[*International Conference on the State of Machine Translation in America, Asia and Europe*. Proceedings of IAI-MT86, 20-22 August 1986, Bürgerhaus, Dudweiler]

Jürgen Kunze

Transfer as a touchstone for Analysis

Zentralinstitut für Sprachwissenschaft
der Akademie der Wissenschaften der DDR
Prenzlauer Promenade 149-152
Berlin
DDR - 1100

TECHNICAL DATA

- experimental research project for automatic analysis of German;

- wordform and syntactic analysis with semantic component;

- dependency structures with subordination relations and paths of action for paradigmatic and selective connections;

- dictionary: about 1000 entries;

- rules: about 350 bundles;

- implemented in LISP on ES 1055 with TSO under OS/MVT

Transfer as a touchstone for Analysis

0. Introduction

The paper consists of three parts. At first I will outline the syntactic model we use as the base of our experimental system. The second part brings a short description of the parser we are developing. It contains three main steps which I will characterize briefly. At the end I want to demonstrate what is expressed in the title. A few examples are presented which show how some essential tasks within the lexical and syntactic transfer may be performed with certain alleviations if they are based on the aforesaid model.

1. An extended version of dependency grammar as a base for analysis

The model we use is a dependency grammar with some additional components. It is highly formalized and has an axiom system as kernel (principle of differentiation). Essential properties may be proved in a strictly mathematical sense. The principle of differentiation is nothing else but an implicit description of the notion "syntactic relation".

I skip here the following questions:

- How to motivate (and define exactly) the subordinations, i.e. the pure dependency structures? This is quite another problem than what has to be accomplished by a parser building up a tree on the base of rules (which contain implicitly these motivations).
- What are the principles for the representation of coordinated sentences (with omitted parts), of ellipses etc.
- How to guarantee that every real syntactic ambiguity may be represented within this formalism?

The last question supposes the application of the model to a concrete language, of course.

1.1. Some general characteristics

The essential features of our model are the following:

(A) Use of subordination relations, which formally appear as labels at the edges of dependency trees. They may be interpreted as signs for the syntactic relation between the two parts (subtree and

tree-context) of the whole tree which one obtains by cutting the edge where this subordination relation appears. The term "syntactic relation" has to be understood in a broader sense, i.e., some semantic relations are involved, too. It is clear that this claim (to represent semantico-syntactic relations) has to be justified by corresponding definitions. This is possible to a high degree, but we skip this here (cf. KUNZE 1975).

(B) On the foundation wall of the pure dependency trees the whole model is borne by four columns, namely paradigmatics, selection, (linear) ordering and subordination relations. The general idea behind this is to express each restriction that is not (or cannot be) expressed by the first three parts, by the last. This principle causes a certain partition (differentiation) of the set of the subordination relations and renders it possible to explain the distinction between any two subordination relations (cf. KLIMONOW 1982).

(G) Paradigmatics and selection are represented by special means - paradigmatic and selective connections and impositions (for impositions see below!). To every such connection belongs a class of paths in dependency trees. We call them "paths of action". They connect pairs of nodes. At the starting node a certain restriction is given (coming from the sememe at this node), it has to be fulfilled by the sememe at the endnode of the path. Thus the paths of action are roads for the transport of restrictions. Their course has to be strictly distinguished from the goods conveyed on them. The first cohere with subordination relations and are a part of syntactic structures, the second not at all.

(D) According to the general idea, the course of these paths is expressed by subordination relations in the following way: We consider a dependency tree that corresponds to a sentence. In such a tree with a subordination relation at every edge there is (independently of the interpretation of the nodes by sememes) only one possibility to draw the paths. This is no axiom or presumption but an assertion that may be proved in a strictly mathematical way.

(E) Linear ordering is represented by special means - another point I skip here.

1.2. Illustrations and examples

In my examples I refer to an older stage of our work when we defined a rather small inventory of paradigmatic and selective connections. A new inventory is under construction together with a set of case relations.

As an example for a paradigmatic connection one may take the congruence between (grammatical) subject and finite verb. It has (in German) two "valid" paradigmatic categories (number and person) and requires the equality of the values at the starting node and the end node. The paths go from the top node of the subject-subtree to the finite verb. And this is the general basic scheme for a paradigmatic connection: Valid paradigmatic categories, paths from ... to ... (expressed in syntactic terms for the "human use", in terms of subordination relations for the formal processing), equality of values for the valid categories at both nodes (what as a practical operation may be performed by taking the common values at both nodes).

