NORDISKE DATALINGVISTIKDAGE

I

KØBENHAVN

9.-10. OKTOBER 1979

Foredrag udgivet af
Bente Maegaard

Institut for anvendt og matematisk lingvistik
Københavns Universitet
København 1979

Udgivet
med støtte fra
Nordisk Kulturfond

INDHOLD

FORORD

De Nordiske Datalingvistikdage 1979 blev afholdt på Københavns Universitet, Institut for anvendt og matematisk lingvistik, den 9. og 10. oktober. Som initiativtager til mødet stod den Nordiske Samarbejdsgruppe for datamaskinel sprogbehandling, og for det praktiske arrangement stod instituttet: demonstrationer: Steen Nielsen, sekretariatshjælp og frokost: Ethel Ussing og Lisbet Larsen, det øvrige: Bente Maegaard.

I mødet deltog ca. 75 deltagere fra Danmark, Island, Norge og Sverige, af hvilke 24 holdt foredrag. Nærværende publikation omfatter skrevne versioner af foredragene. Enkelte foredragsholdere har imidlertid ikke ønsket at indlevere en fuldstændig version; for disses vedkommende er det på forhånd udsendte oplæg medtaget. For ganske få foredrags vedkommende findes hverken oplæg eller fuldstændig version.

Foredragene er trykt i den orden, de forekom i programmet. Bagest i publikationen findes en alfabetisk liste over forfattere.

Ved mødet aftaltes det, at de næste nordiske datalingvistikdage afholdes i Trondheim 1981 og at Eirik Lien vil stå for arrangementet.

Institut for anvendt og
matematisk lingvistik,
Njalsgade 96,
2300 København S

November 1979
Bente Maegaard

NAVF'S EDB-SENTER FOR HUMANISTISK FORSKNING
VILLAVEI 10, POSTBOKS 53
5014 BERGEN-UNIVERSITETET

Oktober 1979

NORSK TEKSTARKIV

Jostein H. Hauge

1. FORHISTORIE

Datamaskinell språkbehandling er kanskje det feltet som har stått mest sentralt siden datamaskinene gjorde sitt inntog i de humanistiske fag på 60-tallet i Norge.

Ved Nordisk institutt, Universitetet i Bergen ble det tidlig bygget opp en avdeling, Prosjekt for datamaskinell språkbehandling, som særlig skulle arbeide med oppgaver av betydning for norsk språkvitenskap og norsk språklig utviklingsarbeid. Senere har NAVF's EDB-senter i Bergen og EDB-tjenestene for de filologiske fag ved universitetene i Oslo og Trondheim vært aktive i arbeidet med språklig databehandling.

NAVF's EDB-senter har arbeidet med flere større språkprosjekter i de siste årene. I tillegg til arbeid med et stort Ibsen-prosjekt er det foretatt revisjon og nyutgivelse på mikrokort av Brown Corpus, og vi avslutter i disse dager det tilsvarende Lancaster/Oslo/Bergen Corpus i samarbeid med dosent Stig Johansson, Britisk institutt, Universitetet i Oslo. Senteret er videre utøvende EDB-organ for organisasjonen International Computer Archive of Modern English. Dette har vi sett på som viktige og lærerike oppgaver, som også har fått stor internasjonal oppmerksomhet, men utførelsen har også kostet oss ganske mye i tid og penger.

I vårt styre har derfor spørsmålet vært reist om ikke tiden nå var moden til å avslutte servicearbeidet for engelsklingvistene og vende øynene mot vårt eget språk, som få andre enn nordmenn kan tenkes å interessere seg for. Når

styret påpekte dette, var det også fordi en generelt kan merke en stadig økende interesse for og et stigende behov for norske språkdata, særlig fra moderne tid.

2. KONFERANSEN OM ET NORSK DATAMASKINELT TEKSTKORPUS

For å få drøftet spørsmål i tilknytning til norske språkdata, inviterte NAVF's EDB-senter i oktober 1978 til en konferanse i Bergen om et norsk datamaskinelt tekstkorpus. På konferansen deltok 32 representater fra universitets- og høyskolesektoren og fra andre miljøer hvor en arbeider med norske språk. I tillegg til norske deltakere var det også invitert gjester fra Sverige og Danmark (Sture Allén, Rolf Gavare og Bente Maegaard). Hensikten med konferansen var i første rekke å få i stand en drøfting av behov for norske språkdata i datamaskinell form og de prinsipper en bør legge til grunn for arbeidet her. Det viktigste innslaget på konferansen kom til å bli plenumsdrøftingene, men forut for disse ble det gitt en serie innlegg om det arbeid med større tekstsamlinger som pågår i Danmark, Sverige og i vårt eget land. Det foreligger en omfattende konferanserapport som gir et detaljert innsyn i det som foregikk på denne konferansen (Et norsk datamaskinelt tekstkorpus. Rapport nr. 2. Februar 1979). I vår sammenheng skal bare følgende punkter nevnes:

1. Det var stor interesse for å få intensivert arbeidet med å skaffe fram, lagre og utnytte tekstmateriale fra moderne norsk i undervisning og forskning.
2. Det er mange ulike brukerinteresser knyttet til dette arbeidet.
3. Det var en stor skepsis mot å lage et norsk standard tekstkorpus.
4. Heller gikk drøftingene i retning av å planlegge et tiltak som hadde karakter av en språkbank. På konferansen, hvor mange knapt nok anerkjente verdien av de tradisjonelle korpus, mente de fleste at det ville være bedre i dag å satse på et tiltak som hadde som hovedmål å samle inn,

tilrettelegge og gjøre bruksklare tekster slik at brukerne selv kan velge ut de tekstene eller tekstdelene de ønsker for ulike spesialoppgaver.

5. Det ble vedtatt å nedsette en gruppe som kunne arbeide videre med de tankene som kom fram på konferansen.

3. PLANLEGGINGSGRUPPEN

Kort tid etter konferansen i Bergen kom det i sving en planleggingsgruppe som skulle arbeide videre med det grunnlagsmateriale som var fremskaffet gjennom konferansen og utforme et tiltak langs de retningslinjer som var trukket opp i plenumsdrøftingene. Følgende personer og institusjoner har vært med i planleggingen:

Jostein H. Hauge, NAVF's EDB-senter
Kolbjørn Heggstad, Nordisk institutt, PDS, UiB
Aagot Landfald, Norsk språkråd
Eirik Lien, EDB-tjenesten for humanistiske fag, Tr.heim
Egil Pettersen, Nordisk inst., UiB (formann)
Jarle Rønhovd, Nordisk inst., UiTrheim
Dagfinn Worren, Norsk leksikografisk institutt

Som en vil se av dette, la en vekt på allerede i utgangspunktet å markere dette tiltaket som et nasjonalt prosjekt og også å sikre at ulike brukerinteresser kom fram i planleggingsarbeidet. I løpet av noen møter høsten 1978 og våren 1979 greidde planleggingsgruppen å konkretisere et tiltak som ble kalt Norsk tekstarkiv.
De planene som planleggingsgruppen kom fram til ble sendt alle deltakerne fra Bergenskonferansen til uttalelse, og det viste seg at det ikke innkom noen merknader til de planene som var lagt fram. Dette velger planleggingsgruppen å tolke som tilslutning til planene, heller enn som manglende interesse for dem. Når vi tror at det er slik, skyldes det det heftige engasjement og den store interesse som ble vist denne saken på konferansen i Bergen.

4. NORSK TEKSTARKIV

Norsk tekstarkiv har som mål å koordinere og øke innsatsen i arbeidet med å samle inn og tilrettelegge tekstmateriale fra moderne norsk til bruk i forskningsog utviklingsarbeid. Tiltaket vil fra starten av bli nasjonalt orientert. En forutsetning for at Norsk tekstarkiv kan resultere i en vitenressurs om norsk språk, er at arbeidet med tekstinnsamling legges opp etter en nasjonal koordinert plan og at materialet tilrettelegges på en standard måte. De datamengder som legges opp, må kunne utnyttes datamaskinelt i alle interesserte miljøer med et minimum av ekstra tilretteleggingsarbeid. Hovedtyngden av de data som samles inn, vil referere seg til moderne norsk skjønnlitteratur og bruksprosa. Det sier seg selv at det her bare kan bli tale om å samle inn en meget liten del av alt som utgis. I størst mulig utstrekning er det derfor ønskelig å oppnå enighet på nasjonal basis om de prioriterte oppgaver. Det er tatt spesielt hensyn til dette ved oppbygging av organisasjonsstrukturen for Norsk tekstarkiv.

Selv om arbeidet med norsk datamaskinelt tekstmateriale bør ses i et langt perspektiv, er det naturlig i første omgang å definere en prosjektramme på 5 år for Norsk tekstarkiv. I løpet av denne perioden bør ulike sider av prosjektet kunne prøves ut. Vi tenker i første rekke på nødvendig metodeutvikling på EDB-siden og organisasjonsformer for samvirke mellom mange EDB-organer. Like viktig vil det bli å få kunnskap om hvilke typer tekstmateriale som bør prioriteres og finne fram til egnede presentasjons- og utnyttelsesformer av tekstdata. Innenfor dette tidsrom vil det bli mulig å gjennomføre ulike typer prosjekter knyttet til spesielle tekstsamlinger. Sist i prøveperioden blir hovedoppgaven å finne fram til permanente organisasjons- og finansieringsformer for et slikt tiltak.

5. ORGANISASJON

Grunnlaget for Norsk tekstarkiv (NT) vil være et formalisert samarbeid mellom PDS, Nordisk institutt, UiB og NAVF's EDB-senter i Bergen, som igjen har EDB-tjenestene ved

HF-fakultetene ved universitetene som sine faste samarbeidspartnere. Siden et tiltak som NT vil ha et betydelig innslag av dataadministrasjon og administrative rutiner generelt, la planleggingsgruppen vekt på å finne fram til et organisasjonsmønster som gir klare ansvarsforhold og også å foreta en markert funksjonsfordeling mellom de ulike medvirkende parter i prosjektet.

6. FAGLIG INNHOLD

Som tidligere nevnt er hovedmålet med NT å kunne intensivere arbeidet med å samle inn, tilrettelegge og presentere data for språkvitenskapelig forsknings- og utviklingsarbeid. Når NT kommer i gang, vil den første oppgaven være å fastlegge et standardformat for tekstmateriale som skal inngå i arkivet. Det vil dessuten være behov for å utvikle program for konvertering og justering til dette standardformatet. De fleste av oss vet at det foreligger en rekke typer utstyr i den grafiske industri, som NT vil forsøke å samarbeide med, og det trengs en god del systemerings- og programmeringsarbeid for å kunne utvikle konverteringsprogram for å ta de datamaskinlesbare tekstene som man kan få fra tykkerier, forlag, aviser etc. på hullbånd eller magnetbånd. Det bør også utvikles et datamaskinelt lagringssytem for tekstmateriale som er effektivt og som kan ta hånd om den stadig økende tekstmengde. Målet er å kunne ta vare på mest mulig informasjon som ligger i tekstene, enten i originalkilden eller i den versjon som den grafiske industri presenterer, men samtidig å ha tekstene slik organisert at hver enkelt bruker kan undertrykke informasjon som er lite interessant i den enkelte forskningsoppgave, og også legge til informasjon som er nødvendig for den aktuelle oppgave. En fordel med et standard dataformat vil være at brukerne av NT alltid vet hvordan de vanligste typer av tekster ser ut, og det vil også ha som fordel at de som arbeider med utviklingsoppgaver på EDB-siden, vil kunne legge til rette sin programutrustning for å kunne behandle data fra NT.

Ved fastsetting av vårt eget standardformat vil vi selvsagt skjele til det som er utført i andre land, ikke minst her i Danmark og i Sverige. Det er etablert en styringsgruppe som

skal lede arbeidet til en medarbeider som skal utgreie spørsmålet om standardformat, og vi håper til slutt å kunne stå med et framlegg som vi kan foreslå som norsk norm innenfor humanistisk databehandling i alle fall.

Det er til i dag ikke foretatt en omfattende vurdering av hvilken ende en skal starte i, dvs. hva slags type tekster en først skal forsøke å få inn i NT. Der er en del planer lansert, f.eks. at NT bl.a. burde ta sikte på å få med all ny norsk skjønnlitteratur eller et utvalg av denne hvert år. Denne oppgaven og mange andre ønsker som kom fram på konferansen i Bergen, må gjennomgås nøye før vi setter i gang. Det bør her nevnes at det i Norge allerede i dag foreligger en ikke ubetydelig del tekstmateriale som ligger tilrettelagt for datamaskinell behandling, i første rekke ved PDS, Universitetet i Bergen, noe ved NAVF's EDB-senter og en god del materiale spredd rundt om ved universitetene, materiale som har vært brukt i konkret prosjektarbeid. Et viktig mål, etter mitt syn, må i første omgang være å få samlet det mest verdifulle av dette materialet og lagt det til rette etter de prinsipper som er nevnt tidligere.

7. STATUS FOR NORSK TEKSTARKIV I DAG

Det er forutsetningen at NT skal komme i gang fra 1980 av. Noe støtte til tiltaket kan trolig gis fra grunnbevilgningen til PDS og NAVF's EDB-senter, men for å komme noe vei er det nødvendig å ha eget personale knyttet til NT. Planleggingsgruppen søkte derfor NAVF om støtte til tiltaket for 1980, bl.a. støtte til å opprette en hel stilling for en vitenskapelig assistent som kan spesialisere seg for oppgavene med NT. Denne søknaden ble innvilget ved Rådets budsjettbehandling i oktober i år. Dermed vil forholdene ligge godt til rette for etablering av Norsk tekstarkiv neste år.

8. ARBEIDET FREMOVER

Det vil derfor i høst bli aktuelt å innkalle det faglige råd for NT for å diskutere de bevilgninger som er gitt til NT, den organisasjonsplanen som er utviklet, og på den bakgrunn

forme planer om den konkrete oppstarting av tiltaket i vinter. Like viktig vil det være å komme fram til en faglig prioritering mellom de ulike arbeidsoppgaver som er stilt opp, d.v.s. velge ut hvilket tekstmateriale en skal ta fatt på i første omgang, og også finne fram til praktiske samarbeidsformer mellom samarbeidspartnere i Bergen, EDB-konsulentene ved universitetene og andre brukermiljøer som har spesielle interesser knyttet til NT, f.eks. Norsk språkråd og Norsk leksikografisk institutt. Det vil i løpet av høsten også bli tale om å forhandle med Norsk språkråd om betalte oppdrag i forbindelse med undersøkelser som Norsk språkråd ønsker utført vedrørende utviklingen av norsk språk slik den fremkommer i litterær prosa fra 1937 til i dag.

Peter Bøgh Andersen:

FANGORN - A LANGUAGE FOR GENERATING COHERENT TEXTS

1. INTRODUCTION AND GENERAL IDEAS

Fangorn is a system that reads descriptions of texts and generates samples of the texts described. It can be used for checking the empirical adequacy of text descriptions: if the output deviates (in some sense) from the corpus intended to be covered by the description, then the description is empirically inadequate.

Since many texts relate narratives about humans acting in purposeful, although conflicting, ways, Fangorn must contain facilities for describing problem-solving algorithms and it can be used for experiments in that area. However, the emphasis is not on efficiency but on simplicity, and it is strongly oriented towards producing readable texts as output. In these respects it deviates from systems such as TALE-SPIN (Meehan(77)).

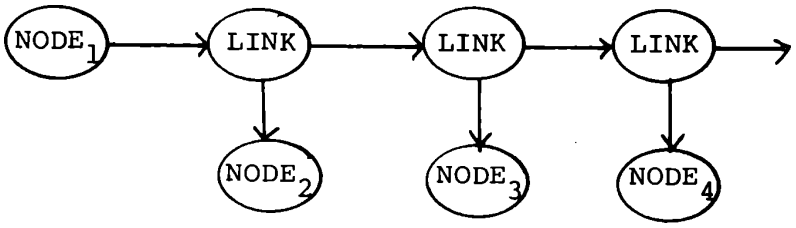
Fangorn is heavily influenced by SIMULA and to a lesser extent by LISP. It is being programmed in SIMULA, but I comtemplate reprogrammingit in a less expensive language when it is debugged.

A Fangorn program is written in F-expressions (akin to LISP S-expressions). They are translated into a connected labelled network with a least upper bound. The network may contain cycles. It consists of NODES linked together with LINKs. Every NODE has certain attributes (variables) and in addition contains a block of actions that are executed when the node is activated. The attributes of a node are called its structure and the actions are called its process. Objects containing a structure and a process are called aggregates. Every node in Fangorn is an aggregate, and every operation Fangorn can perform is a process, that is: it is associated with some node.

Every node has at least two attributes: a variable of type text, name, and a pointer variable, suc. LINKs have no name; instead they contain a pointer variable, val, pointing to a NODE.

The basic network of FANGORN looks like this:

Fig.1



$NODE_1$ is a mother of nodes $NODE_2$, $NODE_3$, $NODE_4$. The latter are daughters of $NODE_1$. Two nodes that are daughters of the same node are said to be sisters with respect to that node.

Note that a node may have several mothers, and that two nodes may be sisters with respect to one node, but not with respect to another one.

Fig.1 may be drawn in a slightly simplified way as fig.2 or fig.3:

Fig.2

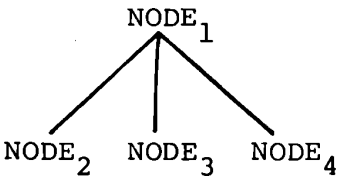
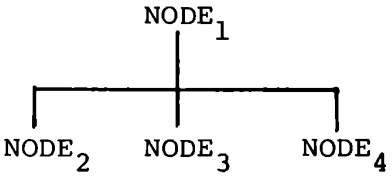


Fig.3



A major principle behind Fangorn is that a network must be able to reproduce itself, if it is to bear a likeness to natural language. We know that a natural language is thus structured that children, when exposed to it, learn it in an amazingly uniform way. We know too that no simple copying operation is involved (copying the contents of the adult brain into the child's brain or the like). Instead, the child's language is built up, stage by stage, and at each stage it is capable of functioning as a language. A grammar does not only produce sentences: it also reproduces itself at the very same time as it produces sentences,

and largely by means of the very same mechanisms it uses to produce those sentences.

For a "programming" language, this has the following consequence: the products it produces must be of the same kind as the program itself, and the operations by means of which it creates its output must be sufficient to create an output that is functionally equivalent to itself.

Of course, any programming language could be fixed to meet these requirements: if it contained a procedure run(file) that would compile and execute the program, written on file, then a program could write another program on file, possibly a copy of itself; run(file) would then compile and execute the program written on file. The difference between Fangorn and our hypothetical language is that the products of Fangorn are "programs" that may be executed without further ado, whereas our hypothetical program produces descriptions of programs that must be translated by very complicated processes before they can be executed.

Suppose that a parent network produced an offspring network in Fangorn: then the parent would be a very skillful educator of the infant, because the infant is structured as any other object that the parent can create and manipulate - the parent may use the same techniques it would use in any other situation when educating the child.

On the other hand, our hypothetical program would be a very poor educator: presumably, the child must be killed (the program must be terminated) and a new version written and compiled (born!) if changes are to be made. We just dont do such things nowadays!

If follows that a Fangorn program may change itself - it can educate itself. And this is obviously a desirable ability: because in many novels the protagonist changes during the narrative: for exampl his problem-solving algorithms may change as a result of successes, failures, new insights, or what have you.

These are the principial reasons why every bit of a Fangorn program is an energetic aggregate, ready to act when requested.

It has certain drawbacks, however. It makes recursive programming exetremely expensive, because every time a "procedure" is called, a copy of the whole procedure must be created. Copying the local variables of the procedure does not suffice, since the procedure may change its body during execution.

Fangorn is a forest in The Lord of the Rings by Tolkien, and since the program treats its network as a collection of trees, I thought that "Fangorn" fitted very well, the novel being one of my favourite books. But there is a little more to it than that: Fangorn is a very peculiar forest, consisting of trees, some of which are alive and move and act, and some of which are asleep and hard to wake. I have always been fascinated by Fangorn, because it contradicts the common idea of the world as consisting of two separate phenomena: things that are dead and passive, and beings that are alive and active. This conception simply does not fit language: a text, for example, is a thing: but it is also a process, influencing the reader in complicated ways, and leading him to conclusions that he may be most unwilling to draw.

In many programming languages, the passive-active dichotomy emerges as the rigid data-statement division, which may be a useful distinction in some areas, but is extremely cumbersome when natural language is concerned. I have sometimes wondered whether this stubborn insistence on the passive-active dichotomy might not be due to an underlying powerful ideology that classifies everything as either being passive, subordinate and willing to undergo manipulation, or as being capable of and entitled to doing the manipulation, with nothing in between. But such a division does not accord with the facts, even if it accords with the wishful thinking of the present potentates.

However that may be, in Fangorn I have tried to obliterate the distinction as far as possible, insisting that every action is performed by some entity that can itself be acted upon, and conversely, that every entity has at least a rudimentary action potential.

2. FLOW OF CONTROL

When a node is activated, it executes its process and then activates one of its daughters defined by its structure. Most nodes conform to the following pattern: they have at most 3 possible outcomes, success, failure and dont know, which correspond to its 3 rightmost daughters in that order. For example, the node BELIEVED may have 3 results: true, false or dont know. A person may believe a sentence, he may believe its negation or he may just dont know. If he believes it, the last-but-two sister is activated, if he disbelieves the last-but-one sister is activated, and if he dont know the last sister is activated.

Some nodes may have only one outcome, succes. For example, the node SET has 3 daughters; it assigns the daughters of its second daughter to its first daughter, and then activates its third daughter, thereby corresponding to the assignment statement.

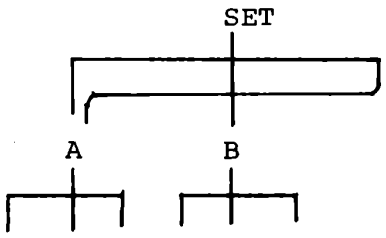
A node may have operands; they are always the youngest daughters. SET has two operands, its first two daughters.

Observe, that the format

(operands) + succes + (failure + (dontknow))

does not prevent a node from playing more than one role at a time. I fig.4, A plays the role of operand and at the same time functions as the succes-node:

Fig.4



The daughters of A are replaced by the daughters of B, and lastly A is activated.

3. MATCHING AND ASSIMILATING NETWORKS

Let A be a node. The network consisting of all nodes accessible from A via mother-daughter relations is called the network defined by A (or dominated by A). Instead of the phrase "the network dominated by A" I will sometimes write just "A" when no confusion results.

The network defined by A is said to match the network defined by B iff there is an isomorphism L from A into a subset of B, preserving

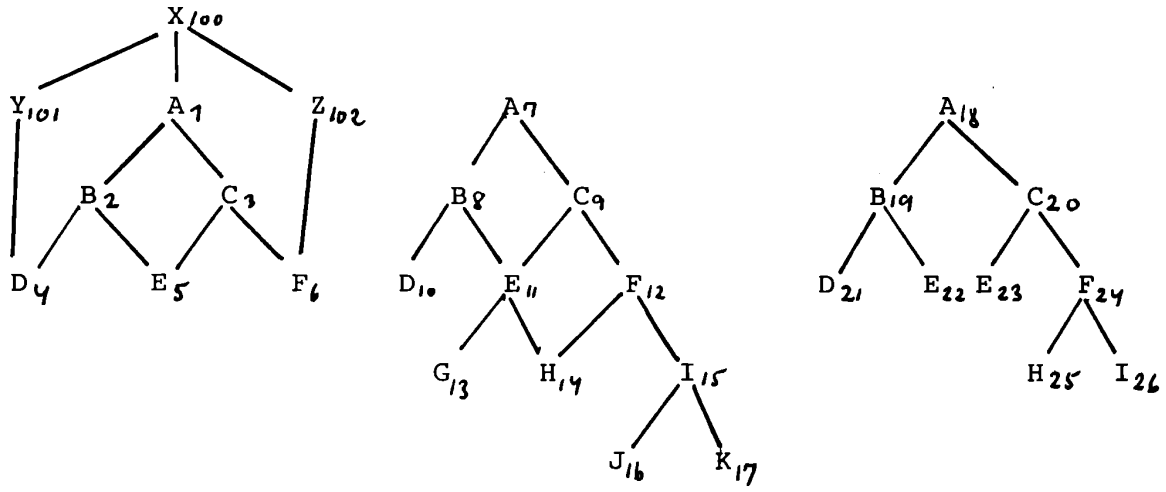
1. names
2. mother-daughter relations
3. sister relations

and such that

4. L(A) = B

Thus, node 1 matches node 7 but not node 18:

Fig.5 Fig.6 Fig.7



because there is an isomorphism L from 1 into 7,

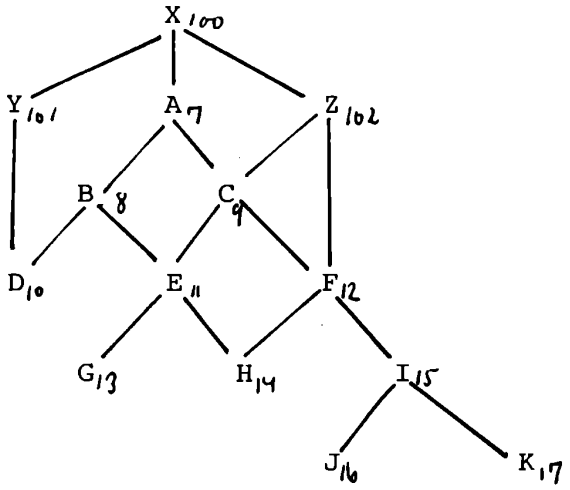
Fig.8

$$L = \{(1,7),(2,8),(3,9),(4,10),(5,11),(6,12)\}$$

If A matches B, then B is said to be an instance of A

If L is an isomorphism and S is any node, then ASSIM(S,L) is the network dominated by S, except that each node A in the domain of L is replaced by L(A). For example, if we assimilate node 1 into node 100 using the isomorphism L i fig. 8, then we get:

Fig.9



Assimilation, using the isomorphism produced by matching, is used in several different ways in Fangorn. Limitations of space prevent me from describing the processes in any detail, but I will give a rough outline. The relevant processes are:

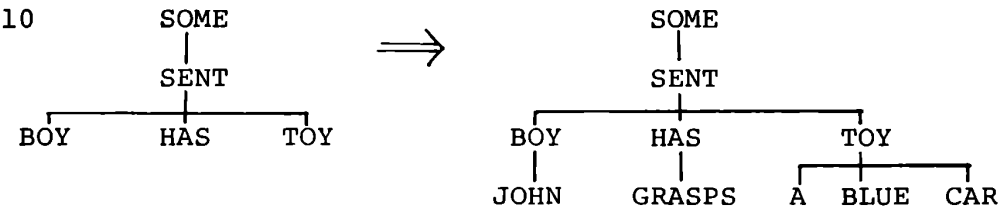
1. instantiation/ binding of variables (node: INSTANCE)
2. transformations in two varieties (node: REORDER)
3. expansions in two varieties (node: GROW and EXPAND)
4. anaphors (not implemented yet)

3.1 INSTANTIATION

In Fangorn it is possible to state conditions. There are four types of conditions, SOME, ALL, EXCEPT and NONE. A condition has exactly one daughter called its proposition. To give an example, SOME requires that at least one instance of its proposition be true at the "time" of the condition.

To instantiate a condition means to replace the proposition by certain sets of instances of the proposition or its negation. Example:

Fig.10



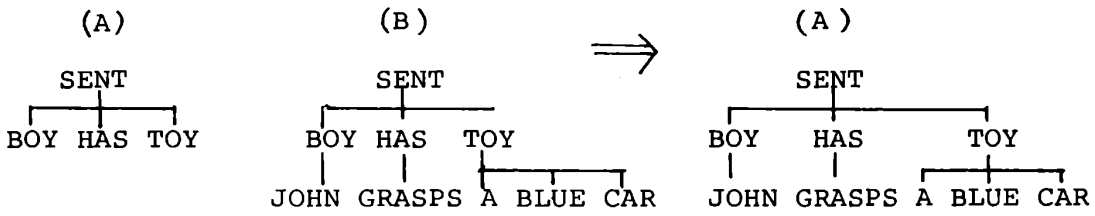
BOY, HAS and TOY may function as variables, to which the values JOHN, GRASPS and A BLUE CAR are assigned.

3.2. EXPANSIONS

We have two expansion atoms, EXPAND and GROW.

GROW is a generalisation of Chomsky's rewrite rules; its process searches a list for a network B that is an instance of another network A. If A matches B, then B is assimilated into A's context:

Fig.11



EXPAND is like GROW in that it searches a list in order to find matching nodes; but in this case B must be a daughter of a node X on the list, and B must match A and not conversely. If B is found, then A is assimilated into X, and X is inserted as the left sister of A. Example:

Fig.12

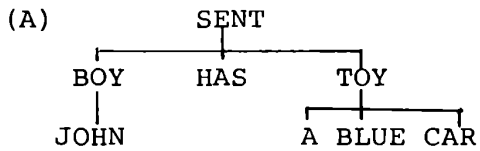
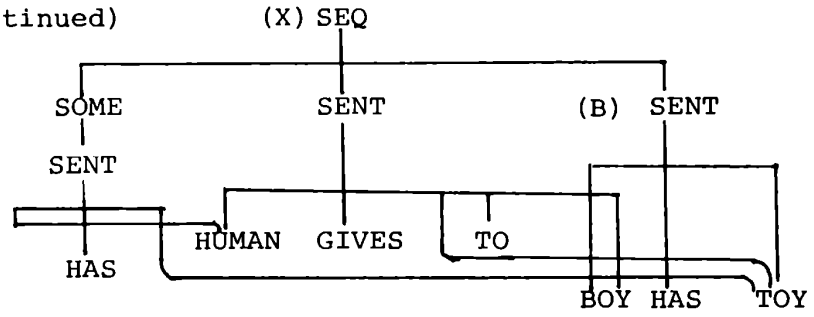


Fig.12 (continued)



Before assimilation the rule reads: if someone has a toy, and he gives it to a boy, then the boy has the toy. After assimilation, the rule reads: if someone has a blue car, and he gives it to John, then John has a blue car.

