An Empirically Based Approach Towards a System Of Semantic Features

Cornelia Zelinsky-Wibbelt

IAI-Eurotra-D
Martin-Luther-Straße 14
D-6600 Saarbrücken

Abstract

A major problem in machine translation is the semantic description of lexical units which should be based on a semantic system that is both coherent and operationalized to the greatest possible degree. This is to guarantee consistency between lexical units coded by lexicographers. This article introduces a generating device for achieving well-formed semantic feature expressions.

1. Intention and procedure

Empirical work with the verbs of the ESPRIT-corpus as well as experience in theoretical semantics, and last but not least, the consulting of the semantic feature inventories of other machine translation systems (METAL, JAPAN, SYSTRAN, SUSY) has resulted in the necessity of an elaboration of the proposal for semantic features made in ELS-3 (EUROTRA LINGUISTIC SPECIFICATIONS). These feature inventories as well as a large amount of already existing, partly fairly traditional work on semantic feature systems of linguistics and philosophy (CHAFE, FRIEDRICH, PALMER, VENDLER), of information sciences (DAHLBERG), and work in the field of cognitive linguistics and artificial intelligence (G.A. MILLER, G.A. MILLER & P.N. JOHNSON-LAIRD, G. LAKOFF, R. LANGACKER, B. COHEN, W.R. GARNER, ATTNEAVE, FREDERIKSEN, MINSKY, CHARNIAK, WINOGRAD, ANDERSON & BOWER, WOODS) and last but not least some recent issues on word semantics (T. BALLMER, W.BRENNENSTUHL, J.BALLWEG, H. FROSCH) have been taken into account in order to meet the requirements of a manageable system of semantic features. This system is intended both to be based on a sensible theory of semantics and to satisfy the special requirements of machine translation in general and of our text type in particular. Moreover, it should be flexible enough to be enlarged and supplemented or changed, whenever this proves necessary on empirical evidence - this last requirement being made possible by the accomplishment of the first.

In trying to meet these requirements the semantic feature inventories at hand have been enlarged, changed, and adapted to our specific purposes and have been merged into one system of semantic features.

2. Comment on the theoretical assumptions made in different machine translation systems with respect to the semantic representation

With respect to the semantic representation which in EUROTRA will be implemented on the interface structure (IS) level of the source and target language it is our first and foremost aim to arrive at a coherent system of semantic features. In order not to start from nothing the above mentioned feature inventories have been consulted. The feature inventories developed for these machine translation systems have different shortcomings which will be briefly commented on in the following.

Since a sufficient definition of how to interpret the features is given in none of the proposals of the above mentioned machine translation systems, we will not comment here on the features themselves included in the proposals. A brief comment, however, is necessary on the general approach, which seems to imply theoretical assumptions (not explicitly mentioned, since neither a theoretical nor a practical usage-based explanation is given) about the organisation and processing of semantic units, for which there is no empirical evidence: neither natural language processing by human beings nor efficiency in automatic processing of natural language gives support to these implied assumptions.

It must be mentioned, however, that this can by no means be considered to be an objective comment, since for an outsider, it is impossible to understand the systematic motivation of these feature inventories for at least one of the following reasons:

- The semantic features are not defined or at least not sufficiently defined in order to make clear their conceptual structure and thus to make clear how they are meant to be used. This is especially true for the EUROTRA proposal, in which semantic features are not defined at all. This is, how-

ever, only a proposal, which has not been applied yet, but is being tested at the moment. But also the SYSTRAN semantic features, as well as those of JAPAN, which have been worked out rather sophisticatedly, are not commented on. The semantic features of METAL are defined, their definition, however, remains rather vague. Even when taking into consideration the examples which are added, the reader does not arrive at a satisfactory understanding.

- The dependencies holding between features are not explained. This is especially true for SYSTRAN, which only gives a list of features referring to arguments. A hierarchical system consisting of two levels of semantic features is defined by METAL, which is far from sufficient. JAPAN is worked out in a more sophisticated way with respect to this problem. Both in METAL and in JAPAN, however, relations between the dominating features are not defined. The EUROTRA proposal gives an enumeration on the second and lowest level of the feature tree, which is just a conglomeration of semantic information, which should be described at different levels, in order to achieve the overall aim of linguistically consistent semantic description.

