Organizing linguistic knowledge for multilingual generation*

Martin Emele, Ulrich Heid, Stefan Momma, Rémi Zajac
Project Polygloss
University of Stuttgart
IMS-CL/IfI-AIS, Keplerstraße 17,
D 7000 Stuttgart 1, Federal Republic of Germany
email: polygloss@informatik.uni-stuttgart.dbp.de

Abstract

We propose an architecture for the organisation of linguistic knowledge which allows to (1) separately formulate generalizations for different types of linguistic information, and (2) state interrelations between partial information belonging to different levels of description. We use typed feature structures for encoding linguistic knowledge. We show the application of this representational device for the architecture of linguistic knowledge sources for multilingual generation. As an example, we describe the use of interacting collocational and syntactic constraints in the generation of French and German sentences.

1 Introduction

1.1 The Problem

The choice of target language realizations in machine translation or in multilingual generation is conditioned by constraints involving different levels of linguistic description for the individual languages. From a descriptive point of view, it is desirable to be able to keep these different levels separate – conceptually as well as in the actual implementation of knowledge sources and representations of linguistic objects; such levels may include, for example, a description of morphological properties, of constituent structure, of predicate-argument structures or functional structures, as well as a description of textual and pragmatic properties.

In generation from semantic representations, constraints on the choice of linguistic realizations depend on properties of the basic elements of lexical and syntactic variants with respect to all these levels; such constraints usually interact in various ways.

Knowledge sources which provide the information necessary for modelling such phenomena should therefore allow for modularization as well as for the declarative formulation of dependencies between information which belongs to different descriptive levels.

1.2 Current Approaches

In research on MT and NL generation, different approaches to both problems, modularization and interaction between the modules, have been proposed.

Although most of these approaches allow for a description of linguistic phenomena at each individual level, it is hard for them to explicitly express *interactions* between levels without using directionality. Usually, adjacent levels are connected by explicit mappings. Conditions acting on nonadjacent levels often cannot be expressed directly: thus, information has to be carried explicitly across levels where it would normally not be stated. Moreover, as the input structure is transformed stepwise, the set of mappings has to be ordered carefully. Other generation researchers, like [DANLOS 1987]:96–99 and [NIRENBURG 1989]:242f have a similar view of the architecture of the linguistic description; most of their solutions to the above problem are based on a heuristic ordering of the mapping and on some loss of strictness in the separation of levels.

In order to alleviate such problems, the use of "codescriptions" has been proposed[1] which allows for statements about the coexistence of partial descriptions belonging to separate descriptive levels. This device makes all relevant information available in one place thus allowing for constraints from different levels to be considered at the same time. Although the use of codescriptions perfectly supports the formulation of interactions between different types of partial information, the current propos-

*Research reported in this paper is partly supported by the German Ministry of Research and Technology (BMFT, Bundesminister für Forschung und Technologie), under grant No. 08 B3116 3.

[1] See. e.g. [FENSTAD ET AL. 1987] who annotate context-free rules of Lexical Functional Grammar with functional as well as semantic descriptions, which allows for the simultaneous construction of f-structures and situation schemata; an application to transfer has been proposed by [KAPLAN ET AL. 1989], where f-structures for different languages are simultaneously built up.

als do not seem to pay enough attention to the separation of different types of information. In many cases, one of the types of information is assigned a predominant role (usually this is c-structure); all other information depends on it, i.e. cannot be expressed without making reference to the dominating information.

In this paper we propose an architecture of knowledge sources for multilingual generation. It is based on linguistic descriptions in the format of typed feature terms following ideas of [Aït-Kaci 1986]. This allows, much like object-oriented systems, for a modularization of knowledge; we can state relations capturing interdependencies between elements of different descriptive levels without losing the possibility to independently formulate generalizations for classes of linguistic objects.

2 Architecture

2.1 The representational device: typed feature terms

The objects used to represent partial linguistic information are **typed feature terms**: i.e. feature terms where each node in the directed graph usually representing an ordinary feature term can be associated with a *type symbol*. For type symbols, the linguist supplies a *feature type definition* which can be a feature term, a feature term with conditions (used to express additional constraints), or a conjunction or disjunction of feature terms.

The system we have implemented compiles a set of feature type definitions, a *feature type system*, into a hierarchy of feature terms which is derived from the hierarchy of type symbols (implicitly defined by the set of feature type definitions) and the usual subsumption relation on ordinary feature terms. In the general case, the hierarchy of feature terms is a multiple inheritance hierarchy.