3.3. TRANSFORMATIONS

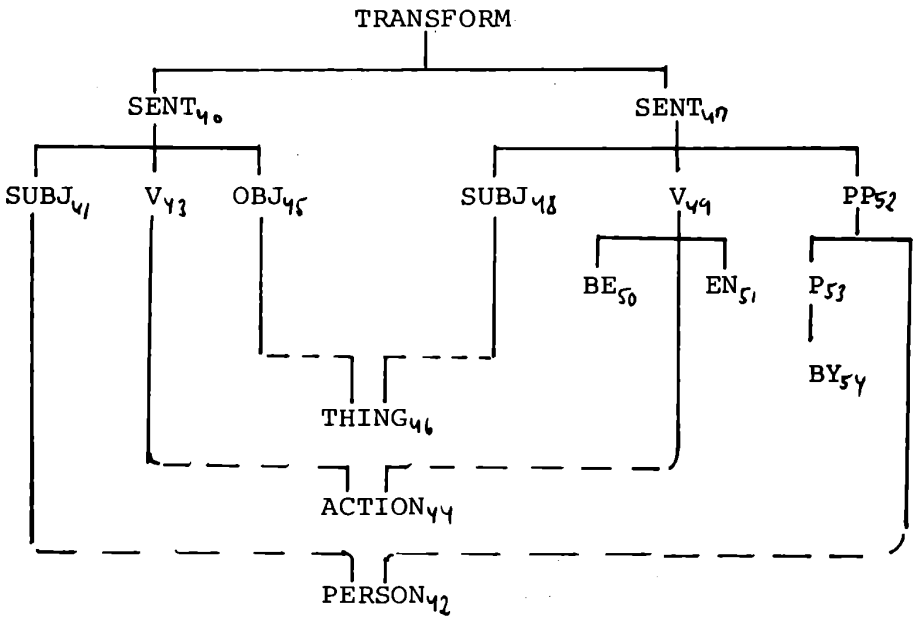
Both varieties, TRANSFORM and MOVE, consist of a structural description part (SD) and a structural change part (SC). A transformation is applicable to any network that its SD matches (actually, we allow two kinds of variables in the SD, the one being the X-variable of transformational grammar).

3.3.1 TRANSFORM

Let a SD match network A, producing isomorphism L. Then A is assimilated into SC using L. And this modified SC is equal to the result of the transformation.

Example: the passive transformation could be formulated:

Fig.13



$SENT_{40}$ is the structural description, and $SENT_{47}$ is the structural change. If fig.13 is applied upon fig.14 we get fig.15 :

Fig. 14

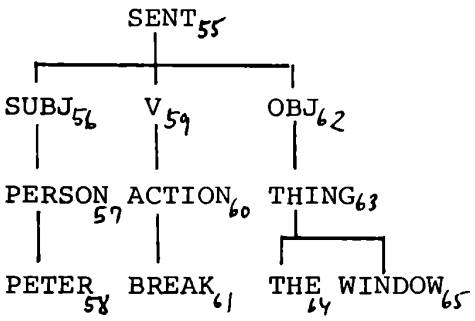
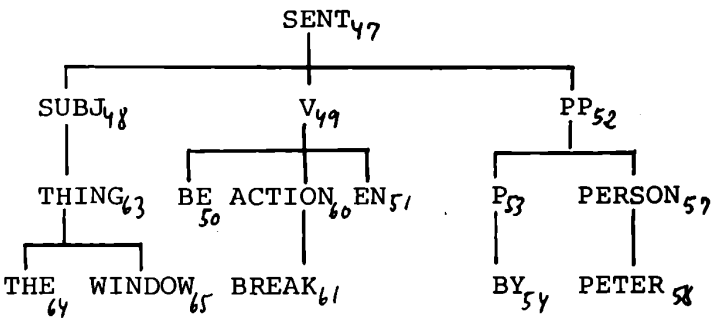


Fig. 15



40 matches 55, and assimilating 55 into 47 gives fig. 15.

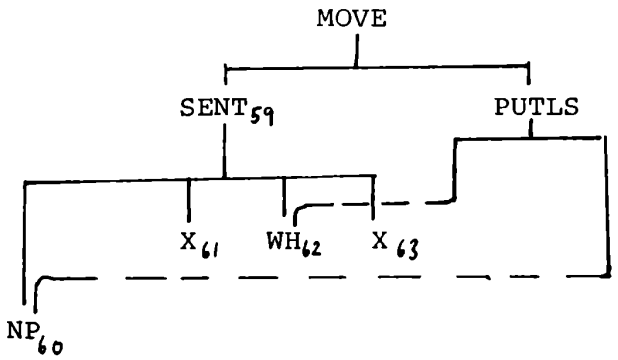
Not all transformations can be represented in this way, and therefore another format is supplied, called MOVE. The SD part is similar to the SD above, but the SC is different. It consists of orders as

ADD (as) L(eft)S(ister)
PUT (as) R(ight)D(aughter)

The first daughter of an order defines the object to be moved, and its second daughter defines the destination.

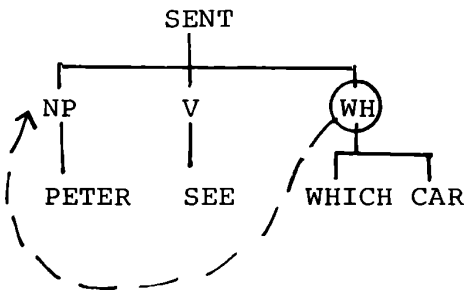
WH-movement could be represented thus:

Fig.16



The node defined by 62 is detached and attached as a left sister of the node defined by 60, eg:

Fig. 17



Expansions and transformations work on plans as well as on sentences: production of sentences and of plans are seen as essentially similar processes, both involving a grammar containing expansion rules and transformational rules, cf. Bøgh Andersen (73, 77 and 78). Thus, sentence production is seen as a special kind of goal-directed behavior, that is: work.

4. A SIMPLE FANGORN PROGRAM

A FANGORN program is a connected network with an upper bound, and the input language is simply a description of this network using parentheses and anaphors. Fig.18 shows a simple FANGORN program, and fig.19 shows how it is described in the input language.

The program contains a list of expansion rules (RSET) and an algorithm for using these rules (the subnetwork dominated by DOWN). The rules are applied on the network dominated by GOAL. Most of the atoms in the algorithm refer to an implicit pointer C, that is: RIGHT moves C to its right sister, ISLASTSIS checks whether C is the last sister, etc.

The algorithm expands the SENT-node under GOAL into a sentence. Initially, the pointer C points to GOAL. DOWN moves it down to SENT, and GROW tries to expand the value of C, that is: SENT. If it succeeds the pointer is moved down to its first daughter, else we check whether C is the last sister. If not, then C is moved to the right and the new pointer value is expanded. If C is the last sister, then we check whether C points to GOAL (ISTOPGOAL). If not, C is set to its mother(UP) and we check whether C is now the last sister. If C points to the top goal we have finished, and the sentence is written on a file called OUTDATA. Then the pointer is moved to the ACTOR node (TOP) and the whole actor is written on a file named PAPER. Then the algorithm stops.

The boxed portion of fig.18 shows a sentence generated by the program.

OUTDATA and PAPER are "channels" connecting the FANGORN program to the file system. OUTDATA and PAPER belong to different types of channels: when a network is written on OUTDATA only its leaves (or terminal nodes) are printed, so OUTDATA is oriented towards accepting natural language texts. When a network is written on PAPER it is translated into the input language, so PAPER can be used for storing and retrieving parts of the FANGORN program.

Fig.18 is very simple and does not generate stories or coherent texts, but facilities for these tasks are present in the program. In other programs, EPIC will have daughters representing the sequence of actions performed by the ACTORs. CAST may contain more than one actor, acting in a pseudo-parallel way. The algorithm in fig.18 may be replaced by algorithms for building and executing plans; in that case, RSET contains means-end rules, and CONT dominates

Fig.18

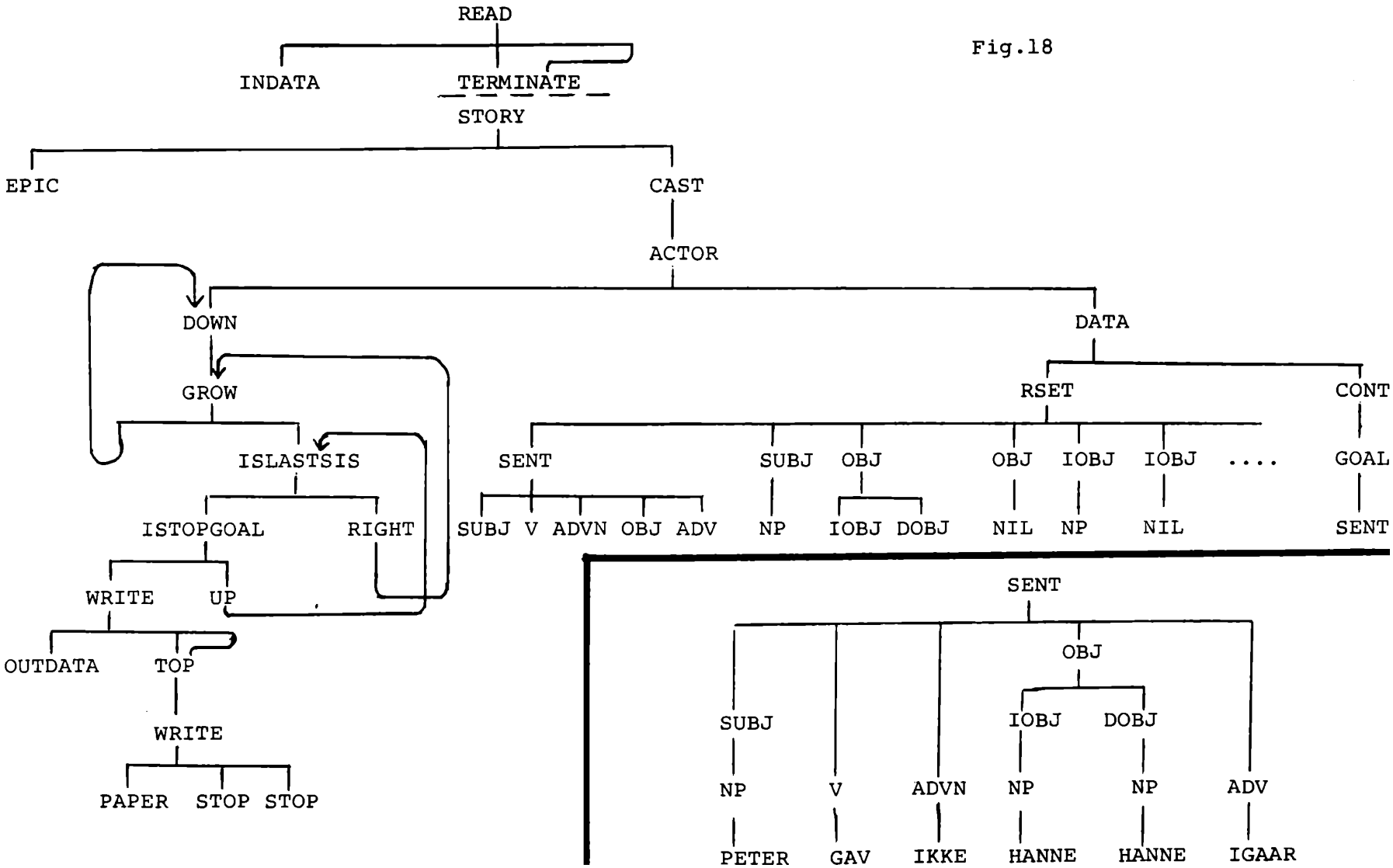

READ
INDATA
TERMINATE
STORY
EPIC
CAST
ACTOR
DOWN
DATA
GROW
RSET
CONT
ISLASTSIS
SENT
SUBJ
OBJ
OBJ
IOBJ
IOBJ
....
GOAL
ISTOPGOAL
RIGHT
SUBJ V ADVN OBJ ADV
NP
IOBJ DOBJ
NIL
NP
NIL
SENT
WRITE
UP
OUTDATA
TOP
WRITE
PAPER STOP STOP
SENT
SUBJ
OBJ
IOBJ
DOBJ
NP
V
ADVN
NP
NP
ADV
PETER
GAV
IKKE
HANNE
HANNE
IGAAR

Fig.19

```
(ACT ACTOR ALL BACKTRACK BELIEVED CAST CURRENT DEEP
 DELETE DOWN EMBRYO EXCEPT FIRSTLEAF GROW HASMOTHER
 INSTANCE ISCOND ISFIRSTSIS ISLASTLEAF ISLASTSIS ISLEAF
 ISPROP ISSEQ ISTOPGOAL LEFT NEGATE NEXTACTOR NONE READ
 REORDER RIGHT SATISFY SET SOME STACK STOP STORY TERMINATE
 TEST TIMEOUT TOP TOPGOAL TRACK TRUE UNSTACK UP WRITE)

(STORY
  EPIC
  (CAST
    (ACTOR
      (0! DOWN
        (1! GROW
          0? DOWN
          (2! ISLASTSIS
            (ISTOPGOAL
              (WRITE
                OUTDATA
                (3! TOP (WRITE PAPER STOP STOP))
                3? TOP
              )
              (UP 2? ISLASTSIS)
            )
            (RIGHT 1? GROW)
      ) ) )
      (DATA
        (RSET
          (SENT SUBJ V ADVN OBJ ADV)
          (SUBJ NP)
          (OBJ IOBJ DOBJ)
          (OBJ NIL)
          (IOBJ NP)
          (IOBJ NIL)
          (DOBJ NP)
          (NP PETER)
          (NP HANNE)
          (NP HUNDEN)
          (NP KATTEN)
          (V GAV)
          (V BRAGTE)
          (ADVN IKKE)
          (ADVN NIL)
          (ADV IGAAR)
          (ADV SIDSTE JUL)
          (ADV NIL)
        )
        (CONT (GOAL SENT))
) ) ) )
```

a network representing the beliefs and goals of the actor.

The implicit pointer C may be moved into the "algorithmic" part of the actor, and the actor may thus change himself.

Also, an actor may contain rules for generating new actors; when they are fully developed, they may be raised from the goal part into the CAST, and start interacting with the older actors.

It is possible to simulate the ontogenetic or phylogenetic development of language by means of a meta grammar that contains an object grammar in its goal part. The meta grammar generates part of the object grammar, activates the object grammar, thereby causing it to produce sentences; control is returned to the meta grammar,which enlarges or transforms the object grammar, again the latter is activated, etc., etc. The output will consist of sentences from encreasingly more complex grammars; the aim is to write a meta grammar that produces a sequence of object grammars whose sentences mirrors the development of linguistic skill in children.

At the time of writing, the FANGORN system is implemented but not debugged. It contains 47 different atoms, but I plan to add approximately 10 new atoms, so the final number will be 55-60.

REFERENCES

Peter Bøgh Andersen: Handlinger og symboler. Elementer af handlingens syntaks (Akademisk Forlag, 1973)

" : Sproget på arbejde (GMT, 1977)

" : "The syntax of texts and the syntax of actions" (in: Pragmalinguistics, ed. J.L.Mey, Mouton,1979)

J. R. Meehan : "TALE-SPIN, an Interactive Program that Writes Stories" (in: 5th Int. Joint Conf. on Art. Int., 1977, p.91 - 98)

Earl D. Sacerdoti : "The nonlinear nature of plans" (in: 4th Int. Joint Conf. on Art. Int., 1975, p.206 - 214)

" : The structure for plans and behavior (Elsevier Computer Science Library, Elsevir, 1977)

John F. Sowa : "Conceptual Graphs for a Database Interface" (IBM Journal of Research and Development, July,1976)

Gert Schmeltz Pedersen: Konceptuelle grafer.

Datamodeller er afbildninger mellem udsnit af virkeligheden, objektsystemer, og databaser. Afbildningen kan opdeles i 3 afbildninger, modellering, fortolkning og opdatering, eller med andre ord, meningsskabelse, meningstildeling og meningsvedligeholdelse.

Databaser er meningsfulde, hvis de giver et sammenhængende og fyldestgørende billede af deres objektsystem.

En datamodels evne til at give meningsfulde databaser afhænger af dens indhold af meningselementer (= semantiske begreber) og dens teoretiske grundlag.

Konceptuel graf-datamodellen er en syntese af en matematisk datamodel, relationsmodellen, og en lingvistisk datamodel, case-grammatikken. Dermed indeholder den såvel et matematisk grundlag og et datalogisk maskineri til operation og inferens på data som en teori for modellering af virkeligheden. Påstanden er derfor, at konceptuel graf-datamodellen kan give særdeles meningsfulde databaser.

Objektsystemer, der kan modelleres med konceptuelle grafer, er for eksempel det naturlige sprog og historier.

— — —

Litteraturhenvisning:
Gert Schmeltz Pedersen: Conceptual Graphs I, DIKU-rapport 78/9, Københavns Universitet 1978,
fås ved henvendelse til
Datalogisk Institut tlf. 01-836466
Sigurdsgade 41
DK-2200 København N

Automatisk orddeling

Hasse Hansson
Datalogisk Institut, Københavns Universitet
Sigurdsgade 41, DK-2200 København N, Danmark

Indledning

Når man skriver på et stykke papir, er man tvunget til at tage en beslutning, når man nærmer sig papirets højre kant: Skal hele det følgende ord flyttes ned på næste linie, eller skal det deles? Vilkårligheden af dette valg understreger, at en orddeling er et nødvendigt onde og ikke i sig selv rummer nogen information i modsætning til fx en indrykket linie, der markerer starten af et nyt afsnit. Det er derfor et rimeligt krav, at informationsindholdet i det delte ord bevares, således at orddelinger, der medfører midlertidig forvirring eller permanente misforståelser, undgås.

På trods af dette har den stadigt stigende anvendelse af automatisk tekstbehandlingsudstyr medført, at megen trykt tekst indeholder et stort antal forkerte orddelinger. Udviklingen er især tydelig inden for avisproduktion, hvor korrekturlæsning ofte foregår før ombrydning, og dermed før de automatiske orddelinger er udført. Nedenstående typiske eksempler på forkerte orddelinger er således fundet i de danske aviser Politiken og Ekstra Bladet.

middelhav-segnene	øjeb-likket
efte-rårskollektion	salg-sapparat
whiskyf-laske	rengøring-spersonalet
hvermand-seje	unds-lap
bes-laglagt	sanitet-stropperne

Da orddelinger tydeligvis er en væsentlig kilde til fejl, bør det undersøges, om orddeling overhovedet er nødvendig. Nedenfor vises to spalter med tilfældig tekst; den første indeholder ingen orddelinger, hvorimod der i den anden er foretaget orddelinger, hvor der er mulighed for det.

I de senere år har edb-baseret udstyr til tekstbehandling vundet en stadig større udbredelse. Indenfor forsknings- og undervisningssektoren kan især fremstillingen af dokumenter som rapporter og manualer lettes betydeligt ved anvendelse af tekstbehandlingsudstyr. Som oftest er sådant udstyr baseret på mini- eller mikrodatamater, men for brugere med terminaler tilsluttet et regnecenter som RECKU er det et naturligt ønske, at terminalerne også kan anvendes til tekstbehandlingsformål.

I de senere år har edb-baseret udstyr til tekstbehandling vundet en stadig større udbredelse. Indenfor forsknings- og undervisningssektoren kan især fremstillingen af dokumenter som rapporter og

manualer lettes betydeligt ved anvendelse af tekstbehandlingsudstyr. Som oftest er sådant udstyr baseret på mini- eller mikrodatamater, men for brugere med terminaler tilsluttet et regnecenter som RECKU er det et naturligt ønske, at terminalerne også kan anvendes til tekstbehandlingsformål.

Det ses, at linie 4 og 7 i øverste spalte indeholder uacceptabelt store mellemrum, og at den nederste spalte er 1 linie kortere end den øverste. Konklusionen må derfor være, at orddeling er nødvendig, dels for at få rimeligt udfyldte linier, dels for at opnå en papirbesparelse.

I perioden februar - september 1979 har jeg derfor i samarbejde med lektor Bente Maegaard fra Institut for anvendt og matematisk lingvistik, Københavns Universitet, udviklet en algoritme til automatisk deling af danske ord. Denne artikel beskriver dels de metoder, som vi har udviklet til brug for algoritmefremstillingen, dels den færdige algoritme. Det forudsættes, at læseren har et godt kendskab til det danske sprog; endvidere vil kendskab til programmeringssproget Pascal eller et andet ALGOL-lignende sprog være ønskeligt.

Regler for dansk orddeling

Målet for algoritmeudviklingen er at fremstille en algoritme, der i så mange tilfælde som muligt deler ordene i overensstemmelse med de regler for dansk orddeling, der er angivet i Retskrivningsordbogen [1]. Formålet med dette afsnit er at analysere reglerne for at klargøre i hvor stort omfang, de lader sig implementere på en datamat.

i) "Sammensatte ord deles efter deres bestanddele, når disse er let kendelige."
På dansk kan sammensatte ord frit dannes ved at skrive de to ord, der danner sammensætningen, uden et adskillende mellemrum, hvorimod man fx på engelsk i stor udstrækning beholder mellemrummet eller anvender en bindestreg. På trods af reglens klarhed er der derfor store vanskeligheder forbundet med maskinelt at afgøre, om et ord er sammensat.

ii) "Afledninger deles ligesom sammensætninger efter deres bestanddele, når disse er let kendelige."
Reglen er en parallel til i); men den er mindre konsekvent, idet det er usikkert, hvornår en afledning er "let kendelig". Antallet af afledningselementer, affikser,[1] er dog ret begrænset, hvorfor denne regel kan implementeres, hvis det er muligt at opstille fuldstændige lister over de affikser, der accepteres som let kendelige.

iii) "For usammensatte ord gælder følgende regler:"
1) "En medlyd mellem to selvlyd skrives sammen med den sidste selvlyd."
Denne regel kan umiddelbart implementeres; men der er dog, som bemærket af Spang Hanssen [2], en tradition for at skrive konsonanten *x* sammen med den første vokal.
2) "Af to medlyd mellem selvlyd går en til hver linie; ..."
Implementation af denne simple regel vanskeliggøres af følgende undtagelser (især b) og c)):
a) "*sk, sp, st* kan gå sammen til ny linie."
Der er tradition for at udnytte denne mulighed, der uden vanskelighed kan implementeres.
b) "I ord, der ikke har tryk på første stavelse, går begge medlyd til den trykstærke

1. Affikser er en fællesbetegnelse for præfikser, infikser og suffikser; men kun præfikser (forstavelser) og suffikser (afledningsendelser) har interesse i denne sammenhæng

stavelse, hvis de danner en lydforbindelse, der kan forekomme i begyndelsen af et ord."
Reglen lader sig ikke implementere, idet en datamat ikke ud fra ordets stavemåde kan afgøre, hvor trykket ligger.

c) "Efter lang selvlyd kan man skrive begge medlyd sammen med den følgende selvlyd, hvis de danner en lydforbindelse, der kan forekomme i begyndelsen af et ord."
Heller ikke denne regel kan implementeres, idet det ikke er muligt på baggrund af ordets stavemåde at afgøre, om en vokal er lang.

3) "Tre eller flere medlyd behandles sådan, at der til den ny linie kun går så mange, som der kan forekomme i begyndelsen af et ord."
Denne regel kan - måske lidt overraskende - implementeres, idet det er muligt at opstille fuldstændige lister over de konsonantforbindelser, der kan forekomme i begyndelsen af et ord.

Regler for orddeling på andre sprog

Efter analysen af reglerne for deling af danske ord er det interessant at undersøge, i hvor stort omfang reglerne er i overensstemmelse for andre europæiske sprog. Nedenfor er derfor anført en simplificeret version af de danske regler:

a) Sammensatte ord deles i deres bestanddele.
b) Afledninger deles efter bestanddele.
For usammensatte ord:
c) Ved 1 konsonant mellem to vokaler skrives konsonanten sammen med den sidste vokal.
d) Ved 2 konsonanter mellem to vokaler går en konsonant til hver stavelse.
e) Ved 3 eller flere konsonanter skrives højst så mange konsonanter sammen med sidste vokal, som der kan forekomme i begyndelsen af et ord.

Det er muligt ved hjælp af reference [1] samt [3] til [7] at undersøge, i hvor høj grad ovenstående fem regler er gældende for dansk, engelsk, fransk, tysk og svensk. Resultatet er angivet i nedenstående skema, hvor + betyder, at reglen gælder, (+) at reglen gælder men ikke følges konsekvent eller at der er undtagelser, og - at reglen ikke gælder.

	dansk	engelsk	fransk	tysk	svensk
a	+	+	+	+	+
b	(+)	+	(+)	-	+
c	+	+	+	+	+
d	(+)	+	(+)	(+)	(+)
e	+	-	+	-	-

Som det fremgår af skemaet, er der ret store uoverensstemmelser - især for de usammensatte ords vedkommende. Man kan derfor ikke gøre sig håb om at udvikle én algoritme, der kan dele ord på to eller flere af ovennævnte sprog.

Metoder til automatisk orddeling

I dette afsnit gennemgås fem forskellige metoder, som vi har udviklet til automatisk orddeling. Grunden til at der er udviklet ikke mindre end fem forskellige metoder er, at problemet på naturlig måde kan opdeles i underproblemer, der hver kræver sin egen metode for at blive løst tilfredsstillende. Ligegyldigt hvilken metode, der anvendes i et givet tilfælde, vil vi kræve, at den skal give en orddeling, der i så høj grad som muligt er i overensstemmelse med de tidligere omtalte regler for dansk orddeling. Metoderne vil i stor udstrækning blive illustreret med

eksempler, i hvilke tegnet - betegner en korrekt orddeling og tegnet ≠ en fejlagtig orddeling.

Affiksgenkendelse

Regel i) kræver, at sammensatte ord deles i deres bestanddele. Som et eksempel kan nævnes ordet *afledningsendelse*, der deles *aflednings-endelse* og ikke *afledning≠sendelse,* som ville være en korrekt orddeling, hvis ordet ikke var sammensat. Det er imidlertid vanskeligt at afgøre, om et ord er sammensat og i givet fald, hvor det er sammensat.

En metode hertil er at genkende et affiks midt i ordet. I førnævnte eksempel kan man således genkende suffikset *nings* og opnå en korrekt deling. Nedenfor vises eksempler på suffikser, som det empirisk har vist sig bør genkendes. Venstre spalte indeholder affikser, som der kan deles efter, medens højre spalte indeholder de affikser, som der kan deles før.

nings-	-agtig
tets-	-bar
ments-	-dom
lig-	-mæssig
som-	-skab

Man skal dog ikke ukritisk forsøge at genkende samtlige affikser, som er nævnt i fx Hansen [8] og Skautrup [9]. Et eksempel er *af,* som korrekt genkendes i ordet *flertals-afgørelse,* medens genkendelsen fører til fejl i ordet *bortsk≠affe.* Ved at afprøve de opstillede affikslister med et omfattende inddatamateriale vil det hurtigt blive klart, hvilke affikser, som med godt resultat kan genkendes.

Regel ii) kræver, at afledninger deles efter deres bestanddele, når disse er let kendelige. Det er ikke muligt at give en præcis definition af, hvornår en afledning er "let kendelig"; men det er dog muligt at opstille en liste med et ret begrænset antal elementer, der accepteres som "let kendelige" afledninger af vide kredse. Det er derfor muligt ved hjælp af affikslister at implementere denne regel således, at der opstår meget få fejl.

Statistiske metoder

Hvis det ikke er muligt at genkende et affiks i ordet, må det deles på baggrund af statistiske resultater. Reglerne for deling af usammensatte ord deler op i tre tilfælde: 1 konsonant, 2 konsonanter og 3 eller flere konsonanter mellem to vokaler. Det er derfor naturligt, at lade programmet foretage den samme opdeling.

I tilfældet 1 konsonant skal der ifølge regel iii) pkt. 1 deles før denne konsonant, og dette påbud kan umiddelbart følges. Den eneste undtagelse er konsonanten *x*, som traditionelt skrives sammen med første vokal.

Matriks for tokonsonantkombinationer

Hvis der optræder to konsonanter mellem to vokaler, skal der ifølge regel iii) pkt. 2 deles mellem dem; men på grund af undtagelse b) og c) kan reglen ikke umiddelbart implementeres. Vi har derfor valgt at analysere et omfattende ordmateriale for empirisk at finde det bedste delepunkt for hver af de 400 mulige konsonantkombinationer. Da analysens kvalitet naturligvis afhænger af det valgte analysemateriale, er det vigtigt, at dette vælges med omhu. Et umiddelbart valg er en ordbog som fx Nudansk Ordbog [10], der findes på maskinlæsbar form. Imidlertid er et ordbogsmateriale uegnet af to grunde: Antallet af sammensatte ord i fx Nudansk Ordbog er forsvindende i forhold til antallet af mulige sammensætninger, og som tidligere nævnt er det netop sammensætninger, der virkelig volder problemer. Den anden grund er, at frekvensen i naturlig tekst af et givet ord fuldstændig ignoreres i ordbogsmaterialet. Dette betyder, at man kan risikere at vælge en orddeling ud fra et antal ord, som forekommer sjældent i naturlig tekst. Ved i stedet at vælge orddelingen ud fra færre, men ofte forekommende ord vil man opnå færre

fejl ved deling af ord i en naturlig tekst.

Vi har derfor valgt at samle et analysemateriale, som indeholder ord fra et stort antal forskellige tekster som vist i nedenstående skema:

Dansk prosa	250.000 ord
Avisartikler	20.000 ord
RECKU-Nyt[2]	50.000 ord
Økonomiske Råds rapport	22.000 ord
Matematisk fagtekst	11.000 ord
Geologisk fagtekst	5.000 ord
I alt	358.000 ord

"Dansk prosa" dækker over tekstprøver fra mange forskellige danske forfattere, hvis bøger er blevet udgivet i de senere år. Materialet indeholder i alt 358.000 ord, hvoraf de 35.000 er forskellige.[3] Til sammenligning kan nævnes, at Nudansk Ordbog indeholder ca. 51.000 ord.