3. A proposal for a EUROTRA semantic feature rule system

3.1. Necessity of a semantic feature rule system

Let us now put forward our conception of the two systems of semantic features with respect to its formalization. We have two grammars, one describing "SITUATION" features, the other one describing "ENTITY" features. Neither of the two systems is strictly hierarchically organized. The hierarchical principle, however, which always defines a refinement of the dominating feature, prevails. Particularly the most general semantic features, such as the "ENTITY" features "CONCRETE/"ABSTRACT", "COUNTABLE"/"MASS", and "NATURAL"/"ARTIFICIAL", and the "SITUATION" features "CONCRETE"/"ABSTRACT", "STATIVE"/"DYNAMIC", and "PUNCTUAL"/ "DURATIVE"/"ITERATIVE", respectively, form pairs or triplets of semantic features. One feature of each of these alternations obligatorily occurs, and the descendents, which specify them, form disjunct sets.[2]

3.2. The basic formalism

Let us now comment informally on our present conception of how the semantic features which we consider necessary so far are related to each other.

We use three operations holding between semantic features in our grammar:

1) Hierarchy is the overall relation defining the derivation of the features.
2) Alternation relates a set of features, only one of which applies.
3) Disjunction relates semantic features obligatorily occurring together. This type of relationship is of course in the minority.

The basic idea is to describe these relations by a context-free rule system, where the rules can for example be of the following form:

(3.1) $X = (A/B)*(C/D)$

The hierarchy here is represented by the sign "=", the alternation by the sign "/", and the disjunction by the sign "*". The interpretation of the rule is the following:

The feature on the left handside of the rule dominates the features appearing on the right handside. A, B, C, and D establish a refinement of X. More precisely, in this example X is specified by a pair of features, the first component of which can be either A or B and the second is either C or D. The subordinate features on the right handside of the rule can get superordinate features themselves on the next level lower down in the hierarchy. The terminal features, that is those features which are not defined for accepting any subordinate features, are represented by the rules

(3.2) $X = 0$
$X = A/0$

Let us exemplify this with the feature "COUNTABLE":

(3.3) COUNTABLE = CATEGORY 1 * CATEGORY 2 * DEFINITION
CATEGORY 1 = INDIVIDUATIVE/PARTITIVE/COMPLEX/COLLECTIVE/PRIVATIVE
CATEGORY 2 = CAUSE/RESULT
DEFINITION = MEASURE/SOCIAL

By this we mean that the feature "COUNTABLE" is represented by three features which always occur together (marked by the operator "*"). Each of these three features again dominates a collection of features only one of which is selected (marked by the operator "/"). The hierarchical relationship itself is implied in the left-to-right-handside associations (marked by the operator "="). Here it is essential to note that every semantic feature is only defined once by one rule. If more than one description exists, all of them are combined by "or". As the "and" relation by definition is prior to the "or" relation, brackets have to be placed around the alternative expression in the opposite case, that is when the "or" relation is prior to the "and" relation.

3.3. The introduction of attributes

So far we have introduced a formal instrument with which we can describe the relations between fea-

tures which are formally possible. In order to describe the actual relationships between features, this formal instrument still has to be restricted. In order to keep the rule system compact, we introduced attributes which are intended to describe important co-occurrence restrictions existing between features in disjunct branches. The existence of a feature activates an attribute called the derived. This attribute effects the restriction of a rule application in a disjunct part of the grammar. There the attribute is called the inherited. In the rule system attributes appear on the left handside of the rule if they are derived, on the right handside if they are inherited. We derive a feature's attribute like that:

(3.4) $X_{[]} = \ldots$

An attribute always gets the name of the semantic feature which causes the attribute, so the derivation can be marked by an empty pair of square brackets.

