AN EXPERIMENT WITH HEURISTIC PARSING OF SWEDISH

Benny Brodda
Inst. of Linguistics
University of Stockholm
S-106 91 Stockholm SWEDEN

ABSTRACT

Heuristic parsing is the art of doing parsing in a haphazard and seemingly careless manner but in such a way that the outcome is still "good", at least from a statistical point of view, or, hopefully, even from a more absolute point of view. The idea is to find strategic shortcuts derived from guesses about the structure of a sentence based on scanty observations of linguistic units in the sentence. If the guess comes out right much parsing time can be saved, and if it does not, many subobservations may still be valid for revised guesses. In the (very preliminary) experiment reported here the main idea is to make use of (combinations of) surface phenomena as much as possible as the base for the prediction of the structure as a whole. In the parser to be developed along the lines sketched in this report main stress is put on arriving at independently working, parallel recognition procedures.

The work reported here is both aimed at simulating certain aspects of human language perception and at arriving at effective algorithms for actual parsing of running text. There is, indeed, a great need for fast such algorithms, e.g. for the analysis of the literally millions of words of running text that already today comprise the data bases in various large information retrieval systems, and which can be expected to expand several orders of magnitude both in importance and in size in the foreseeable future.

I BACKGROUND

The general idea behind the system for heuristic parsing now being developed at our group in Stockholm dates more than 15 years back, when I was making an investigation (together with Hans Karlgren, Stockholm) of the possibilities of using computers for information retrieval purposes for the Swedish Governmental Board for Rationalization (Statskontoret). In the course of this investigation we performed some psycholinguistic experiments aimed at finding out to what extent surface markers, such as endings, prepositions, conjunctions and other (bound) elements from typically closed categories of linguistic units, could serve as a base for a syntactic analysis of sentences. We sampled a couple of texts more or less at random and prepared them in such a way that stems of nouns, adjectives and (main) verbs - these categories being thought of as the main carriers of semantic information - were substituted for by a mere "-", whereas other formatives were left in their original shape and place. These transformed texts were presented to subjects who were asked to fill in the gaps in such a way that the texts thus obtained were to be both syntactically correct and reasonably coherent.

The result of the experiment was rather astonishing. It turned out that not only were the syntactic structures mainly restored, in some few cases also the original content was reestablished, almost word by word. (It was beyond any possibility that the subjects could have had access to the original text.) Even in those cases when the text itself was not restored to this remarkable extent, the stylistic value of the various texts was almost invariably reestablished; an originally lively, narrative story came out as a lively, narrative story , and a piece of rather dull, factual text (from a school text book on sociology) invariably came out as dull, factual prose.

This experiment showed quite clearly that at least for Swedish the information contained in the combinations of surface markers to a remarkably high degree reflects the syntactic structure of the original text; in almost all cases also the stylistic value and in some few cases even the semantic content was kept. (The extent to which this is true is probably language dependent; Swedish is rather rich in morphology, and this property is certainly a contributing factor for an experiment of this type to come out successful to the extent it actually did.)

This type of experiment has since then been repeated many times by many scholars; in fact, it is one of the standard ways to demonstrate the concept of redundancy in texts. But there are several other important conclusions one could draw from this type of experiments. First of all, of course, the obvious conclusion that surface signals do carry a lot of information about the structure of sentences, probably much more than one has been inclined to think, and, consequently, it could be worth while to try to capture that information in some kind of automatic analysis system. This is the practical side of it. But there is more to it. One must ask the question why a language like Swedish is like this. What are the theoretical implications?

Much interest has been devoted in later years to theories (and speculations) about human per-

ception of linguistic stimuli, and I do not think that one speculates too much if one assumes that surface markers of the type that appeared in the described experiment together constitute important clues concerning the gross syntactic structure of sentences (or utterances), clues that are probably much less consiously perceived than, e.g., the actual words in the sentences/utterances. To the extent that such clues are actually perceived they are obviously perceived simultaneously with, i.e. in parallel with, other units (words, for instance).

The above way of looking upon perception as a set of independently operating processes is, of course, more or less generally accepted nowadays (cf., e.g., Lindsay-Norman 1977), and it is also generally accepted in computational linguistics that any program that aims at simulating perception in one way or other must have features that simulates (or, even better, actually performs) parallel processing, and the analysis system to be described below has much emphasis on exactly this feature.