Given an arbitrary linguistic object represented by a feature term which is only partially specified, the linguist is interested in obtaining the most precise description of this object according to a grammar (specified as a feature type system). Given such a term, the interpreter computes the set of most specific feature terms which are derived from it by applying feature type definitions: each member of the solution set is subsumed by (i.e. is more specific than) the initial term. For this derivation, the interpreter of the system uses only two basic operations[2], **typed unification** of feature terms, and **rewriting** based on unifying substitutions of feature terms.

[2]A more detailed description of these operations can be found in [Emele/Zajac 1990a]

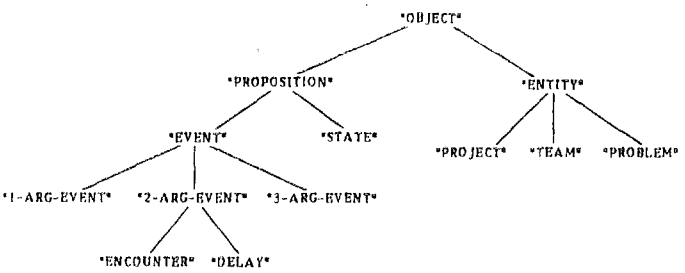

Figure 1: A part of the conceptual hierarchy

2.2 Linguistic Objects

Descriptions of linguistic objects may contain pragmatic, semantic, and syntactic information. We organize the linguistic information as a hierarchy of classes of linguistic objects, where we assume that for each of the levels, it is possible to define classes of basic objects, classes of structures these objects can be parts of, and interrelations between objects of different levels and between structures of different levels[3].

For example, basic objects of a conceptual description are *concepts*, those of a lexical description are *lexical units* (words as well as multi word lexemes). Semantic structures may, for example, define the temporal structuring of states of affairs; syntactic structures define well-formed phrase structures and functional structures. Relations between concepts and lexemes define possible ways of lexicalizing conceptual information in a given language. Relations between semantic and syntactic structures describe possible syntactic realizations of semantic descriptions.

For the remainder of this paper, we isolate the part of the architecture concerned with the relations between conceptual and syntactic descriptions, abstracting away from other types of information.

3 Relating conceptual and syntactic knowledge

3.1 A conceptual hierarchy

Figure 3.1 shows a simplified version of the upper part of a hierarchy of concepts. It distinguishes between entities and propositions (events, states etc.), the latter being subclassified according to their number of arguments. The lower parts of the hierarchy (not shown in the figure), may contain domain spe-

[3]Our semantic description is largely based on work in Discourse Representation Theory [Kamp/Reyle 1990]; the syntactic representation follows the lines of Lexical Functional Grammar

cific concept classifications. As to the level of detail down to which the conceptual description proceeds, it seems useful to stop decomposing concepts at a level where none of the languages to be treated has more specialized lexemes available. This allows to treat cases where languages differ with respect to the specialization level down to which they have lexemes. French and Italian, for example, do not have lexemes for the concept of "transport-in-the-air", which makes it necessary to use the lexeme denoting a generic *transport*-action and to separately realize (in a prepositional phrase) the instrument: *transporter des fleurs en avion de Nice a Berlin; trasportare fiori in aereo, da Nizza a Berlino.* English and German, on the other hand, not only have lexemes for the generic concept, but also for the specialized one, namely *to fly sth.* and *etw. fliegen*, as in *fly flowers from Nice to Berlin; Blumen von Nizza nach Berlin fliegen.* We therefore introduce a concept *fly*, defined as a subclass of *transport* with *airplane* as an instrument.

3.2 A lexical hierarchy

In analogy to the semantic classification, we introduce a hierarchy of syntactic objects classified according to their syntagmatic properties; this allows for immediate access to information relevant for the realization of each lexeme.

We use a hierarchy of subcategorization types where "monosemous readings" of lexical units are classified. As an example, we discuss some characteristics of the organization of the French verbal subcategorization hierarchy. Although basically using functional structures of LFG as syntactic representations, we do not only use LFG's grammatical function labels as a basic vocabulary[4]. We also specify, among others, the phrasal category of the complement (NP, AP, finite or infinite clause, etc.), pronominalization possibilities (e.g. *le, la, les* vs. *en, y* vs. *lui, leur*, etc.), and the presence/absence of prepositions. This decomposes grammatical functions according to the distinctions relevant for their definition. Similar procedures have been proposed within LFG's lexical mapping theory, e.g. by [BRESNAN/KANERVA 1989]. Our representation would also allow, without major changes, the construction of syntactic representations in the format of a different unification-based grammatical theory, if this theory makes use of the same types of elementary distinctions as the classification used here.