The derived attribute appears in the right-handside context as inherited attribute e.g. like that

(3.5) $Y = \ldots A^{[X]} \ldots$

With the above mentioned example this would look as the following:

(3.6) CAT 2 = CAUSE/RESULT
$\text{CAUSE}_{[]} = 0$
$\text{RESULT}_{[]} = 0$

The inherited attribute can also be assigned to a feature expression. In this case it would apply to every feature within this expression. Moreover, instead of a single attribute, an expression of attributes can appear. In an attribute expression the above mentioned operators "or" and "and" can appear and in addition the negation operator "not" (represented by the sign \). With the introduction of attributes the generation mentioned above has to be modified: the rule (3.4) states that the feature X is derived and has to be registered so that it can be used in the relevant disjunct feature context which may also be dominated by Y as described in rule (3.5).

We have therefore to extend the above mentioned example (3.6) by the following rule:

(3.7) $\text{DIRECTION} = \text{SOURCE}^{[\text{CAUSE}]}/\text{GOAL}^{[\text{RESULT}]}$

With the definition of these rules it was proved that on the one hand the formalism is powerful enough to represent all the above mentioned phenomena and on the other hand it is still simple enough so that changes necessary in later stages may be accomplished without too much cost.

Our rule system is based on the definition of the semantic features. So far we have defined 87 features for the description of "ENTITIES" and 87 features for the description of "SITUATIONS".

3.4. The use of the formalism

Our grammar is intended to be a generating system. It will be used as input for an automatic procedure. For every lexical unit this can be used to generate the list of semantic features which semantically describe the lexical unit sufficiently. Our notion of sufficiency arises from our goal of automatic disambiguation. The automatic procedure leads the lexicographer through the system in the right way, so that the correct list of semantic features is generated for each lexical entry. This procedure makes use of the rule system in order to produce menus which show the alternatives valid in each actual state. In general the list of semantic features which describes a lexical unit contains only the terminal features which are generated, since the dominating non-terminals can be deduced. This is, however, not valid for features and their derivates, which appear on more than one right handside of a rule (named critical rule), i.e. the resulting terminal features do not give an unambiguous specification of the lexical unit. In these cases we add the non-terminal feature from the left handside of the critical rule to the feature list, which then gives us an adequate feature spectrum. It is, however, possible to take into account redundancies by taking other dominating features into the list as well. This would possibly lead to a more efficient translation process.

This output of our generating system will be the input to our dictionary and in later stages of the translation process, precisely on the interface-structure level, is intented to be used for disambiguation and other strategic purposes in the process of semantic analysis and synthesis.

It follows that all lists of features which can be generated by the grammar make up the set of all possible semantic descriptions which may describe concepts referring to our object world.

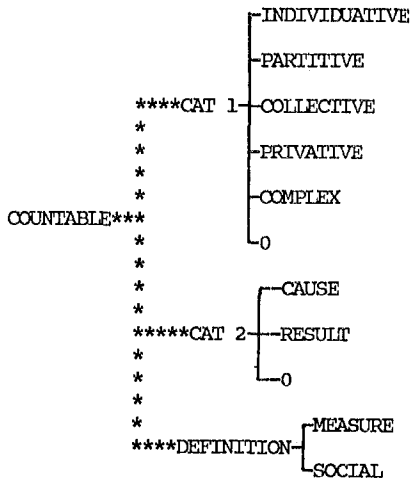
Moreover, this rule system may be used in synthesis in order to find out dependencies holding between semantic features. We think that this will be necessary for semantic generalizations in the target language. This domain, however, has not yet been worked out.

A side effect of the automatic processing of the rule system is the generation of a graphic representation which has a treelike form.

The graphic representation of this rule system has proved to be very useful. In this graphic representation the axiom is the root and every rule is represented by a subtree as shown in Figure 1.