Another common saying nowadays when discussing parsing techniques is that one should try to incorporate "heuristic devices" (cf., e.g., the many subreports related to the big ARPA-project concerning Speech Recognition and Understanding 1970-76), although there does not seem to exist a very precise consensus of what exactly that would mean. (In mathematics the term has been traditionally used to refer to informal reasoning, especially when used in classroom situations. In a famous study the hungarian mathematician Polya, 1945 put forth the thesis that heuristics is one of the most important psychological driving mechanisms behind mathematical - or scientific - progress. In AI-literature it is often used to refer to shortcut search methods in semantic networks/spaces; c.f. Lenat, 1982).

One reason for trying to adopt some kind of heuristic device in the analysis procedures is that one for mathematical reasons knows that ordinary, "careful", parsing algorithms inherently seem to refuse to work in real time (i.e. in linear time), whereas human beings, on the whole, seem to be able to do exactly that (i.e. perceive sentences or utterances simultaneously with their production). Parallel processing may partly be an answer to that dilemma, but still, any process that claims to actually simulate some part of human perception must in some way or other simulate the remarkable abilities human beings have in grasping complex patterns ("gestalts") seemingly in one single operation.

Ordinary, careful, parsing algorithms are often organized according to some general principle such as "top-down", "bottom-to-top", "breadth first", "depth first", etc., these headings referring to some specified type of "strategy". The heuristic model we are trying to work out has no such preconceived strategy built into it. Our philosophy is instead rather anarchistic (The Heuristic Principle): Whatever linguistic unit that can be identified at whatever stage of the analysis, according to whatever means there are, is identified, and the significance of the fact that the unit in question has been identified is made use of in all subsequent stages of the analysis. At any time one must be prepared to reconsider an already established analysis of a unit on the ground that evidence against the analysis may successively accumulate due to what analyses other units arrive at.

In next section we give a brief description of the analysis system for Swedish that is now under development at our group in Stockholm. As has been said, much effort is spent on trying to make use of surface signals as much as possible. Not that we believe that surface signals play a more important role than any other type of linguistic signals, but rather that we think it is important to try to optimize each single subprocess (in a parallel system) as much as possible, and, as said, it might be worth while to look careful into this level, because the importance of surface signals might have been underestimated in previous research. Our experiments so far seem to indicate that they constitute excellent units to base heuristic guesses on. Another reason for concentrating our efforts on this level is that it takes time and requires much hard computational work to get such an anarchistic system to really work, and this surface level is reasonably simple to handle.

II AN OUTLINE OF AN ANALYZER BASED ON THE HEURISTIC PRINCIPLE

Figure 1 below shows the general outline of the system. Each of the various boxes (or subboxes) represents one specific process, usually a complete computer program in itself, or, in some cases, independent processes within a program. The big "container", labelled "The Pool", contains both the linguistic material as well as the current analysis of it. Each program or process looks into the Pool for things "it" can recognize, and when the process finds anything it is trained to recognize, it adds its observation to the material in the Pool. This added material may (hopefully) help other processes in recognizing what they are trained to recognize, which in its turn may again help the first process to recognize more of "its" units. And so on.

The system is now under development and during this build-up phase each process is, as was said above, essentially a complete, stand-alone module, and the Pool exists simply as successively updated text files on a disc storage. At the moment some programs presuppose that other programs have already been run, but this state of affairs will be valid just during this build-up phase. At the end of the build-up phase each program shall be able to run completely independent of any other program in the system and in arbitrary order relative to the others (but, of course, usually perform better if more information is available in the Pool).

In the second phase superordinated control programs are to be implemented. These programs will function as "traffic rules" and via these systems one shall be able to test various strategies, i.e. to test which relative order between the different subsystems that yields optimal result in some kind of "performance metric", some evaluation procedure that takes both speed and quality into account.

The programs/processes shown in Figure 1 all represent rather straightforward Finite State Pattern Matching (FS/PM) procedures. It is rather trivial to show mathematically that a set of interacting FS/PM procedures of the type used in our system together will yield a system that formally has the power of a CF-parser; in practice it will yield a system that in some sense is stronger, at least from the point of view of convenience. Congruence and similar phenomena will be reduced to simple local observations. Transformational variants of sentences will be recognized directly - there will be no need for performing some kind of backward transformational operations. (In this respect a system like this will resemble Gazdar's grammar concept; Gazdar 1980.)