The direction of the paths has some essential meaning for definitions, but afterwards (e.g. for parsing) it is unimportant. The classes of paths are not always so simple: For the congruence between a noun/pronoun and a reflexive pronoun one has rather complicated paths (^{a} = starting node, ^{e} = endnode):

(1) ^{a}Er fürchtet ^{e}sich.

(2) ^{a}Er soll ^{e}sich gefürchtet haben.

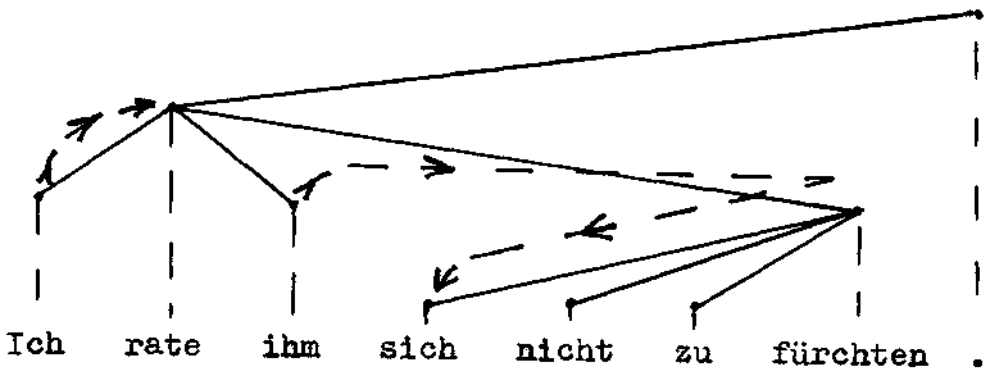
(3) Ich lasse ^{a}ihn ^{e}sich fürchten.

(4) Ich rate ^{a}ihm ^{e}sich nicht zu fürchten.

• • •

The term "complicated paths" refers to the fact, that in the trees these paths have to follow the edges as in (5).

(5)



A complete path has to be composed of fragments of the type

- "edge down" (fürchten - sich)
- "edge up" (does not appear in (5) for the reflexive congruence, but for the finite verb congruence: Ich - rate)
- "bridge" (ihm - fürchten).

The fact that starting nodes and endnodes are not connected straight and the paths have to "follow the edges" renders it possible to realize another principle: The fragments of paths become parts of syntactic rules (which are bundles, i.e. tree-like structures of depth one in case of projectivity). During parsing the fragments for the several connections are concatenated to complete paths. So one can handle even "long-distance-connections" by syntactic rules.

It is again subject to a mathematical proof, that the paths are not only uniquely determined by the whole tree with subordination relations at the edges (as said above), but that furthermore the paths may be effectively constructed from fragments that are stored in the rules. It can also be shown, that in this process no confusion can arise and that only the actual paths join completely. In (5) the path ihm - sich is dissected into two fragments that appear in two quite different syntactic rules.

It is clear that for a language like German the inventory of paradigmatic connections has narrow natural limits (it covers the different types of congruences and governments). This is not true for the selective connections. Here one has much more "candidates". The essential distinction on the formal side between paradigmatic and selective connections is the replacement of the "category-value"-mechanism by the application of selective features, which is a stronger means. It can be pointed out that the "category-value"-mechanism is not sufficient for expressing semantic constraints. Furthermore there are even purely grammatical restrictions that cannot be represented by this mechanism, e.g. the inflection of attributive adjectives in German:

(6) ein alter Mann, eines alten Mannes, ...

(7) der alte Mann, des alten Mannes, ...

We illustrate the selective mechanism by the connections Actor and Patient. Here the classes of paths of action are much richer than in the paradigmatic case. For Actor the class contains among others the following paths:

(8) Der ^{e}Onkel ^{a}sucht einen Bleistift.

(9) Ich lasse den ^{e}Onkel einen Bleistift ^{a}suchen.

(10) Ich rate dem ^{e}Onkel einen Bleistift zu ^{a}suchen.

