

**Suomen kielitieteellisen yhdistyksen julkaisuja
Publikationer utgivna av Språkvetenskapliga
Föreningen i Finland**

3.

**Papers from the Conference
on General Linguistics**

Selli 29.-30.8.1979

**Publications of The Linguistic Association of Finland
Turku 1979**

PREFACE

The papers in this volume were read at the Conference on Quantitative and Computational Linguistics arranged by The Linguistic Association of Finland at Seili, August 29-30, 1979. A simple format was chosen in order to avoid delays in the publication of the papers.

Sincere thanks are due to the Council of Humanities (Valtion humanistien toimikunta) for a grant that facilitated publication of the volume.

Copies of the volume can be ordered from Suomen Kielitieteellinen Yhdistys, c/o Kaisa Häkkinen, Fennicum, Turun Yliopisto, SF-20500 TURKU 50, Finland.

Kaisa Häkkinen Fred Karlsson

ISBN 951-95263-2-3
ISSN 0357-0622
Turun yliopiston offsetpaino 1979

Fred Karlsson
 AUTOMATIC MORPHOLOGICAL SEGMENTATION OF FINNISH WORD FORMS

1. This paper is a preliminary report on a project called FINMRF conducted in collaboration by Benny Brodda and the present author; cf. Brodda & Karlsson (forthcoming) for a more comprehensive presentation. Our main Fragestellungen are these: To what extent is it possible to provide an algorithm for automatic morphological segmentation of Finnish word forms? What kinds of morphological ambiguities do the surface strings contain? What theoretical implications does work of this type have? Due to the well-known complexity of Finnish morphotactics (cf. Karlsson (1976, 1977, 1979)), we have at this initial stage restricted our analysis to cover inflectional, possessive, and cliticized morphs. Our segmentation procedure does not (yet?) cope with compounds or derivational morphemes.

The segmentation rules are formulated in the BETA system constructed and implemented by Benny Brodda at Språkförlaget Skriptor, Stockholm (cf. Brodda (1977, 1979, forthcoming)). Brodda (e.g. 1979) has applied BETA to the automatic analysis of Swedish morphology with considerable success, and the present project is to be conceived as an experiment with a language of a different structural type. BETA is a programming language organized as a substitution grammar: a string is read and rewritten as another string, provided certain conditions obtain. The BETA rules are slightly modified Turing rules. In particular, BETA operates with a system of internal states, which can be referred to in the substitution rules. The general format of a BETA rule is shown in (1) (cf. e.g. Brodda (1977)).

- (1) X Y LC RC SC RS MV MD

X is the string to be substituted, Y the rewritten string. LC and RC (Left and Right Context Condition) express restrictions on the segment immediately left and right of X, respectively. SC (State Condition) specifies the state the system must be in for the rule to be applicable, RS (Resulting State) specifies the state the system proceeds to when the rule has been applied. MV (Move) indicates where the following operation is to be performed (e.g. left of the string being handled), and MD (Mode) is a rule parameter used for resolving ambiguities when several rules are applicable to the same string. LC and RC can be defined so as to comprise any segments; if a rule is applicable e.g. only when vowels occur left of a string identified as a potential morph, one postulates a set (called e.g. '32') comprising l, e, ä, y, ö, u, o, a, and the label

'132" can then be used as LC in the appropriate rule(s). In the same way, SC's may be defined so as to cover any desired set of states. All LC's, RC's, and SC's are defined under DEFSET. Sets of segments and sets of states may be summed under the same heading. (2) is an excerpt from DEFSET.