The control structures later to be superimposed on the interacting FS/PM systems will also be of a Finite State type. A system of the type then obtained - a system of independent Finite State Automatons controlled by another Finite State Automaton - will in principle have rather complex mathematical properties. It is, e.g., rather easy to see that such a system has stronger capacity than a Type 2 device, but it will not have the power of a full Type 1 system.

- - -

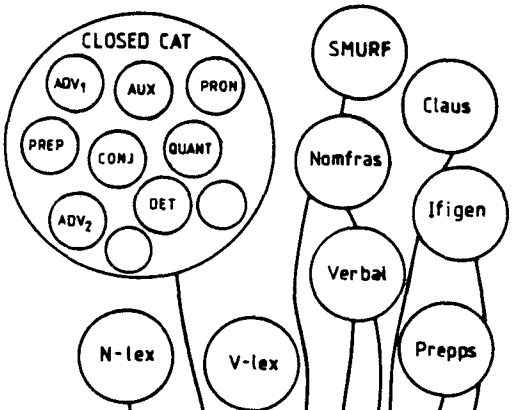
Now a few comments to Figure 1

The "balloons" in the figure represent independent programs (later to be developed into independent processes inside one "big" program). The figure displays those programs that so far (January 1983) have been implemented and tested (to some extent). Other programs will successively be entered into the system.

The big balloon labelled "The Closed Cat" represents a program that recognizes closed word classes such as prepositions, conjunctions, pronouns, auxiliaries, and so on. The Closed Cat recognizes full word forms directly. The SMURF balloon represents the morphological component (SMURF = "Swedish Murphology"). SMURF itself is organized internally as a complex system of independently operating "demons" - SMURFs - each knowing "its" little corner of Swedish word formation. (The name of the program is an allusion to the popular comic strip leprechauns "les Schtroumpfs", which in Swedish are called "smurfar".) Thus there is one little smurf recognizing derivational morphemes, one recognizing flectional endings, and so on. One special smurf, Phonotax, has an important controlling function - every other smurf must always consult Phonotax before identifying one of "its" (potential) formatives; the word minus this formative must still be pronounceable, otherwise it cannot be a formative. SMURF works entirely without stem lexicon; it adheres completely to the "philosophy" of using surface signals as far as possible.

NOMFRAS, VERBAL, IFIGEN, CLAUS and PREPPS are other "demons" that recognize different phrases or word groups within sentences, viz. noun phrases, verbal complexes, infinitival constructions, clauses and prepositional phrases, respectively. N-lex, V-lex and A-lex represent various (sub)-lexicons; so far we have tried to do without them as far as possible. One should observe that stem lexicons are no prerequisites for the system to work, adding them only enhances its performance.

The format of the material inside the Pool is the original text, plus appropriate "labelled brackets" enclosing words, word groups, phrases and so on. In this way, the form of representation is consistent throughout, no matter how many different types of analyses have been applied to it. Thus, various people can join our group and write their own "demons" in whatever language they prefer, as long as they can take sentences in text format, be reasonably tolerant to what types of "brackets" they find in there, do their analysis, add their own brackets (in the specified format), and put the result back into the Pool.



Of the various programs SMURF, NOMFRAS and IFIGEN are extensively tested (and, of course, The Closed Cat, which is a simple lexical lookup system), and various examples of analyses of these programs will be demonstrated in the next section. We hope to arrive at a crucial station in this project during 1983, when CLAUS has been more thoroughly tested. If CLAUS performs the way we hope (and preliminary tests indicate that it will), we will have means to identify very quickly the clausal structures of the sentences in an arbitrary running text, thus having a firm base for entering higher hierarchies in the syntactic domains.

The programs are written in the Beta language developed by the present author; c.f. Brodda-Karlsson, 1980, and Brodda, 1983, forthcoming. Of the actual programs in the system, SMURF was developed and extensively tested by B.B. during 1977-79 (Brodda, 1979), whereas the others are (being) developed by B.B. and/or Gunnel Källgren, Stockholm (mostly "and").