Ved analysen af tokonsonantkombinationer viser det sig, at man kan finde sikre delepunkter for langt de fleste kombinationer, fx:

-bl: *øje-blik*
r-n: *tor-ne*
zz-: *jazz-orkester*

Der er dog enkelte kombinationer, der er meget usikre, som fx *dr:*

d-r: *yd-re*
-dr: *a-dresse*

Endelig er der en række kombinationer, som er umulige både som initial- og som finalkombinationer. Forekommer en sådan kombination midt i et ord, kan man med sikkerhed dele mellem de to konsonanter, fordi ordet da må være sammensat. Som et eksempel kan nævnes kombinationen *df:*

d-f: *flod-forurening*

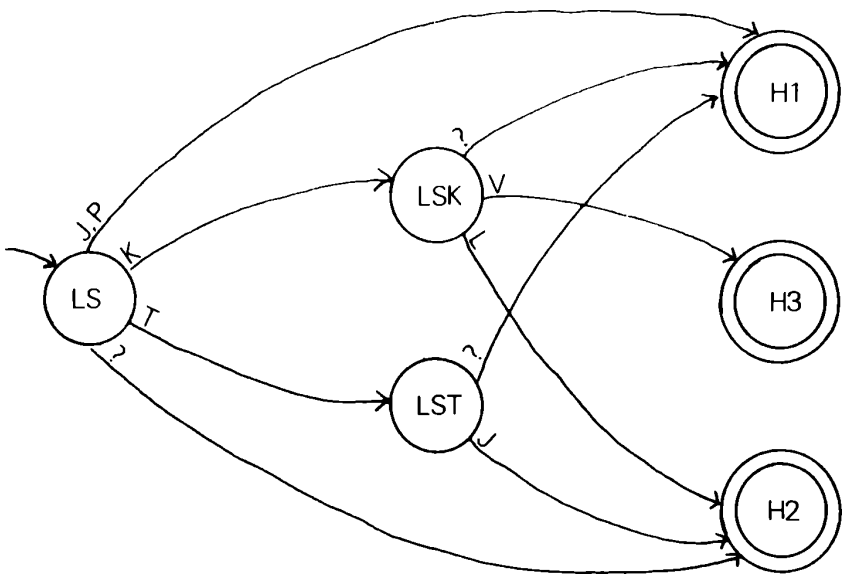
Resultatet af analysen er en 20 × 20 matriks, der for hver konsonantkombination angiver, om der skal deles til venstre, mellem eller til højre for de to konsonanter. Endelig indeholder matricen oplysninger om de kombinationer, der er usikre; anvendelsen af denne oplysning vil blive forklaret senere.

Tilstandsautomat

Ved kombinationer af tre og flere konsonanter er mulighederne så talrige, at det ikke er hensigtsmæssigt på tilsvarende måde at opstille en matriks i tre eller flere dimensioner. Dette skyldes dels sådanne matricers mange elementer, dels at de fleste konsonantkombinationer er utænkelige i praksis og derfor uinteressante. I stedet har vi ud fra analysen af materialet opstillet en endelig tilstandsautomat, hvor man - populært sagt - kun registrerer de kombinationer, som man er interesseret i. For at begrænse automatens størrelse indeholder den ikke de konsonantkombinationer, som man kan dele korrekt ved at anvende den tidligere omtalte tokonsonantmatriks på de to konsonanter umiddelbart foran den anden vokal. Et eksempel er ordet *overblik,* hvor matricen giver delingen *-bl* og dermed *over-blik.*

Der er dog stadig mange kombinationer, som må klares ved en tre- eller endda firkonsonantanalyse. Mange kombinationer forekommer kun i sammensætninger, hvorfor mange sammensatte ord bliver delt korrekt netop på grund af automaten. Nedenfor vises den lille del af automaten, der finder en orddeling i alle konsonantkombinationer, der indledes med konsonanterne *ls.*

2. RECKU-Nyt er en brugerorientering, der udsendes ca. 10 gange årligt af RECKU, Det Regionale Edb-center ved Københavns Universitet.
3. Forskellige bøjningsformer af samme ordstamme regnes i denne forbindelse for forskellige ord.



Tilstandene med dobbeltring - på tegningen H1, H2 og H3 - kaldes *sluttilstande*. Når en sådan nås, er analysen færdig, idet H1 betegner en deling efter den første konsonant, H2 en deling efter den anden konsonant etc. Ved at følge ord som ***hals-klud, bol-sje, detail-specifikation, elsk-værdig, formåls-tjenlig*** og ***galskab*** gennem automaten ses, hvorledes vidt forskellige ord - såvel sammensatte som usammensatte - analyseres og deles korrekt.

Fonetisk kontrol

Dansk orddeling følger ligesom fx fransk et fonetisk princip, hvorfor den konsonantkombination, der forekommer efter delestregen, skal kunne forekomme i begyndelsen af et ord. Når man derfor ved en af de ovenfor nævnte metoder har fundet en orddeling, undersøges om kombinationen er lovlig. Er dette ikke tilfældet, flyttes bindestregen mod højre, indtil kravet er opfyldt. Denne *fonetiske kontrol* forhindrer især de fejl, der opstår ved forsøg på genkendelse af et affiks, hvor den genkendte tegnfølge i det givne ord viser sig at indgå i en ganske anden sammenhæng. Et eksempel er ordet *angsten,* hvor præfikset *an* genkendes og ordet derfor deles *an≠gsten.* Den fonetiske kontrol sikrer, at *gs* ikke forekommer efter delestregen, og delingen rettes derfor til *ang-sten.*

Undtagelsesordbog

Vi vil i det følgende sammenfatte de hidtil omtalte metoder under betegnelsen de *algoritmiske metoder.* Selvom disse metoder forfines, vil der stadig findes ord, der ikke deles korrekt. For alligevel at kunne dele sådanne ord rigtigt opbygges en *undtagelsesordbog,* der indeholder alle de ord, som erfaringsmæssigt giver anledning til fejl. På grund af det store antal opslag i en sådan ordbog er det vigtigt, at den organiseres således, at søgning er effektiv; specielt skal søgning efter ord, der ikke findes i ordbogen, være effektiv. Dette opnås ved at benytte en såkaldt hashteknik. For en nærmere beskrivelse af teknikken henvises til Filsystemer og databaser [11].

Som et kuriosum kan nævnes, at selv med anvendelse af en undtagelsesordbog vil homografer[4)] kunne føre til forkerte orddelinger. Et eksempel er *vandrende,* som deles *van-drende,* hvis det er præsens participium af verbet *vandre,* medens det deles *vand-rende,*

4. Homografer er forskellige ord, der har fælles stavemåde.

hvis det er et substantiv. Vi har hidtil kun beskæftiget os med analyse af enkelte ord; men dette problem kan kun løses ved en syntaktisk analyse af hele den sætning, hvori ordet indgår.

Den komplette algoritme

Formålet med dette afsnit er at vise, hvorledes de fem metoder, som er omtalt i forrige afsnit, kan kombineres til en komplet algoritme. Desuden omtales nogle detaljer, som ikke er af principiel art og derfor ikke har været omtalt tidligere. Algoritmen, der i denne udformning læser ord fra en datakilde og finder samtlige mulige delepunkter i disse, er skitseret nedenfor i et Pascal-lignende sprog.

```
 1 while more_words( source ) do
 2 begin
 3    safe_hyphen( word );
 4    dictionary_lookup( word );
 5    search_prefix( word );
 6    while more_syllables( word ) do
 7    begin
 8       search_affixes( syllable, found );
 9       if not found then
10       begin
11          case consonants of
12             0: if divisible( syllable ) then save_hyphen;
13             1: save_hyphen;
14             2: matrix_hyphen( syllable );
15             otherwise
16             begin
17                finite_state_automata( syllable, found )
18                if not found then matrix_hyphen( syllable );
19             end
20          end
21       end
22    end
23 end.
```

I linie 3 findes de *sikre delesteder,* hvorved vi forstår forekomster af tegnene '-', ',' eller '/'. Det forekommer os indlysende, at et ord som *dansk-tysk* skal deles efter bindestregen; men på trods heraf genkender mange orddelingsalgoritmer ikke bindestreger, hvilket naturligvis fører til besynderligt udseende delinger. Grunden til at der deles efter et komma er, at der i visse tekster ikke er indsat et blanktegn efter kommaet. Ligeledes anvendes skråstregen undertiden til at danne sammensatte ord som fx *input/output.*

I linie 4 slås ordet op i undtagelsesordbogen, og hvis det findes, registreres de delepunkter, som ordbogen angiver.

I linie 5 søges efter et præfiks i en særlig præfiksliste. Dette skyldes, at det ved analysen af affikser viste sig, at der findes mange præfikser, som med fordel kan genkendes først i et ord men ikke midt i et ord. For ikke at miste disse ret sikre orddelinger samledes sådanne præfikser i præfikslisten.

Algoritmen har i linie 3 - 5 arbejdet på hele ordet; men i linie 6 - 22 findes en løkke, som anvendes til at finde delepunkter mellem de enkelte stavelser. Betingelsen for at der findes et delepunkt mellem to stavelser i denne løkke er, at der ikke allerede i linie 3 - 5 er registreret et delepunkt mellem de pågældende stavelser.

I linie 8 gennemløbes to affikslister, hvor den ene indeholder de affikser, som der skal deles efter, medens den anden består af de affikser, som der skal deles før. Hvis et sådant affiks ikke

genkendes, antages stavelsen at være en del af et usammensat ord, og der foregår derfor en opsplitning efter antallet af konsonanter ved hjælp af CASE-sætningen i linie 11.

Det tilfælde, hvor der ikke findes nogen konsonanter mellem de to vokaler, der afgrænser stavelsen, behandles i linie 12. Et sådant *vokalsammenstød* kan altid deles i oprindeligt danske ord; men i mange fremmedord betegner de to vokaler en diftong, som ikke kan deles. For eksempel kan hverken *ea* eller *au* deles i ordet *niveau.* Der er derfor opbygget en lille tabel, der angiver mellem hvilke vokaler, en deling er tilladelig.

Tilfældet 1 konsonant klares umiddelbart i linie 13, medens tilfældet 2 konsonanter løses ved opslag i tokonsonantmatricen. Ved tre eller flere konsonanter anvendes tilstandsautomaten; indeholder den ingen løsning, anvendes tokonsonantmatricen på de to konsonanter umiddelbart før den anden vokal.

Nøjagtighed

Den konstruerede algoritme er mere kompleks end tilsvarende algoritmer, som vi har fået kendskab til. Det er derfor interessant at undersøge, om resultatet er så meget bedre, at den forøgede programkompleksitet kan betale sig. Nedenfor vises de fejlprocenter, som opnås med den færdige algoritme.

Uden ordbog	1.8 % fejl
Med ordbog	1.2 % fejl

Fejlprocenten uden ordbog er målt på hele analysematerialet (358.000 løbende ord), medens fejlprocenten med ordbog er målt på tilfældige avisartikler, da det oprindelige analysemateriale er blevet brugt ved opbygning af undtagelsesordbogen.

Tallene virker måske ikke umiddelbart imponerende, men der er to faktorer, som skal med i vurderingen. For det første er resultatet klart bedre end tilsvarende, kommercielt udviklede programmer. Politiken hævder således, at deres program giver en fejlprocent på 3, og leverandører af tekstbehandlingsprogrammel er stolte af at markedsføre algoritmer, der har en fejlprocent på 2; men vores algoritme er med 1.2 % næsten dobbelt så god. For det andet indeholder undtagelsesordbogen på nuværende tidspunkt kun ca. 1000 ord, og en fuldt udbygget ordbog vil bevirke, at fejlprocenten falder til et niveau, hvor fejlene hovedsagelig skyldes nye sammensatte ord.

Anvendelse i tekstbehandlingsprogrammel

Jeg har i nogle år beskæftiget mig med udvikling af et tekstbehandlingssystem kaldet PHOTODOC, der anvendes ved fremstilling af sats på RECKU's[5] fotosætter. Ved dette arbejde viste der sig et stort behov for en algoritme til at dele danske ord, og det er grunden til, at jeg begyndte at udvikle orddelingsalgoritmen. Imidlertid er en fleksibel grænseflade mellem orddelingsprogrammet og det øvrige tekstbehandlingssystem en forudsætning for, at man får det fulde udbytte af orddelingsalgoritmen. I det følgende beskrives derfor, hvorledes algoritmen kan indbygges i et tekstbehandlingssystem, således at en høj grad af fleksibilitet opnås.

Parametre til orddelingsalgoritmen

Da automatisk orddeling er en kilde til fejl, bør orddelingsalgoritmen ikke aktiveres, hvis man uden orddeling kan opnå en linie af typografisk acceptabelt udseende. Man kan ved hjælp af følgende fire parametre styre dels antallet af delinger, dels udseendet af disse:

1) Aktiv eller passiv
2) Relativ spildfaktor

5. RECKU er en forkortelse for Det Regionale Edb-center ved Københavns Universitet.

3) Sikkerhedsniveau
4) Minimum antal tegn før og efter delestreg

Første parameter giver mulighed for at hindre, at orddelingsalgoritmen aktiveres. Dette har især betydning, hvis kildesproget ikke er dansk, da algoritmen er uanvendelig til andre sprog end dansk.

Anden parameter, den *relative spildfaktor,* angiver den procentdel af linlen, der skal være ubrugt før orddelingsalgoritmen aktiveres. Angives en stor spildfaktor, fås få orddelinger men til gengæld også en "løs" og dermed uøkonomisk sats; omvendt vil man ved en lille spildfaktor få en "tæt" sats med mange delinger. Den bedste spildfaktor afhænger af personlig smag; men en værdi mellem 5 og 10 procent giver sædvanligvis gode resultater. Det er væsentligt, at spildfaktoren er relativ: En lang linie giver bedre mulighed for at fordele en vis uudnyttet plads end en kort linie.

Ved omtalen af tokonsonantmatricen blev det nævnt, at visse kombinationer bliver registreret som usikre; men også nogle få konsonanter i enkonsonantforbindelser samt enkelte kombinationer i tilstandsautomaten er registreret som usikre. Ved at angive et *sikkerhedsniveau* for orddelinger kan man undgå, at ord deles i usikre delepunkter, og man kan ligeledes vælge kun at dele ord i *sikre* delepunkter, hvortil regnes forekomster af bindestreg, komma eller skråstreg i ordet. Sidste mulighed har især betydning for engelsk, hvor man på denne måde har mulighed for at dele alle de sammensætninger, der er dannet ved hjælp af en bindestreg.

Sidste parameter angiver det minimale antal bogstaver, der skal findes såvel før som efter delepunktet. Der er almindelig enighed om, at to bogstaver må være minimum; men et minimum på tre tegn foretrækkes undertiden.

Interaktiv kontrol

Da orddelingsalgoritmen som nævnt har en fejlprocent på ca. 1.2, må man ved store tekster regne med at skulle rette nogle forkerte orddelinger i den ombrudte tekst. Imidlertid har man ved hjælp af en *interaktiv kontrol* mulighed for at rette forkerte orddelinger i det øjeblik, de opstår. Således kan man ved kun én programudførelse opnå en ombrudt tekst, der i hvert fald med hensyn til orddelinger er helt korrekt. Nedenfor vises et eksempel på en sådan interaktiv kontrol. Den kursiverede tekst er brugerens inddata, medens alt andet udskrives af datamaten.

@pho★to.doc file.elt
PHOTODOC 4R1 RLIB 73R1 Wednesday, 1979 November 14, at 12:27:46
@hyphen
★ Hyphen control is active ★
TEKSTBE H★ANDLING ->
UD ★STYR ->
TRYKFÆRDIG ★DIG ->*tryk færdig*
TY ★POGRAFISKE ->*typo*
Hyphen beyond maximum point - try again!
TY ★POGRAFISKE -> *t*
FACILITE T★ER ->*favili*
Misspelling - try again!
FACILITE T★ER ->*facili*
MU L★IGHED ->
0 ERRORS AND 0 WARNINGS
PROOF LIST HAS BEEN SENT TO PR2
TIME USED: 9.318 SECONDS.
END PHOTODOC

PHOTODOC præsenterer alle orddelinger for brugeren, idet der indsættes et blanktegn på det sted i ordet, hvor datamaten mener, at ordet skal deles. Stjernen indsættes for at markere den del af ordet, der maksimalt kan stå før delestregen. Det første delte ord i eksemplet, *tekstbehandling,* vil datamaten således dele *tekstbe-handling,* og delestregen kan ikke indsættes senere end efter *h.* Accepteres orddelingen, trykker brugeren blot på vognreturknappen, i modsat fald skrives ordet med et blanktegn indsat på det sted, hvor ordet ønskes delt. Et eksempel er vist ved ordet *trykfærdig,* som brugeren ønsker delt *tryk-færdig.* Af effektivitetsgrunde kan den del af ordet, der skrives efter blanktegnet, dog udelades.

Endelig vises eksempler på de to fejl, som brugeren kan begå. I det første tilfælde forsøger man at indsætte en delestreg til højre for maksimumspunktet, og i det andet tilfælde staves ordet forkert ved rettelsen. I begge tilfælde udskrives en fejlmeddelse, hvorefter brugeren har mulighed for at rette fejlen.

Konklusion

Det er lykkedes at udvikle en orddelingsalgoritme, som er klart bedre end de hidtil kendte algoritmer til deling af danske ord. Algoritmen har vist sin anvendelighed i praksis og giver ifølge brugerudsagn en signifikant besparelse i forhold til manuel orddeling.

Efter min mening er en væsentlig årsag til de gode resultater samarbejdet med en lingvist, der bevirkede, at vi under algoritmeudviklingen betragtede både lingvistiske og datalogiske aspekter. Anvendelsen af lingvistik sikrede dels, at vi fx gennem affiksgenkendelse ikke ignorerede ordenes sproglige struktur, dels at vi opbyggede et dækkende analysemateriale. Anvendelsen af datalogi førte til en programstruktur, der nøje afspejler de regler for dansk orddeling, som har været vort arbejdsgrundlag. Under udviklingen har det derfor været muligt at omskrive eller forfine et enkelt modul uden at berøre programmets øvrige moduler, hvorfor det har været tidsmæssigt overkommeligt at eksperimentere med forskellige løsningsmuligheder.

Dette arbejde er således et eksempel på de gode resultater, der kan opnås gennem et konstruktivt, tværfagligt samarbejde.

Referencer

[1] Retskrivningsordbogen. Udgivet af Dansk Sprognævn, 1955.

[2] Henning Spang-Hanssen: Orddeling ved linieskifte - er reglerne tidssvarende? Artikel i SAML, nr. 5 1979, p. 73 - 93.

[3] Poul Steller og Knud Sørensen: Engelsk grammatik, København 1966.

[4] Der Sprach-Brockhaus, Wiesbaden 1966.

[5] Poul Høybye: Fransk grammatik, København 1966.

[6] Skrivregler. Tekniske Nomenklaturcentralens publikationer nr. 37, Stockholm 1976.

[7] C. Hansen-Chrisensen og Sven Brüel: Oversigt over tysk grammatik, København 1979.

[8] Åge Hansen: Moderne dansk, København 1967.

[9] Peter Skautrup: Det danske sprogs historie, bd. 3, København 1968.

[10] Nudansk Ordbog, København 1969.

[11] Bratsbergsengen, Høfstad og Wibe: Filsystemer og databaser, Trondheim 1974.

Gustav Leunbach

Om automatisk orddeling.
Forslag til en undersøgelse.

I flere systemer til automatisering af trykning og tekstfremstilling i det hele taget indgår automatisk orddeling for at undgå for store forskelle i linjelængderne, og især i diskussionen om læseligheden af avistryk spiller de af og til bizarre virkninger af denne orddeling en vis rolle; aviserne hævder - med rette - at det kun drejer sig om få procent af samtlige orddelinger, men det er ting som falder i øjnene.

De anvendte algoritmer er vistnok forretningshemmeligheder, og en opgave kunne være at forsøge at dechiffrere dem udfra den producerede sats. Men det forslag jeg nu fremsætter går i en anden retning, nemlig at sammenligne forskellige algoritmers virkning på den samme simulerede eller naturlige tekst.

Kvaliteten af en orddelingsalgoritmes produktion er sammensat af to komponenter: akceptabiliteten af de foretagne delinger og variationen i linjelængde. Der er et komplementært forhold mellem disse to; hvis man slet ikke deler må man akceptere den maximale variation i linjelængder; omvendt kan man slutte enhver linje ved position nr. 60 hvis man til gengæld finder sig i at ord deles hvor som helst.

Akceptabiliteten af foretagne delinger kan ikke bedømmes efter faste retningslinjer, men må i nogen grad bero på skøn, f.ex. er retskrivningsordbogens angivelse af tilladte delinger ikke fuldt algoritmiserbar. En graduering af akceptabiliteten er meget vel tænkelig, men jeg foreslår at et ulige antal bedømmere blot får til opgave at svare ja eller nej til hver foretagen deling og at flertallet afgør. Hvis graden af overensstemmelse er for lille kan man overveje at sende alle bedømmerne til omskoling.

Variabiliteten af linjelængder er et numerisk kriterium, men ikke endimensionalt, idet en algoritme både producerer mange linjer med lidt under standardlængden og enkelte som afviger stærkt; en rimelig sammenvejning kunne være det gennemsnitlige "tab" i forhold til standardlængden kombineret med et absolut forbud mod at overskride denne.

Udover algoritmers produktion kan de også bedømmes på deres kompleksitetsgrad, deres "længde". Der kan sikkert skabes enighed om at en algoritme som består i en opregning af alle tilladte delinger af alle ubøjede og bøjede ord i et korpus er for lang; omvendt kan man få en ultrakort og ubrugelig algoritme ved at sige at ethvert ord kan deles efter hvert tredje bogstav. Den første indeholder for mange konstanter,

den anden for få; en opregning af hvilke bogstavsymboler der angiver vokaler vil de fleste nok betragte som et tilladeligt sæt konstanter. Hvis man ønsker at præcisere "længden" af algoritmer skal man vel tænke på antallet af linjer i en kodning i et vist assemblersprog.

Det giver et stort spild at måle en algoritmes produktion på de ord som tilfældigvis kommer til at afslutte linjerne i en normal sætning af en tekst, og mit forslag indebærer en simulation på grundlag af en sandsynlighedsbetragtning for hvilke ord som bliver udsat for at skulle deles.

Grundmaterialet er en liste over ordformer med deres frekvenser i et korpus sorteret efter ordlængde, således at 1-bogstavs ord (f.ex. ordformerne i, ø og å) tilsammen har en frekvens på a_1, 2-bogstavs ord tilsammen a_2 osv. Lad os antage at der længste ordform har 27 bogstaver. Sandsynligheden for at det næste ord har længden j kan da til enhver tid sættes til

$$b_j = a_j/(a_1+a_2+\ldots+a_{27}).$$

(Virkningen af at ord ikke kommer i tilfældig rækkefølge er svær at beregne, men er antagelig lille for de variable det her drejer sig om.)

Hvis der på et givet tidspunkt er k positioner tilbage på linjen (medregnet det mellemrum som skal følge efter sidste ord) vil der da være sandsynligheden

$$c_k=b_k+b_{k+1}+\ldots+b_{27}$$

for at det næste ord ikke kan stå der og altså må søges delt. Det kan da beregnes at sandsynligheden for at delingsbehovet opstår netop når der er k positioner tilbage er omtrent proportional med c_k, og sandsynligheden for at det ord der skal deles er af længden j bliver en konstant gange $c_k \cdot b_j$ (for hvert $j \geq k$).

Simulationen går herefter ud på at der ved tilfældige tal trækkes er tal k med sandsynligheden c_k gange normeringskonstanten og derefter et ord blandt alle der mindst fylder k bogstaver med sandsynligheder bestemt af ordenes frekvens.

Dette ord søges så delt sådan at højst k-2 bogstaver står før delestregen (denne skal også kunne være på linjen); tabet bliver k-2-i hvis i bogstaver står før delestregen og k hvis deling ikke lader sig gøre. Resultatet af en algoritme udtrykkes dels ved fordelingen af tabene på værdierne mellem 0 og 27, dels ved procenten af uakceptable delinger efter bedømmernes flertalsafgørelse. Forskellige algoritmer

kan bedømmes på det samme sæt af ord-antal-par; de foretagne delinger kan præsenteres for bedømmerne i randomiseret orden, og de tilfælde hvor alle algoritmer har givet samme deling behøver slet ikke bedømmes, medmindre man foruden sammenligningen også ønsker at bestemme algoritmernes akceptabilitetsniveau.

Jeg har selv en algoritme parat som jeg tilbyder til sammenligning med andres. Det korpus jeg disponerer over er måske for lille, men andre kan vel levere et bedre.

Geir Berge

En datastruktur for ordbøker for naturlige språk

1. Innledning

Det arbeidet som skal refereres her hadde som mål å konstruere en tjenlig datastruktur for konstruksjon av ordbøker for naturlig språk på små datamaskiner. Arbeidet ble gjort for noen år tilbake av George M. Gillow og undertegnede, og vi anvendte en PDP 11/20. I dag kan derfor våre resultater ha interesse i forbindelse med mikromaskinanvendelser på samme område.

1.1 Designkrav

- Systemet skal hurtig og enkelt kunne laste en ordbok. Hurtig skal her forsås slik at tid for lasting, t_l, i størst mulig grad burde være en linjær funksjon av grunnlagstekstens størrelse, dvs. antall tokens, N_{token}.
- Ingen preprosessering av grunnlagsteksten skal være nødvendig (f.eks. sortering).
- Egenskaper ved naturlige språk skal i størst mulig grad formulere resterende krav. (Dette er gjort i tabellen på neste side).

EGENSKAP	→	DESIGNKRAV
Stor variasjon i ordlengde		Variabel lengde record format
En liten del av typene utgjør en stor del av teksten		Systemet skal ta spesielt hensyn til høyfrekvens-typer

⇓

(Hovedparten av høyfrekvenstypene vil dekkes av den første delen av tekstgrunnlaget.)

og

(Å øke tekstgrunnlaget for en ordbok vil i hovedsak gi opphav til lavfrekvens-typer)

Typer vil svært sjelden bli strøket fra ordboken	"Garbage collection" trenger ikke være inkludert i systemet

Behovet for oppslag av enkeltord vil være nær proposjonalt med de enkelte typers forekomstfrekvens i tekstbasen, og vi vil gjerne at systemet gir priviligert, hurtig aksess til slike typer.

Til slutt skal nevnes at systemet også bør gi gode muligheter for listing av større eller mindre deler av ordboken i alfabetisk orden.

Vi fant ikke at noen standard filstrukturer med tilhørende aksessmetoder oppfylte våre designkrav, og vi konstruerte derfor systemet som blir beskrevet i det følgende.

2. Løsningsforslag

Forslaget er implementert med types og tokens i form av grafiske ord, men det er intet til hinder for at det kan fungere på andre nivå. Videre konstruerte vi et enkelt-aksess system, men alle programdeler er utviklet med enkel utvidelse til multiaksess system for øye.

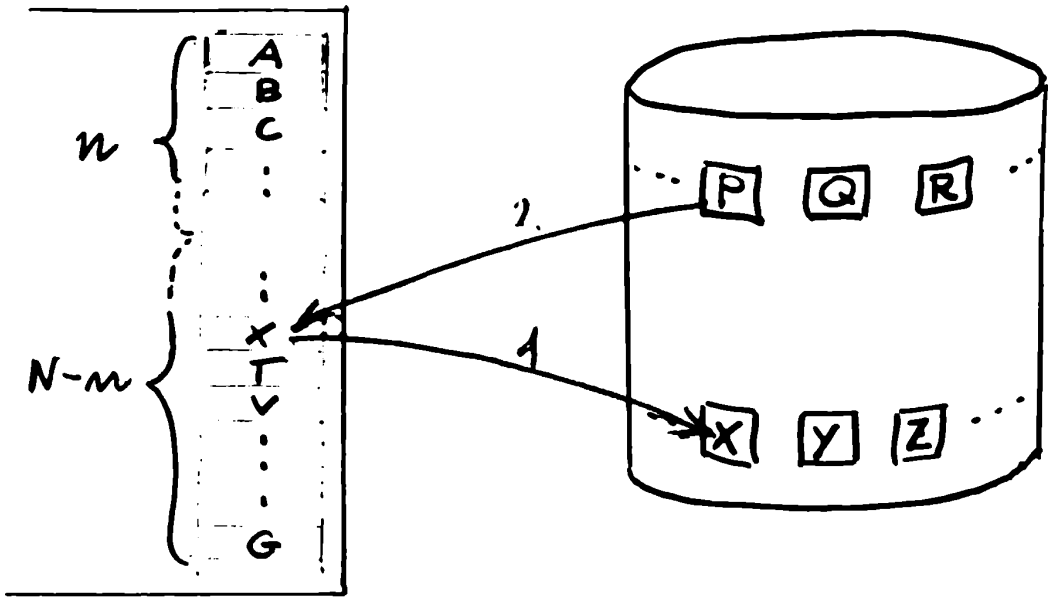
2.1 Paging systemet

Hele ordboken tenkes delt i et antall sider (pages). I hurtiglageret avsettes et område med plass til N slike. Hvert delområde med plass til en side kaller vi en nisje (slot). n av de N ($n < N$) nisjene inneholder residente sider, de resterende N-n sider som kan rulles ut på ytre lager (platelager) for å bli erstattet av andre. Dette skjer når systemet vil aksessere en side som ikke er i noen av de N nisjene i hurtiglageret. Se fig. 1.

For å avgjøre hvilken side som skal rulles ut fra hurtig-lageret, er det til hver side knyttet en "historikkvariabel" hvis verdi avspeiler sidens bruksfrekvens. Den av de N-n sidene som har lavest verdi for denne variabelen rulles ut. På denne måten oppnås at de sidene som oftest refereres har minst sannsynlighet for å bli rullet ut.

Av og til er det ønskelig å "låse" en side temporært fra å bli rullet ut etter regelen over, og dette kan gjøres fra systemprogrammet.

Det kan også nevnes at bare dersom en side er forandret fra den ble hentet inn til hurtiglageret blir den virkelig overført til ytre lager ved utrulling, en detalj som øker systemets effektivitet.



Hurtiglager Ytre lager

P forespørres og rulles inn,
X ut.
X er den side blant de N-n som har lavest verdi for 'historikk'.

Fig. 1

2.2 Generelt om systemet

Recordformat for elementene i ordboken fremgår i fig. 2. De enkelte betegnelsene

SCT: Neste elements seksjons-(side) nr.
DISPL: Adresse innen side for ditto.
SHC: Peker til neste logiske element innen denne side, hvis satt:
√ : Frekvens for dette element (type)
LGTH: Lengde av datafeltet.