This double classification which uses categorial as well as pronominalization information, allows for the independent formulation of generalizations coded in the definition of the respective subcategorization classes; for each pronominalization type, including predicative and oblique complements, a range of possible syntactic realizations is described. For each "monosemous reading" of a verbal lexeme we can thus describe in detail a set of synonymous syntactic construction variants.

The lexical entry for enthousiasmer specifies, for example, that both its complements may only be realized as NPs:

(1) enthousiasmer =

$$\text{2-f-verb} \begin{bmatrix} \text{PRED:} & \text{"enthousiasmer"} \\ \text{1ST-COMPL:} & \text{subj} \wedge \text{nominal-phrase} \\ \text{2ND-COMPL:} & \text{obj} \wedge \text{nominal-phrase} \end{bmatrix}.$$

enthousiasmer inherits from the class 2-f-verb of verbs taking two complements and selects as a first complement a nominal phrase of type subj, and as a second complement an NP of type obj. subj represents the whole range of possible structures which can appear as first complements; obj also defines a set of realization variants with the same pronominalization behaviour, namely an NP, affirmative or interrogative clause or infinitivals with or without *à* or *de* as a preposition.

(2) 2-f-verb =

$$\text{verb-class} \begin{bmatrix} \text{1ST-COMPL:} & \text{subj} \\ \text{2ND-COMPL:} & \left\{ \begin{array}{l} \text{obj} \vee \text{iobj} \vee \\ \text{en-obj} \vee \text{y-obj} \vee \\ \text{predi} \vee \text{obl} \end{array} \right\} \end{bmatrix}.$$

(3) subj = nominative $\wedge$
$(\{\text{nominal-phrase} \vee \text{affirmative} \vee \text{infinite}\})$.

3.3 Relating conceptual and syntactic knowledge

3.3.1 Types of relations

From what we said about semantic decomposition above, it follows that, for each concept, we assume there exists at least one lexeme in one of the languages; usually there will be more than one. This is captured by a relation ρ on pairs of partial semantic and partial syntactic structures. The relation is specified as a feature term with top level labels SEM and SYN. The fact that many of the interactions between semantic and syntactic descriptions can be expressed for whole classes of structures is reflected by a modular and hierarchical specification of this relation. For example, there is a relation ρ-entity$\leftrightarrow$NP which connects objects of type *entity* with noun phrases, or the relations ρ-proposition$\leftrightarrow$NP and ρ-proposition$\leftrightarrow$VP which relate *proposition* type

[4] E.g., grammatical functions may be assigned to complements with different internal structure (e.g. NPs or subclauses, described as ⟨OBJ⟩ects).

objects with NPs (whose head is a nominalization) or VPs (e.g. infinitival complements, that-clauses, etc.), respectively.

Since verbs are classified according to their number of complements, it is possible to define relations between predicates and lexical classes, e.g. for predicates with two arguments and their two-place lexical counterparts (ρ-2arg).

On the lexical level, relations between concepts and lexemes allow for the specification of lexical material (single lexemes, as well as e.g. support verb constructions etc.) available for a given concept; the lexical types represent a range of construction variants which can be formed with the lemma. So, for example ρ-enthuse relates the concept *enthuse* with the verbal lemma enthousiasmer discussed in (1). The concept *enthuse* is defined as taking a *proposition* as its first argument. The lemma enthousiasmer, however, selects only nominal-phrase type first complements. Consequently, subject clauses or subject infinitives are ruled out for this verb, and among the realization possibilities for *proposition*s, only nominalizations can be used.

3.3.2 Using the relations in generation

The use of the architecture of knowledge sources for different descriptive levels is best shown in an application where the interaction of constraints from different levels has to be treated: in the following example, collocational and syntactic (subcategorization) constraints interact.

For the realization of a simplified conceptual structure like (4), different English, French and German collocations such as (5) – (11) can be used;

(4) *ENCOUNTER* $\begin{bmatrix}\text{ARG1: *TEAM*}\\ \text{ARG2: *PROBLEM*}\end{bmatrix}$

such as:

(5) *E: the team encounters a problem;*

(6) *E: the team comes across a problem;*

(7) *F: le groupe rencontre un problème;*

(8) *F: le groupe bute contre un problème;*

(9) *F: le groupe se heurte à un problème;*

(10) *G: das Team stößt auf ein Problem;*

(11) *G: das Team trifft auf ein Problem.*

The verbs used in (5) to (11) are collocationally preferred.