III EXPLODING SOME OF THE BALLOONS

When a "fresh" text is entered into The Pool it first passes through a preliminary one-pass-program, INIT, (not shown in Fig. 1) that "normalizes" the text. The original text may be of any type as long as it is regularly typed Swedish. INIT transforms the text so that each graphic sentence will make up exactly one physical record. (Except in poetry, physical records, i.e. lines, usually are of marginal linguistic interest.) Paragraph ends will be represented by empty records. Periods used to indicate abbreviations are just taken away and the abbreviation itself is contracted to one graphic word, if necessary; thus "t.ex." ("e.g.") is transformed into "tex", and so on. Otherwise, periods, commas, question marks and other typographic characters are provided with preceding blanks. Through this each word is guaranteed to be surrounded by blanks, and delimiters like commas, periods and so on are guaranteed to signal their "normal" textual functions. Each record is also ended by a sentence delimiter (preceded by a blank). Some manual post-editing is sometimes needed in order to get the text normalized according to the above. In the INIT-phase no linguistic analysis whatsoever is introduced (other than into what appears to be orthographic sentences).

INIT also changes all letters in the original text to their corresponding upper case variants. (Originally capital letters are optionally provided with a prefixed "=".) All subsequent analysis programs add their analyses in the form of lower case letters or letter combinations. Thus upper case letters or words will belong to the object language, and lower case letters or letter combinations will signal meta-language information. In this way, strictly text (ASCII) format can be kept for the text as well as for the various stages of its analysis; the "philosophy" to use text input and text output for all programs involved represents the computational solution to the problem of how to make it possible for each process to work independently of all other in the system.

The Closed Cat (CC) has the important role to mark words belonging to some well defined closed categories of words. This program makes no internal analysis of the words, and only takes full words into account. CC makes use of simple rewrite rules of the type "PÅ => ePÅe / (blank)__(blank)", where the inserted e's represent the "analysis" ("e" stands for "preposition"; PÅ = "on"). A sample output from The Closed Cat is shown in illustration 2, where the various meta-symbols also are explained.

The simple example above also shows the format of inserted meta-information. Each identified constituent is "tagged" with surrounding lower case letters, which then can be conceived of as labelled brackets. This format is used throughout the system, also for complex constituents. Thus the nominal phrase "DEN LILLA FLICKAN" ("the little girl") will be tagged as "nDEN+LILLA+FLICKANn" by NOMFRAS (cf. below; the pluses are inserted to make the constituent one continuous string). We have reserved the letters **n**, **v** and **s** for the major categories nouns or noun phrases, verbs or verbal groups, and sentences, respectively, whereas other more or less transparent letters are used for other categories. (A list of used category symbols is presented in the Appendix: Printout Illustrations.)

The program SWEMRF (or SMURF, as it is called here) has been extensively described elsewhere (Brodda, 1979). It makes a rather intricate morphological analysis word-by-word in running text (i.e. SMURF analyzes each word in itself, disregarding the context it appears in). SMURF can be run in two modes, in "segmentation" mode and "analysis" mode. In its segmentation mode SMURF simply strips off the possible affixes from each word; it makes no use of any stem lexicon. (The affixes it recognizes are common prefixes, suffixes - i.e. derivational morphemes - and flexional endings.) In analysis mode it also tries to make an optimal guess of the word class of the word under inspection, based on what (combinations of) word formation elements it finds in the word. SMURF in itself is organized entirely according to the heuristic principles as they are conceived here, i.e. as a set of independently operating processes that interactively work on each others output. The SMURF system has been the test bench for testing out the methods now being used throughout the entire Heuristic Parsing Project.

In its segmentation mode SMURF functions formally as a set of interactive transformations, where the structural changes happen to be extremely simple, viz. simple segmentation rules of the type "P=>P-", "S=> -S" and "E=>-E" for an arbitrary Prefix, Suffix and Ending, respectively, but where the "job" essentially consists of establishing the corresponding structural descriptions. These are shown in Ill. 1, below, together with sample analyses. It should be noted that phonotactic constraints play a central role

in the SMURF system; in fact, one of the main objectives in designing the SMURF system was to find out how much information actually was carried by the phonotactic component in Swedish. (It turned out to be quite much; cf. Brodda 1979. This probably holds for other Germanic languages as well, which all have a rather elaborated phonotaxis.)