Ordboken har altså form av en listestruktur, og denne bygges i sin helhet under lasteprosessen. For hver ny token gjennomsøkes listen for å avgjøre om det er en ny type eller ikke. I førstnevnte tilfelle settes denne inn i listen, ellers økes bare frekvens for angjeldende type med 1. Se fig. 3.

Av fig. 3 fremgår det hvordan SHC'ene kan øke hastigheten i søkeprosessen vesentlig; v.h.a. disse går vi aldri mer enn en gang gjennom hver side.

Record Format

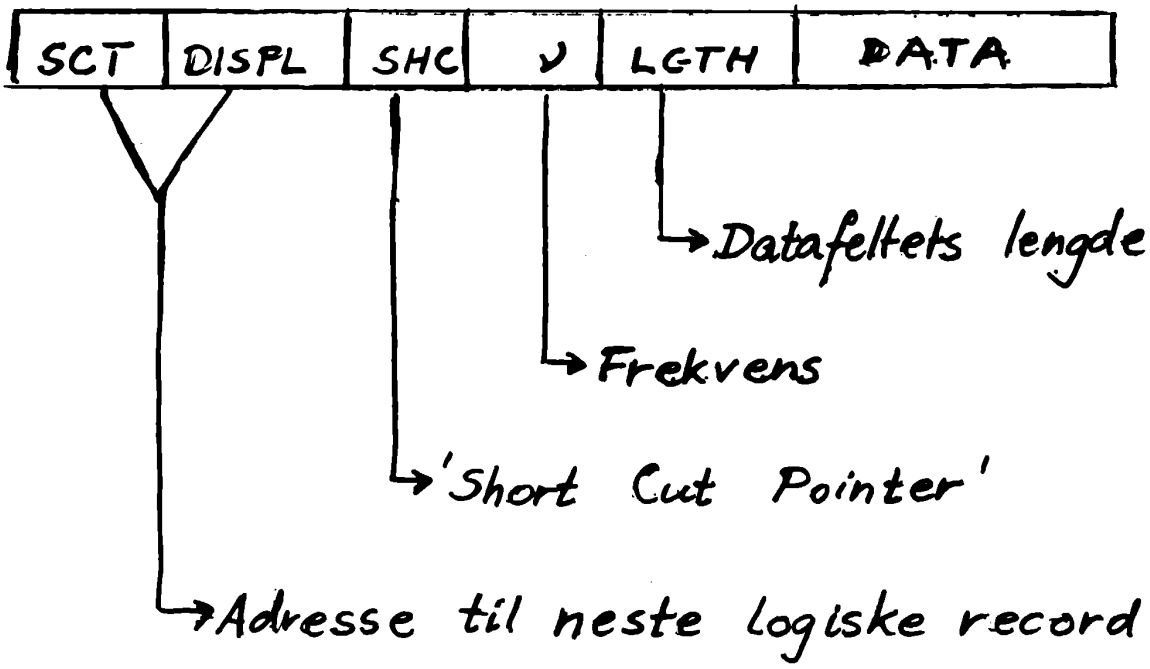


Fig. 2

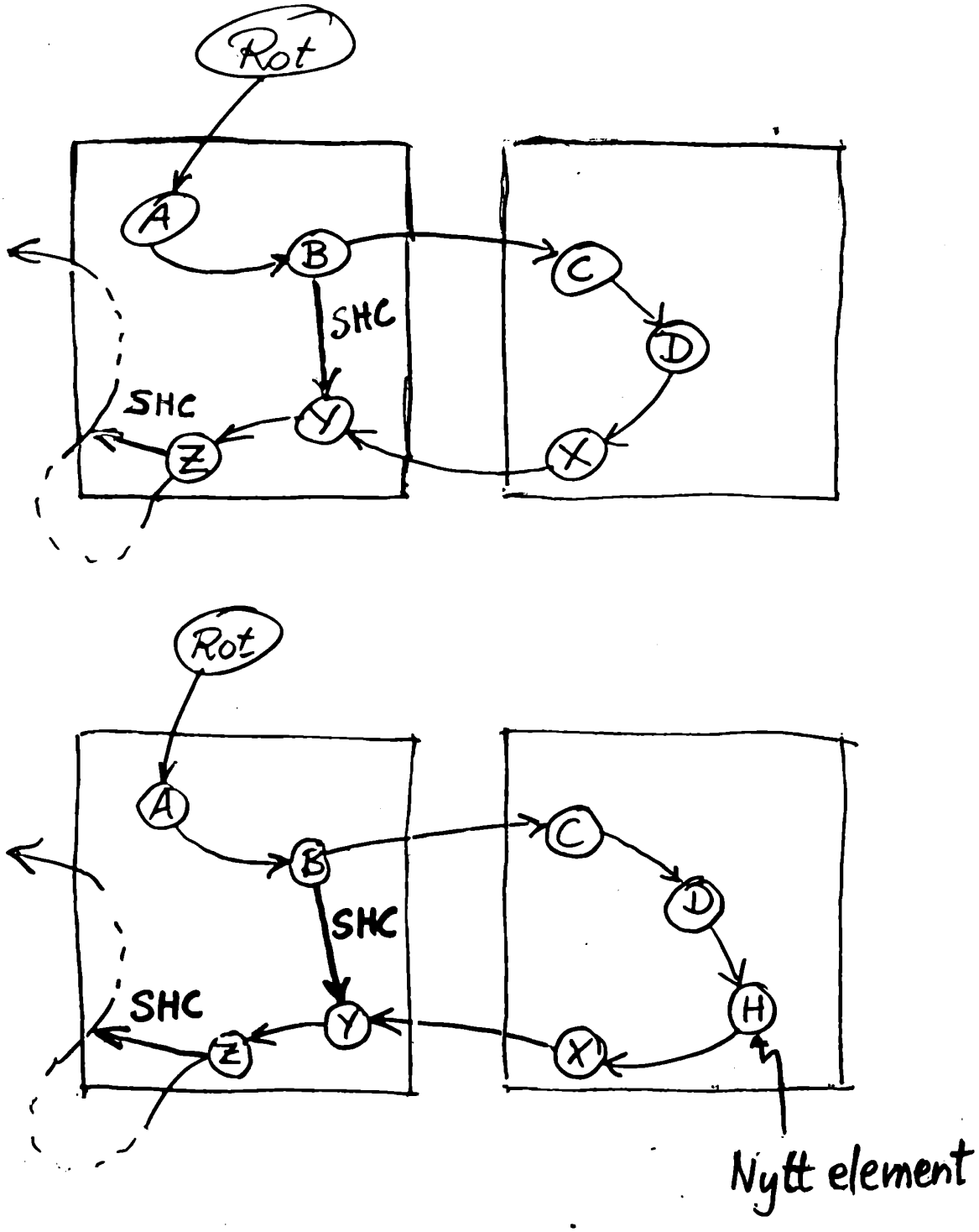


Fig. 3. Listebygging

2.3 Laste-strategi

Sidene kan under lasteprosessen være i ulike "tilstander":

- Nyeste side er den som sist er tatt i bruk.
- Backup siden inneholder logisk forgjenger til det elementet vi prøver å plassere.
- Nåværende side inneholder logisk etterfølger til samme.

Fig. 4 anskueliggjør dette.

Strategi for oppfylling av sidene, dvs. for plassering av nye typer var som vist under fig. 4.

For å gi en indikasjon på hvordan dette virker, henviser vi til fig. 5. Reglene anvendt for de nye elementene vil her være:

Regel 1 for C,
" 2 " D,
" 3 " M,
" 4 " B,

Hensikten med lastegrensen (se preferanselisten) som bare anvendes når en side er i tilstand "nyeste", er å øke sjansene for at regel 1 kan anvendes og derved for lange, ubrutte tråder innen en side. Dette medfører mindre inn-ut rulling og mindre overhead forbundet med short-cut pekerne.

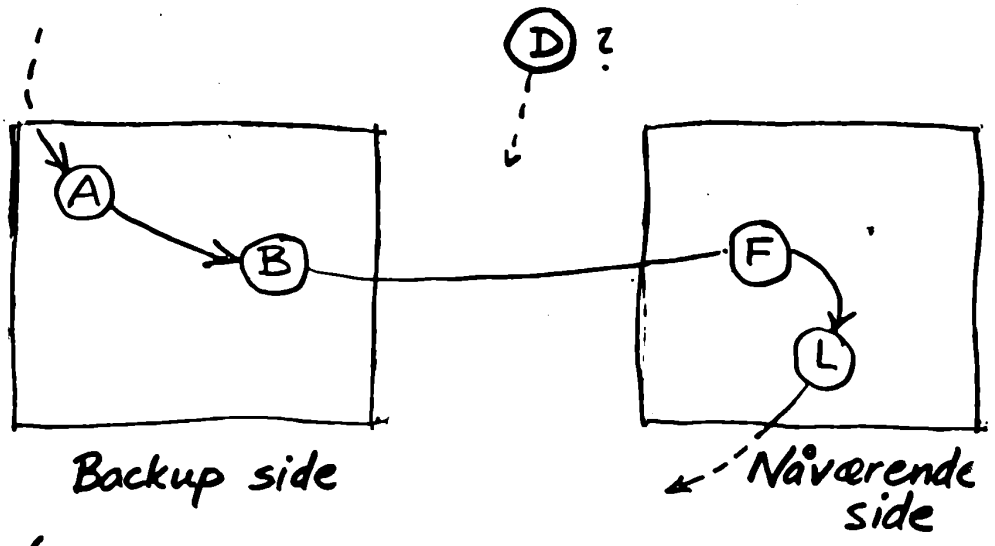
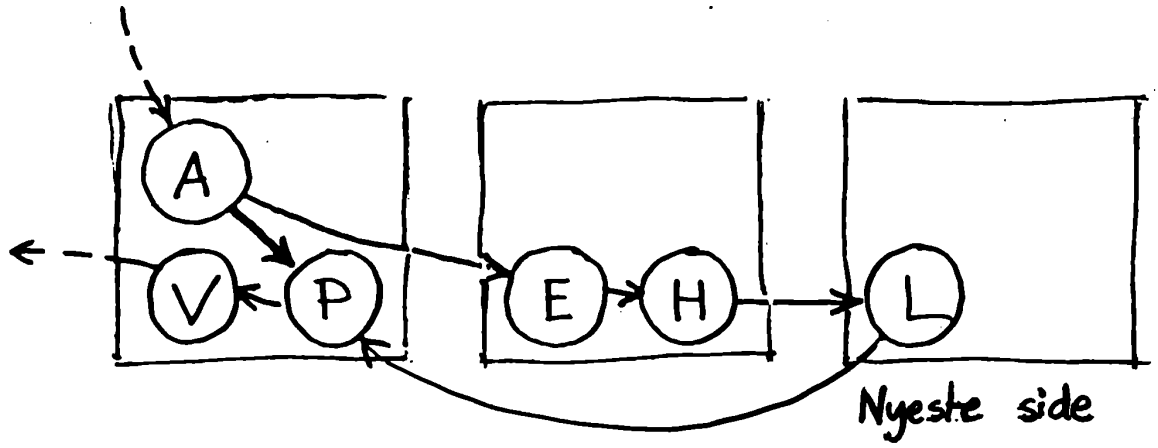
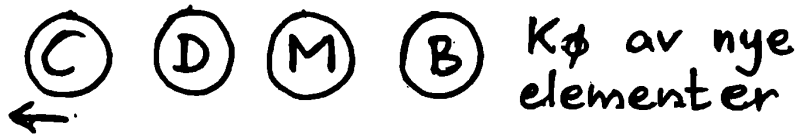


Fig. 4. Sidetilstander relativt (D).

Preferanseliste for plassering av element:

1. Backup, hvis plass
2. Nåværende, hvis plass
3. Nyeste, hvis "lastegrensen" ikke er nådd
4. Ta i bruk en ny nyeste





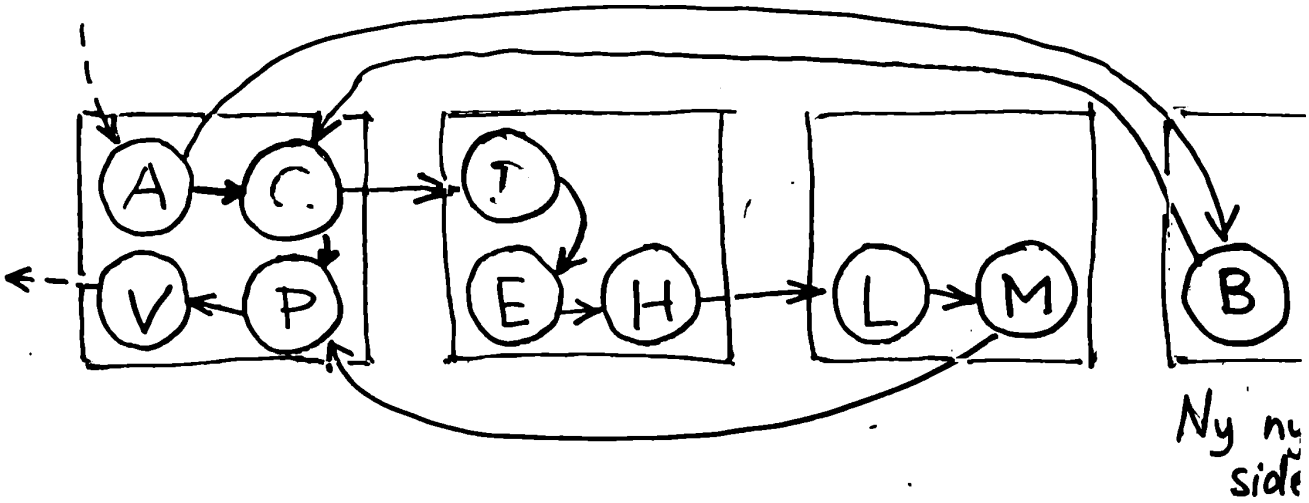


Fig. 5. Innsetting av nye elementer.
(Lastegrense = 0.5)

Dette vil imidlertid komme i konflikt med vårt ønske om å plassere flest mulig høyfrekvenstyper i de første sidene (spesielt i de residente), både for å få rask aksess til disse typene, og fordi de første sidene i stor grad bør virke som en indeks for resten av ordboken. For å oppnå det siste bør disse sidene fylles med så mange element som mulig spredd best mulig alfabetisk.

Vi valgte derfor å la fyllingsgraden variere som funksjon av sidenummer som antydet i fig. 6.

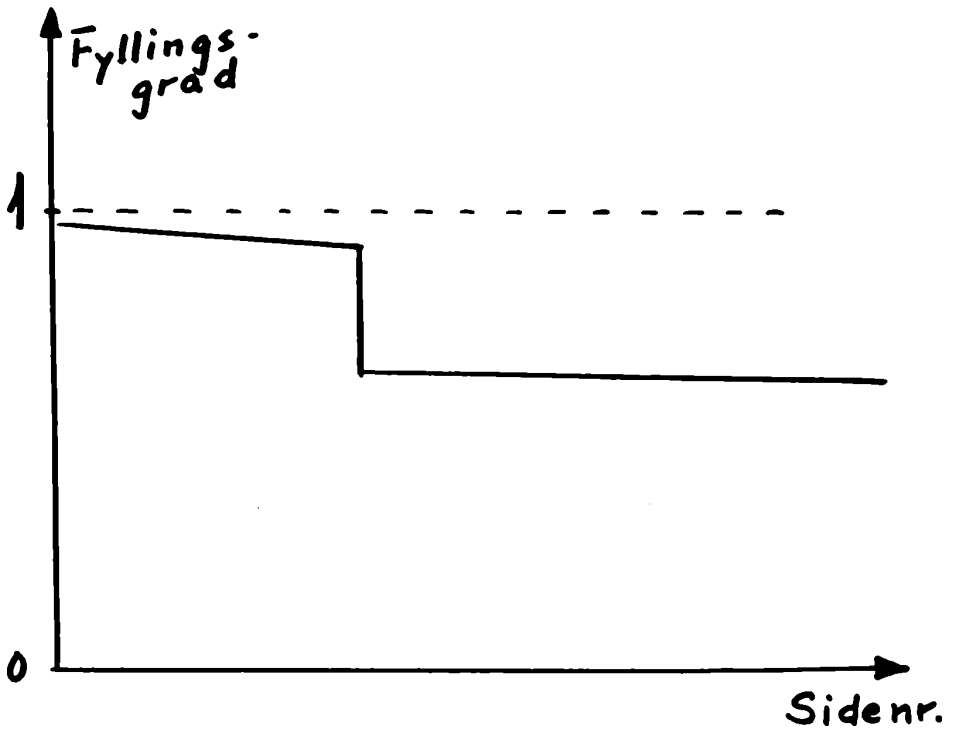


Fig. 6. Fyllingsgrad som funksjon av side nummer

3. Eksperimenter og resultater

Som tekstbase ble Garman & Worse av A. Kielland benyttet. Denne inneholder omtrent 65.000 grafiske ord og gir opphav til en ordbok med ca. 10.500 ord.

Vi fant at vår teknikk virkelig ga mange høyfrekvensord i de første sidene, og at teknikken med lastegrense og plassering av typer i henhold til preferanselisten førte til lange uavbrudte tråder i de senere sidene.

En del forsøk ble gjort for å finne et "godt" parametersett for systemet. Her skal bare effekten av partiell fylling av nyeste side påpekes. Allerede med fyllingsgrad litt under en fikk vi resultater som vist i fig. 7, og svært lave fyllingsgrader ga ikke ytterligere forbedringer.

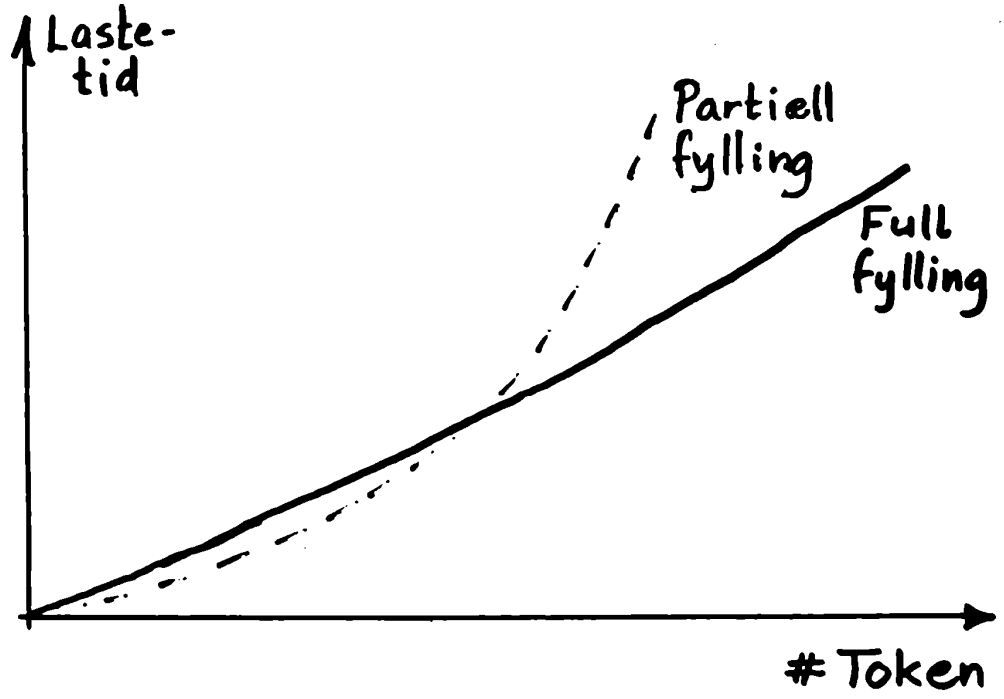


Fig. 7. Effekt av partiell fylling av "nyeste side"

Som vi ser av fig. 8 er vårt krav om en lastetid som er nær proposjonal med antall elementer i tekstbasen godt oppfylt. Fig. 9 viser en del andre data fra lasteprosessen, og vi ser at:

- TID/TOKEN vokser relativt lite og fremfor alt jevnt ettersom ordboken øker i omfang.
- PR/TOKEN som er sidereferanser pr. token forklarer mye av systemet "pene" adferd idet denne bare vokser fra 2,66 ved $M_{token} = 2 \cdot 10^4$ til 3.0 ved $M_{token} = 6{,}5 \cdot 10^4$ dvs. ved 11 %, mens ordboken i samme intervall øker fra ca. $4{,}1 \cdot 10^3$ types til ca. $10.5 \cdot 10^3$ types, altså med 156 %.
- Hvis vi ser på oppslagstid/token med full ordbok, var denne ca. 0.05 s., noe vi var vel tilfreds med, utstyret vi brukte tatt i betraktning. Brukes ordboken bare for oppslag, reduseres denne tiden vesentlig fordi vi slipper alle "skriveoperasjoner" knyttet til plassering av nye types. Senere tester indikerte et fall på omkring 60 % som bringer oppslagstiden ned på 0,02 s.

Alt i alt mener vi disse resultatene er oppmuntrende, og at en slik eller lignende design egner seg godt for konstruksjon og bruk av ordbøker på mindre datamaskiner som ikke tilbyr gode databasesystem eller tilsvarende avansert software.

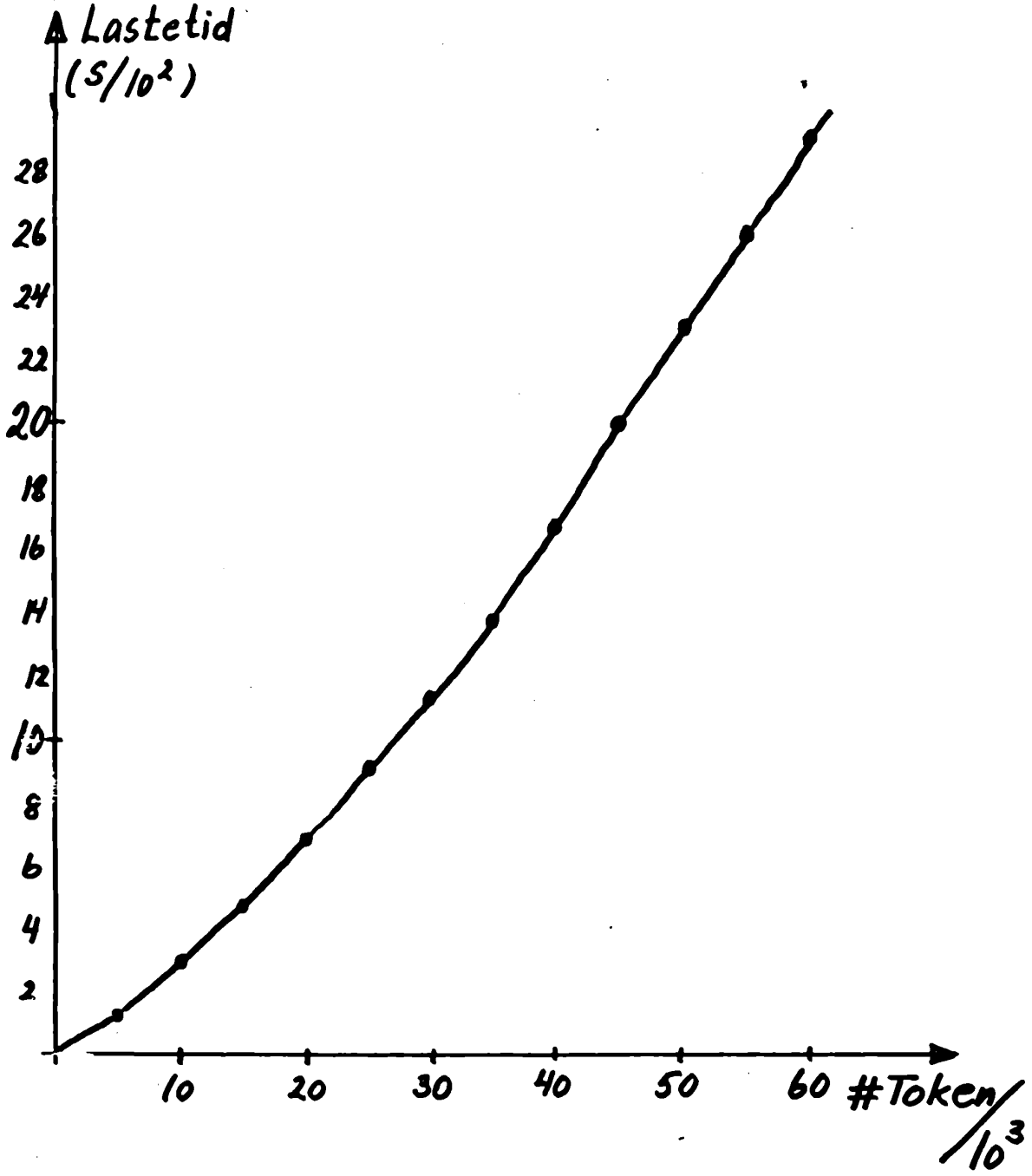


Fig. 8. Lastetid som funksjon av # token

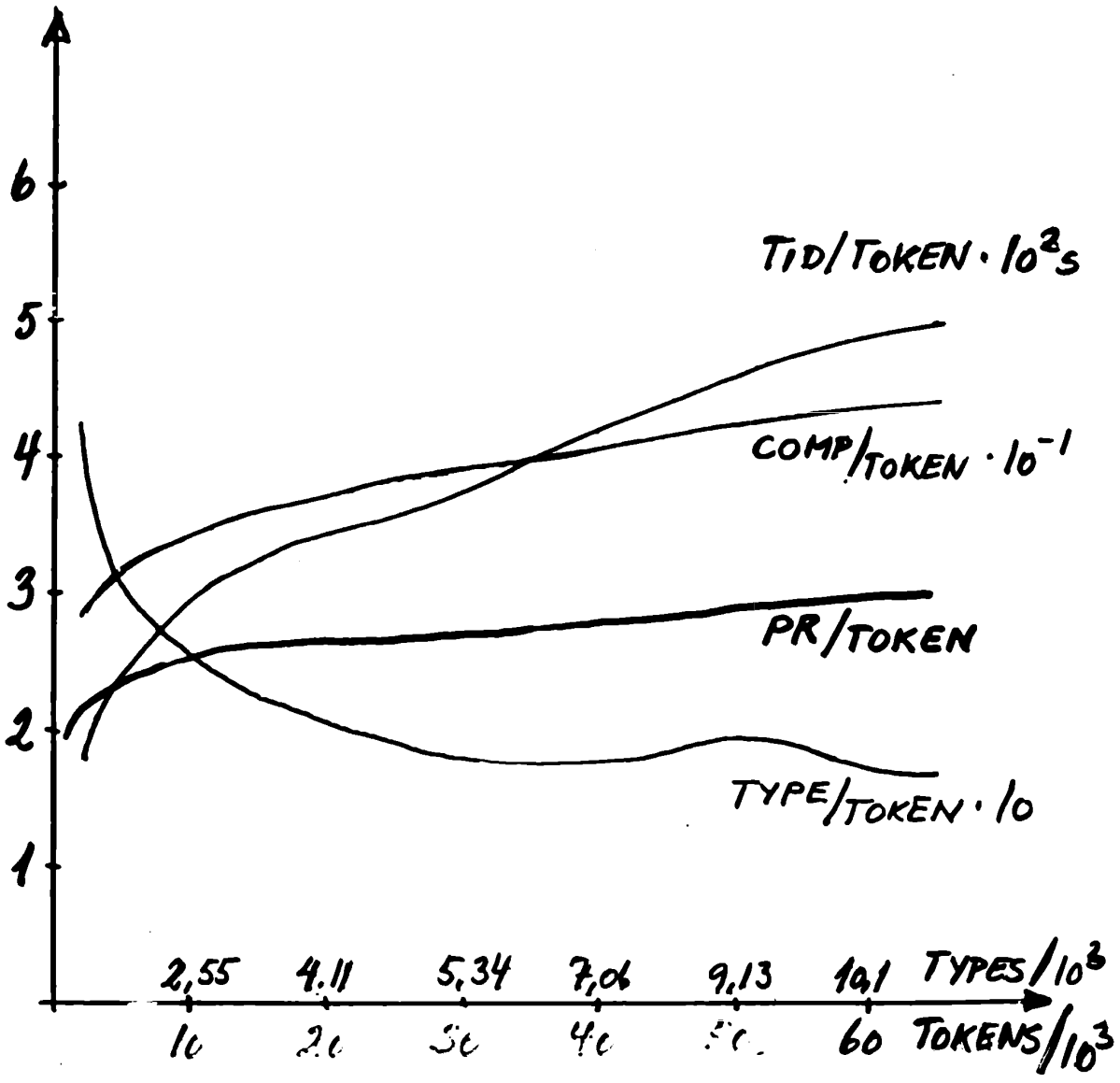


Fig. 9. Statistikk fra lasting av ordbok.

Østasiatisk Instituts EDB-arbejde, 1979.
Eric Grinstead.

Japansk poesi (waka): Test of small sample (600 poems).
Output to paper tape and Katakana type-head.
Japanese can now be typed directly on the data screen, so enquiries are being made to have output in Japan.
Researcher: Yoichi Nagashima.

Kinesisk maleri, 1949-1979. The changes in China are seen in the paintings reproduced in periodicals, albums, etc. 4000 titles (will go to 7000 eventually), adaptable to MARC fields.

Researcher: Simon Heilesen.

Binomer i kinesisk: The "word" in Chinese is usually a binome, i.e., two Chinese characters. A large dictionary includes all classical references, all well-known poetry, placenames, pen-names, etc. The project is designed as a statistical survey, but could save time in looking up characters, as modern Chinese is now written with simplified characters, and systems of arranging the characters go out of date.

The unit entry is: First character, second character, and location. "Fire" is the 99th compound with "heart".

心 (hjerte)	火 (ild)	(location)
10295	18850	99

The first two numbers can be placed in a single 36-bit word and the location used for additional information gained from statistical work. I would like to add a figure for rank even before the whole dictionary is coded.

As the dictionary I am using is a Chinese-Japanese dictiona the results are of interest to the Japanese, and some cooperation is envisaged.

Coded to date: One-third of the dictionary (12 vols).
Final total: ca. 288,000 units.

Researcher: Eric Grinstead.

PROJEKTET ENGELSKT TALSPRÅK

Cecilia Thavenius

Projektet Engelskt talspråk i Lund, eller som det heter på engelska, Survey of Spoken English (SSE), är ett dotterprojekt till Survey of English Usage vid University College London. Projektet stöds sedan 1975 av Riksbankens Jubileumsfond. Projektledare är Professor Jan Svartvik.