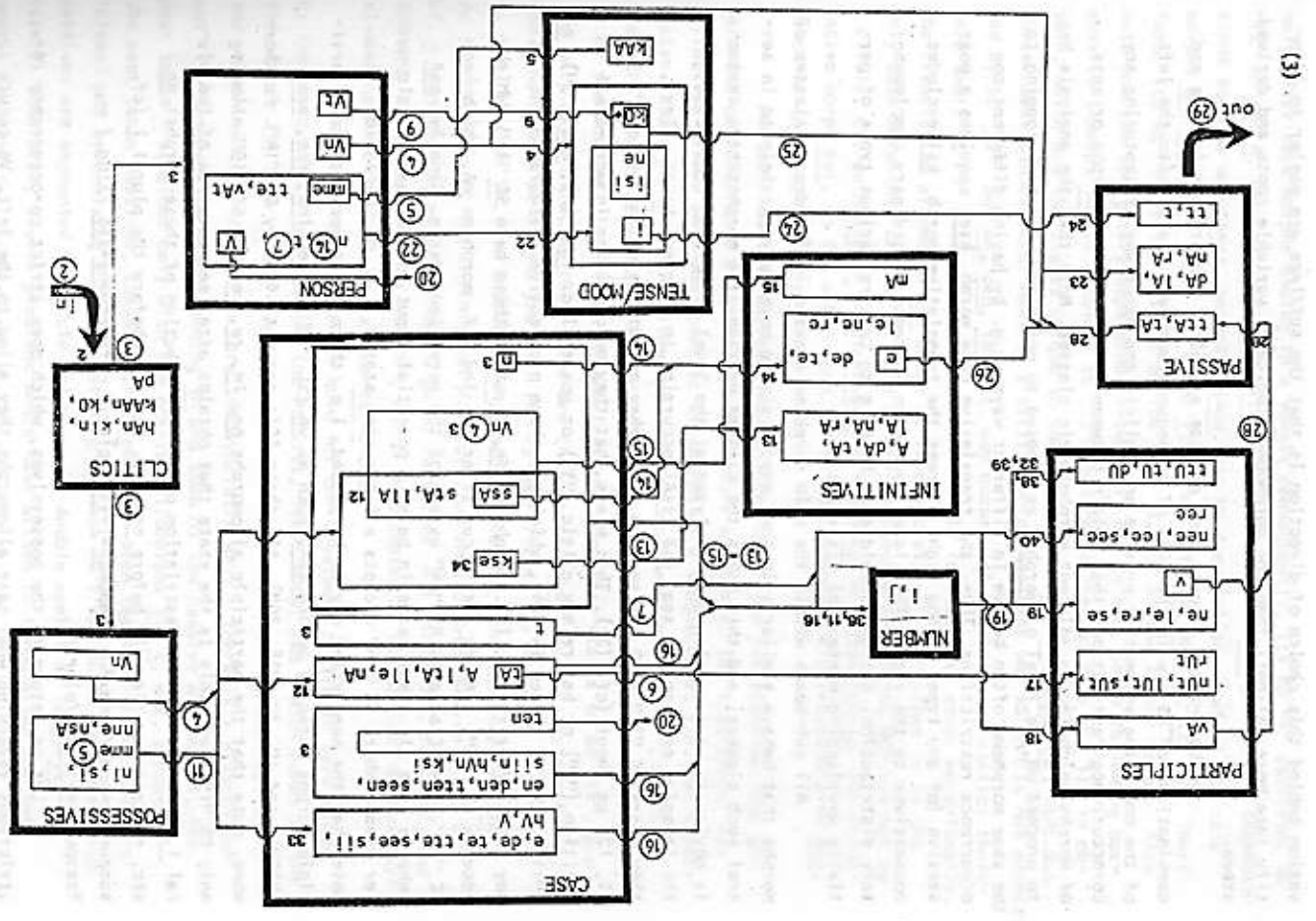
(2)

DEFSET	#	LC	RC	SC
11 1	32	45	#	' ?
21 2	32	45	#	' ?
31 2-3	65-93	96-124		
41 4	43	65-93	96-124	
51 5	B	C	D	P
61 6	G	H	J	K
71 7	L	H	N	P
81 8	Q	R	S	T
91 9	V	X	Y	Z
101 10				
111 2-5	11	12		
121 2-5	11	12		
131 2-5	13			
141 14				
151 4	14	15		
161 6	13	16		
171 2	3	6	17	
181 2-4	7	11-18		
191 19				

E.g. set '18" denotes the states '2-4, 7, 11-18" when used as SC, and the segments 2, 3 when used as LC and RC.

