luxemburg.

[Copeland et al. (Eds.), 1991] C. Copeland, J. Durand, S. Krauwer and B. Maegaard, 1991. **The EUROTRA Linguistic Specification** and **The EUROTRA Formal Specifications**. Studies in Machine Translation and Natural Language Processing, Vol. 1 and 2, CEC, Luxembourg.

[Dalrymple et al., 1991] M. Dalrymple, S. M. Shieber and F. C. N. Pereira, 1991. Ellipsis and higher-order unification. Linguistics and Philosophy 14(4).

[Gerzymisch-Arbogast, 1994] H. Gerzymisch-Arbogast, 1994. **Übersetzungswissenschaftliches Propädeutikum**. UTB 1782, Francke Verlag, Tübingen, Germany.

[Hovy and Knight, 1993] E. Hovy and K. Night, 1993. Motivating Shared Knowledge Resources: An Example from the Pangloss Collaboration. In: Proceedings of the IJCAI workshop on Shared Knowledge, Chambery, France.

[Kay, 1984] M. Kay, 1984. Functional Unification Grammar: A Formalism for Machine Translation. In: *Proceedings of COLING-84*, Stanford, CA.

[Knight, 1994] K. Knight, 1994. Buildling a Large-Scale Knowledge Base for Machine Translation. In: Proceedings of AAAI-94.

[Jackendoff, 1990] R. Jackendoff, 1990. **Semantic Structures**. MIT Press, Cambridge, Massachusetts.

[Nirenburg et al., 1992] S. Nirenburg, J. Carbonell, M. Tomita and K. Goodman, 1992. **Machine Translation: A knowledge-based Approach**. Morgan Kaufmann Publishers, San Mateo, California.

[Pustejovsky, 1991] J. Pustejovsky, 1991. The Generative Lexicon. In: Computational Linguistics, 17(4).

[Ripplinger, 1994] B. Ripplinger, 1994. Concept-based Machine Translation and Interpretation. In: Proceedings of Cranfield Conference 'MT - Ten Years On'.

[Ripplinger et al., 1994] B. Ripplinger, J. Schütz, G. Talbot, 1994. Terminology and Extra-Linguistic Knowledge, ET-10-66 Final Report. European Commission, Luxembourg.

[Schütz, 1994] J. Schütz, 1994. **Terminological Knowledge in Multilingual Language Processing**. Studies in Machine Translation and Natural Language Processing, Volume 5, Office for Official Publications of the European Communities, Luxembourg.

[Sowa, 1991] J. F. Sowa, 1991. Towards the Expressive Power of Natural Language. In: J. F. Sowa (ed.). **Principles of Semantic Networks: Explorations in the Representation of Knowledge**. Morgan Kaufmann Publishers, San Mateo, CA.