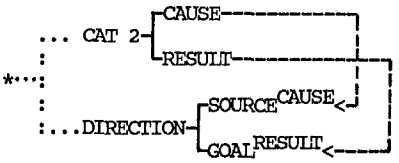
If a node generates a derived attribute, an empty pair of brackets appears as a subscript of the

feature denoting the node. If the node inherits another attribute or an attribute expression, the corresponding names or expressions are represented as superscripts of the feature. Figure 2 represents the attributes and their effect in an evident way. It makes clear that the order of the disjunctive nodes (branching underneath the "and" node) is most essential for the application of the attributes. This representation in the form of a tree has proved to be a very transparent way of illustrating the structure of the set of semantic features.



The arcs represented by asterisks correspond to a disjunctive expression, the arcs represented by a solid line correspond to a conjunction.

Figure 1 The feature "COUNTABLE" as an example of disjunctive branching



The dotted line which relates the features "CAT 2" and "DIRECTION" represents that both features are dominated by the same node higher up in the hierarchy. The dashed line shows the derivation and inheritance of attributes.

Figure 2 Illustration of the derivation and inheritance of the features "CAUSE" and "RESULT"

4. The linguistic motivation for the specific make-up of our rule system of semantic features

Now that we have given a description of the formalism that we have made use of in order to describe the existing relationships between semantic features, let us explain at this point, why semantic features are organized like that, since it is less the formalism, but rather empirical evidence and linguistic knowledge by means of which we arrive at exactly this organization.
Although it is essential to know that there are no inherent but only context-dependent features[3], apart from the features "NATURAL" and "ARTIFICIAL", the basis of our system of semantic features is an as objective as possible a definition of each feature itself. This definition is based on the criteria of prototypicality by means of which we abstract from our experience. Moreover, the criteria of prototypicality resulted in disjunct and alternative feature sets, which are described by our generative rule system. This means that the whole system is based on how we categorize concepts. The general process of refinement into different subfeatures on which our systems are based depends on the principle of focusing different areas of the superconcept and thus imaging different subconcepts.

As one general characteristic of the system we stated above that the alternative branchings are in the majority, since in most cases the system defines a refinement of superordinate features into subordinate features. As the other general characteristic we stated the disjunctive branching of the root node. We can explain this "and" relationship between the dominating features of the system by how we conceive of our environment. According to gestalt psychology this proceeds at least according to the following two principles holding for the perception of "ENTITIES" and "SITUATIONS" respectively. These two principles correspond to the "and" relationships dominated by the root.

On the one hand the definition of concepts depends on whether our conceptualization of "ENTITIES" or "SITUATIONS" is based more or less directly or indirectly on our sensory perception. The former case in which concepts are abstracted directly on the basis of sensory perception holds for "CONCRETES", the latter case of indirect conceptualization holds for "ABSTRACTS".[4]

In the case of "CONCRETES" no higher order information processing takes place, because there are no parts for which an elaborate substructure has to be reconstructed. Moreover the perceptual properties remain fairly constant between exemplars, so that they are easy to reidentify. Just the opposite holds for "ABSTRACT" concepts.

On the other hand we either define concepts by their outlines or by their inner configuration.[5]

In the former case, in which our definition depends on the more or less sharp outlines of the "ENTITIES" or "SITUATIONS", we conceptualize "COUNTABLES" in the case of "ENTITIES". In the case of "SITUATIONS" we conceptualize "PERFECTIVE" "ACTIONS" or "EVENTS", that means, either "SITUATIONS" for which a terminal phase is expected, which holds for "ACCOMPLISHMENTS", or "SITUATIONS" which are just in the terminal phase, which holds for "ACHIEVEMENTS" and "EVENTS".[6] In this case the boundary of the concept can be defined in terms of a terminal point or phase of the situation.
In the latter case in which the outlines are indistinct, we define "ENTITIES" by means of their inner configuration as different subcategories of "MASS".[7] Correspondingly we define "SITUATIONS" as different subcategories of "IMPERFECTIVES", if the situation is focused without reference to its terminal point or phase, that means as either "ACTIVITY" or "PROCESS" respectively or as "STATIVE".