BETA works on the basis of phonological (graphic) information only, plus the states, which can be interpreted as 'traces' of identified morphotactic classes (cf. (3)). No higher-level syntactic parsing is involved, neither does the segmentation involve any lexicon (apart from an exception list containing about 200 frequent grammatical free morphemes that the rules would segment incorrectly; e.g. *vain* 'only' should not be segmented **va-in* as if it were a past tense first person verb form). The input to the rule system consists of running texts; cf. below.

2. The flow chart on the following page shows our analysis of Finnish morphotactics. This is the linguistic substance underlying our BETA rules; the rules are an implementation of the flow chart. The thick-line boxes are the basic morphotactic classes (excluding derivational morphemes). Encircled figures are RS's (output states), e.g. the resulting state is 3 after a clitic segmentation, 19 after a number 1 segmentation, and 25 after a conditional 1st segmentation. 2 is the input state to the whole segmentation procedure, 20 the RS of nominal segmentation paths and 29 the RS of verbal segmentation paths. Note that the segmentation proceeds from left to right. The main linguistic moti-



vation behind this choice of direction is that the suffixes are easier to identify (are more salient) than the morphophonologically variable roots and derived stems.

The bold-faced figures such as 6, 2a are SC's. Proper formulations and combinations of RS's with the SC's of subsequent morpheme classes (to the left of the one being segmented) provide an explicit and exact way of capturing any co-occurrence restrictions that obtain between single adjacent morphs or sets of morphs belonging to adjacent morphotactic classes. Note that the analysis has to proceed on the level of morphs, as opposed to morphemes. Morphs belonging to the same morpheme often behave in different ways, e.g. by having different co-occurrence restrictions. Thus, the translative case morph kse requires a possessive (of any type) to the right whereas the translative morph ksi excludes possessives to the right. This is of course an instance of ordinary complementary distribution, but one should also be able to capture various types of partially overlapping contextual distribution.

All sub-boxes within the main morpheme classes in (3) denote classes of morphs that behave similarly in some way. A single morph may participate in several such classes, and this makes the classes overlap. The morphotactic system is obviously very complex when analyzed at the level of morphs. Consider e.g. the inessive case morphs ssa, ssā (ssā) occurring in four sub-boxes. First, they share SC '12' with stā and llā, i.e. they accept any of the states '2-5, 11, 12' as input (cf. (2)). This means that they occur either in word-final position ('21) or before any clitic ('31) or possessive: ŷn ('141), mmē ('151), nī ('111). The RS of ssā is '141', which has three prospective leftwards paths. It may occur after the 2nd Inf. morphs e, de, te etc.; these have SC '141', which accepts RS '141'. ssā may also occur after the 3rd Inf. morpheme mā, which has SC '151'. This SC accepts RS '141' as one of the permitted states. Finally, ssā behaves like all other cases in being a potential input to the morphotactic number class, where SC '136' accepts all the case states, i.e. '4, 6, 13-16'. Also note that the number box may be by-passed, i.e. there might occur certain participle morphs directly before ssā, such as vā (SC '181), nee, lee, ree, see ('101'), and lū, dū ('138, 391'). All these SC's are satisfied by RS '141'. Furthermore, note that the participle allomorphs ne, le, re, se have SC '191' allowing only RS '191', and this is the state that obtains after segmentation of the plural l. This is an exact description of the distribution of these morphs: ne etc. never occur directly before ssā etc. but only before the plural l (cf. saapu=nee=ssa /sing./ ~ saapu=ne=l=ssa /plur./ ~ *saapu=ng=ssa /sing./ ~ *saapu=nee=l=ssa /plur./).