NOMFRAS is the next program to be commented on. The present version recognizes structures of the type

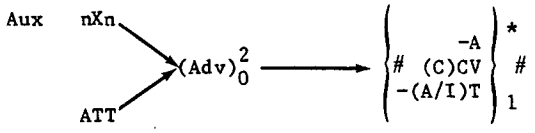
$$\text{det/quant} + (\text{adj})_0^n + \text{noun};$$

where the "det/quant" categories (i.e. determiners or quantifiers) are defined explicitly through enumeration - they are supposed to belong to the class of "surface markers" and are as such identified by The Closed Cat. Adjectives and nouns on the other hand are identified solely on the ground of their "cadences", i.e. what kind of (formally) ending-like strings they happen to end with. The number of adjectives that are accepted (n in the formula above) varies depending on what (probable) type of construction is under inspection. In indefinite noun phrases the substantial content of the expected endings is, to say the least, meager, as both nouns and adjectives in many situations only have 0-endings. In definite noun phrases the noun mostly - but not always - has a more substantial and recognizable ending and all intervening adjectives have either the cadence -A or a cadence from a small but characteristic set. In a (supposed) definite noun phrase all words ending in any of the mentioned cadences are assumed to be adjectives, but in (supposed) indefinite noun phrases not more than one adjective is assumed unless other types of morphological support are present.

The Finite State Scheme behind NOMFRAS is presented in Ill. 2, together with sample outputs; in this case the text has been preprocessed by The Closed Cat, and it appears that these two programs in cooperation are able to recognize noun phrases of the discussed type correctly to well over 95% in running text (at a speed of about 5 sentences per second, CPU-time); the errors were shared about 50% each between over- and undergenerations. Preliminary experiments aiming at including also SMURF and PREPPS (Prepositional Phrases) seem to indicate that about the same recall and precision rate could be kept for arbitrary types of (non-sentential) noun phrases (cf. Ill. 6). (The systems are not yet trimmed to the extent that they can be operatively run together.)

IFIGEN (Infinitive Generator) is another rather straightforward Finite State Pattern Matcher (developed by Gunnel Källgren). It recognizes (groups of) nonfinite verbs. Somewhat simplified it can be represented by the following diagram (remember the conventions for upper and lower case):

IFIGEN parsing diagram (simplified):



where "Aux" and "Adv" are categories recognized by The Closed Cat (tagged "g" and "a", respectively), and "nXn" are structures recognized by either NOMFRAS or, in the case of personal pronouns, by CC. (It should be worth mentioning that the class of auxiliaries in Swedish is more open than the corresponding word class in English; besides the "ordinary" VARA ("to be"), HA ("to have") and the modals, there is a fuzzy class of semi-auxiliaries like BÖRJA ("begin") and others; IFIGEN makes use of about 20 of these in the present version.) The supine cadence **-(A/I)^T** is supposed to appear only once in an infinitival group. A sample output of IFIGEN is given in Ill. 3. Also for IFIGEN we have reached a recognition level around 95%, which, again, is rather astonishing, considering how little information actually is made use of in the system.

The IFIGEN case illustrates very clearly one of the central points in our heuristic approach, namely the following: The information that a word has a specific cadence, in this case the cadence **-A**, is usually of very little significance in itself in Swedish. Certainly it is a typical infinitival cadence (at least 90% of all infinitives in Swedish have it), but on the other hand, it is certainly a very typical cadence for other types of words as well: FLICKA (noun), HELA (adjective), DENNA/DETTA/DESSA (determiners or pronouns) and so on, and these other types are by no comparison the dominant group having this specific cadence in running text. But, in connection with an "infinitive warner" - an auxiliary, or the word ATT - the situation changes dramatically. This can be demonstrated by the following figures: In running text words having the cadance -A represents infinitives in about 30% of the cases. ATT is an infinitive marker (equivalent to "to") in quite exactly 50% of its occurences (the other 50% it is a subordinating conjunction). The conditional probability that the configuration ATT ..-A represents an infinitve is, however, greater than 99%, provided that characteristic cadences like -ARNA/-ORNA and quantifiers/determiners like ALLA and DESSA are disregarded (In our system they are marked by SMURF and The Closed Cat, respectively, and thereby "saved" from being classified as infinitives.) Given this, there is almost no overgeneration in IFIGEN, but Swedish allows for split infinitives to some extent. Quite much material can be put in between the infinitive warner and the infinitive, and this gives rise to some undergeneration (presently). (Similar observations regarding conditional probabilities in configurations of linguistic units has been made by Mats Eeg-Olofson, Lund, 1982).