MATERIAL

Projektets material är den s.k. 'London-Lund Corpus'. Den består av ca 1/2 miljon ord av talad engelska i ett flertal olika situationer. Materialet är inspelat, ortografiskt transkriberat, och prosodiskt och paralingvistiskt analyserat under ledning av Professor Randolph Quirk vid University College i London. De s.k. texterna av engelskt talspråk består av följande kategorier:

Material with origin in speech (100 "texts")

A Monologue (24)

- Prepared (but unscripted) oration 6
- Spontaneous
  - oration 10
  - commentary
    - sport 4
    - non-sport 4

B Dialogue (76)

- Conversation
  - surreptitious
    - intimate 24
    - distant 10
  - non-surreptitious
    - intimate 20
    - distant 6
  - telephone
    - intimate 10
    - distant 6

Talarnas identitet är skyddad, så vi har inte tillgång till deras namn. Däremot finns en del information om dem såsom yrke, ålder, kön och typ av relation mellan talarna. Men den djupare typ av socio-ekonomisk information, som t.ex. finns för Montrealkorpusen av talad franska saknas för detta material. När insamlandet av London-Lund korpusen påbörjades för ca 20 år sedan, hade man ännu inte börjat diskutera vikten av sådan kunskap för konversationsanalys.

Den prosodiska och paralingvistiska analysen, som helt har utförts vid Survey of English Usage i London, har varit oerhört tidskrävande. En timmes inspelat tal har tagit ca 80 timmar i anspråk för analysen. När analysen avslutats har materialet blivit liggande i London. Givetvis har det blivit utnyttjat för forskning av lingvister, som kommit till London och suttit där en tid och arbetat med materialet. Men tillgängligheten och användbarheten är inte särskilt stor när man tänker på det enorma arbete, som ligger bakom.

SYFTE

Projektet Engelskt talspråk har tre huvudsyften:

1 att med hjälp av dator göra materialet tillgängligt och sprida det till alla intresserade

2 att förse materialet med grammatisk 'tagging' för att möjliggöra en intressantare sortering, som går ovanför en ren lexikalisk ordnivå, och ge större kunskap om det engelska talspråkets grammatik

3 att producera forskning och forskningsresultat på grundval av materialet för att klargöra hur talspråket ser ut; flera doktorsavhandlingar är under arbete, och en del mindre studier har gjorts; dessutom finns planer på en lärobok i engelskt talspråk för skolor och universitet i Sverige, och en större talspråksgrammatik för internationellt bruk

ARBETSGÅNG

Efter fyra års intensivt arbete har vi nu kommit så långt att vi dels har ett färdigt magnetband med konversationstexterna, dels en bok, datasatt efter bandet, som beräknas komma ut i slutet av 1979. (A Corpus of English Conversation, ed. by Jan Svartvik and Randolph Quirk, Lund Studies in English, Lund: Gleerups/Liber, 1979). Bandet och boken omfattar alltså inte hela talspråksmaterialet, utan endast vad vi ansett vara den intressantaste delen, nämligen 'face-to-face conversation', inspelad med dold mikrofon. Detta material omfattar ca 170.000 ord, vilket utgör 34 prosodiskt markerade texter à 5000 ord. Resten av korpusen kommer att finnas tillgänglig på magnetband.

Vi fick materialet från London i form av maskinskrivna A6 kort. Vi gick först igenom det stora materialet flera gånger, redigerade bort en del paralingvistisk notation, som av tekniska skäl var svår att ta med, och påbörjade en ännu inte avslutad korrespondens med SEU i London för att kontrollera inkonsekvenser eller vad vi misstänkte var rena skrivfel. Själva har vi inte haft möjlighet att göra en sådan kontroll, eftersom vi inte har haft tillgång till ljudbandsmaterialet, och dessutom hållit på den principen, att eftersom analysen gjorts i London måste ändringar i denna komma därifrån. Men de allra flesta av våra frågor har visat sig relevanta, och vi har på detta sätt kunnat rätta ett par tusen fel i det ursprungliga materialet.

Därefter har vi kört in hela materialet på magnetband från bildskärmsterminal, som är ansluten till Lunds Datacentral. Datacentralen har sedan producerat utskrifter åt oss. Dessa utskrifter har sedan kontrollästs mot original av minst tre och i de flesta fall fyra personer. Sedan har materialet rättats efter varje korrekturläsning via bildskärmsterminalen. Varje text har körts ut och kontrollästs sex gånger, och vi hoppas nu att vi kommit så nära fulländningens stadium som möjligt. Men det har alltså tagit fyra år. F.n. håller Liber

Tryck i Stockholm på med datasättning efter magnetbandet. Programmering och provkörningar är nu klara, och vi håller på med korrekturet. Boktexten kommer att få detta utseende:

> A 45 ΔOld and Middle ₐEnglish GRAPHÓLOGY■ 47 «or ‖something ☆like ΔTHÀT■ 48 ‖you SÉE■»☆

B 49 ☆‖WÉLL■ 50 ‖you give☆ them the ₐLÒT {‖you SÉE■} ■ ☆ · ☆ 51 ‖that's the ☆☆PÓINT■☆☆ 52 «and» ‖make sure that there's ₐSÒMETHING■ 53 [ə:] ‖fairly ₐCLÒSELY RELÁTED■

A 54 ☆‖[m̀]■☆ 55 ☆☆‖[m̀]■☆☆

> B 56 «to ‖what they've STÙDIED■»

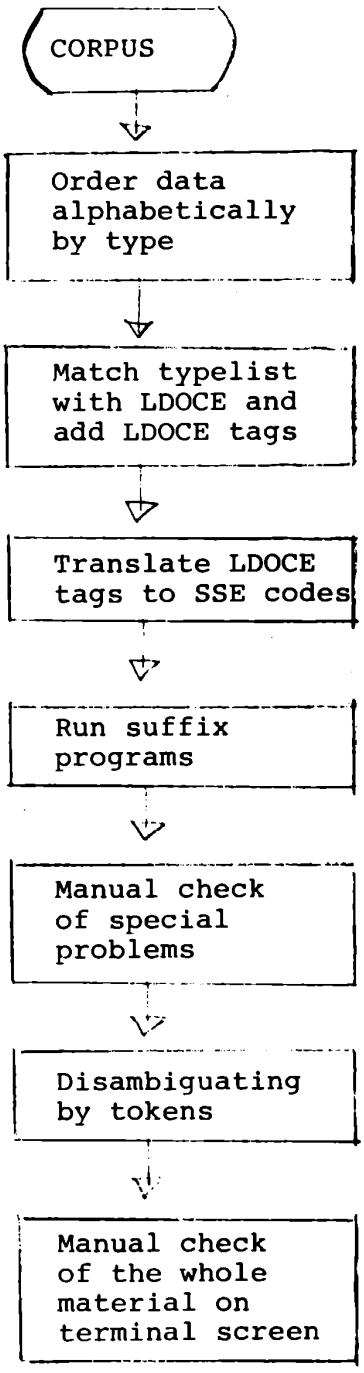
A 57 «it's ‖just 'one ΔQUÈSTION that they have to do ÍSN'T it■»

B 58 ‖well there were [ə] ΔÓNE■ 59 or ‖TWÒ we've ☆got on THÉRE■☆ 60 ‖you SÉE■

A 61 ☆‖yes ΔÍ SÈE■☆ · 62 ‖YÉS■ · 63 ‖YÉS■ - 64 [ə:m] · ‖one ÒTHER thing SÁM■ - 65 [ə:m] - ‖DEΔLÀNEY■ - 66 a CA‖NÁDIAN■ 67 ☆«who» ‖graduated☆

B 68 ☆«[ə] ‖WHÉRE did you put those THÌNGS■☆ 69 ‖just one» · ‖let me put this in my BÀG■ 70 «or» I'll «‖walk AₐWÀY with▷out it■» - - -

Vårt andra huvudsyfte var att förse det prosodiskt markerade materialet med grammatisk tagging. Där har vi efter ett långt förarbete och många och långa diskussioner kommit fram till följande procedur av ordklassanalysen. Longman Dictionary of Contemporary English från 1978, det s.k. LDOCE, finns på ett magnetband, som vi erhållit från Longmans. Vi kommer nu att utnyttja LDOCE-bandet för att få en så automatisk märkning som möjligt av vårt material. Det intressanta med detta tillvägagångssätt består i att man kan ta reda på hur långt man kan komma med automatiska metoder. Enkelt beskrivet blir gången följande:



Word class analysis

Materialet sorteras alltså först i en alfabetisk ordtypslista. Därefter matchas ordtypslistan med LDOCE och får LDOCE's grammatiska information, som sedan översätts till våra egna koder. Sedan körs suffixprogram på ord som slutar på -s, -er, -est, -ed, -ing. Vi kontrollerar sedan vissa problematiska kategorier manuellt i ordtypslistan. Därefter går maskinen över till att arbeta med den löpande texten och disambiguerar i de fall ord har fått mer än en 'tag'. Disambigueringsprogrammen grundar sig på sådant som position och statistik. Allra sist sker en manuell interaktiv kontroll och bearbetning vid terminal.

Efter ordklassanalys går vi över till 'phrase-tagging'. Tillvägagångssättet blir då att maskinen går igenom varje s.k. 'tone unit' i det löpande materialet, och söker från höger till vänster inom 'tone uniten' efter fraserna i följande ordning: verbfras, adverbfras, adjektivfras, nominalfras, prepositionsfras. Även 'phrase-tagging' fasen avslutas med en manuell kontroll vid terminalen. I mån av tid och ekonomiska resurser ska sedan en analys av satsfunktioner utföras.

Eirik Lien.

DEMONSTRASJON AV PP*TT - EN PROGRAMPAKKE FOR KVANTITATIV TEKSTANALYSE

Nordiske datalingvistikdage,
København 9. - 10. oktober 1979

En god del av de kvantitative analyser og beregninger som en kan få en datamaskin til å hjelpe seg med, er svært like fra prosjekt til prosjekt. Det vil i de fleste tilfeller dreie seg om å lage listeprodukter, slik som alfabetisk ordliste (forlengs og baklengs sortert), frekvensordliste og konkordans. Sett fra en datamaskins synspunkt ligger forskjellen mellom to tekster i at de har forskjellig lengde.

Denne oppdagelsen har mange gjort, og det har igjen ført til at det rundt omkring ved humanistiske forskningsmiljøer er blitt laget en del program som produserer slike lister. Disse er gjerne i utgangspunktet laget for en spesiell tekst, men nettopp fordi forskjellen til en annen tekst ligger bare i lengden av dem, har en gjort programmene mer fleksible ved å bygge inn denne forskjellen ved hjelp av parametre. Etter hvert som behovet for nye analyser har økt, har en med utgangspunkt i det en har, lagt til nye program som fortsetter videre der det gamle programmet stanset.

Universitetet i Trondheim er ingen unntakelse. Ganske snart etter at jeg startet edb-tjenesten, så jeg behovet for slike program. I stedet for å starte med et bestemt prosjekt, startet jeg med å tenke etter hvilke analyser det kunne være behov for - og kunne derfor planlegge hele strukturen under ett.

Nærværende programpakke er derfor ment som et eksempel på hvordan systemet kan struktureres - den gjør alldeles ikke krav på å være genial og enestående. Jeg er fullstendig klar over at tilsvarende programpakker fins det mange andre steder, programpakker som gjør minst like god jobb.

Programpakken slik den er nå, gir listeprodukter på tre nivåer; ordnivå, stavingsnivå og grafemnivå. På ordnivå kan en få disse listeproduktene:

- initialalfabetisk ordliste med frekvens
- finalalfabetisk ordliste med frekvens
- frekvensordliste

- konkordans
- lister basert på ordlengde

På stavingsnivå:
- alfabetisert stavingsliste med frekvens
- frekvenssortert stavingsliste
  Initiale og finale stavinger kan, med parameter, i begge disse programmene få spesielle tilleggstegn som markerer denne funksjonen og dermed skiller dem fra andre, ellers like stavinger.
- liste over stavingstyper, representert ved sifre, nemlig antall bokstaver i første konsonantgruppe, antall vokaler i stavingskjernen og antall konsonanter i andre konsonantgruppe.

På grafemnivå:
- "konkordans" - eller klic-liste ("key-letter-in-context")
- opptelling av hvilke tegnkombinasjoner som fins i teksten.
- frekvenssortering av disse kombinasjonene

Ved et referansesystem som går både på side og linjenummer og kapittel, avsnitt, periode og ordnr. i perioden, har en direkte referanse tilbake til teksten og kan f.eks. kontrastere utvalgte kapitler mot hverandre.

Systemet er hovedsaklig laget for satsvise kjøringer ("batch"). Den delen av programpakken som deler ordene opp i stavinger, er imidlertid laget for interaktiv bruk slik at en kan korrigere de forslagene til oppdeling programmet gjør.

Systemet er i laget slik at en med parametre kan velge hvilke tegn en vil oppfatte som skilletegn. En har også mulighet for å justere utskriften ved noen av listene og gi ekstrainformasjoner i enkelte tilfeller.

Programpakken er skrevet i NU ALGOL og går foreløpig bare på Universitetet i Trondheims UNIVAC-anlegg. Dette programmeringsspråket gjør at det ikke er særlig distribusjonsvennlig. Det har vært i drift siden høsten 1975, men har i år fått en del nye tillegg.

Videre planer

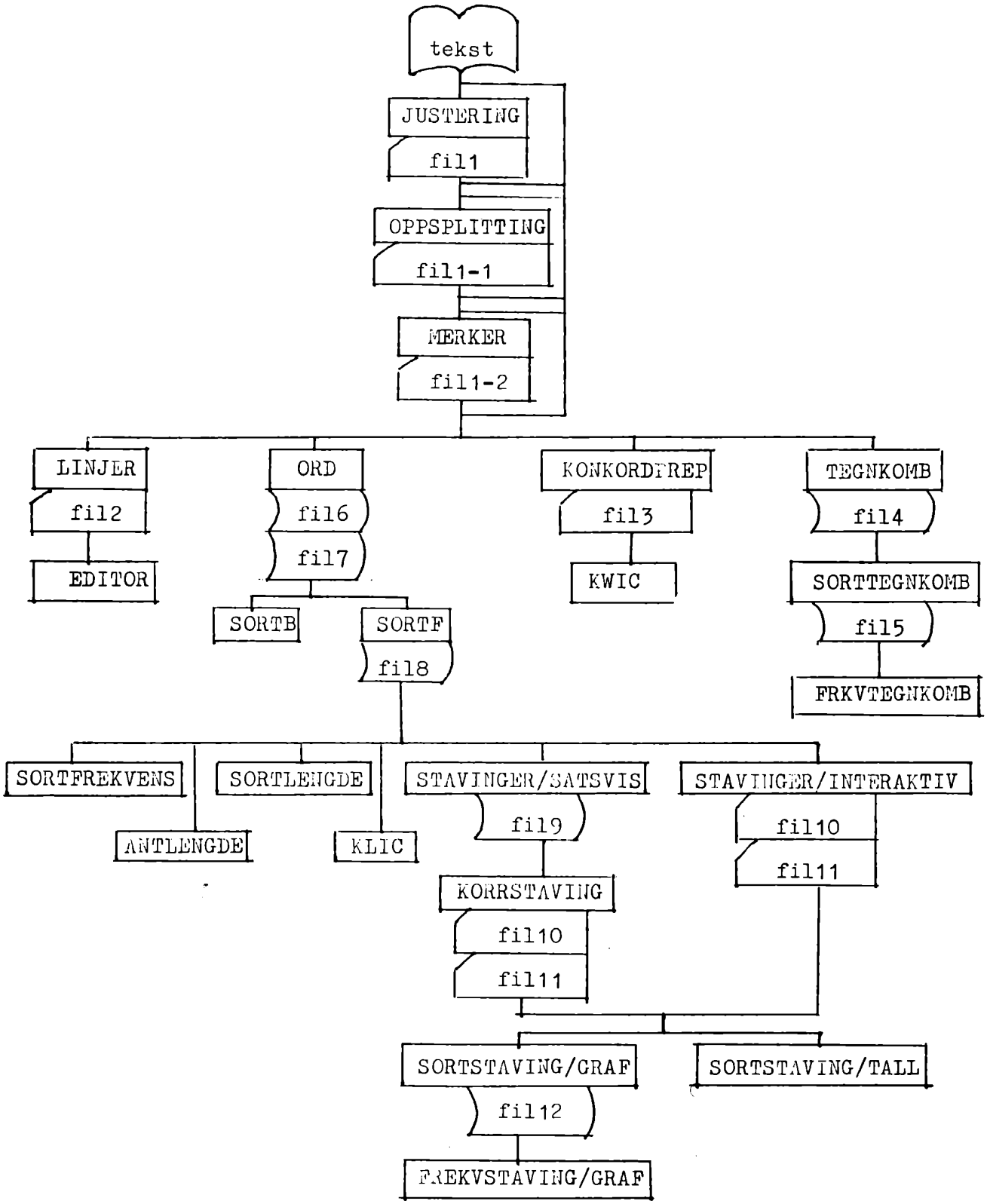
Det kan bli aktuelt å skrive om programmene til PASCAL, som ser ut til å bli det programmeringsspråket som nå slår igjennom.

Konkordansprogrammet bør bli mer fleksibelt ved at konteksten kan varieres, f.eks. til å omfatte den perioden ordet står i. Jeg har også planer om å arbeide inn program for analyse på frasenivå. Dessuten bør programmene for lemmatisering komme med i en slik programpakke. Likeså bør muligheten for å behandle merket ("tagget") tekst være med. Derimot vil nok analyse på setningsnivå ligge et godt stykke fram i tida.

Den metoden som er brukt, nemlig å ta vare på resultatene på filer, gjør det enkelt å innarbeide nye program fordi disse kan arbeide videre fra de filene som allerede er generert.

Trondheim 1979-09-17

Appendiks 1
Strukturen i PP*TT



EL/1979-10-03

BIBELTEKST ON-LINE

P R O S J E K T R A P P O R T

Nr. 1

for Bibel-prosjektet

12. november 1979

Per-Bjørn Pedersen Michael Gillow

Bjørn Ride

1. INNLEIING

Våren og sommaren 1976 vart det ved Nordisk institutt, Prosjekt for datamaskinell språkbehandling (PDS), Universitetet i Bergen arbeidd med dei nye omsettingane av Det nye testamentet for å gjere dei datamaskinelt tilgjengelege. Arbeidet baserte seg på trykkeriband og målet var å lage KWIC-konkordansar og ordlister av eit nytt og viktig materiale til språkvitskapleg bruk. Det Norske Bibelselskap (DNB) hadde gjeve si velsigning til dette og late oss få kopiar av trykkeribanda - eit avgjerande grunnlag for i det heile å tilrettelegge materialet innan rimeleg tid. DNB gjorde seg nytte av våre produkt ved å ta ut selektive lister til arbeidsgrunnlag for fastlegging av den endelege språkdrakta for nyomsettingane av heile Bibelen.

Hausten 1978 kom heile Bibelen i den nye omsettinga både på bokmål og nynorsk. Det reiste med eit gong spørsmålet om nye bibelordbøker. Frå språkvitskapleg hald vart det uttrykt sterke ønske om å gjere dette store og aktuelle språktilfanget tilgjengeleg for nærare studium og etterrøking. For båe oppgåver ville det ut frå 1976-løynslene vere naturleg å gå vegen om datamaskinene for å gjere tilfanget tilgjengeleg og brukeleg. Dette førte til at det vart gjord avtale mellom DNB og PDS om gjennomføring av eit liknande prosjekt for heile Bibelen som det som vart gjort med Det nye testamentet i 1976.

Dette arbeidet kom i gang fyrst i april i år og dei fyrste resultata ligg føre.

2. PLANLEGGING

Listene våre i 1976 viste at kvart nytestamente innehaldt om lag 170 000 laupande ord (tokens). Eit grovt overslag for heile Bibelen gjev eit tilfang på nær 3/4 million ord.

Med våre 76-røynsler ville det bli ein enorm jobb å lage konkordansar og lister av ymse slag for ein heil bibel med det programsystemet vi då bygde opp. Eit program-system som vandra langs tradisjonelle KWIC-vegar og som var like effektivt som det meste tilsvarande så langt vi kunne dømme. Preet ved 76-systemet var at vi visste det fungerte. Vi kunne såleis komme i gang med produskjon nokså snart medan eit nytt opplegg ville krevje grundig planlegging og vurdering og ei uviss utviklingstid. Med dei store datamengdene dette nye prosjektet innebar fann vi det likevel viktig å leite nøye og grurdig etter nye vegar å gå.

Fleire modellar vart skisserte av arbeidsgruppa: Michael Gillow, Per-Bjørn Pedersen og Bjørn Fide med leiar for PDS - Kolbjørn Heggstad - som formell leiar.

Det einaste som i utgangspunktet var klårt var at vi måtte lage eit fleksibelt og dynamisk system for lesing av trykkeriband. Allereide i 1976 hadde vi sysla med tankar kring denne problematikken. Det vi då laga til vart eit einfelt lese-/styre-program med ein utskriftsdel og ein primitiv tilstandsgrammatikk. Sjølv i den primitive forma systemet fekk i 1976 hadde det vist seg som tenleg og brukeleg. Denne delen fann vi difor nyttig å byggje vidare på i vårt nye arbeid. Tilstandsgrammatikken er blitt utvikla til ein svært fleksibel og allmenn reiskap som også - etter vårt syn - er lettfatteleg og bruksvenleg. Men den gjev høve til å byggje så kompliserte strukturar og mønster at det vart heilt naudsynt å byggje inn i systemet gode og fleksible avlusings- og utprøvings-mekanismar.

For brukar-sida såg vi det viktig å lage eit interaktivt system som i stor grad skulle vere parameter-styrt og som skulle gje høve til å lage fleire ulike typar lister ut frå same grunndata, med eit absolutt krav om gode høve til å produsere KWIC-konkordansar etter behov utan fyrst å måtte lage fullstendige konkordansar for heile materialet.

Med slike krav for eit ny-opplegg var det sjølvsagt nærliggjande å tenkje på å bruke database-systemet DMS1100. Men ved vårt universitet var dette utelukka med dei vilkår det er for bruk av DMS1100 i dag. Vi måtte difor sjå oss om etter noko anna. Vårt val vart då å byggje vårt eige skreddarsydde database-system ved hjelp av såkalla MSAM-filar som er tilg-

jengeleg frå og med COBOL-74.

Det var allereide i utgangspunktet semje om å bruke COBOL som programmeringsspråk fordi det var det språket som troikaen hadde best fellles kjennskap til.

Som nemnt ovanfor tok arbeidet for alvor til i april. I fyrste fase - frå april og fram til juli - gjekk utviklinga av systemet for lesing av trykkeri-data og fastlegginga av brukar-delen parallelt. Det vesentlege av mai månad vart brukt til å skissere og prøve ut tre-fire ulike modellar. Modellane bygde alle på MSAM-filar, men i ulik grad og omfang. Det vart gjort prøver med splitting av data i fleire og færre filar, lagring av ekte data eller i ei eller anna form for numerisk representasjon, skiljing av indeksstruktur og datastruktur etc.

Modellprøvingane gav som resultat at vi i fyrstninga av juni stod att med ein modell som syntes rimeleg og brukeleg for store datamengder. Den baserte seg på MSAM-filstruktur, men alt samla i ein fil - både data og indeks, og med data i ei blanding av ekte og numerisk representasjon.

Dei prøvene vi gjorde i juni viste at plassbehovet ville vere om lag ein tredjepart av det vårt gamle KWIC-system trong, og då hadde vi data som kunne bere i seg all informasjon frå trykkeri-bandet, men lagre denne informasjonen berre ein gong og såleis gje gode høve til andre utlistingstypar enn KWIC.

Å lage ein fullstendig KWIC-konkordans av materialet ville - frå våre prøver - truleg koste om lag det same ressursmessig som i vårt 76-system, men selektive konkordansar og andre typar lister ville koste svært mykje mindre. Dellister og seleksjon er det som brukarane spør etter. Dette var også den slags ønske som DNB fyrst og fremst hadde. Dessutan ytra både dei og språkfolk sterke ønske om at systemet måtte gje høve til tagging av tekster. I systemet er også dette bygt inn, og utan at det går på kostnad av plass eller tid for oppretting av grunndata. For kvart laupande ord er der høve til å setje på fleire millionar taggar.

Eit viktig, men hittil unemnt krav for eit eventuelt nytt programopplegg var at det burde vere bra høve til oppdatering/neddatering av data. Også dette har vi greidd å passe inn i systemet utan at det har gått på kostnad av tid eller plass for oppretting av grunndata.

Av dei mindre gode sider ved systemet kan nemnast: frekvensen for eit ord kjem vi fram til ved å trikse litt under lasting av data og under produksjon. Dette skuldast den måten MSAM-filane blir bygde opp på. Vi kunne ha gjort frekvensen for kvart einskilt ord lett og direkte tilgjengeleg, men det ville ha kosta etter måten mykje i tid og plass under op-

pretting av grunndata. Vi gjekk difor frå ein slik måte og ordnar det ved litt fusk. Ein annan litt bakvend ting som for brukaren vil verke uvant er at ved utlisting vil dei orda komme fyrst som er blitt lagra sist. Dette kan vi nokolunde lettvint bøte på ved å laste grunndata på ein litt annan måte, noko som vil koste lite eller ingenting.

3. OVERSYN OVER PROGRAM-SYSTEMET

Som ein vil sjå av diagrammet i Tillegg 1 har vi funne det naturleg å dele det i to: den delen som les trykkeribånd og den delen som legg opp utdata frå trykkeribanda i den endelege MSAM-strukturen og som driv produksjon på grunnlag av MSAM-data. Fyrste delen har vi kalla CROMWELL-delen og den andre EINSTEIN-delen. Heile systemet har vi nemnt MONOLITT fordi det endelege lagringsformatet baserer seg på berre ein fil i motsetnad til dei fleste andre modellane vi laga. Data-formatet ut frå CROMWELL har vi kalla GRANAT (GRafisk ANAlysert Tekst), og lagringsformatet BRUTE (BRUks TEkst).

Noko nærare inn på namngjevinga skal vi ikkje gå her, men den ville kunne fortelje ein del om korleis vi har tenkt og arbeidd. Alle namn spring direkte ut frå idear og røynsler undervegs.

3.1. CROMWELL-DELEN

Diagrammet i Tillegg 2 syner hovudkomponentane.

Dei to faste delane er styreprogrammet CROMWELL og parsar-programmet PURITAN. I tillegg finst dei applikasjonsavhengige: FORM og grammatikken. Grammatikken fortel korleis trykkeribandet skal tolkast og FORM tek seg av skrivinga av GRANAT, dvs. kva data frå trykkeribandet som skal gå vidare til grunndata.

I CROMWELL blir løsinga av 'råteksten' styrt ved hjelp av ein kommandostruktur. I denne strukturen er det bygt inn gode høve til avlusing/debugging. I Tillegg 3 er eit oversyn over dei kommandoar som no finst i CROMWELL. Kommandoar for testing av grammatikken vantar til dels, men skal byggjast inn.

3.1.1. PURITAN og grammatikk

Systemet inneheld eit analyseprogram for 'råteksten' som t.d. kan finnast på eit trykkeriband med alt det grafiske 'rusket' som er med der. Det er dette programmet som vi har kalla PURITAN. Det er eit generelt analyseprogram, som er styrt av ein grammatikk. Det leverer frå seg dei tekststykka som grammatikken definerer. I Bibel-prosjektet er det grafiske ord. Tekststykka vert leverte med den koden som grammatikken bestemmer. Dei syntagmatiske eigenskapane som tekststykket har kan takast vare på i ein bit-streng som vert levert med kvart tekststykke frå analyseprogrammet. I Bibel-

prosjektet er tekstegenskapane til dei grafiske orda t.d.: stor bokstav, hermeteikn (føre og etter), komma og spørjeteikn.

Grammatikken er sett saman av eit tal med tilstandstabellar. Talet på tilstandstabellar, dvs. talet på paragrafar i grammatikken, vil skifte etter kor innfløkt råteksten er, og etter kva analysestrategi ein må fylgje. I Bibel-prosjektet greier det seg med godt under 100 paragrafar.

Kvar grammatikk-paragraf er sett saman av liner. Logisk sett har kvar kode i råteksten si grammatikkline i alle paragrafane. Grovt sett har ei grammatikkline fylgjande format:

1) Inn-kode
2) Inn-attributt
3) Ut-kode (=< 2 teikn)
4) Ny tilstandstabell
5) Ut-attributt

'Inn-kode' er den verdien som finst i 'råteksten', kode A. 'Inn-attributt' er eigenskapar ved denne kodeverdien som påverkar sjølve analyseprogrammet, PURITAN, t.d. varsel om retur eller at kodeverdien skal brukast som 'inn-kode' ein gong til for ein ny grammatikk-paragraf eller kontroll av 'push-down' stakk. 'Ut-kode' er den verdien (dei verdiane) som finst i 'uttekst', kode B. 'Ny tilstandstabell' er nummeret til den grammatikk-paragrafen som skal brukast av PURITAN for neste kode. 'Ut-attributt' påpverkar ikkje PURITAN, men vert levert vidare i bit-strengen som attributt til 'uttekst'.

3.1.2. Nokre eigenskapar ved grammatikkar

Grammatikken er uavhengig av programmet. Med nye grammatikkar kan vidt ulike råtekst-format analyserast utan at program treng skrivast om. Grammatikken blir lagra som ein eigen datapost i ein fil og er ein av mange mulege grammatikkar som PURITAN kan gjøre seg bruk av i same analysen.

Grammatikken er ein tilstands-automat med ymse tillegg. Kvar paragraf kan representerast på to måtar:

1) 'FULL'-tilstand, som er ein indeksert tabell.
2) 'KORT'-tilstand, som er ein søkje-tabell.

FULL/KORT er heilt ut transparent for PURITAN.

Nye grammatikk-paragrafar er 'billege' å skrive fordi det er høve til indirekte representasjon av alle felt i grammatikk-linene, bortsett frå 'inn-kode'. Programmet finn då ikkje direkte dei verdiane det skal bruke, i inneverande grammatikk-paragraf, men finn i staden opplysning om kva paragraf verdien skal stå i.

Gjennom ein 'push-down'-stakk som er styrt av inn-attributta, kan vi i grammatikken bruke grammatikk-paragrafar som subrutinar.