Let us return to the possessives, which obey strict co-occurrence restrictions regarding what case allomorphs they allow to the left. ŷn occurs

only after case morphs ending in a vowel; consequently, its RS '141' satisfies SC '121' of tā, A, lā, mā, ssā, stā, llā as well as '141' of kse permitting segmentations such as maata=ŷn, men=ŷ=ksen. But ŷn does not occur after the case morphs en, den, ten etc., nor after l, n, nor after e, de, te, tte etc. Therefore, its RS '141' is not allowed by the SC's '13, 39' of the latter morphs. Also note how the obligatory dependency between e, de, te etc. and a following possessive (except ŷn) is handled. SC '1331' requires either of RS's '15, 11' to obtain, which are the states unique to the possessives nī, sī, mme, nee, mā. But these cases cannot occur next to a clitic, which explains why RS '131' is not permitted by SC '1331'. But en, den, ten etc. may occur with a subsequent clitic: therefore their SC '131' allows RS '131', but also RS '21' in case they occur word-finally. The translative allomorph kse needs an idiosyncratic SC '141' since it is the only case morph that can be followed by any possessive (including ŷn) but not by a clitic, cf. auto=ksen=ŷn ~ auto=ksen=l ~ *auto=ksen=l ~ *auto=ksen=kin. The translative morph ksi, on the other hand, is in complementary distribution with kse, which is captured by SC '131'.

Now consider the infinitives, in particular the restrictions that govern the co-occurrence of infinitival functors and cases. The traditional 1st infinitive allows only one specific allomorph of a single case, i.e. the translative allomorph kse. This allomorph leads up to the unique RS '141', which is the only case RS accepted by the SC '131' allotted to the 1st Inf. functors. The SC of the 2nd Inf. is '141' which requires the permitted states to be '141', and the only cases allotted this state are the inessive ssā and the instructive n. And finally the SC of the 3rd Inf. is '151' permitting the RS's '14, 15', i.e. the illative ŷn, the inessive ssā, the elative stā, and the adessive llā. As for the RS's of the infinitival functors, we should note that only the 2nd Inf. e permits the segmentation to go on leftwards. All other Inf. functors lead OUT (RS '1291'), but e is in RS '1261', which is accepted by the SC '1281' of the passive morphs to the left (e.g. sano=ttā=e=ssa).

Analogous considerations hold for our treatment of finite verb forms. As the connection lines between the PASSIVE, TENSE/MOOD, and PERSON boxes in the flow chart (3) show, quite intricate restrictions obtain here, too. Imperative and person endings combine in very idiosyncratic ways, there is a discontinuous link between the person ending ŷn and some of the passive allomorphs, the past tense l combines only with the passive allomorphs tt, t, etc. Except for the past tense passives, these restrictions have been resolved in the ordinary 'clean' way by devising a proper matching between RS's and SC's. The past tense passives are segmented by a rule performing a double operation (e.g. tti → tt=l). Otherwise, severe problems would have arisen in formulating the segmental contexts for the past tense and passive rules.

If FINNRF had access to a root (stem) lexicon, we would certainly receive a correctness level in excess of 90%. For example, of the incorrect segmentations just cited, *vi^h-den*, *vu^o-den*, and *katsom^h-sta* would be rejected, since there are no potential roots (stems) *vi*, *vu^o*, *katsomu*. The string *filosofaan*, on the other hand, cannot be resolved even using lexical information, as it might derive from either one of the noun lexemes *filosofi* 'philosopher' or *filosofia* 'philosophy'.

The computer facilities developed in connection with BETA by Benny Brodda also provide the opportunity of printing out alphabetical lists of the analysis of all running words with indications of their text frequencies. (12) shows what word forms there are in Leikola's book that belong to some lexemes related to *hyv^g* 'good'. The figures denote frequencies and the "u" in front of *hyvln* that this word belongs to the exception list. Every single morph boundary in (12) is correct; there are no over- or undergenerations.