IV REFERENCES

Brodda, B. "Något om de svenska ordens fonotax och morfotax", Papers from the Institute of Linguistics (PILUS) No. 38, University of Stockholm, 1979.

Brodda, B. "Yttre kriterier för igenkänning av sammansättningar" in Saari, M. and Tandefelt, M. (eds.) Förhandlingar rörande svenskans beskrivning - Hanaholmen 1981, Meddelanden från Institutionen för Nordiska Språk, Helsingfors Universitet, 1981

Brodda, B. "The BETA System, and some Applications", Data Linguistica, Gothenburg, 1983 (forthcoming).

Brodda, B. and Karlsson, F. "An experiment with Automatic Morphological Analysis of Finnish", Publications No. 7, Dept. of Linguistics, University of Helsinki, 1981.

Gazdar, G. "Phrase Structure" in Jacobson, P. and Pullam G. (eds.), Nature of Syntactic Representation, Reidel, 1982

Lenat, D.P. "The Nature of Heuristics", Artificial Intelligence, Vol 19(2), 1982.

Eeg-Olofsson, M. "En språkstatistisk modell för ordklassmärkning i löpande text" in Källgren, G. (ed.) TAGGNING, Föredrag från 3:e svenska kollokviet i språklig databehandling i maj 1982, PILUS 47, Stockholm 1982.

Polya, G. "How to Solve it", Princeton University Press, 1945. Also Doubleday Anchor Press, New York, N.Y. (several editions).

APPENDIX: Some computer illustrations

The following three pages illustrate some of the parsing diagrams used in the system: Ill. 1, SMURF, and Ill. 2, NOMFRAS, together with sample analyses. IFIGEN is represented by sample analyses (Ill. 3; the diagram is given in the text). The samples are all taken from running text analysis (from a novel by Ivar Lo-Johansson), and "pruned" only in the way that trivial, recurrent examples are omitted. Some typical erroneous analyses are also shown (prefixed by **).

In Ill. 1 SMURF is run in segmentation mode only, and the existing tags are inserted by the Closed Cat. ^A and ^E in word final position indicates the corresponding cadences (fullfilling the pattern "..V^M^A/E", where M denotes a set of admissible medial clusters).

The tags inserted by CC are: a=(sentence) adverbials, b=particles, d=determiners, e=prepositions, g=auxiliaries, h=(forms of) HA(VA), i=infinitives, j=adjectives, n=nouns, K=conjunctions, q=quantifiers, r=pronouns, u=supine verb form, v=verbal (group).

(For space reasons, Ill. 3 is given first, then I and II.)

Ill. 3: PATTERN: aux/ATT^(pron)^(adv)^(adv)^inf^inf^...:

```
..PLOCKNINGEN eEFTER..ikATTk+iHAi+uGÅTTui FRAMÅTBÖJD HELA DAGEN..
  ..  rDETr vVARv ORIMLIGT ikATTk+iFINNAi qETTq KÄDSTRECK eIe ..
             rJAGr gSKAg aBARAa iHJÄLPAi eTILLe ikATTk+iSEi ..
                   - rDETr gKANg iLIGGAi qENq KARL INUTI?
                       gSKAg rVIr iVÅGAi VIPPEN?
                 - rVIr gKANg aINTEa iGÅi HEM eMEDe SKAMMEN ...
...ORNA vHÖLLv SIG FÄRDIGA ikATTk+iKASTAi eOMe NÄRSOMHELST.
        rDEr gVÅGADEg aÄNTLIGENa iLYFTAi ePÅe rDETr.
       gSKAg rNIr aNÖDVÄNDIGTVISa iGÖRAi NÅT eMEDe rDENr, kSÅk
..rVIr hHADEh aÄNNUa aINTEa uHUNNITu iFÅi eUPPe POTATISEN.
..BECKMÖRKRET eMEDe ikATTk+iFÖRSÖKAi+iFÅi BALLONGEN FYLLD.
eMEDe VÄTGAS eFÖRe ikATTk+iKUNNAi+iHÅLLAi SEJ OPPE.
          SKOGEN, LANDEN gTYCKTESg iSTÅi STILLA eUNDERe OSS.
 rDENr hHADEh MISSLYCKATS eIe ikATTk+iNÅi SITT MÅL.
  *** qENq ÅS gVÅGADEg iKVINNORNA+STANNAi .
```