The third "and" relationship of our rule systems cannot be explained by the same cognitive principle both for "ENTITIES" and "SITUATIONS", though it is obligatory for both. Only the obligatoriness of the situational "and" relationship can be made evident by cognitive principles. This third "and" relationship of "SITUATIONS" is the perception of their procedural characteristics, which is precisely the "AKTIONSART". Depending on whether it is "PUNCTUAL" or "DURATIVE" or "ITERATIVE", the "Aktionsart" combines in a definite way with aspect, which can either be "PERFECTIVE" or "IMPERFECTIVE". Now, both "PERFECTIVES" and "IMPERFECTIVES" can take the subcategory "CAUSATIVE" whereas the other subcategories of both aspects branch into disjunct feature sets, the refinement being defined by the "is" relationship and by the inheritance of attributes. Here the manifold branching of the "PERFECTIVE" aspect into "MUTATIVE", "INCHOATIVERE","REVERSATIVE", "RESULTATIVE" and also "CAUSATIVE" and the inheritance of the semantic features "ACHIEVEMENT", "ACCOMPLISHMENT", and "EVENT" are remarkable, whereas the "IMPERFECTIVE" aspect, apart from the possibility of taking the subcategory "CAUSATIVE" only inherits the features "PROCESS", "ACTIVITY" and "STATIVE". This is the reason for sympathizing with GALTON (1964.140f.), who pleads for considering the "PERFECTIVE" as the unmarked aspect since it "represents our normal scheme of arranging our perceptions". In using the "IMPERFECTIVE" we create an artificial stability by stopping the procedure of situations and thus making them timeless, whereas the procedural arrangement within time is usually considered as the unmarked case of "SITUATIONS".

With "ENTITIES" the third "and" relationship which branches from the root of our grammar is the alternation between "NATURALS" and "ARTIFICIALS".

We have thus shown how on the basis of empirical work two systems have grown independently of one another, one for "ENTITY" features and one for "SITUATION" features, which both have the same number of disjunctive arcs descending from the root node. And what is even more interesting and corroborates our systems is the fact that two of the three disjunctive arcs of both systems can be explained by the same cognitive principles, which also are obligatory in the process of conceptualization.

Notes

1. I want to express special thanks to Angelika MUELLER-v.-BROCHOWSKY for programming the grammar and for valuable suggestions.
2. This conclusion is not our private impression. A look into the literature on semantic feature networks shows that they are generally organized like this: the dominating nodes of the network are related by disjunction, whereas the features lower down in the network are rather related by alternation; that is, they are more strictly hierarchically organized (cf. e.g. WOODS 1975)
3. Especially in order to cope with the manifold semantic problems when coding lexical units one cannot ignore this fact. BARSALOU (1982) has tested and verified the existence of two types of concepts: context-independent and context-dependent concepts associated with verbal expressions. The results of his investigation make him conclude, that context-dependent properties have a major role in the definition and establishment of meaning, as they are also responsible for changes in the accessibility of context-independent properties (cf. ebd. p.92).
4. This definition of "CONCRETE" matches the GIBSONIAN theory of "direct" perception.
5. This principle again holds for "ENTITIES" and "SITUATIONS" respectively. Among "ENTITIES", there are e.g. tables, books, knives, wars for which we can image rather definite and clear outlines, by means of which they are limited against their environment, either as "CONCRETES" by a definitely shaped limitation of material or as "ABSTRACTS" by the limitation of a definite phase structure of a "PROCESS" or "ACTION". In English the possibility of pluralization indicates that thus conceptualized entities are "COUNTABLES". Among situations there are "DYNAMIC" "SITUATIONS" like She wrote a letter yesterday or The avalanche rolled down the mountain, which are also imaged as having a definitely limited phase structure, that means as a "PROCESS" or "ACTION" occurring in a definite order and ending in a definite, i.e. expected way. This should explain how we image "COUNTABLE" "ENTITIES" and "DYNAMIC" situations by the same cognitive principle.
The opposite of such a definite and sharp limitation towards the environment is the imagination of an amorphous mass, which is less precisely defined for its inner configuration and thus not at all for any definitely shaped limitation. This is the case with "MASS" entities like the "CONCRETE" substances water and gold or

abstract "SITUATIVES" like information, inflation. This is also the case with "ACTIVITIES" and "PROCESSES" like Yesterday she painted or The mast was shaking in the wind and even more so with "STATIVES" like During the week she gets up at seven or This mast shakes in the wind.