(12)	1	HYVE=I=TA#*	2	HVYXKSY=NYT#*
	1	HYV=I=EN#*	1	HVYAKSY=TTY#*
	1	HYV=I=LE#*	2	HVYAKSY=TY=J=K#*
	65	+HYVIN#*	1	HVYAKSY=TY=NA#*
	6	HYV=I=N=KIN#*	1	HVYAKSY=TTA=ISI=IN#*
	1	HYVINVOINN=I=LE=IHE#*	1	HVYAKSY=TTA=V=I=LE#*
	1	HYVINVOINIT#*	1	HVYAKSY=TTA=VA#*
	2	HYVINVOINIT=A#*	2	HVYAKSY=TTA=AN#*
	1	HYV=I=NX#*	1	HVYAKSY=VAT#*
	3	HYV=I=A#*	3	HVYAKSY=X#*
	2	HVYVYODE=NA#*	1	HVYX=LE#*
	1	HVYVYTE=EN#*	2	HVYX=LLX#*
	1	HVYVYT=TX#*	2	HVYX=LTX#*
	6	HYVX#*	6	HVYX=NA#*
	3	HVYA=K3=EN#*	26	HVYX=NSX#*
	9	HYVX=KSI#*	1	HVYXNTAHTOINEN#*
	1	HVYAK3IKKATTTX=E=NA#*	1	HVYXNTAHTOISUDEL=LLA=AN#*
	1	HVYAKSIKXKXTTDB=OH#*	1	HVYXNTEKIJU=I=STX#*
	1	HVYAKSY#*	2	HVYX=VA#*
	1	HVYAKSY=I#*	1	HVYX=VA#*
	1	HVYAKSY=I=VXT#*	1	HVYX=VA#*
	1	HVYAKSY=MIKEN#*	2	HVYX=STI#*
	1	HVYAKSY=K=I#*	6	HVYX=VA#*
	1	HVYAKSY=HA=AN#*	14	HVYX=VA#*
	1	HVYAKSYNTX=K#*	1	HVYX=VA#*
	4	HVYAKSYNTX=K#*	3	HVYX=VA#*

Naturally, lists like (12) are methodologically very useful in checking where the rule system still leaks.

6. Finally, I shall touch upon some theoretical implications of our work and briefly hint at some possible avenues for continuing the research.

As noted in the previous section, our segmentation procedure has achieved an 85-90% level of success; Brodda's rule system for Swedish is even more efficient. How is this possible? The first conclusion to be drawn is that

the phonological (graphemic) string is very rich in information and contains abundant clues to morphological structure, and consequently to syntactic structure and semantic interpretation as well. Several present-day theories of language perception, recognition, and comprehension (especially the so-called "Active Theories" of speech perception, cf. Clark & Clark (1977)) lay most of the interpretive burden directly on syntax and semantics, as it were. Our results do not conflict with these views, but they certainly stress the importance of the information coded into the sound substance.

Another conclusion of our experiment is that there seems to be a fairly systematic difference in phonological structure between root ends and suffix ends. One should, in fact, speak of different canonical structures. If there were no systematic phonological differences between root and suffix ends, the outcome of our segmentation experiment should revolve around 50% with a lot of overgenerations where roots and stems have been incorrectly "invaded" by the suffix rules (e.g. *vu^o-den* instead of *vuode-n*). Such errors do of course occur, but they are relatively infrequent.

What are the phonological characteristics of suffixes and are there differences between different morphotactic classes as well? In order to elucidate these matters, we postulate the following three basic binary features.

- (13) a. Does the suffix close the preceding syllable?
- b. Are the vowels primarily open, i.e. a, ä, ö?
- c. Are the consonants primarily dental, i.e. t, d, s, l, r, n?

There is no doubt that these features are the fundamental ones for syllable structure, vowels, and consonants, respectively. Next we apply these features to the morphotactic classes in the flow chart (3). This yields the matrix (14), where a,b,c denote the features of (13), "u" a positive value, "-" a negative value, and "u" that no clear answer can be given. It should be stressed that the pluses and minuses depict tendencies only, the structure of a prototypical representative of the class in question.