Ill. 1: SMURF - PARSING DIAGRAM FOR SWEDISH MORPHOLOGY

PATTERNS ("Structural Descriptions"): Structural changes

1) ENDINGS (E): $X \cdot \begin{Bmatrix} /S \\ V \cdot M_e \end{Bmatrix} \cdot \underline{E} \# ;$ E => =E

2) PREFIXES (P): $\begin{Bmatrix} \# \\ \# P> \\ X \cdot V \cdot F\ (s) \end{Bmatrix} \cdot \underline{P} \cdot V \cdot X ;$ P => (-)P>

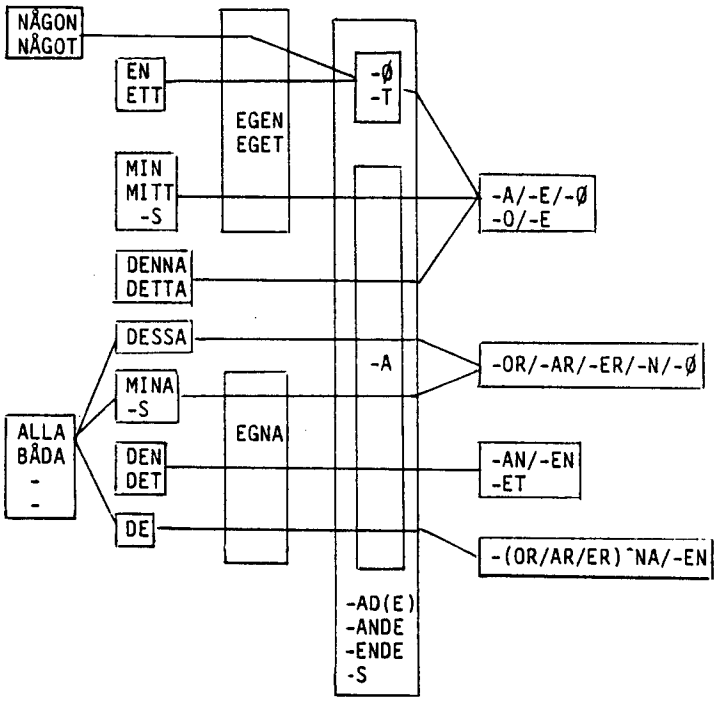
3) SUFFIXES (S): $X \cdot V \cdot F \cdot \underline{S} \cdot \begin{Bmatrix} (s) \cdot I \cdot V \cdot X \\ E\# \\ \# \end{Bmatrix};$ S => /S(-)

where I = (admissible) initial cluster, F = final cluster, M_e = morpheme internal cluster, V = vowel, (s) the "gluon" S (cf. TIDNINGSMAN), # = word boundary, (=,>,/,-) = earlier accepted affix segmentations, and ·, finallay, denotes ordinary concatenation. (It is the enhanced element in each pattern that is tested for its segmentability).