(11) Ich beauftrage den ^{e}Onkel einen Bleistift zu ^{a}suchen.

(12) Die ^{a}Suche/Das ^{a}Suchen eines Bleistifts durch den ^{e}Onkel ...

(13) Die ^{a}Suche/Das ^{a}Suchen des ^{e}Onkels nach einem Bleistift ...

(14) Die Bleistift^{a}suche des ^{e}Onkels ...

(15) Der einen Bleistift ^{a}suchende ^{e}Onkel ...

(16) Ein Bleistift wird vom/durch den ^{e}Onkel ^{a}gesucht.

(17) Der ^{e}Onkel lief einen Bleistift ^{a}suchend umher.

(18) Der ^{e}Onkel ist mit dem ^{a}Suchen/der ^{a}Suche eines Bleistifts beschäftigt.

(19) Ich beobachtete den ^{e}Onkel beim ^{a}Suchen/bei der ^{a}Suche eines Bleistifts.

The same sentences provide us with examples for paths for the Patient-connection (put the ^{e} at Bleistift-). It should be noted that in (8) to (11) all paths for Actor are different (as types), whereas this is not the case for Patient.

Details about selective connections and the use of selective features can be found in KUNZE 1982.

By comparing the paradigmatic and selective connections one finds out in general, that per selective connection there are more paths (= path-types), the paths are longer and (in some intuitive sense) more complicated than in the paradigmatic case. If one intends to handle only simple "one-edge-paths" like that for the subject-finite-verb-congruence, these aspects are of no special interest. But what about (8) - (19)?

Another important principle is, that for each selective connection the class is closed under transformations: If a path-type is contained in a class, so all its transformational variants are contained, too. (9) to (19) may be considered as variants of (8).

Attached with this property of the classes of paths is another question, which I can only mention here: In order to have some real advantage from this treatment of transformations one has to establish an analogous principle on the level of wordforms. This means that e.g. the verbal noun "die Suche" has to be connected to the verb "suchen". Then it is possible to perform the transition

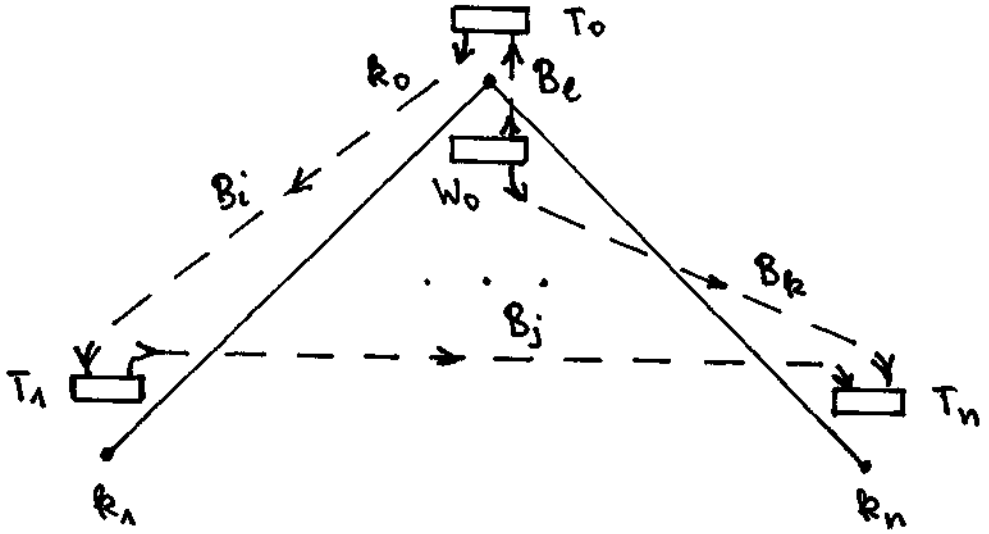
(20) Der Onkel sucht ...

(21) Die Suche des Onkels ...

on the syntactic level (subject becomes genitivus subjectivus) and on the wordform level (finite verb becomes verbal noun) in parallel with the invariant Actor-restriction.

The whole model permits to introduce a structure for the informations assigned to wordforms: They have five components (paradigmatics, selection and ordering according to the general structure of the model, one component related to word classes, and another related to valency and government). The same components are suitable to describe properties of subtrees.