The following statements relate the concept *encounter* with French and German lexemes, implicitly keeping track of the collocational restrictions by listing only the possible collocates of *problem* and *Problem*, respectively:

(12) ρ-f-11 =
$\begin{bmatrix}\text{SEM: *encounter* [ARG2: *problem*]}\\ \text{SYN: \{rencontrer} \vee \text{buter} \vee \text{se\_heurter}\}\end{bmatrix}$.

(13) ρ-g-11 =
$\begin{bmatrix}\text{SEM: *encounter* [ARG2: *problem*]}\\ \text{SYN: stossen\_auf}\end{bmatrix}$.

A simplified subcategorization description for the French and German verbs shows that they fall roughly into two classes, transitive and intransitive[5]: French *rencontrer* (7) is described as transitive, taking a subject and an object and being passivizable. French *buter contre* (8) and *se heurter à* (9), as well as German *stoßen auf* or *treffen auf* (10), (11), are described as taking a subject and a prepositional object, and disallowing passivization:

(16) rencontrer =
$\{\text{s-o-v} \vee \text{s-by-v}\}$ [PRED: "rencontrer"].

(17) buter =
s-p-v $\begin{bmatrix}\text{PRED: "buter"}\\ \text{POBJ: [PCASE: "contre"]}\end{bmatrix}$.

(18) se_heurter =
s-p-v $\begin{bmatrix}\text{PRED: "se\_heurter"}\\ \text{POBJ: [PCASE: "a"]}\end{bmatrix}$.

(19) stossen_auf =
s-p-v $\begin{bmatrix}\text{PRED: "stossen"}\\ \text{POBJ: [PCASE: "auf"]}\end{bmatrix}$.

The problem is that collocationally correct realizations of (4) in German are only possible with verbs which do not have passive forms, like *stoßen auf* or *treffen auf* (10), (11).

If we want to generate French and German from a predicate-argument structure like (19), we still want to use a general relation describing the relation between a RESTR of an *entity* and its syntactic realization variants, namely a relative clause or an embedded participle. This relation carries the constraint, however, that the verb which the participle is to be derived from be transitive. Taking the *French* collocations discussed in (7)–(9), the following realizations for (19) exist:

- participle:

  (20) *un problème rencontré a retardé un projet;*

[5] We simplify considerably and abstract away, for the purpose of this paper, from the question of language-specific elements in the definition of syntactic classes like "transitive" or "intransitive" and take passivization possibility as a distinctive criterion in all languages.

(19) *DELAY* [ARG1: [1] *PROBLEM* [RESTR: *ENCOUNTER* [ARG1: *NULL* / ARG2: [1]]] / ARG2: *PROJECT*]

(20)
[SEM: *DELAY* [ARG1: [1] *PROBLEM* [RESTR: *ENCOUNTER* [ARG1: *NULL* / ARG2: [1]]] / ARG2: *PROJECT*]
SYN: S-O-V [PRED: "verzoegern"
SUBJ: N [PRED: "Problem"
REL-ADJ: S-P-V [PRED: "stossen"
TOPIC: [rel] [PRED: "pro" / RELPRO: + / PCASE: "auf"]
SUBJ: N [PRED: "man"]
POBJ: [rel] PRO]]
OBJ: N [PRED: "Projekt"]]]

- passive relative clause:

  (21) *un problème qui a été rencontré a retardé un projet;*

- active relative clause with impersonal subject (*on*):

  (22) *un problème qu'on a rencontré a retardé un projet;*

  (23) *un problème contre lequel on a buté a retardé un projet;*

  (24) *un problème auquel on s'est heurté a retardé un projet.*

Since in *German* there is no collocation with the same syntactic structure as *rencontrer un problème*, i.e. where the collocate is a transitive verb, only an active relative clause with an impersonal subject (*man*) is possible for (19):

(25) *Das Problem, auf das man stößt, verzögert das Projekt.*

4 Conclusion

In this paper we proposed an architecture for the organization of linguistic knowledge allowing for both (1) the separate formulation of generalizations for different types of linguistic information, and (2) the use of relations to state correspondences between partial information pertaining to different levels of linguistic description. Typed feature terms are used for encoding linguistic knowledge; the TFS system incorporates a multiple inheritance mechanism, which allows to minimize redundancy within "specialized" knowledge sources, like e.g. the hierarchy of subcategorization types. On the other hand, no interfacing problems between different levels of linguistic description arise, due to the use of one and the same data structure for representing all these levels for a given linguistic object. In addition, instead of explicitly controlling complex interactions, the relational approach allows to constrain realization choices in generation as a result of the simultaneous application of distributed linguistic constraints.