6. Refer to LANGACKER 1984. For the differentiation of "ACTION" "SITUATIONS" into "ACTIVITY", "ACCOMPLISHMENT", and "ACHIEVEMENT" refer to VENDLER who has introduced this classification. For the distinction between "PROCESS" and "EVENT" cf. e.g. BRANSFORD & McCARRELL for their criteria. See also LYONS 1977.483 and MILLER & JOHNSON-LAIRD 1976.85ff.

7. Refer to LANGACKER 1984

References

ANDERSON,R.C. & G.H. BOWER
1980 Human Associative Memory. Hillsdale.

ATTNEAVE,F.
1972 Representation of Physical Space. In: MELTON & MARTIN 1972.

BACHE,C.
1985 Verbal Aspect.

BALLMER,T. & BRENNENSTUHL,W.
1982 Lexical Analysis and Language Theory. In: EIKMEYER & RIESER 1982.

BALLMER,T. & BRENNENSTUHL,W.
1982 An Empirical Approach to Frametheory: Verb Thesaurus Organisation. In: EIKMEYER & RIESER 1982.

BARSALOU,L.W.
1982 Context-dependent and Context-independent Information in Concepts. Memory and Cognition 10(1).82-93.

CHAFE,W.L.
1971 Meaning and the Structure of Language. Chicago.

COHEN,B.
1984 Models of Concepts. Cognitive Science 8.27-58.

DAHLBERG,I.
1982 ICC-Information Coding Classification - Principles, Structrure and Application Possibilities. International Classification 9(2).87-93.

GARNER,W.R.
1972 Information Integration and Form of Encoding. In: MELTON & MARTIN 1972.

EIKMEYER,H.J. & H.RIESER (Eds.)
1982 Words, Worlds, and Contexts. New Approaches to Word Semantics. Berlin.

FREDERIKSEN,C.H.
Semantic Processing Units in Understanding Text.

FRIEDRICH,P.
1974 On Aspect Theory and Homeric Aspect. Bloomington.

GIBSON,J.J.
1977 Affordances: In: R.SHAW & J.BRANSFORD (Eds.). Perceiving, Acting, and Knowing. Hillsdale.

LAKOFF,G.
1982 Categorization and Cognitive Models. Linguistic Agency of the University of Duisburg (previously Trier). D-4100 Duisburg.

LANGACKER,R.
1983 Foundations of Cognitive Grammar 1,2. Linguistic Agency at the University of Dusiburg (previousl Trier). D-4100 Duisburg.

LANGACKER,R.
1984 Topics in Cognitive Grammar. Lectures 1-8,MS.

LYONS,J.
1977 Semantics. Oxford.

MELTON,A.W. & E.MARTIN (Eds.)
1972 Coding Processes in Human Memory. Washington.

MILLER,G.A. & P.N.JOHNSON-LAIRD
1976 Language and Perception. Cambridge.

MINSKY,M.
1975 A Framework for Representing Knowledge. In: P.H. WINSTON (Ed.). The Psychology of Computer Vision. New York.

PALMER,F.R.
1974 A Linguistic Study of the English Verb. London.

ROSCH,E.
1978 Principles of Categorization. In: E.ROSCH & B.B.LLOYD (Eds.).Cognition and Categorization. Hillsdale.

VENDLER,Z.
1967 Linguistics in Psychology.Ithaca.

WINOGRAD,T.
1972 Understanding Natural Language. Cognitive Psychology 3.1-191.

WOODS,W.A.
1975 What's in a Link: Foundations for Semantic Networks. In: D.G.BOBROW & A.COLLINS (Eds.). Representation and Understanding: Studies in Cognitive Grammar. New York.