A bundle (of depth one) is a structure consisting of one top node k_o, no or some depending nodes k_1, ..., k_n, subordination relations R_1, ..., R_n at the edges, at the top node conditions W_o for wordforms (to be an interpretation of this node) and a resulting subtree description T_o, at the depending nodes conditions T_1, ..., T_n for subtrees (theoretical realization of a generative bottom-up-principle), fragments of paths for paradigmatic and selective connections B_i and restrictions for linear ordering. This concept can



be justified theoretically, it is subject to practical modifications and adaptations. A bundle is a rather complex structure and contains implications between edges, too! For verbal forms e.g. normally one bundle describes the whole "government-behaviour"! For details see REIMANN 1982.

2. The parser

After the red-tape-discussion of the theoretical foundations we turn now to the more unpleasant question of parsing. There are different strategies for the realization of such a task. Our principle is (in spite of limited resources) to develop a general procedure in the sense that the methods have some general character. But the procedure is also general in the sense that no particular application is intended or taken into consideration till now. The parser has purely experimental status and is at first a means for checking some theoretical positions and practical assumptions. It undergoes constant modifications,

The whole parser consists of three procedures, which I will outline now.

2.1. The ATN-analysis

At first sight it looks a bit curious to rely on ATN for an analysis with dependency trees as output. This is however no ideological question but a matter of practical advantages. The use of bundles in an early stage of analysis causes some problems (in direction to "combinatorial explosion"), that make the whole parsing procedure inefficient. They are connected with the fact that bundles are "too complicated" for the "simple" subordination of articles under nouns or of an infinitive under an auxiliary. So we apply an ATN as a limited tool for a limited purpose: The ATN has to solve the following tasks:

- segmentation of the sentence into clauses;
- segmentation of clauses into simple groups, i.e. noun groups with left attributes, prepositional groups and some others;
- segmentation of complex verbal groups into simple verbal groups;
- treatment of simple verbal groups.

The two last points may be illustrated by

(22) er hat gesehen werden können
(complex verbal group)

(23) er gesehen werden + können hat
(two simple verbal groups)

(24) er $\text{sehen}_{\text{Inf.Pass.Präs.}}$ + $\text{können}_{\text{Akt.Perf.Ind.}}$

Thus each simple verbal group (tense-, mood- und active/passive-variants) becomes one node in the tree, which represents this verbal group in a compressed form. In this way German becomes a language like Latin, some problems of unprojectivity and subordination (e.g. adverbials) lose their purpose (for details see KÜSTNER 1985).

2.2. The local analysis

The result of the ATN-analysis is the input of the next step, which is mainly based on the subordination relations. It has the aim to find out possible subordinations between nodes. The criteria are all properties connected with subordination relations as single units:

- word classes at the two nodes to be checked for subordination;
- paradigmatic connections with paths of action that pass along only one edge (e. g. congruence subject-finite verb);
- paradigmatic and selective impositions.

Impositions are another type of restrictions than connections: Certain values or features are "demanded" by certain subordination relations (case of subject = nominative, prepositional groups as local adverbials have to fulfil certain restrictions, e. g. *während des Hauses, *neben dem Abend).

The output of the second step of analysis is a (rather chaotic) graph. The edges are marked by subordination relations. The trees will arise from these graphs by cancellation of edges.

2.3. Analysis by bundles

Because the bundles are a rather complex type of rules it is advisable to apply them only when some basic problems have been solved before. Our strategy is to use bundles as a filter for the elimination of possible subordinations in order to obtain trees that represent the actual syntactic structure(s)

of the given sentence.

The main criteria (expressed in the bundles) are:

- mutual exclusion of subordination relations under a node;
- implications between subordination relations (Ich schenke ein Buch. *Ich schenke dir. —> Ich schenke dir ein Buch.);
- long-distance-connections (fragments of paths are joined, then the connections are checked), one-edge-connections are considered, too;
- restrictions for linear ordering,
- projectivity (if suitable).

It should be noted that some criteria depend on the wordforms at the nodes (cf. the schenken - example).

Another type of criteria is based on tree-conditions (e.g. no node is subordinated twice).