(14)	CASE	NUMB.	TENSE/ MOOD	POSS.	PERS.	PASS.	INF.	PART.	CLIT.
a.	+	±	-	±	±	-	-	-	-
b.	+	-	-	-	-	+	+	-	±
c.	+	±	-	+	+	+	+	+	-

Some clear tendencies emerge. A prototypical case suffix closes the syllable and consists of open vowels and dentals (exceptions: many of the genitive plural allomorphs like *en*, *den*, *ten*, the translative *ksi*, *kse*). A clitic has two basic characteristics: it never closes the syllable and contains no (initial) dentals. Particle morphs don't close the syllable (exception: *tu*, *ty*), they contain open vowels and prefer dentals (exception: *va*, *vg*). Passive and infinitive

itive morphs seem to be almost identical: they don't close the syllable (exception: ttä, tt) and prefer open vowels and dentals. Person and possessive endings also turn out to be almost alike: some of them close the syllable, the vowels are non-open (exceptions: vät, nä) and the consonants are dental. Tense and mood morphs never close the syllable and generally don't contain open vowels.

The number morphs are fairly indeterminate, but it might be noted that i is non-open as are the vowels in the morphotactically equivalent tense and mood markers.

These are tendencies only, but even so, they have a discriminatory function which in part lies behind our 85-90% result. Of course the contextual specifications of the BETA rules express these restrictions. Roots don't often end in the ways specified by (14).

Next, one should ask what the function of these differences is. Why are root and suffix ends different and why do several of the morphotactic classes have phonological characteristics of their own? One obvious hypothesis is that such differences facilitate perception and comprehension of utterances, i.e. certain phonological features occurring in certain positions of the word are a direct signal of certain syntactic functions and/or meanings. In other words, the suffixal canonical structures serve as indexes of the relevant functions and meanings: e.g. a closed syllable towards the end of the word is a strong indication of there being a case or person ending (at any rate not a tense, mood, or clitic), etc. It is known from several studies of language acquisition that suffixes are highly salient (more salient than prefixes) for children acquiring their native tongue (e.g. Slobin (1966)). A high proportion of the morphosyntactic information relevant to the interpretation of the meaning of single words is located at the end of the word, and therefore it is perceptually profitable for the word end to be easily identifiable. The canonical structures might thus be conjured to serve the perceptual strategies. - On the other hand, we don't know what the relative importance of lexical and grammatical meaning is in language comprehension, and therefore one should not overemphasize the point just discussed.

Results of this type show one of the benefits of such explicit analyses as are required in computer studies: these analyses might eventually lead to new qualitative insights as well. In our study, this holds for the intricate morphotactic relations uncovered by the flow chart (3) and the structural and possibly even perceptual salience of the basic classes of bound morphemes.

BETA yields morphologically segmented strings as outputs, but does not perform higher-level syntactic or semantic analysis. It is obvious, however, that the internal state structure is equivalent to a syntactic description. Consider an infinitive form like sano=tt=ssa_{1,4}=ni₁ 'when I'm saying', where the

resulting state of each morph has been marked. RS '26' uniquely defines the 2nd infinitive, '14' the two cases inessive and genitive (Instructive) and '11' possessives other than Vn. In principle, one could supply the states with a semantics stating the grammatical category and/or meaning of the morph in question. Another way of developing a syntactic parser for Finnish would be to construct a system of syntactic interpretation rules taking the segmented BETA output as input. Given morphologically segmented strings, quite a few of the morphs provide direct hints of syntactic structure, either because they are unambiguous (i.e. there are no grammatical homonyms) or because they can be identified using contextual information (neighboring morphs) even though there are homonyms.

Of the 192 morphs included in our analysis, some 40% are unambiguous (cf. Brodda & Karlsson (forthcoming) for details). These include the clitic kin 'even', the possessive ni 'my', the inessive case ssa ~ ssä 'in', the mood isi 'conditional', the participle tu ~ ty, etc. Once these morphs have been segmented, there are no problems in identifying their function and meaning.

Most of the morphs are ambiguous with one or several other morphs: e.g. Vn can be the illative case (or rather one set of allomorphs of the illative morpheme), a 3rd person possessive, or a finite passive person ending; mne can be a 1st person possessive or a 1st person finite verbal ending, etc. But many of these homonyms can in fact be resolved by simple interpretation rules utilizing available information on the morph context. Consider Vn: it is the passive person ending when it occurs to the right of the passive stem-forming morph (e.g. sano=tt=i=ni), it is interpretable as a possessive only when occurring to the right of a case ending terminating in a vowel (e.g. taloss=an), and it is a representative of the illative case only when it directly follows the non-inflected stem (taloen).