På grunn av 'KOPT'-tilstandar og høvet til å bruke indirekte representasjon kan grammatikken også brukast til å kjenne att strenger. I Bibel-prosjektet er denne eigenskapen brukt til å kjenne att strengene i kontroll-språket på trykkeribandet. Strengene vert kjende att ved si særlege vandring gjennom paragrafane i grammatikken. Dette gjev t.d. grammatikk-skrivaren høve til å la andre aksjonar i PURITAN vere avhengige av heile teiknsekvensar. Ved å nytte denne eigenskapen saman med mekanismen for 'push-down'-stakk kan vi få ein avgrensa parameterstyrt utmating ved t.d. å gje ei gyldig attkjenning av t.d. strengen 'AB' dersom attkjenninga tok til i paragraf C, men ikkje gyldig dersom ho tok til i t.d. paragraf D.

Gjennom 'inn-attributta' kan grammatikk-skrivaren også la ein og same kodeverdi komme att som 'inn-kode' så mange gonger som han ønskjer. Han har dermed ein kraftig ein-til-fleire mekanisme til rådvelde.

Gjennom grammatikken kan også utmatings-handsaminga vørta utsett, slik at vi i røynda oppnår ein 'look-ahead' mekanisme.

3.2. EINSTEIN-DELEN

Heile dette programsystemet er fast og gjev ikkje rom for applikasjons-programmering. Det kan nærast samanliknast med ein editor.

EINSTEIN er i store drag bygd opp som CROMWELL: styreprogram med kommando-struktur.

Eit oversyn over EINSTEIN-kommandoane er å finne i Tillegg 3.

4. SAMANFATNING

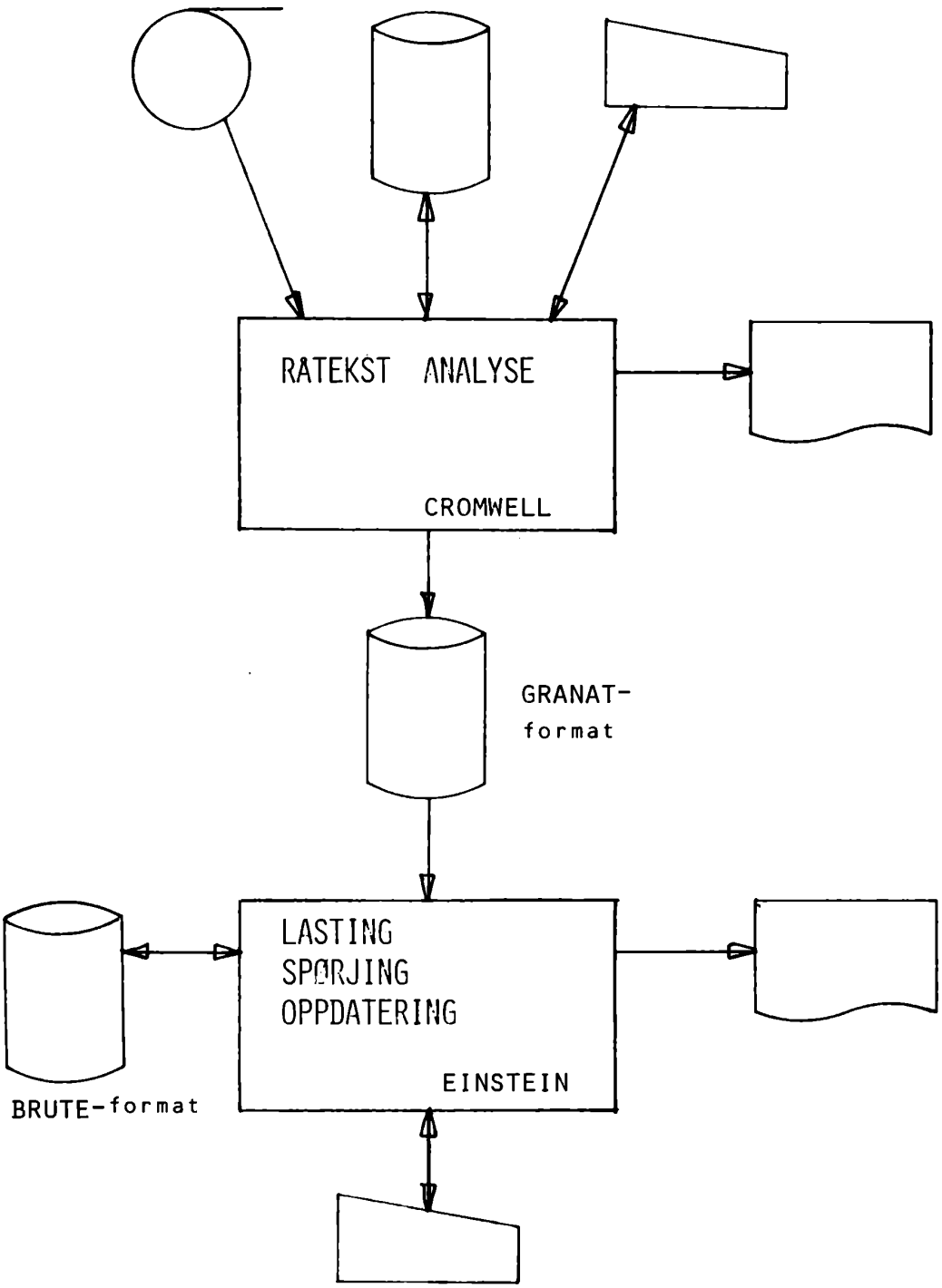
Program-systemet MONOLITT er no i full bruk i arbeidet for DNB med dei nye bibelordbøkene. Til no har vi lagt til rettes heile Det nye testamentet (bokmål). Som det går fram av Tillegg 4 er vårt nye system mindre tidkrevande og svært mykje mindre plasskrevande enn vårt gamle KWIC-system og når det dessutan har vist seg å fungere svært godt i produksjon, så meiner vi systemet tilfredsstiller ganske godt dei krav vi og andre stilte i utgangspunktet for at det skulle vere verd å ta fatt på ei nyutvikling.

Utviklingsfasen kan heller ikkje seiast å ha vore unormalt lang. Fyrste del av arbeidet byrja i april. Frå juli og i nær på to månader stod arbeidet nesten stilt på grunn av sommarferie og installasjon av ny maskin ved Universitetet. Frå september av og fram til no har det praktiske og daglege arbeidet blitt gjort av Bjørn Eide, men med regelmessige samtalar og konsultasjonar. Alt i alt kan arbeidet til no reknast å ha krevd seks-sju mann-månader, men av dei må eit par reknast som meir eller mindre spilt tid grunna dårlege og til dels urette og unøyaktige opplysningar frå trykkeriet om data og kodar på trykkeribandet, noko som førte oss på ville vegar i utforminga av grammatikken.

Mike Gillow har vore hjernen bak modellbygginga og utviklinga av det noverande opplegget medan Per-Bjørn Pedersen og Bjørn Eide har hatt hovudarbeidet med programskriving og uttesting.

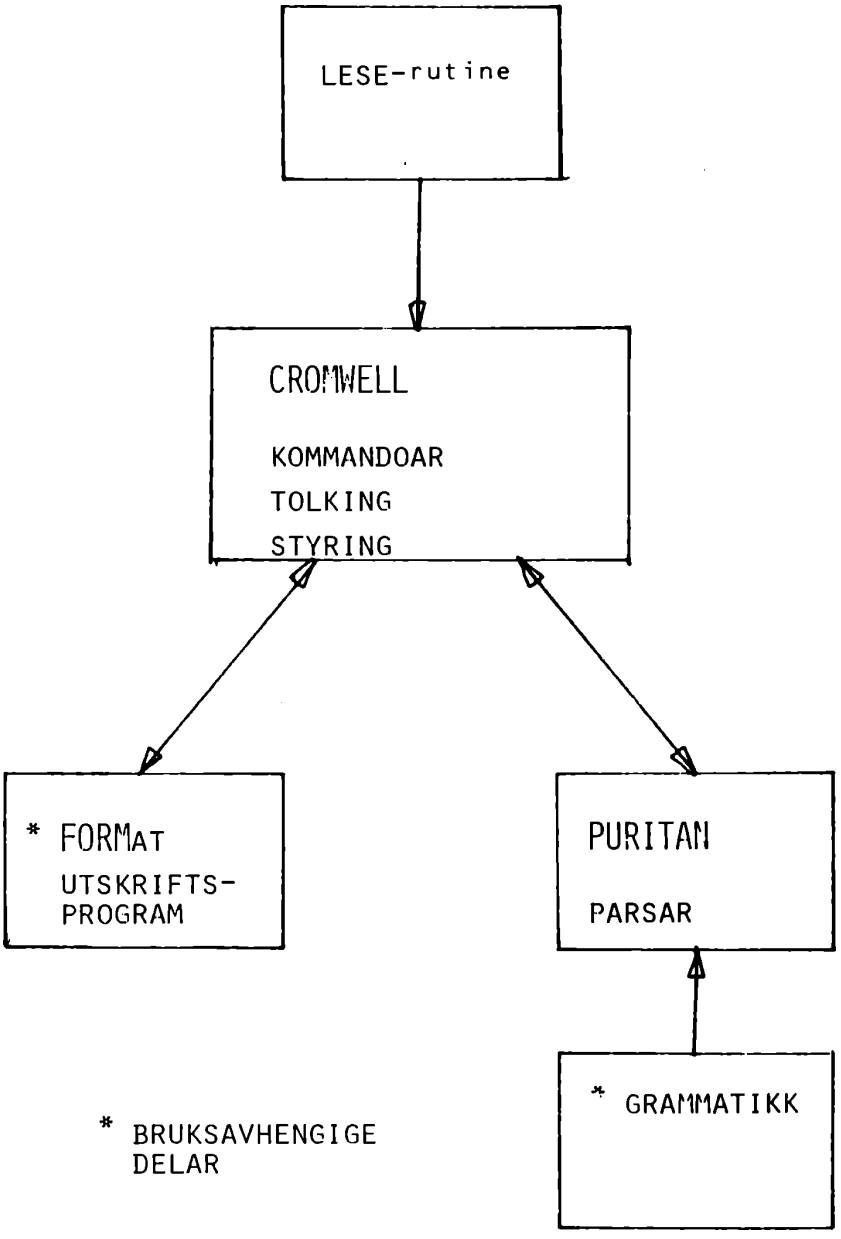
Men framfor alt må dette sjåast på som eit lagarbeid der vi alle tre har bryt våre idear og tankar mot kvarandre i seine nattetimar og tidlege morgonstunder. Ein ikkje uviktig del av vår filosofi og drivkraft for arbeidet har vore at vi alle skulle kjenne alle delar av systemet slik at vi skulle kunne bruke, utvikle og kritisere kvarandres program. Men for ein slik arbeidsmåte krevst det ein viss grad av stardardisering og felles arbeidsmetodar. Det har difor vore gledeleg og inspirerande å sjå at også slike tankar har kunna gjennomførast. Vi meiner også her å ha vandra langs vegar som ikkje er heilt vanlege, men som burde kunne gje gode og fruktbare resultat også i andre programmeringsoppgåver.

5. Tillegg 1



OVERSYN OVER KOMPONENTANE I MONOLITT

6. Tillegg 2



OVERSYN OVER KOMPONENTANE I CROMWELL

7. Tillegg 3

KOMMANDOAR I CROMWELL

INIT	initialiser TESTFIL
DDAT	lage testdata on-line (i desimalkode)
LIST	liste noverande TESTbuffer (i talformat)
SAVE	lagre buffer i TESTFIL
REST	motsett av SAVE
HIV	slette ein post i TESTFIL
FDAT	overføre røynlege datablokker til TESTFIL
STYR	nystille inn-peikar og tilstand
REF	nystille tekstreferanse
SET	styre avlusingsinformasjon
RUN	produksjon
UTOR	handsame eitt ord frå PURITAN

K O M M A N D O A R I E I N S T E I N

KWIC	lage KWIC-konkordans
FREK	lage frekvens-ordliste
VERS	liste tekstebetar
FIL	tilordnar datafilen
FORM	gjev nye verde til format-parameter
INDX	lagar ordindeks
LAST	lastar ny data

8. Tillegg 4

Samanlikning mellom vårt gamle KWIC-system og vårt nye MONOLITT-system.

Med det gamle systemet trongst det rundt 30 køyringar og kring 400 min. maskintid for å lage ein fullstendig KWIC. Denne KWICen hadde eit plassbehov på rundt 6000 spor. Med MONOLITT vil tilsvarande data få plass i litt over 2000 spor, faktisk då med ein heil del meir opplysningar lagra enn etter gamle metoden. For å laste eit heilt NT med MONOLITT trengst kring 220 min. maskintid. Då har vi rett nok ikkje ein KWIC-konkordans på bordet. Vi har enno ikkje gjort nøyaktige overslag over kva mskintid som trengst for å køyre ut ein full KWIC med MONOLITT, bl.a. av den grunn at den slags utlistingar er der sjeldan eller aldri spørsmål etter, men dei prøver vi har gjort viser at det går svært snøgt å skrive ut selektive KWICar, og det peikar i retning av at MONOLITT truleg vil vere i stand til å produsere ein full KWIC på den maskintida som står att før vi har teke ut like mykje maskinressursar som det gamle systemet.

Suzanne Hanon
Romansk Institut
Odense Universitet

Ordbøger, konkordanser og lemmatisering.

1. Indledning.

Ordbøger og de fleste konkordanser har mange træk til fælles: bortset fra det faktum, at de giver oplysninger om betydningen og (eller) brugen af ord, repræsenterer de hver resultatet af en operation, som vi kan kalde med en fælles betegnelse lemmatiseringen. Lemmatiseringen er en manipulation af teoretisk-praktisk karakter, som gør det muligt at forvandle en løbende tekst (parole) til en ordnet liste af kanoniske former (langue). Disse kanoniske former kalder man stikord, når vi har at gøre med ordbøger, mens vi bruger lemma (lemmaer, lemmata), når vi omtaler samme enheder i konkordanser (de såkaldte 'lemmatiserede' konkordanser). Ordbøger og konkordanser kan således betragtes som ordnede lister af kanoniske former, men det materiale til indsamling af disse kanoniske former er principielt meget forskelligt: lemmata stammer traditionelt fra de løbende ord i en begrænset tekstenhed (f.eks. et værk eller et forfatterskab), mens det materiale, der ligger til grund for udvælgelse af stikord, stammer fra større tekstenheder, eller mindre vel afgrænsede tekstenheder (f.eks. en større periode som nudansk eller oldfransk). Lemmatiseringen består i at klassificere og reducere visse ordformer til principielt "mindre komplicerede" ordformer. Denne operation forudsætte en konvention af morfologisk og semantisk art. Lemmatiseringen kan i datamatisk sprogbrug for så vidt også godt betragtes som en generel beskrivelse af de adresser, under hvilke man kan finde de enkelte ordformer.

Nærværende artikel har som formål at studere en række ordbøger og en række lemmatiserede konkordanser og udlede deraf de kriterier, som har været herskende for at lemmatisere eller udpege en for til stikord. Det materiale, som er brugt til dette studium, omfatter især franske værker, i særdeleshed ordbøger og konkordanser om oldfransk eller om oldfranske forfattere. Enkelte eksempler er dog taget fra danske ordbøger og fra nufransk materiale. Der bliver også sat spørgsmålstegn ved, om lemmatiseringen i det hele taget er en nødvendig operation. Artiklen forsøger sluttelig at opstille en liste over

de betingelser, der skal være til stede, for at lemmatiseringen (om nogen) kan være tilfredsstillende.

2. Kanoniske former og konvention.

De fleste grupperinger af ordformer under en fælles nævner støtter sig til en almindelig accepteret konvention: f.eks. serier af ordformer som book og books hører sammen med fællesnævneren BOOK, men også work, works og worked med fællesnævneren WORK. Dette kaldes af Ladislav Zgusta i Manual of Lexicography "word-and-paradigm" modellen, og denne antagelse af en "word-and-paradigm model" er grundlæggende for ordbogforfatteres virke og alt andet leksikografisk arbejde. "When using this model (and for the lexicographer, it is really the best one), the lexicographer assumes that the lexical meaning of every single series of forms such as those quoted above remains the same, while only the different grammatical categories are expressed by different forms...", og længere hen ang. de 1350 mulige former af verbet på sanskrit: "It is only natural that the lexicographer cannot indicate all these forms in his dictionary. For every paradigm, i.e. for every inflected word and for the totality of its forms, he indicates only one form that is considered basis: this is the so-called canonical form of a word. The canonical form represents the whole paradigm...". (Zgusta, op.cit., p. 119)

For at forene ligeledes chevaux og cheval under CHEVAL, meilleur, bonnes, bons, bon til BON , vais, irai, allant til ALLER må vi bruge samme antagelse. Man kan hermed konstatere, at traditionen uvægerligt rangerer flertal under ental, feminimum under maskulinum, bøjede verbalformer under infinitte former (helst infinitiv), komparativ (og superlativ) under den ikke komparerede form osv. At konventionen afviger somme tider fra sprog til sprog behøver ikke at undre os, konventionen udspringer bare fra en tradition, og det er derfor ikke mærkeligt i sig selv at skulle søge en verbalform under infinitiv på fransk eller på dansk, men under første person ental indikativ præsens for latin eller græsk. "The guiding principle for the choice of the canonical form is that it should be as good a starting point for the construction of the other forms of the paradigm as possible. In the majority of languages, the canonical forms are by now fixed by tradition and the lexicographer will do well not to abandon the tradition unless the reasons for a change are really imperative...". (Zgusta, ibid. p. 12o) Problemerne opstår virkeligt, når man ikke kan finde

nogen mulig "kandidat" i ental til en flertalsform for eksempel, eller en maskulinum form til en feminimum form osv. Hvad skal man stille op med følgende ord, hvis ikke de selv skal udpeges til kanoniske former:

Fransk:	vacances	VACANCE?	VACANCES?
	fastes (subst)		
	fiançailles		
Engelsk:	trousers		
Dansk:	bukser		
Fransk:	enceinte (adj.)		

Har man desuden lov til at sætte som kanonisk form en lidet brugt form som GÉSIR (for mere almindeligt forekommende former som gisant, gît, og ligeledes TISTRE for tissu? Og i fald man går med til at samle under ét en række forskellige rødder som vais, allai, irai under ALLER hvad er der som forhindrer én i at samle je me rappelle og je me suis souvenu under SE SOUVENIR? Zgusta forestiller sig i disse tilfælde, at man udnævner en form til at være en kanonisk ad hoc form ("a canonical ad hoc"). Disse betragtninger sammen med en række andre gør, at det at finde kanoniske former eller lemmata byder på ikke så få problemer, fo ikke at sige faremomenter (især i skabelse af ens egne "ad hoc" former

3. Den vanskelige norm.

Hvis en konkordansforfatter strengt taget kan lade være med at lemmatisere sin konkordans, og dermed vælger at give sine brugere de rå resultater af konkordansen, kan man ikke gøre det samme, når det drejer sig om ordbøgerne. Selv om forfatterne til ordbøgerne er bundne til at finde kanoniske former, er de på en måde bedre stillet, efters de mange gange selv kan vælge deres egne kriterier for normen: For eksempel viser forfattere til den franske ordbog Le Petit Robert en d frihed, når de ca. 10 år efter udgivelsen af deres ordbog tillader si at omfatte slang, tabuord m.m. i 2. udgave. Således siger Alain Rey i forordet af 1977-udgaven (side xviii) "Quant aux mots tabous, argotiq ou malsonnants, il nous a semblé que l'évolution des moeurs devait êt reflétée, ne serait-ce que pour répondre aux besoins de compréhensio les plus évidents".. og senere: "Le Petit Robert s'est donné comme ob jectif de combler les lacunes les plus criantes de l'éducation offi-

cielle". Her er det privilegium, som moderne forfattere har, når de behandler et sprogtrin, de selv kender, fordi de selv oplever sproget og dets mulige skiften men også tidens skiftende moralske begreber. Det forholder sig anderledes, når en ordbog eller en konkordans afspejler en tilstand, som hører til historien, og hvor man kan have svært ved at genetablere ord-paradigme-modellen, for eksempel fordi man mangler belæg på visse ordformer. Da lemmatisering eller kanonisering af visse former i forhold til andre altid i sidste analyse hviler på et skøn, kan man selvfølgelig vælge andre kriterier end de, som er tilgængelige fra teksterne selv. Man vil umiddelbart have en tendens til at bearbejde et materiale,som afspejler et ældre sprogtrin, med de kriterier, som gælder for det nyere sprogtrin (denne tendens er meget markant, hvis man oven i købet tilhører gruppen, som taler det sprog på et nyere trin end det, man studerer). En omvendt procedure, som man ofte iagttager hos franske forskere, består i at "gøre teksten ældre", fordi man f.eks. sammenligner med latin eller andre mulige etymologiske former.

For at etablere normen for en tekst af ældre sprogtrin er der dog følgende andre teoretiske muligheder:

(1) Teksten indeholder teoretisk alle mulige former, det vil sige også de former, man vil udnævne til kanon. Teksten må derfor være af et vist omfang, men sprogforskeren vil stadig skulle sortere og etablere værdinormer for kanoniske former og lemmata.
(2) Teksten er ikke selvtilstrækkelig, men der eksisterer gode beskrivelser af samme sprogtrin: pålidelige morfologiske og syntaktiske beskrivelser kan bruges til at udlede former, som ikke er fundet i teksten.
(3) Teksten er ikke selvtilstrækkelig, og der findes ikke gode morfologiske eller syntaktiske beskrivelser af samme sprogtrin. Der findes til gengæld ordbøger, som afspejler samme sprogtrin. Her vil det være oplagt at følge disse ordbøger, vel at mærke hvis disse redskaber er pålidelige. Selv om de skulle være pålidelige, skal man altid gardere sig for mulige forvrængninger, der kan stamme fra forskellige opfattelser af den ydre virkelighed, en anden kultur, forskellige leksikografiske normer osv.

Der forbliver dog en række problemer i etaperne (2) og (3), fordi man principielt introducerer elementer, som ikke er til stede i den oprindelige tekst. Og hvad værre er, man skal vogte sig for,

at den udledte form ikke bliver en re-konstruktion, en tænkt form.

4. En komparativ undersøgelse af 5 konkordanser over oldfransk.

Materiale: 3 konkordanser fra det leksikologiske laboratorium ved Liège Universitet, som er etableret af C. Dubois, M. Dubois-Stasse, A. Fontaine-Leauve, G. Lavis og M. Graitson.
(a) Chretien de Troyes: Philomena (udg. 1969-70)
(b) Les Chansons de Blondel de Nesle (udg. 1971)
(c) Chretien de Troyes: Guillaume d'Angleterre I-II (udg. 1974)
samt 2 andre konkordanser, den ene over oldfransk, den anden over en nyere tekst (15. årh.), som er etableret af et andet belgisk team, 'Service de linguistique française de l'université de Gand' med følgende medarbejdere: A. Boone, H. Cuyvers-Lippens, G. de Poerck, R. van Deyck-Bauwens, D. Willems, R. Zwaenepoel-Dhanis.
(d) Le Charroi de Nîmes (udg. 1970)
(e) François Villon (udg. 1974)
(Jeg henviser til bibliografien for de fuldstændige titler.)

Konkordanserne a, b og c fra Liège er blevet etableret med samme grundprincipper: ordformerne er blevet klassificeret alfabetisk under deres kanoniske former eller lemmata, og disse er også ordnet alfabetisk. Disse lemmata (herunder også grafierne af ord (et vigtigt element i ældre sprogtrin)) er hentet i Tobler-Lommatzschs Altfranzösisches Wörterbuch, en stor ordbog på 10 bind over oldfransk, men hvis udgivelse endnu i 1979 ikke er færdiggjort.

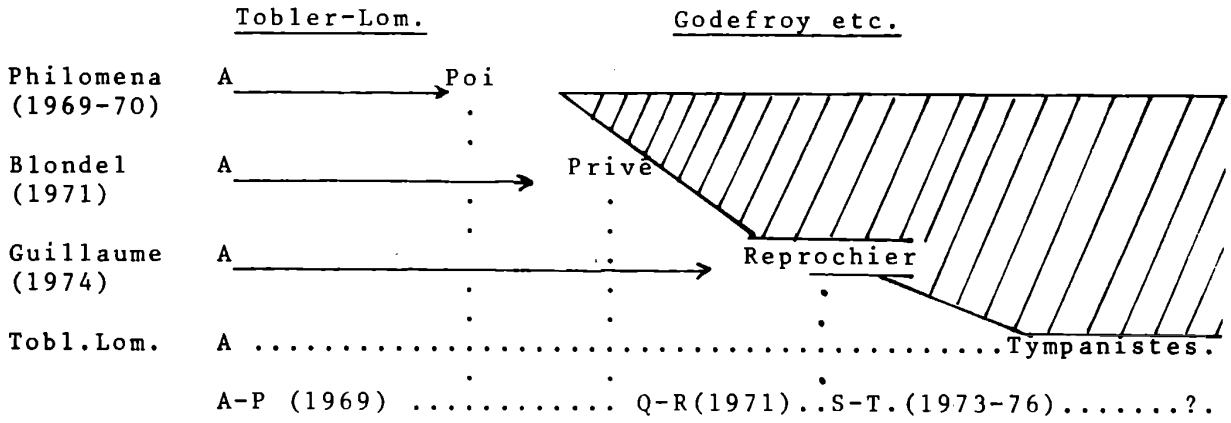


Fig. 1. Udviklingen af det lemmatiserede ordforråd iflg. Tobler-Lommatzsch i 3 konkordanser over oldfransk.

Fig. 1 viser fordelingen af lemmata ordnet efter Tobler-Lommatzsch for de 3 konkordansers vedkommende, efterhånden som tiden går. Forfatterne fra Liège har selvfølgelig måttet supplere deres kildemateriale og har derfor måttet bruge andre forhåndenværende ordbøger, f.eks. Fr. Godefroys Dictionnaire de l'ancienne langue française et de tous ses dialectes du IX^{e} siècle au XV^{e} siècle (8 bind + 4 tillægsbind), Flûtres Table des noms propres avec toutes leurs variantes figurant dans les romans du moyen âge (CESCM, Poitiers, 1962), Walther von Wartburgs Französiches Etymologisches Wörterbuch, m.m.

En sammenligning af ordet QUE i de tre konkordanser viser, at man for Philomenas vedkommende har måttet bruge Godefroys ordbog, og får derefter 2 lemmata; for Blondel får man tre lemmata for QUE, stadig væk med Godefroy som kilde, mens man for Guillaume, som er lemmatiseret ifølge Tobler-Lommatzschs ordbog, får 5 forskellige lemmata.

Konkordanserne (d) og (e) er også blevet lemmatiseret på direkte eller indirekte vis med Tobler-Lommatzsch som forbillede. Konkordansen (e) om Villons forfatterskab introducerer en såkaldt "dobbelt lemmatisering" ("entrée" eller "lemme" siden "forme de base" og endelig "occurrences"). Det mest bemærkelsesværdige er, at man for at lemmatisere en tekst fra det 15. århundrede bruger kriterier fra et ældre sprogtrin. I Villons tid er f.eks. kasus bortfaldet, mens det trin, Tobler-Lommatzsch beskriver, er det såkaldte klassisk oldfransk, som bl.a. endnu byder på et to-kasus-system.

5. En undersøgelse af 4 oldfranske ordbøger og deres valg af lemmata eller stikord.

Bortset fra Tobler-Lommatzsch og Godefroy vil jeg også omtale Greimas' Dictionnaire de l'ancien français jusqu'au milieu du XIV siècle samt Grandsaignes d'Hauterives Dictionnaire d'ancien français. Sidstnævnte ordbøger er snarere "skole"-udgaver end reelt videnskabelige ordbøger.

De ord, der er afprøvet, er følgende:

(1) verbet finirs ækvivalenter i oldfransk. Det drejer sig om at vide, om man vælger fenir, finir eller finer, som er 3 parallelle former i oldfransk.

(2) om den etymologiske dublet faillir/falloir svarer til 2 eller kun til et verbum i oldfransk.

(3) hvilken form er valgt som kanonisk form for craindres repræsentanter i oldfransk (6 eller flere forskellige varianter ifølge Kr. Nyrops Grammaire historique).

	fenir	finir	finer
Tobl. Lom.	FENIR ←	(FINIR) (cf. FENIR)	FINER
Godefroy	(FENIR) (cf. FINIR) →	FINIR fenir	FINER finyer
Greimas	(FENIR) (cf. FINIR) →	FINIR, FENIR	FINER
Grandsaignes	(fenir)	0	FINER

Fig. 2. Fordelingen af de former, der betyder finir i 4 ordbøger over oldfransk.

Fig. 2 viser fordelingen af former for finir.

	falir, fallir, faillir	faloir, falloir
Tobl. Lomm.	FALIR	0
Godefroy	FAILLIR fallir, falir, falyr	0
Godefr.Compl.		FALLOIR
Greimas	FAILLIR	FALOIR
Grandsaignes	0	0

Fig. 3. Fordelingen af formerne af dubletten: faillir/falloir i 4 ordbøger over oldfransk.

Fig. 3 viser, hvilke valg af kanoniske former, der er foretaget for faillir/falloir.

	criembre	criendre	craindre	cremer	cremir	cremoir
Tobl. Lom.	CRIEMBRE	0	0	cremer v.Godefroy	v.criembre	v.criembre
Godefroy	0	0	0	CREMER cromer	CREMIR	CREMOIR cremmoir
Godefr.Compl.	CRIEMBRE	0	CREINDRE	0	0	0
Greimas	CRIEMBRE, CREIMBRE V. CREMER	0	0	CREMER,IR,OIR CRIEMBE(?)		
Grandsaignes	CRIEMBRE	0	0	CREMER	CREMIR,OIR	

Fig. 4. Fordelingen af mulige infinitivformer af "craindre" i 4 ordbøger over oldfransk.