2.4. The final result

At the end one obtains per sentence one or more dependency trees. Each edge is marked by a subordination relation, each node interpreted by the (actual) sememe-description and the tree contains the paths of action for the paradigmatic and selective connections.

It is no feature of inconsistency, if we represent the trees per sentence in a "resulting graph". This is a natural factorization and avoids some combinatorial disadvantages. The resulting graph contains proper nodes and edges (of the dependency trees) and besides subsidiary nodes and edges (to achieve the branchings).

The paths of action per tree form a net (not a tree!). They may be used as a semi-product for semantic networks (see JUNG 1982).

3. Some aspects of the transfer

At first we consider two typical tasks of the lexical transfer. For the translation of the English verb "to mean" into German one has at least the following equivalents:

(25) I mean to do it: beabsichtigen

(26) I mean you to do it: wollen, wünschen

(27) I mean your father: meinen

(28) This word means ...: bedeuten

One could list even some more equivalents, but then a discussion would arise whether and how they overlap each other.

The actual choice among these four possibilities may be formulated by certain criteria. They use two levels, namely syntax (valency) and semantics (features of actants): We outline them in a quasi-formal and simplified way:

(29) Actor (HUM.BEING) & infinitive-complement:

- beabsichtigen

(30) Actor (HUM.BEING) & Ad-construction

- wollen

(31) Actor (HUM.BEING) & Patient-object

- meinen

(32) Actor (~HUM.BEING)

- bedeuten

In a syntactic model which has a certain minimal degree of adequacy it is possible to formulate these criteria, e.g. as syntactic environments of the node of "mean" with conditions for some other nodes. But this is only the first half of the solution: The method of environments appears rather inconvenient, if one takes into account their transformational variants: The list to be checked in order to find out the actual equivalent becomes very long. If one has at hand the paths of action in the structure obtained by analysis, the situation looks a bit better: The classes of paths are closed under transformations, so e.g. the question "Is the Actor of 'mean' a HUM. BEING?" may be put and answered rather easily, one need not bother about variants and looks only for the path of action in the tree, which leads immediately to the decisive node.

The next question we touch is the translation of compounds. In many cases a German compound has to be expressed by an English syntactic construction, and one has e.g. to choose the correct preposition, if the equivalent has the form "$Noun_1$ Preposition $Noun_2$". In languages like Russian the choice of the prepositions and cases is more closely related to the semantic relation between the nouns than in English. In many cases, e.g. if $Noun_2$ is a deverbativum, one can rely on selective connections to find out

the actual semantic relation:

(33) Kinderarbeit
Actor-connection fulfilled, possibly subjective relation

(34) Kinderernährung
Patient-connection fulfilled, possibly objective relation

For the syntactic transfer I place a proposal that seems a bit tricksy, but sometimes it may help perhaps. A well-known expedient in case of unsolved (syntactic) ambiguities in the source language is to translate them, i.e. to find a target formulation with an analogous ambiguity. This requires a rather intelligent transfer. One facility is given by the resulting graph (cf. 2.4.): It is (if there remains some ambiguity after analysis) a complex representation of the different readings. If the transfer works in such a way that it tries not to divide this graph into two (or more) disjunct target-graphs, one gets at the end at least on the syntactic level the desired result!

KUNZE 1975
J. Kunze, Abhängigkeitsgrammatik.
Studia grammatics XII, Berlin 1975

KLIMONOW 1982
G. Klimonow, Zum System der Unterordnungsrelationen im Deutschen.
Automatische Analyse des Deutschen, Berlin 1982, pp. 65-174

REIMANN 1982
D. Reimann, Büschel als syntaktische Regeln.
Automatische Analyse des Deutschen, Berlin 1982, pp. 175-192

KUNZE 1982
J. Kunze, Probleme der Selektion in der automatischen Satzanalyse.
Automatische Analyse des Deutschen, Berlin 1982, pp. 223-256

JUNG 1982
U. Jung, Einige Zusammenhänge zwischen syntaktischen Strukturen und semantischen Netzen. Explizite Beschreibung der Sprache und automatische Textbearbeitung, Prag 1982, pp. 52-73

KÜSTNER 1985
A. Küstner, Verbgruppenanalyse, interner Bericht