It seems clear, then, that the parsing mechanisms that have to be developed if one wants to analyze Finnish syntax by automatic methods are somewhat different and perhaps simpler, in some sense, than those developed e.g. for English. This is due to the richness of Finnish surface morphology, or, more precisely, the relative transparency of Finnish morphosyntax.

REFERENCES

Brodda, Benny 1977. "BETA-systemet: En sammanfattning." Nordiska datalingvistikdagar 1977, Rapporter från Språkdata, Göteborgs universitet, Institutionen för språklig databehandling, 3, Göteborg 1977, pp.20-26.

----- 1979. Något om de svenska ordens fonotax och morfotax: Iakttagelser med utgångspunkt från experiment med automatisk morfologisk analys. Papers from the Institute of Linguistics, University of Stockholm, 38, Stockholm 1979.

----- (forthcoming) The BETA System. Institute of Linguistics, University of Stockholm.

Brodda, Benny & Fred Karlsson (forthcoming) Automatic analysis of Finnish word forms. Institute of Linguistics, University of Stockholm.

Clark, Herbert H. & Eve V. Clark 1977. Psychology and language. Harcourt Brace Jovanovich, New York.

Karlsson, Fred 1976. Finskans struktur. Liberförlaget, Lund.

----- 1977. "Ergisid morfologian teorian ajankohtaisista ongelmista." Sanaajalka 19, pp.26-56.

----- 1979. Finsk grammatik.² Suomalaisen Kirjallisuuden Seura, Helsinki.

Stobin, Dan 1966. "Grammatical transformations and sentence comprehension in childhood and adulthood." Journal of Verbal Learning and Verbal Behavior 5, pp.219-227.

Jarmo Korhonen

DEPENDENZSYNTAX UND AUTOMATISCHE DATENVERARBEITUNG

1. Zur Kodierung von Strukturmodellen

Zu den wichtigsten Aufgaben der Dependenzsyntax gehören die Ermittlung der syntaktischen Valenzbeziehungen des Satzes und die Erarbeitung von abstrakten Strukturmodellen verschiedenen Umfangs, die auf diesen Beziehungen beruhen. Werden in den Sätzen die Valenz des Prädikats mit einer finiten Verbform als unentbehrlichem Bestandteil, die Valenz der infiniten Verbformen und die des prädikativen Adjektivs untersucht, können entsprechend finite, infinite und adjektivische Strukturmodelle unterschieden werden. Die finiten Modelle, die in der Regel "Satzmodelle" genannt werden, lassen sich nach dem Genus des Verbs in aktive und passive Modelle einteilen. Eine ähnliche Einteilung ist auch bei den infiniten Strukturmodellen möglich, wobei entscheidend ist, ob die als Valenzträger fungierende infinite Verbkonstruktion im Aktiv oder Passiv steht. Die Strukturmodelle können ausserdem in Haupt- und Nebenmodelle untergliedert werden: Zu den ersteren werden die finiten Modelle, zu den letzteren neben den infiniten Modellen die adjektivischen Modelle gezählt, für die die Einteilung in aktive und passive irrelevant ist.

Trotz der drei verschiedenen Valenzträgergruppen sind für die Kodierung der Satzmodelle nicht mehrere Verfahrenswesen nötig, sondern dafür eignet sich ein einheitliches Prinzip. Gemäss diesem Prinzip werden für die konstitutiven Glieder der Modelle, d. h. für die valenzbedingten Bestimmungen oder Ergänzungen, zwei Kodezahlen verwendet. Die eine Zahl bezieht sich auf die Hauptklasse, die die Anzahl der Ergänzungen und die Art des Modells im Sinne der Unterscheidung von Aktiv und Passiv zum Ausdruck bringt, die andere auf die Unterklasse, die die morphosyntaktische Form und Funktion der Ergänzungen ausdrückt. Die finiten und infiniten Modelle werden im Hinblick auf Aktiv und Passiv so abgegrenzt, dass für die eine Art der Modelle