Fig. 4 viser den interessante fordeling af de mange former afledt af latin trĕmere. Iøjnefaldende ændringer som det at skifte meningen mellem selve ordbogen og dens supplement (Godefroy), "Parti pris" mod det moderne hos Tobler-Lommatzsch og samtlige andre, bortset fra Godefroy Compl. Det allerinteressanteste består i valget af "primus inter pares"-formen, når der findes mange sideformer. Tobler-Lommatzsch er den eneste, som vælger CRIEMBRE fremfor de andre former. En større undersøgelse har siden vist mig, at denne form ikke har noget belæg hos Tobler-Lommatzsch selv, men heller ikke i hans kilder. (Jeg henviser til en anden artikel, jeg har skrevet om emnet i en lidt modificeret form, i "Proceedings from the Conference on General and Jewish Lexicography, Delaware, 1979)(i trykken).

Man kan derfor sætte et spørgsmålstegn ved denne form for kilder, når samtlige under punkt 4 omtalte konkordanser har gjort Tobler-Lommatzsch's principper til deres, hvad angår valg af lemmata.

6. Undersøgelse af typisk "kvindelige" begreber i nogle moderne ordbøger.

(a) ordet enceinte (gravid).

Dette ord er brugt som kanonisk ad hoc form i både 1968- og 1977-udgaver af Petit Robert. Desuden også i TLF's frekvensordbog. I dette tilfælde kan man sige, at man ved at udvælge formen enceinte til kanonisk ad hoc følger den metalingvistiske virkelighed.

(b) ordet infirmière i en række franske ordbøger. Gougenheims Dictionnaire fondamental giver som stikord infirmier, og derunder skal man finde infirmière (jvfr. épicier-épicière, employeur-employeuse osv.). Dette faktum er interessant i en ordbog, som skulle afspejle de 3500 hyppigste ord! Petit Robert udgaver fra 1968 og 1977 giver samme præsentation af ordet infirmière. Man skal søge det under infirmier. Man skal dog bemærke, at samtlige eksempler henviser til en kvindelig sygeplejerske, mens 2 modeksempler er af hankøn.

(c) ordet sage-femme (jordmor). Dette ord forekommer som stikord i le Petit Robert. En interessant henvisning leder læseren til accoucheuse, som til gengæld ikke findes som stikord. Ordet findes, men under stikordet accoucheur.

(d) ordet sygeplejerske og dets behandling i Nudansk ordbog. Dette ord er blevet studeret udelukkende for at se, om der var forskel i opfattelse mellem 2 kulturer. Der er desuden det spændende, at dette ord på dansk kan siges at tilhøre et af de få paradigmer, som udtrykker hunkøn på dansk, jvfr.:
lærer-lærerinde/arbejder-arbejderske.
Man kan på dansk også udtrykke køn med følgende midler: dvs. ved at danne komposita: han-hun vs hun-hund, mandlig kollega vs kvindelig kollega.
Ang. sygeplejerske er det af interesse at se, under hvilket stikord formen bliver givet.
Nudansk ordbog, som er en tætpakket ordbog på 2 bind, har åbenbart pladsmangel, og dens præsentation ser ud til at være præget deraf. Stikordet for sygeplejerske er sygemelde, som også fungerer som stikord for sygemelding, sygepasser, sygepleje, sygeplejer, -plejerske, -seng. Man skal også her bemærke, at et vigtigt begreb bliver givet under dets afledning og ikke som første valg.

Valget af eksempler i denne artikel, men også andre steder er meget lærerig: "forstanderinde for hospitalets sygepleje.." Sygeplejerske indgår som definerende element i definitionen af ordet sygehjælper, et ord som er medtaget noget før under stikordet syge- er endnu mere afslørende: sygehjælper,"om person, der har mindre uddannelse end en sygeplejerske, og som yder lettere sygepleje". Alt dette viser, at den semantiske sfære åbenlyst er knyttet til det kvindelige, mens det rent lingvistiske valg opereres på gængse mandlige kriterier.

7. Konklusion.

En række iagttagelser om henholdsvis konkordanser, ordbøger over ældre sprogtrin, og ordbøger over nyere sprogtrin viser, at den leksikografiske tradition byder på en række dårligt definerede områder, hvad angår valget og kriterierne for valget af den såkaldte makrostruktur. Man skal derfor her som konklusion forsøgsvis opstille en række betingelser, som burde være til stede for at sikre, at en eventuel lemmatisering er forsvarlig:

(1.) Den kanoniske form X repræsenterer ordformerne y, z, osv., hvis og kun hvis x, y, x er en del af samme paradigme.

(2.) For at være i samme paradigme må x, y og z nødvendigvis være samtidige i tid (samme periode).

(3.) Og samtidige i rummet (én og samme tekst, ét og samme sprogniveau, samme stil, samme dialekt osv.).

(4.) X må være mere repræsentativt end y, x osv. Det vil sige, at y, z, osv. må være lette at uddrage fra X.

(5.) X's frekvens må være mindst lige så høj som y's, z's (X må altså ikke være en teoretisk, rekonstrueret, ikke belyst eller mindre brugt form af y, z osv.).

(6.) Den lingvistiske enhed X må samtidig være relateret til den ekstralingvistiske virkelighed, og den må repræsentere denne virkelighed bedst muligt.

BIBLIOGRAFI

1. Leksikografiske problemer.

Baldinger, Kurt: Introduction aux dictionnaires les plus importants pour l'histoire du français, Klincksieck, Paris, 1974.

Brackenier, R.: "Index et Concordances d'auteurs français modernes. Etudes critiques" Travaux de linguistique 3 (1972) 1-43, et 4 (1975) 1-61.

Dubois, Jean et Claude: Introduction à la lexicographie: le dictionnaire, Larousse, Paris, 1971.

Hanon, S.: Compte rendu de J.J. Duggan: A Concordance of the Chanson de Roland et G. de Poerck et al. Le Charroi de Nîmes I-II, Revue Romane, VIII, (1973) 421-423.

Hanon, S.: "Chronique. Colloque sur l'analyse des corpus linguistiques. Problèmes et méthodes de l'indexation maximale" Cahiers de Lexicologie, 23, (1973) 117-124.

Hanon, S.: "Mots dans le texte, mots hors du texte: réflexions méthodologiques sur quelques index et concordances appliquées à des oeuvres françaises, italiennes ou espagnoles." Revue Romane, XII, (1977), 272-296.

Hanon, S.: Compte rendu de Rika van Deyck et Romana Zwaenepoel: François Villon Oeuvres d'après le manuscrit Coislin, textes, variantes et concordances. Revue Romane, XIII, (1978), 340-344.

Hausmann, Franz Josef: Einführung in die Benutzung der neufranzösischen Wörterbücher, Niemeyer, Tübingen, 1977.

Matoré, Georges: Histoire des dictionnaires français, Larousse, Paris, 1968.

Nyrop, Kr.: Grammaire historique de la langue française, tome II, Gyldendal, Copenhague, 1960.

Quemada, Bernard: Les dictionnaires du français moderne 1539-1863, Didier, Paris, 1968.

Rey, Alain: "Les Dictionnaires: forme et contenu". Cahiers de Lexicologie, VII, (1965), 65-102.

Rey, Alain: Le lexique: images et modèles du dictionnaire à la lexicologie. Colin, Paris, 1977.

Rey-Debove, Josette: Etude linguistique et sémiotique des dictionnaires français contemporains, Mouton, The Hague, Paris, 1971.

Wagner, R.L.: Les vocabulaires français I-II, Didier, Paris, 1967-1970.

Zgusta, Ladislav: Manual of Lexicography, Mouton, Academia, The Hague, Paris, Prague, 1971.

2. Konkordanser.

Blondel de Nesle: Les Chansons de Blondel de Nesle. Concordances et index établis d'après l'édition. L. Wiese par G. Lavis. Traitement automatique: C. Dubois. Publications de l'Institut de Lexicologie française de l'Université de Liège. Faculté de Philosophie et Lettres de l'Université de Liège, (1971).

Chanson de Roland: A Concordance of the Chanson de Roland, compiled by Joseph L. Duggan. Ohio State University Press, 1969.

Charroi de Nîmes: Le Charroi de Nîmes, chanson de geste, par G. de Poerck, concordances, R. van Deyck, texte et variantes, R. Zwaenepoel, traitement automatique. Tome I-II. Textes et traitement automatique. Librairie- Editions Mallier, Saint-Aquilin-de-Pacy (Eure) 1970.

Chrétien de Troyes: Philomena. Concordances et index établis d'après l'édition C. de Boer, par C. Dubois, M. Dubois-Stasse et G. Lavis. Faculté de Philosophie et Lettres de l'Université de Liège. Publications de l'Institut de Lexicologie française de l'Université de Liège, (1970).

Chrétien de Troyes: Guillaume d'Angleterre. Concordances et index établis d'après l'édition M. Wilmotte par M. Dubois-Stasse, A. Fontaine-Lauve. Traitement automatique: C. Dubois, M. Graitson, I-II. Faculté de Philosophie et Lettres de l'Université de Liège. Publications de l'Institut de Lexicologie française de l'Université de Liège. (1970).

Joachim du Bellay: Joachim du Bellay, La Deffence et Illustration de la Langue Francoyse. Concordance établie par Suzanne Hanon. Traitement automatique: Poul Bonne Jørgensen et Ulf Hagen Køllgaard. Etudes romanes de l'Université d'Odense, vol. 6, Odense University Press, 1974.

Villon: François Villon: Oeuvres d'après le manuscrit Coislin. Rika van Deyck: Texte, variantes et concordances. Romana Zwaenepoel: Traitement automatique. I-II, Mallier, Saint-Aquilin-de-Pacy 1974.

3. Ordbøger.

de Gorog, Ralph: Lexique Français moderne - Ancien français. University of Georgia Press, Athens, 1973.

Dictionnaire alphabétique et analogique de la Langue française par Paul Robert, Société du Nouveau Littré, Paris, 1968.

idem, 2e ed. Paris, 1977.

Dictionnaire des Fréquences. Vocabulaire littéraire des XIXe et XXe siècles. Centre de Recherche pour un Trésor de la Langue française, I-IV. C.N.R.S.- T.L.F. Didier, Nancy, 1971.

Dictionnaire fondamental de la Langue française par Georges Gougenheim, Didier, Paris, 1958, nelle édition revue et augmentée.

Diez, Friedrich: Etymologisches Wörterbuch der Romanischen Sprachen Bonn, Heilbronn, Leipzig, 1887-1889 et Georg Olms Verlag, Hildesheim, New York, 1969.

Gamillscheg, Ernst: Etymologisches Wörterbuch der Französischen Sprache, Heidelberg, Carl Winther, 1969.

Godefroy, Frédéric: Dictionnaire de l'ancienne langue française et de tous ses dialectes du IX^{e} au XV^{e} siècle (I-VIII + Compl. I-IV), Vieweg, Paris, 1881-1902.

Grandsaignes d'Hauterive, R.: Dictionnaire d'ancien français. Moyen âge et Renaissance, Larousse, Paris, 1947.

Greimas, A.J.: Dictionnaire de l'ancien français jusqu'au milieu du XIV^{e} siècle, Larousse, Paris, 1969.

Mayer-Lübke, W.: Romanisches Etymologisches Wörterbuch, (REW) 4. auflage, Heidelberg, C. Winther, 1968.

Nudansk ordbog udgivet med støtte af Undervisningsministeriet I-II. Politikens Forlag, København, 1964 red. Lis Jakobsen et. al.

Tobler-Lommatzsch: Altfranzösisches Wörterbuch (I-X), Franz Steiner Verlag, Wiesbaden, 1915-1976.

Wartburg, Walther v.: Französisches Etymologisches Wörterbuch, (FEW), Eine Darstellung des galloromanischen Sprachschatzes, Band XIII, 2. Teil, Basel, Zbinden Druch und Verlag AG, 1965.

BETA-systemet som verktyg för lingvistiska experiment - morfologisk analys enligt ytkriterier.

Benny Brodda
Institutionen för lingvistik, Stockholms Universitet

1. Inledning.

Detta att få en dator att rent mekaniskt utsegmentera strängar av typ '-ade' i 'hoppade', '-er' i 'sitter', '-orna' i 'flickorna', etc, är naturligtvis ytterligt trivialt ur datamaskinell synpunkt, och redan de första mer seriöst inriktade datalingvistiska projekten vid mitten på 50-talet hade avsevärt mer sofistikerade metoder än så för morfologisk analys. Inom vissa tillämpningsområden av typ dokumentsökning har dock "trunkering", som metoden kallas, fått viss användning - och icke utan framgång.

Naturligtvis är det inte alltid ändelser som man på så vis får utsegmenterade. "Analyser" av typ 'sp-ade', och 'sk-orna' kan utgöra exempel på hur absurda resultat man kan få om trunkeringen tillämpas helt mekaniskt, och även om det är lätt att rensa bort sådana uppenbara absurditeter återstår många, många fall där man inte får rätt resultat. Bara ett exempel: huru göra med '-er' i orden i satsen 'En mager neger niger sönder alla bönder', där '-er' utgör ändelse i blott två av orden och dessutom verbändelse i det ena ordet och substantivändelse i det andra?

Dessa exempel torde räcka för att föra trunkeringsmetoden helt åt sidan i varje seriöst projekt som syftar till något så när fullständighet i analysen.

Men, frågar man sig, varför skall man ha fullständighet? Ja, svaret måste bero på vad man skall ha analysen till, och är syftet att man vill lära sig något om språket, är det inte alls säkert att fullständighet i sig är särskilt eftersträvansvärd. I själva verket kan det ur teoretisk synpunkt vara avsevärt intressantare

att få reda på att en rent mekanisk trunkering trots allt ger uppåt 70-procentig korrekt (eller rimlig) analys av löpande text, än att man med manår av insatser kan visa att en datamaskin med tillgång till ett omfattande lexikon kan analysera ord, säg till 98 % korrekt. Den senare siffran säger egentligen ingenting om språket, utan bara om det dataprogram som används, och det är sedan 50-talet bekant att det är ungefär den nivån man uppnår.

Trunkeringsexperiment leder dessutom fram mot intressanta nya experiment, experiment som kan belysa viktiga aspekter av språket. Låt oss till att börja med sätta upp den mera begränsade målsättningen att försöka höja kvaliteten på analysen genom att lägga in diverse kontextvillkor o dyl för utsegmenteringen av ändelser och andra affix. Här följer några exempel på typer av sådana villkor:

'-te' i 'vante' kan inte vara analogt med '-te' i 'köpte' av fonologiska skäl; med några ytterligt få undantag måste preteritumändelsen '-te' föregås av en fortiskonsonant. Analyser av typ 'undanb-er' och 'o-rdentlig' är omöjliga av fonotaktiska skäl; utsegmenteringen kvarlämnar otillåtna konsonantkombinationer. 'ning' i 'baningenjör' kan inte utgöra suffix (analogt 'betalningsbalans') av morfotaktiska skäl; ord på '-ning' räknas som tunga och får normalt foge-s i förledsställning.

Om man nu anstränger sig maximalt för att utnyttja ytkriterier av den art jag ovan antytt, hur pass bra kan man då få analysen? Svaret på denna fråga beror i hög utsträckning på vad man menar med "riktig analys", vilket man kan kvalifiera på många olika sätt, men grovt taget kan man faktiskt få analysen att närma sig 90-procentig korrekthet och väl det.

Nedan skall jag redogöra för några experiment av det slaget; ett som jag själv utfört på svenskt material, och ett på finskt som jag utfört tillsammans med Fred Karlsson, Åbo. Dessutom skall jag beröra ett experiment initierat av Tore Janson, Stockholm, på latinskt material.

I någon mening blev dessa experiment överraskande bra (åtminstone överraskade de mig), och man kan fråga sig om detta har några

språkvetenskapliga implikationer. Den numera gängse teorin om språkperception är att parallellprocessning utnyttjas i hög grad vid tolkning av de språkliga signalerna. Hjärnan utnyttjar språklig information av vad slag som helst, varhelst denna påträffas - semantisk, syntaktisk, fonetisk, etc - och den gör det utan särskilt förutbestämd ordning, dvs varken enligt något "bottom-to-top" eller 'top-to-bottom" förfarande. En rimlig hypotes är, att ytsignalerna (de grammatiska morfemen) sammantagna ger starka indicier om de syntaktiska relationerna i yttrandena. Med experiment av det slag jag ovan beskrivit kan man komma åt frågan om hur mycket information som faktiskt finns i ytsignalerna (och, som sagt, det tycks finnas ganska mycket).

En annan iakttagelse tycker jag mig ha kunnat göra vid dessa experiment. I de språk som jag hitintills varit med om att undersöka med dessa metoder - svenskan och finskan utförligast, men också preliminärt ungerskan (turkiskan hoppas jag kunna komma igång med under året) - så tycks totalmängden information i ytsignalerna vara ungefär densamma på den morfologiska nivån, men det tycks som om denna informationsmängd byggs upp på fundamentalt olika sätt i de olika språken. Med det programsystem - BETA - som jag arbetat med, kan jag lätt plocka in och plocka ut den fonologiska komponenten, den fonotaktiska komponenten, etc, ur analysen, och genom att jämföra körningar med resp utan dessa olika komponenter kan jag få en uppfattning om vilken "börda" varje komponent för sig bär, och det verkar som om t ex svenskan och finskan avsevärt skiljer sig här. I svenskan tycks fonotaxen spela en betydande roll men vara av tämligen perifer betydelse i finskan där i stället morfotaxen bär motsvarande börda.

2. Något om programsystemet BETA.

Programsystemet BETA finns relativt utförligt beskrivet i förhandlingarna från de första nordiska datalingvistdagarna (Brodda, -77) och några av de bevekelsegrunder jag hade vid själva utformningen av systemet finns berörda i Brodda -79a, varför jag inte här behöver gå in på några närmare diskussioner om systemet som sådant. (Kortfattat kan systemet beskrivas som ett regelinterpreterande

system, som accepterar en enkel typ av kontextkänsliga regler, påbyggda med en tillståndsmekanism à la Turingmaskin.) Låt mig dock nämna en viktig sida av systemet som kanske inte framgår av ovan nämnda rapporter.

Om man vill åstadkomma ett tämligen generellt system för lingvistisk databehandling, så bör man ha klart för sig att varje sådant system måste komma att innebära en kompromiss mellan många i och för sig önskvärda men i dagens läge inkompatibla egenskaper. Visst vore det bra om man i samma system kunde utföra transformationer både framlänges och baklänges, slå i stora lexika, bygga semantiska nätverk, simulera kognitiva processer, etc, etc, och detta dessutom med ett system som är snabbt och billigt i drift och helst så enkelt att vem som helst kan lära sig det på någon timme. Visst, men var finna denna underbara cigarr? Jag tycker alltför många system sett dagens ljus där man antingen inte haft klart för sig att ovan uppräknade saker kanske inte går att förena, eller där man inte klargjort vilken aspekt av språklig databehandling man försökt optimera.

Huvudmålsättningen med BETA-systemet var att enkla, ytnära, strängmanipulationer skulle kunna utföras mycket enkelt och billigt, och på ett för utbildade lingvister något så när "transparent" sätt. Systemet skulle dock vara så pass generellt att det i princip skulle vara möjligt att åstadkomma vilken slags analys som helst. Resultatet blev - som ovan nämnts - ett regelinterpreterande system, där kravet på generalitet tillgodoses genom att alla sorters "action" i systemet skall kunna styras av reglerna.

Det exakta regelformatet bestämde jag efter en analys av vad som verkligen behövs och - minst lika viktigt - vad som inte behövs för simulering av tämligen ytnära processer av fonologisk eller morfologisk natur (typ "morfologisk analys utan lexikon"), och det är alltså den typ av analys som systemet bör vara optimerat för. Naturligtvis går det också att skriva t ex TG-regler o dyl i BETA; dock är systemet inte särskilt bekvämt för sådant.

Den enskilda konstruktionsdetalj som mest bidragit till att göra systemet kraftfullt är den mekanism enligt vilken kontext- och

tillståndsvillkoren utvärderas. Man kan bilda arbiträra klasser av tecken och tillstånd (under direktivet DEFSET (= DEFine SET)) och sedan använda (namnen på) dessa klasser i reglernas villkorsuttryck. Denna mekanism har gjort att enskilda språkliga enheter mycket sällan behöver anges mer än en gång i en regeluppsättning trots att enheterna som sådana kan förekomma i de mest skilda omgivningar och funktioner.

En annan mycket viktig egenskap hos systemet är att icke-deterministiska situationer kan omhändertas på ett mycket "geschwind" sätt. Vanligtvis utvärderas reglerna i ett regelsystem strikt disjunktivt - den första tillämpbara regeln i en viss given situation tillämpas varefter systemet fortsätter med nästa steg i enlighet med direktiven i den tillämpade regeln. Reglerna kan dock märkas så att de tillåtes verka konjunktivt; om två eller flera olika regler samtidigt är tillämpbara på samma situation så tillämpas också dessa regler den ena efter den andra, men på så sätt att efter varje sådan regeltillämpning så sparas hela arbetssträngen i en "jobbkö", arbetssträngen återställes därefter i det skick den nyss hade, och nästa tillämpbara regel kan nu få verka på samma ursprungssituation.

Detta förfarande att omhänderta icke-deterministiska situationer innebär förstås att sådana får en multiplikativ effekt, men genom att administrationen av den interna kön är mycket enkel och snabb så får man normalt ändock rimliga analystider på måttligt långa strängar; således kan nämnas att en sträng som var approximativt 500-ambiguös (enligt en enkel CF-grammatik) kunde analyseras i sin helhet på ca 30 sek (på en PDP 11/34).

Vad beträffar den tekniska uppbyggnaden av systemet har jag bl a ansträngt mig att få det mycket modulärt och därmed flexibelt och användbart för andra saker än ren strängmanipulation av här beskriven typ. I Brodda -79b demonstreras hur BETA-systemet kan utnyttjas som ett mycket avancerat excerperingshjälpmedel.

3. Tre experiment med automatisk morfologisk analys.

Jag skall nu gå över till att presentera några faktiska analyser

utförda med BETA-systemet. Det rör sig om tre experiment som alla hade en till det yttre likartad målsättning, nämligen automatisk utsegmentering av morfer i löpande text utan användande av (stam)lexikon, ja, i ett av dem - det på latin - saknades lexikon över huvud taget; morferna definierades där helt och hållet genom ett strukturvillkor.

Alla dessa morfologiska experiment är eller kommer att bli presenterade i sin helhet i andra sammanhang, och det är i vilket fall som helst uteslutet att på den plats som här står mig till förfogande göra annat än en kort presentation och ge några få exempel. Det gemensamma med dessa tre experiment är att de varit avsedda att belysa teoretiska frågeställningar beträffande respektive språks morfologi; i det latinska experimentet hypoteser av språkhistorisk karaktär, i de svenska och finska för att belysa hypoteser om ordstrukturer och ordperception. Det finska experimentet kan också förmodas få betydande pedagogisk tillämpning.

Ett annat gemensamt drag i dessa tre experiment är att regelsystemen är framtagna och uttestade på grundval av omfattande testkörningar på relativt stora datamängder (30-50.000 ord löpande text). Jag har i olika sammanhang, när jag vid presentation av BETA-systemet betonat att jag lagt stor vikt vid att få systemet både lättarbetat och snabbt, fått höra den invändningen att detta med snabbhet är ointressant så länge man håller på med uttestandet av teoretiska modeller. Detta är något jag inte vill hålla med om. Det är väldigt lätt att sätta upp snygga och prydliga modeller som behandlar ett litet antal, väl valda exempel på ett mycket elegant sätt, men det är då också väldigt lätt att man låter lura sig av modellens elegans att tro att språket i sin helhet är på det sätt som modellen visar. Min definitiva erfarenhet är att det är först då man låter modellen konfronteras med ett realistiskt språkligt material som man får någon uppfattning om modellens bärighet. Därför är det inte ointressant, vare sig ur teoretisk eller praktisk synpunkt för den som håller sig inom ramen för normala forskningsanslag att man kan köra igenom ganska stora material flera gånger.

De regelsystem som framtagits för de tre experimenten är kolossalt

olika till hela sitt sätt att verka. När det gäller latinexperimentet är detta kanske inte så underligt, eftersom uppgiften där var så annorlunda till sin karaktär, men när det gäller det svenska resp. finska experimentet kan man fråga sig om denna deras olikhet är en ren slump, eller om det svarar mot en verklig skillnad vad beträffar deras ordstrukturer. Den skillnad jag talar om gäller inte den rent ytliga skillnad som ligger i att analysen i det ena fallet går väsentligen från vänster till höger inom ordet och i det andra fallet tvärtom. I stället rör det sig om en mer fundamental skillnad i hela den strategi med vilken orden attackeras enligt de båda regelsystemen. De valda strategierna kändes naturliga och självklara både när det gällde svenskan och när det gällde finskan. Jag är personligen helt övertygad om att det är nödvändigt att analysera svenskan och finskan med helt olika algoritmer, och att detta svarar mot att orden i de bägge språken har fundamentalt olika uppbyggnad.

De svenska och finska regelsystemen representerar i sitt nuvarande skick naturligtvis många månaders arbete, men ett arbete som väsentligen kunnat inriktas på "subject matter", på diskussioner om och funderingar kring ordstrukturerna i de aktuella språken. Själva programmeringen - i den mån den går att urskilja från språkanalysen - har i vardera fallet tagit någon vecka i anspråk. För latinexperimentet åtgick det ganska exakt en dags programmeringsinsats för experimentets slutförande.

I appendix I visar jag för vart och ett av språken en fullständig analys av ett ord och ett analyserat textavsnitt.

4. Ett system för morfologisk analys av svenska.

Det svenska systemet kan karakteriseras som att affixens avskiljande sker i enlighet med en uppsättning starkt interaktiva transformationer (i transformationsteorins mening). Den strukturella förändringen är i varje enskilt fall mycket enkel. Ett prefix P utsegmenteras med en regel av typ "P → P-", ett suffix S med en regel av typ "S → -S" och analogt för en ändelse E, där "-" representerar den införda segmentgränsen.

Fig. 1: Mönster för

1. Ändelser: $X \left\{ \begin{matrix} \hat{} V \hat{} Mc \\ -S \end{matrix} \right\} \hat{} \underline{E} \#$

2. Prefix: $\left\{ \begin{matrix} \# \\ \# P- \\ X \hat{} V \hat{} E \hat{} (s) \end{matrix} \right\} \hat{} \underline{P} \hat{} I \hat{} V \hat{} X$

3. Suffix: $X \hat{} V \hat{} E \hat{} \underline{S} \left\{ \begin{matrix} \hat{} (s) \hat{} I \hat{} V \hat{} X \\ \hat{} E \# \\ \# \end{matrix} \right\}$

Villkor (mönster) som måste vara uppfyllda för utsegmentering av svenska affix. "^" betecknar konkatenation, { } betecknar alternativ. Övriga symboler förklaras i texten.

Schemat (fig. 1 ovan) representerar de strukturella villkoren för dessa transformationer, och det är i villkoren all utnyttjad information finns inbyggd. Man kan se dessa villkor - liksom strukturella villkor för transformationer i allmänhet - som ett slags mönster som skall vara uppfyllda för att transformationerna i fråga skall få tillämpas. I varje sådant mönster är det det framhävda elementet som så att säga "står i tur" att utsegmenteras. (I regeln förekommande "-" markerar tidigare utsegmenterade affix.)

Dessa transformationer är interaktiva i den meningen att output från en av dem kan bli input till hela transformationsuppsättningen igen (inklusive den nyss tillämpade regeln). De är också interaktiva i den meningen att två eller flera kan vara tillämpbara på samma situation. Ett exempel på en sådan konfliktsituation kan utgöras av ordet 'benet'; om prefixregeln tillåtes verka före ändelseregeln ges en analys 'be-net', medan om den omvända regelordningen tillämpas erhålles (den korrekta) analysen 'ben-et'. Jag håller för närvarande på med en utvärdering av hela systemet för att se vilka

regelordningar som bör byggas in i systemet för erhållande av en "over all" optimering. (Jfr diskussion i Brodda, -79c.)

De i mönstren ingående symbolerna I och F betecknar tillåtna initiala resp. finala konsonantklustrar och Me sådana mediala klustrar som kan förekomma omedelbart före en ändelse. V betecknar vokal. Dessa symboler representerar den fonotaktiska information som finns inbyggd i systemet. Ett av huvudsyftena med den aktuella undersökningen var att ta reda på hur pass mycket information som ligger förborgad i den fonotaktiska komponenten i svenskan; jag är inte färdig med utvärderingen av dessa experiment ännu, men mina undersökningar hitintills tyder på att det är avsevärt mer än vad jag från början trodde. Ill. 1, App. 1, visar hur väl fonotaxen "mejslar ut" prefixet 'be-' i ord börjande på 'be' och samplade ur ett antal Ivar Lo-Johansson noveller. Systemet är så gjort att jag lätt kan "plocka in" och "plocka ut" t ex den fonotaktiska informationen i reglerna, och ill. 2 visar en jämförelse mellan hur systemet behandlar ett potentiellt prefix 'o-' utan resp. med den fonotaktiska komponenten inkopplad.

I ill. 3 visas analysgången för ett tämligen intrikat fall, nämligen ordet 'bearbetningsbehov'. Raden märkt "11" representerar input, den märkt "22" output och mellanliggande rader, märkta "44", visar diverse mellanresultat, nämligen varje situation före det att en regel skall tillämpas på ordet. Till vänster har maskinen själv skrivit ut ett "protokoll", en anvisning om vilken delsträng systemet "ser". Asterisken ("dot") ute i strängen visar exakt var regeln skall till att tillämpas. Kommentarerna till höger har jag skrivit till för att förklara arbetsgången. Observera särskilt behandlingen av de tre förekomsterna av strängen 'be'. Observera också att analysen på det hela taget fortskrider från vänster till höger.

I det svenska textsamplet används "-" för att markera prefix, "/" suffix och "=" ändelse.

5. Ett system för morfologisk analys av finska.

Det finska experimentet kan kort beskrivas som följer: Finskan är

som bekant ett agglutinerande språk, man "staplar" långa räckor av ändelser i ordsluten; ända upp till 7 ändelser efter varandra är fullt möjligt och 4 à 5 är inte ovanligt. Nu är det bekant att det råder starka morfotaktiska och i viss utsträckning fonologiska begränsningar för vilka ändelser som kan kombineras med vilka andra (jfr Karlsson, -78), men genom systemets komplexitet är det veterligt ingen som i detalj försökt sig på att explicit beskriva alla dessa kopplingar. Att åstadkomma en sådan totalbeskrivning var ett av syftena med denna undersökning. Naturligtvis var inte detta det enda syftet, men utrymmet här räcker inte alls till för att gå in på dessa saker. Experimentet som sådant och dess praktiska och teoretiska konsekvenser kommer att presenteras utförligt i Brodda-Karlsson, -80.