The TFS system has been implemented in Common-LISP by Martin Emele and Rémi Zajac [Emele/Zajac 1989a], on Symbolics, TI Explorer and VAX. Sample grammars have been documented in [Emele 1988] and [Zajac 1989b]. The specification of the knowledge sources is currently under way.

5 Acknowledgements

Our proposal has benefited from discussions with Stanley Starosta, Sergei Nirenburg and our colleagues at the IMS, whom we would like to thank for comments on previous versions of this paper.

References

[Aït-Kaci 1986] Hassan Aït-Kaci: "An Algebraic Semantics Approach to the Effective Resolution of Type Equations." in: *Theoretical Computer Science*, Vol. 45, p. 293-351

[Bäuerle 1988] Rainer Bäuerle: *Ereignisse und Repräsentationen*, (Stuttgart: IBM Germany), 1988, [LILOG-Report, 43],

[Bateman et al. 1989] John Bateman, Bob Kasper, Johanna Moore, Richard Whitney: "The Penman Upper Model – 1988. Penman Development Note", ms. (Los Angeles: ISI), May 1989

[BRESNAN/KANERVA 1989] Joan Bresnan, Jonni M. Kanerva: "Locative Inversion in Chichewa: A Case Study of Factorization in Grammar" in: *Linguistic Inquiry*, Vol. 20/1, p. 1-50, 1989

[DANLOS 1987] Laurence Danlos: *The Linguistic Basis of Text Generation*. Cambridge University Press.

[EMELE 1988] Martin C. Emele: "A Typed Feature Structure Unification-based Approach to Generation" in: *Proceedings of the WGNLC of the IECE 1988*, (Japan: Oiso University) 1989

[EMELE/ZAJAC 1989A] Martin Emele, Rémi Zajac: "RETIF: A Rewriting System for Typed Feature Structures", (Kyoto) 1989, [ATR Technical Report TR-I-0071]

[EMELE/ZAJAC 1989B] Martin Emele, Rémi Zajac: "Multiple Inheritance in RETIF", (Kyoto) 1989, [ATR Technical Report TR-I-0114]

[EMELE/ZAJAC 1990A] Martin Emele, Rémi Zajac: *Semantics for Feature Type Systems*. Internal paper. IMS. University of Stuttgart.

[EMELE/ZAJAC 1990B] Martin Emele, Rémi Zajac: "Typed Unification Grammars.", in: Proceedings of COLING-90, 1990, this volume

[FENSTAD ET AL. 1987] Jens Erik Fenstad, Per-Kristian Halvorsen, Tore Langholm, Johan van Benthem: *Situation, Language and Logic.*, (Dordrecht: Reidel) 1987

[HALVORSEN/KAPLAN 1988] Per-Kristian Halvorsen, Ronald Kaplan: "Projections and Semantic Description.",in: *Proceedings of the International Conference on Fifth Generation Computer Systems*, (Tokyo) 1988

[KAMP/REYLE 1990] Hans Kamp, Uwe Reyle: *From Discourse to Logic*, (Dordrecht: Reidel): to appear

[KAPLAN ET AL. 1989] Ronald M. Kaplan, Klaus Netter, Jürgen Wedekind, Annie Zaenen: "Translation by Structural Correspondences" in: *Proceedings of the 4th Conference of the ACL, European Chapter, Manchester, 10-12 April 1989*, 1989

[KAY 1984] Martin Kay: "Functional Unification Grammar: a formalism for machine translation". *Proceedings of Coling-84*. Stanford.

[NIRENBURG 1989] Sergei Nirenburg: "KBMT-89. Project Report.", ms. (Pittsburgh, Pa: Center for Machine Translation, Carnegie Mellon University) 1989

[POLLARD/SAG 1987] Carl J. Pollard, Ivan A. Sag: *Information-Based Syntax and Semantics*, Vol.1, Fundamentals, (Chicago: University Press) 1987 [CSLI Lecture Notes, No. 13]

[ZAENEN 1988] A. Zaenen: *Lexical Information in LFG — An Overview*, ms. (Palo Alto: Xerox PARC) 1988

[ZAJAC 1989A] Rémi Zajac: "A Relational Approach to Translation", (Stuttgart: IMS), internal paper, submitted to Third International Conference on Theoretical and Methodological Issues of Machine Translation of Languages (Austin, 1990)

[ZAJAC 1989B] Rémi Zajac: "A Transfer Model Using a Typed Feature Structure Rewriting System with Inheritance.", in: *Proceedings of the 27th Annual Meeting of the ACL-89* (Vancouver, Canada) 1989