BÄGG´E vDROGv . REP=ET SLINGR=ADE MELLAN STEN=AR , FÖR>BI TALLSTAMM=AR , MELLAN RÖD´A LINGONTUV=OR eIe GRÖN IN>FATT/NING.
qETTq STORT FÖRE>MÅL hHADEh uRÖRTu ePÅe SIG BORT´A eIe SLÄNT=EN . FÖRE>MÅL=ET NÄRM=ADE SIG HOTFULL´T . dDETd KNASTR==ADE eIe SKOG=EN . - SPRING .
BÄGG´E SLÄPP=TE kOCHk vSPRANGv . rDEr LÅNG´A KJOL=ARNA VIRV1=ADE eÖVERe O<PLOCK=ADE LINGONTUV=OR . BÄGG´E KVINNO=RNA hHADEh STRUMPEBAND FÖR>FÄRDIG=ADE eAVe SOCKERTOPPSSNÖR=EN , KNUT=NA NEDAN>FÖR KNÄN´A .
aFÖRSTa bUPPEb ePÅe qENq ÅS VÅG=ADE KVINNO=RNA STANN´A . rDEr vSTODv kOCHk STRÄCK=TE ePÅe HALS=ARNA . qENq FRÄN UT>DUNST/NING eAVe SKRÄCK SIPPR=ADE bFRAMb . rDEr vHÖLLv BE>SVÄRJ/ANDE HÄND=ERNA FRAM>FÖR SIN´A SKÖT=EN .
- dDETd vSERv STORT kOCHk eRUNTe bUTb , vSAv dDENd KORT´A eOMe FÖRE>MÅL=ET . dDETd vÄRv aVÄLa aINTEa qNÅGOTq IN>UT>I ?
- dDETd gKANg LIGG´A qENq KARL IN>UT>I ? dDETd vVETv rMANr aVÄLa kVADk rHANr vGÖRv eMEDe OSS .
- rJAGr TYCK=TE dDETd RÖR=DE ePÅe SEJ . gSKAg rVIr iVÅGAi VIPP=EN ? - JA ? gSKAg rVIr iVÅGAi VIPP=EN ?
BÄGGE vSMÖGv SIG ePÅe GLAPP´A KNÄN UT>FÖR BRANT=EN . kNÄRk rDEr NÄRM=ADE SIG rDEr FLÄT=ADE POTATISKORG=ARNA eMEDe LINGON kSOMk vSTODv ePÅe LUT eVIDe VARSIN TUVA , vVARv rDEr aREDANa UT>OM SIG eAVe SKRÄCK . oDERASo SANS vVARv BORT´A .
- PASS ePÅe . rVIr KANHÄND´A aINTEa vTÖRSv NÄRM=ARE ? vSAv dDENd MAGR´A HUSTRUN .
- rVIr gKANg aINTEa GÅ HEM eMEDe SKAMM=EN aHELLERa . rVIr gMÅSTEg aJUa iHAi BÄRKORG=ARNA eMEDe .
- JAVISST , BÄRKORG=ARNA .
kMENk kNÄRk rDEr uKOMMITu bNERb eTILLe STÄLL=ET I<GEN uVARTu rDEr NYFIK=NA . rDEr vDROGSv eTILLe FÖRE>MÅL=ET eIe

Ill. 2: NOMFRAS - FS-DIAGRAM FOR SWEDISH NOUN PHRASE PARSING

quant + det + "OWN" + adj + noun




```
                - PYTT , vSAv nDEN+LÅNGAn .
                  kVADk vVARv NU nDET+DÄRn kATTk VARA RÄDD eFÖRe ?
   nDET+OMFÅNGSRIKA+,+SIDENLÄTTA+TYGETn .
              nDEn GJORDE nEN+STOR+PACKEn eAVe dDETd .
 eMEDe SIG SJÄLVA eOMe kATTk nDET+HELAn aINTEa uVARITu qETTq ..
.. nDET+HELAn aINTEa uVARITu nETT+DUGGn FARLIGT .
              nDET+FÖRMENTA+KLÄDSTRECKETn vVARv kDÅk SNOTT FLE..
 .. GRÖN eMEDe HÄNGBJÖRKAR kSOMk nALLAn FYLLDE FUNKTIONER .
.. MODERN , nDEN+LÅNGA+EGNAHEMSHUSTRUNn kSOMk uVARITu eIe SKO..
STORA BOKSTÄVER nETT+SVENSKT+FIRMANAMNn .
            ePÅe nDEN+ANDRA+,+FRÅNVÄNDAn , vSTODv ORDEN ..
    nDETn vVARv nEN+LUFTENS+SPILLFRUKTn kSOMk hHADEh uRAMLAT..
kOCHk nDEN+ANDRA+EGNAHEMSHUSTRUNS+ÖGONn VATTNADES eAVe ÖMSOM
                nETT+STORT+MOSSIGT+BERGn HÖJDE SIG eMOTe SKYN..
.. SIG eMOTe SKYN eMEDe nEN+DISIG+MÅNEn kSOMk qENq RUND LYKTA ..
                    eVIDe nDET+STÄLLEn kDÄRk LANDNINGSLINAN ..
  SÄGA HONOM kATTk nALLA+DESSA+FÖREMÅLn aÄNDÅa aINTEa FÖRMÅD..
..ARNA kSOMk nEN+AVIGT+SKRUBBANDE+HANDn .
                kSOMk nEN+OFORMLIG+MASSAn VÄLTRADE SIG BALLONG..
                    - nEN+RIKTIG+BALLONGn gSKAg VARA FYLLD eMEDe..
**nDETn aINTEa vLÅGv nNÅGON+KROPP+GÖMDn INUNDER .
*** TVÅ kSOMk BÄRGADE nDEN+TILLSAMMANSn .
```