Den slutgiltiga morfotaktiska modell som experimentet ledde fram till redovisas i schemat (fig. 2) nedan. Som synes representerar schemat ett Finite State diagram där ordets stam tänkes stå någonstans till vänster i figuren. Schemat representerar sedan de möjliga ändelsekombinationerna från vänster till höger. Enkliterna är alltså de ändelser som - om de finns med - står längst bak i orden.

Varje "låda" i figuren representerar en naturlig kategori i den meningen att morfer som är inneslutna i en låda har i någon mening likartade distributionella egenskaper och tillför ordet likartad typ av semantisk information. Som synes bildar inte lådorna någon hierarkisk struktur och därför kan man säga att t ex kasus ssA (inessiv) på ett sätt liknar stA (ellativ), på ett annat sätt Vn (illativ) och på ett tredje sätt n (genitiv) etc.

Analysgången i systemet sker väsentligen från höger till vänster i figuren: man börjar alltså att "strippa" enkliter, och om ordet är ett nomen så kanske man hamnar i possessivlådan, varefter man vandrar in i kasuslådan som i sin tur kan leda till numeruslådan etc (observera att varje kategori i stort sett är fakultativ, man kan t ex komma direkt från enklit till numerus). Detta nomen visar sig kanske sedan ha varit ett participavlett verb, och man kan alltså "ramla in" i någon av verblådorna, t ex passiv.

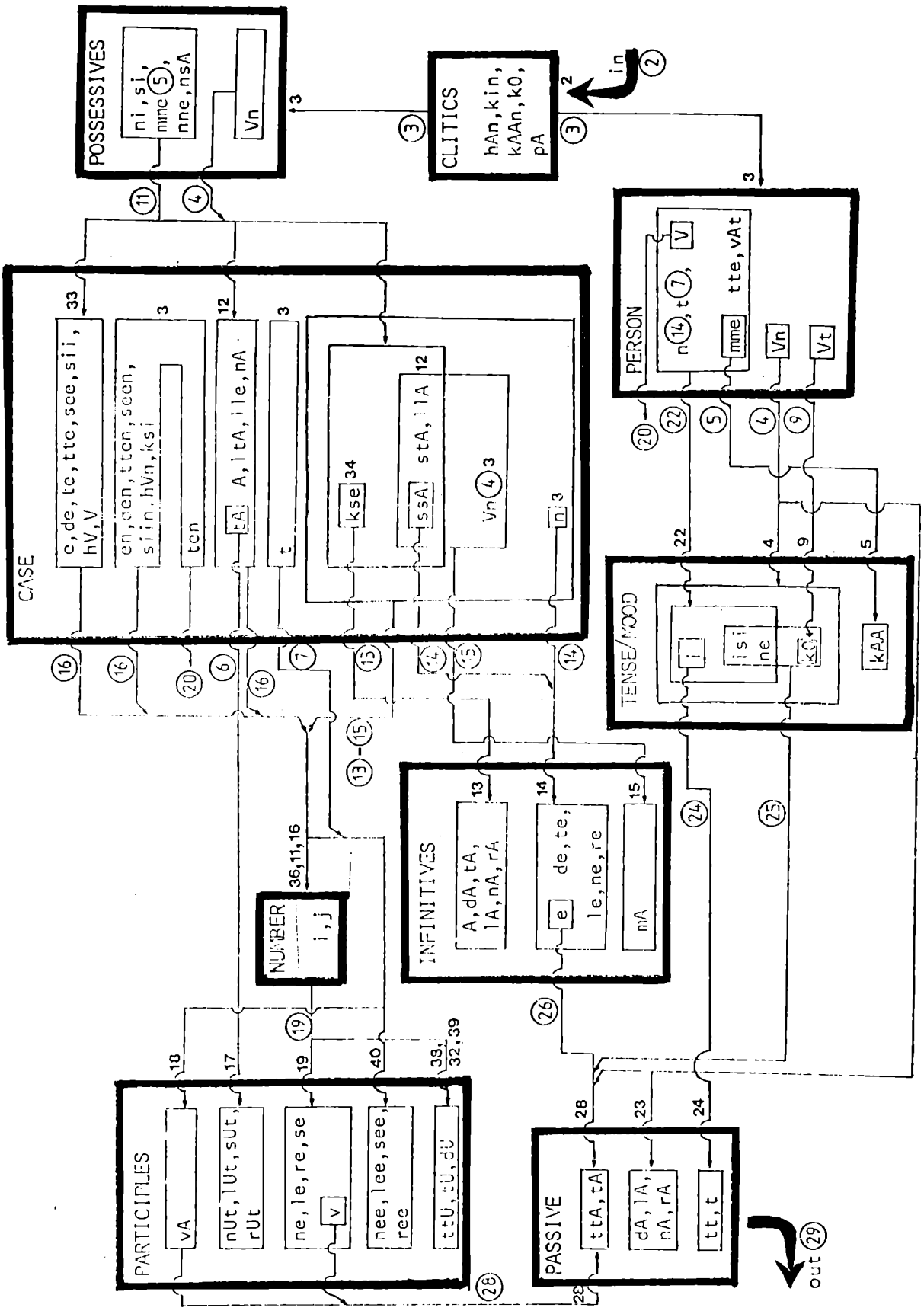


Fig. 2 Analysschema för finskan.

Siffrorna i figuren refererar till tillstånd och tillståndsvillkor i det regelsystem som utarbetats för att implementera schemat.

Siffra till vänster om låda representerar (namnet på) det utgående tillstånd som råder efter det att ett motsvarande element utsegmenterats, siffra till höger om låda är (namnet på) det villkor som måste vara uppfyllt för att man skall få passera in i lådan. Således innebär t ex villkoret 13 att något av tillstånden 2, 3 eller 13 skall råda (framgår ej av fig. 2), dvs villkoret för att infinitiverna A, dA, tA,... skall få antas föreligga är att antingen ordslut (tillstånd 2), enklit (3) eller kasus kse (13) "föregått" (i höger-till-vänster-ordningen) segmentet i fråga. Som jag tidigare påpekat är denna frikoppling mellan tillstånd och villkor en av de saker som starkast bidragit till att göra BETA till ett programmeringsspråk snarare än bara ett sätt att skriva kontextkänsliga regler.

6. Ett experiment med latinsk morfologi.

Det latinska exemplet tar jag med här för att få tillfälle att demonstrera hur ett relativt enkelt men ändock icke trivialt BETA-regelsystem kan se ut. Regelsystemet är framtaget i samarbete med docent Tore Janson, Stockholm, för att testa en hypotes om ändelsernas fonotaktiska struktur i latinet. Undersökningen som sådan och dess teoretiska implikationer finns presenterade i Janson, -79.

Det systemet är mycket enkelt att beskriva: det "strippar av" längsta möjliga sekvens av stavelser från ordslut med strukturen $(V1 + C1)^k$ (V1 + C2) där V1 är en delmängd av latinets vokaler, och C1 och C2 vissa genom uppräkning definierade konsonanter och konsonantkombinationer (sammanlagt ett 10-tal). Dock skall åtminstone en vokal lämnas kvar ute i ordet. Analysen går till så att stavelserna segmenteras ut en efter en från ordslutet räknat tills dess strukturvillkoret inte längre är uppfyllt eller tills dess man kommit till ordbörjan. I det senare fallet lämnas den "vänstraste ändelsen" tillbaka till stammen, och i övrigt tages alla utom den första av de införda segmentgränserna bort.

För att få bort en del ovidkommande svårigheter gjordes ett första

pass från vänster till höger där vokalen AE skrevs om till Ä, QU till QW och AU till AW (detta för att få varje grafisk vokal att beteckna en fonologisk vokal), samt avskiljdes de enklitiska konjunktionerna -QUE och -VE; dessa är alltså de enda morfem som är instoppade i systemets "lexikon". Detta sista kan vara särskilt värt att poängtera; systemet gav - med hänsyn tagen till den minst sagt magra information som fanns inlagd i det - en nästan orimligt bra analys. Enligt Jansons beräkningar erhölls 65 à 70-procentigt korrekt ändelseanalyserad text (beräknat på löpande text), ett resultat som naturligtvis stödde den uppställda hypotesen om ändelsestrukturen i latinet mycket starkt.

Det regelsystem som bifogas som illustration (Appendix II) är förenklat därhän att det förutsätter att ett ord i taget matas in.

Litteraturlista:

Brodda, B., 1977. BETA-systemet: en sammanfattning, i Nord. datalingvistdagar i Göteborg, okt 1977, Martin Gellerstam (red.). Rapport no 3 från Språkdata, Göteborg 1977.

Brodda, B., 1979a. BETA - en kort presentation. SAML 5, Köpenhamn 1979, Inst för Anv. og Mat. lingvistik, Köpenhamn 1979, även i COMPILING, maj 1969.

Brodda, B., 1979b. Något om excerperingsprogram i allmänhet och ett exempel i Konkordanser: Föredrag från 2:a svenska symposiet i språklig databehandling i Lund 1979, Thavenius-Oreström (red.). SSE-projektet, Lunds universitet 1979.

Brodda, B., 1979c. Något om de svenska ordens fonotax och morfotax - iakttagelser med utgångspunkt från experiment med automatisk morfologisk analys. PILUS nr 38, Inst för lingvistik, Stockholms Universitet, dec 1979.

Brodda, B. och Karlsson, F., 1980. An experiment with automatic morphological analysis of Finnish (prel. titel). PILUS feb 1980.

Janson, T., 1979. Mechanisms in Language change (kap 4), Acta Universitatis Stockholmiensis (Studia Latina), Stockholm 1979.

Karlsson, F., 1978. Finsk Grammatik. Soumalaisen Kirjallisuuden Seura, Helsingfors 1978.

Appendix 1: datorillustrationer.

Ill 1

1 BE-MÄKTIG=ADE
1 BE-MÄSTRA
1 BE-MOT/ANDE
2 BENGTSSON
1 BENKOT=ORNA
1 BE-RED'D
3 BE-RED'DE
1 BERGGRUND=EN
1 BERGKLACK
3 BE-ROD'DE
1 BE-RO/ENDE
1 BERSÅ
1 BE-RUS=ADE
2 BE-RÄKN-ADE
2 BE-RÄTTA
1 BE-RÖMT
1 BE-SKREV
1 BE-SKRIV=NA
1 BE-SLUT
1 BE-SLUTA
1 BE-SLUT=AT
1 BE-SLUT=EN
1 BE-SLUT=NA
1 BE-SLUT/SAM
1 BE-SPARING=AR
1 BE-STIGA
6 BE-STOD
1 BE-STODS
1 BE-STRÅL=ADE
1 BE-STÄLL-SAMMA

Ill 2

1 O'M-TAL=ADES
1 OM-TAL=ADES

1 O'M-TAL=ATS
1 OM-TAL=ATS

1 O'M-VÄXLING
1 OM-VÄXLING

2 O-RDENT=LIG
2 ORDENT/LIG

3 O-RDENT=LIGT
3 ORDENT/LIGT

4 O-RDNING
4 ORDNING

1 O-RDSPRÅK
1 ORDSPRÅK

1 O-RKESLÖSA
1 ORKESLÖSA

1 O-RMFAR=AN
1 ORMFAR=AN

1 O-RSAK=EN
1 ORSAK=EN

(Utskrifterna ovan kommenteras på sid 9.)

Ill. 3. Analysgång: svenska

				Kommentarer
11	*BEARBETNINGSBEHOV			
44	BE	BE-	*BEARBETNINGSBEHOV	pröva BE-!
44	A	A	BE-*ARBETNINGSBEHOV	följs av A! (ok)
44	R	R	BE-A*RBETNINGSBEHOV	notera F-kluster!
44	BE	-BE	BE-AR*BETNINGSBEHOV	tillåter internt -BE!
44	BE	BE-	BE-AR-*BETNINGSBEHOV	pröva BE-!
44	T	T	BE-AR-BE-*TNINGSBEHOV	följs av I-kluster? (nej)
44	-		BE-AR-BE*-TNINGSBEHOV	sudda införda - -!
44	-		BE-AR*-BETNINGSBEHOV	
44	E	E	BE-ARB*ETNINGSBEHOV	notera vokalpassage!
44	T	T	BE-ARBE*TNINGSBEHOV	notera F-kluster!
44	NING	/NING	BE-ARBET*NINGSBEHOV	pröva /NING!
44	NING	NING	BE-ARBET/*NINGSBEHOV	
44	S	´S-	BE-ARBET/NING*SBEHOV	följs av S! (ok)
44	BE	BE-	BE-ARBET/NING´S-*BEHOV	tillåter BE-!
44	H	H	BE-ARBET/NING´S-BE-*HOV	följs av I-kluster? (ja)
44	O	O	BE-ARBET/NING´S-BE-H*OV	notera vokalpassage!
44	V	V	BE-ARBET/NING´S-BE-HO*V	notera F-kluster!
22	BE-ARBET/NING´S-BE-HOV*			färdig!

Ill. 4. Analysgång: finska

11	*JÄTETTÄISIINKIN			"det skulle också lämnas"
44	Ä	Ä´	J*ÄTETTÄISIINKIN	notera stamvokal!
44	´	´	JÄ*´TETTÄISIINKIN	och dess plats!
44	KIN	=KIN	JÄ´TETTÄISIIN*KIN	notera enklit KIN!
44	N	N	JÄ´TETTÄISII*N=KIN	fortsätt åt vänster!
44	IN	=IN	JÄ´TETTÄISI*IN=KIN	notera "4:e person"!
44	I=	=I=	JÄ´TETTÄIS*I=IN=KIN	pröva =I!
44	S	S	JÄ´TETTÄI*S=I=IN=KIN	fortsätt åt vänster!
44	IS=I=	=ISI=	JÄ´TETTÄ*IS=I=IN=KIN	ompröva ISI!
44	Ä	Ä	JÄ´TETT*Ä=ISI=IN=KIN	fortsätt åt vänster!
44	T	T	JÄ´TET*TÄ=ISI=IN=KIN	fortsätt åt vänster!
44	TTÄ	=TTÄ	JÄ´TE*TTÄ=ISI=IN=KIN	notera passiv TTÄ!
44	E	E	JÄ´T*E=TTÄ=ISI=IN=KIN	fortsätt åt vänster!
44	T	T	JÄ´*TE=TTÄ=ISI=IN=KIN	fortsätt åt vänster!
44	´	´	JÄ*´TE=TTÄ=ISI=IN=KIN	stryk stamvokalmärke!
22	JÄTE=TTÄ=ISI=IN=KIN*			färdig!

Ill. 5. Analysgång: latin

11	*LEGATUS#			"utsänd (nominativ)"
44	#	#	LEGATUS*#	gå åt vänster från ordslut
44	S	S	LEGATU*S	S tillhör C2, fortsätt!
44	U	U	LEGAT*US	U tillhör V1, segmentera!
44	=	=	LEGAT*=US	kolla segmentgräns!
44	T=	T=	LEGA*T=US	T tillhör C1, fortsätt!
44	A	A	LEG*AT=US	A tillhör V1, segmentera!
44	=	=	LEG*=AT=US	kolla segmentgräns!
44	G	G	LE*G=AT=US	G tillhör ej C1, gå höger.
44	=	=	LEG*=AT=US	notera 1:a segmentgräns!
44	=		LEG=AT*=US	stryk 2:a segmentgräns!
22	LEG=ATUS*			färdig!

111. 6. Exempel på analyserad löpande text.

Textexempel 1. Svenska. (- markerar prefix, / suffix, = ändelse.)

DET ÄR MÖJ/LIG=T ATT DESSA FRÅG=OR ÄR AV UNDER-ORDNAD BE-TYD/ELSE
IN-OM SATSGRAMMATIK=EN, OM MAN NÄM/LIGEN KAN AN-TA ATT DET RÅD=ER
ÖVER-ENSSTÄMM/ELSE MELLAN DE REGL=ER SOM GENERER=AR EN SATS HOS
EN TAL=ARE OCH DE REGL=ER SOM EN LYSSN=ARE AN-VÄND=ER FÖR ATT
AV-GÖRA OM EN SATS ÄR KORREKT. DET ÄR MER TVEK/SAM=T OM MAN KAN
AN-TA SAM-MA ÖVER-ENSSTÄMM/ELSE MELLAN DE REGL=ER HOS TAL=AREN
SOM KOD=AR ETT VISS´T KOGNITIVT INNE-HÅLL SPRÅK/LIG=T OCH DE
REGL=ER HOS LYSSN=AREN SOM TOLK=AR DETTA. DET GÖR ATT
TEXTLINGVISTIK=EN AR-BE-TAR UT-I-FRÅN ANDRA FÖR-UT-SÄTT/NING=AR
ÄN SATSLINGVISTIK=EN. SKILL/NAD=EN MELLAN AV-SÄND=ARE OCH
MOTTAG=ARE TAR SIG DÄR ANDRA UT-TRYCK OCH MAN KAN INTE BARA
BE-TRAKTA DEN ENA SOM OM-VÄND/NING=EN AV DEN ANDRA. DET BLIR
NÖDVÄND=IGT ATT KLARGÖRA UR VILK=EN SYNPUNK´T MAN BE-TRAKT=AR
TEXTEN.
DET FINNS EN MÄNG´D O-LIKA KOMMUNIKA/TION´S-MODELL=ER UPP-STÄLLDA.
ALLA INNE-HÅLL=ER ÅT-MINSTONE AV-SÄND=ARE, MED-DEL/ANDE OCH...

Textexempel 2. Finska. (≠ markerar felaktigt införd segmentgräns.)

TÄMÄ=N KOKOELMA=N KIRJOITUKSE=T +O=VAT PARI=A KOLME=A LUKU=UN
OTTA=MA=TTA SYNTY=NEE=T VI≠I=DEN VIIME VUO≠DE=N AIKA=NA. ERÄÄ=T
NI=I=STÄ ON JULKISTE=TTU LEHDISTÖ=SSÄ, ERÄÄ=T RADIO=SSA, ERÄÄ=T
ESITELM=I=NÄ. AIHEPIIRI ON VERRA=TE=N KIRJAVA: MUKA=NA ON
ENSIN=NÄ=KIN TIEDEPOLITIIKKA=AN JA TIETEENFILOSOFI≠A=AN
LIITTY=V=I=Ä KIRJOITUKS=I=A, TOISE=KSI OPPIHISTORIALLIS=I=A
PAKINO=I=TA, KOLMANNE=KSI LUONNONTIETEILIJÄ=N
MAAILMANKATSOMU≠STA JA ETIIKA=A SIVUA=V=I=A ARTIKKELE=I=TA JA
NELJÄNNE=KSI VIELÄ AINEISTO=A, +JO=TA VO=ISI NIMITTÄ=Ä
YMPÄRISTÖTIETEELLIS-FUTUROLOGISE=KSI. ERI AIHEALUE=I=DEN
VÄL≠I=LLÄ +EI +KUITENKAAN OLE JYRKK=I=Ä RAJO=J=A. KAIKK=I=A +TAI
A≠I≠NA=KIN USEIMP=I=A KIRJOITUKS=I=A YHDIS≠TÄ=VÄ KÄSITE ON 'TIEDE'
+TAI 'LUONNONTIEDE'. OLE=N +TOISINAAN KÄYTTÄ=NYT NÄ=I=TÄ SANO=J=A...

Textexempel 3. Latin. (≠ markerar felaktigt införd segmentgräns.)

HOSP≠ES, QWI NIHIL SUSPIC=ARETUR, VER=ITUS, NE QWID IN IPS=O SE
OFFEND=ERETUR, HOMIN=EM SUMM=A VI RETIN=ERE COEP=IT.
IST=E, QWI HOSP≠ITIS RELINQWEND=I CAWS=AM REP≠ERIRE NON POSS=ET,
AL≠IA SIB=I RATION=E VIAM MUN=IRE AD STUPR=UM COEP=IT;
RUBR=IUM, DELIC≠IAS SUAS, IN OMN=IBUS E≠IUS MON=I REB≠US AD≠IUTOREM
SUUM ET CONSC≠IUM, PAR=UM LAWT=E DEVERS=ARI DIC=IT;
AD PHILOD≠AMUM DEDUC=I IUB=ET.
QWOD UB=I EST PHILOD≠AMO NUNT≠IATUM, TAMETS≠I ER=AT IGN≠ARUS,
QWANT=UM SIB=I AC LIB≠ERIS SUIS IAM TUM MAL=I CONST≠ITUERETUR,
TAMEN AD IST=UM VEN=IT;
OSTEND=IT MUN=US ILLUD SUUM NON ESSE;
SE, CUM SUÄ PART=ES ESS=ENT HOSP≠ITUM RECIPIEND=ORUM, TUM IPS=OS
TAMEN PRÄT=ORES ET CONSUL=ES, NON LEG=ATORUM ADSECUL=AS, RECIP=ERE
SOL=ERE.
IST=E, QWI UN=A CUPID≠ITATE RAP=ERETUR, TOT=UM ILL=IUS POSTUL=ATUM
CAWS=AM QWE NEGLEX=IT;
PER VIM AD E=UM, QWI RECIP=ERE NON DEB=EBAT, RUBR≠IUM DEDUC=I
IMPERAV=IT.

```
TLST:              LATIN.X12       DATUM :07-OKT-79       PAGE:    1
```



```
  LATIN.X12       T. JANSON - B. BRODDA  OCT - 77
  PAR: 33 2

  DEFTYP
  2: 32-64
  1: #
  3: 48-57
  4: 65-127


  DEFSET
  1: 1
  2: 1 2
  3: 3 4
  4: 4
  5: 5
  6: 5 6
  7: 6 7
  10: 32-64
  11: 65-127
  12: B C D F G H J K L M N P Q R S T V W X Z
  13: A O U E I Å Y
  14: B C D F G H J K L M N P Q R S T V W X Z . ' , ? # 32
  15: I U


  FORMAT(4A1,X,4A1,5I4)


  !X!   Y!   LC  RC  SC  RS  MV

  !AVDELA NÅGRA ENKLITIKA I 1:A PASSET
  QUE!  QWE  11  10   2   3   2
  VE!   VE!  11  10   2   3   2


  ! GÖER NÅGRA ENKLA SUBST I FÖRSTA PASSET
  AE!   Ä!    0   0   2   0   5
  QU!   QW!   0   0   2   0   5
  AU!   AW!   0   0   2   0   5

  ! BÖRJA BACKA VID ORDSLUT
  #!    #!   11   0   2   3   2

  ! TILLBAKS VID ORDBÖRJAN IGEN. SUDDA 1:A '='
  #!    #!    0   0   4   5   5

  ! BACKA ÖVER KONSONANT SÅ LÄNGE TILLST 3 GÄLLER
  B!    B!   14   0   3   0   2
  C!    C!   14   0   3   0   2
  D!    D!   14   0   3   0   2
  F!    F!   14   0   3   0   2
  G!    G!   14   0   3   0   2
  H!    H!   14   0   3   0   2
  L!    L!   14   0   3   0   2
  M!    M!   14   0   3   0   2
  N!    N!   14   0   3   0   2
  P!    P!   14   0   3   0   2
  Q!    Q!   14   0   3   0   2
  R!    R!   14   0   3   0   2
```

```
S|    S|      14    0    3    0    2
T|    T|      14    0    3    0    2
V|    V|      14    0    3    0    2
W|    W|      14    0    3    0    2
X|    X|      14    0    3    0    2
Z|    Z|      14    0    3    0    2

{ OEVER TILL VOK VID FINALT ANDELSEKLUSTER (=  2)
M|    M|      11   10    3    0    2
R|    R|      11   10    3    0    2
S|    S|      11   10    3    0    2
T|    T|      11   10    3    0    2
NT|   NT|     11   10    3    0    2

{ INFOR MARKOR VID ANDELSEVOK.  (=  1)
A|    =A|      0   14    3    4    3
E|    =E|      0   14    3    4    3
I|    =I|      0    0    3    4    3
O|    =O|      0   14    3    4    3
U|    =U|      0    0    3    4    3
Ä|    =Ä|      0   10    3    4    3

{ Ä ÄR INGEN ANDELSEVOKAL. FÄRDIG|
Ä|    Ä|      0    0    3    6    5
Y|    Y|      0    0    3    6    5

{ AGERA PÅ =
=|    |      15   13    4    0    2
=|    =|     12   13    4    0    2
=|    =|     13   13    0    7    5
=|    |      10   13    4    6    5
=|    |       0    0    5    6    5
=|    =|      0    0    6    7    5
=|    |       0    0    7    0    5

{ BACKA OVER MEDIALA ANDELSEKLUSTRAR  (  1)
B=|   B=|    13    0    4    0    2
M=|   M=|    13    0    4    0    2
R=|   R=|    13    0    4    0    2
T=|   T=|    13    0    4    0    2
NT=|  NT=|   13    0    4    0    2
SS=|  SS=|   13    0    4    0    2
ST=|  ST=|   13    0    4    0    2

{ NÄHÄ, DET GICK INTE. FÄRDIG
B|    B|      13    0    3    6    5
C|    C|      13    0    3    6    5
D|    D|      13    0    3    6    5
F|    F|      13    0    3    6    5
G|    G|      13    0    3    6    5
H|    H|      13    0    3    6    5
L|    L|      13    0    3    6    5
M|    M|      13    0    3    6    5
N|    N|      13    0    3    6    5
P|    P|      13    0    3    6    5
Q|    Q|      13    0    3    6    5
R|    R|      13    0    3    6    5
```

R|	R|	13	0	3	6	5
S|	S|	13	0	3	6	5
T|	T|	13	0	3	6	5
V|	V|	13	0	3	6	2
X|	X|	13	0	3	6	2
Z|	Z|	13	0	3	6	2
J|	J|	0	0	3	6	5
K|	K|	0	0	3	6	5
W|	W|	0	0	3	6	5

SPRÅKDATA
Rolf Gavare
1979-09-10

AUTOMATISK LEMMATISERING UTAN STAMLEXIKON

Några synpunkter tio år efteråt

En tillbakablick

För precis tio år sedan, hösten 1969, genomfördes det första stora lemmatiseringsarbetet vid Språkdata. Det var den bearbetning som kom att bilda grundmaterialet till Nusvensk frekvensordbok 2. Två år senare gjordes en lemmatisering av Svenska psalmboken, med oförändrad programvara men med lite andra förutsättningar, vilket jag skall återkomma till. Trots att såväl datortekniken som de datalingvistiska metoderna utvecklats starkt under de år som gått, har erfarenheterna från det nämnda projektet fortfarande stor aktualitet. Jag skall först belysa några av de speciella omständigheter som gällde vid de båda nämnda körningarna, för att sedan skissera en modern implementering av interaktiv lemmatisering, som en vidareutveckling av den gamla modellen.

Grunddragen i lemmatiseringsmodellen framgår av Staffan Hellbergs bifogade artikel Computerized Lemmatization without the Use of a Dictionary. Sammanfattningsvis krävs enligt denna modell att de komponenter som skall ingå i lemmat uppfyller de tre huvudkriterierna: (1) alla enheterna skall ha en identisk stam, (2) ändelserna skall tillhöra samma paradigm och (3) eventuella ordklass/lemmabeteckningar skall vara lika och dessutom passa till samtliga komponenters ändelser. Några speciella förutsättningar som gällde det aktuella projektet bör hållas i minnet. Algoritmen var utformad för att användas på ett alfabetiskt sorterat, tidigare homografseparerat material. (Att det var alfabetiskt ordnat betyder förstås inte att de enheter som skall grupperas samman behövde stå intill varandra.) Det material som redovisas i Nusvensk frekvensordbok omfattade 1 miljon löpande ord, vilket gav 103 000 olika graford och 112 000 homografkomponenter (cirka 30 % av de olika graforden var homografa). Nära 97 % av homografkomponenterna placerades automatiskt i rätt lemma. Av de rätt avgränsade lemmana fick 85 % också rätt rubrik med uppslagsform, ordklass- och lemmabeteckning. Svenska psalmbokens 8500 olika graford (av c. 110 000

löpande ord) var inte homografseparerade och gav därför ett sämre resultat – omkring 80 % av de oseparerade orden hamnade i rätt lemma (d.v.s. rätt för åtminstone någon homografkomponent av det grafiska ordet). Ytterligare skäl till att psalmbokslemmatiseringen gav sämre utfall är att paradigmtabellerna var utformade med tanke på sannolikheter för uppträdande i modern svenska och framför allt, ju större ordmaterialet är, ju fler olika böjningsformer är belagda inom varje lemma, vilket i sin tur ger en större säkerhet vid såväl sammanföringen av böjningsformer som etableringen av en lemmarubrik. Algoritmen är inte utformad för lemmatisering av enstaka ord.

Bland de tekniska förutsättningarna märks särskilt att vi vid den aktuella tidpunkten saknade direktaccessminne (skivminne). I stället fick magnetbandstationer utnyttjas för lagring av arbetsfiler, vilket också i viss mån kom att återspeglas i programlogiken. Programutrymmet var också begränsat. Någon möjlighet till interaktion via terminaler fanns inte. Trots de praktiska begränsningarna får man ändå anse att den automatiska lemmatiseringen lyckades väl.

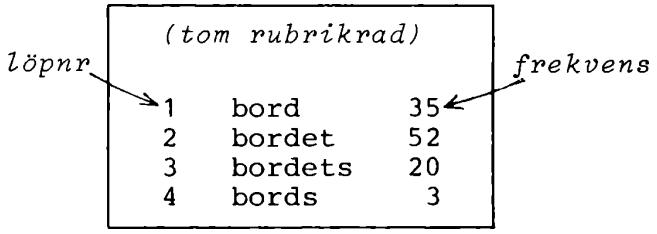
Dagsläget

Jag skall nu övergå till att skissera huvuddragen av hur en lemmatisering (inklusive homografseparering), byggd på dessa principer, skulle kunna gå till idag.

1. Antag att vi har ett stort, obearbetat textmaterial (på grafordsnivå). En första åtgärd blir att ta fram en komplett konkordans i radskrivarutskrift (eller på mikrokort). Om möjligt skall konkordansen också vara tillgänglig i datorn för sökning direkt från terminalen. Både versaler och gemena tecken skall återges i utskriften för att underlätta bedömningen av egennamn, initialförkortningar och versal i meningsbörjan. Konkordansen skall vara ordinärt alfabetiskt sorterad på de alfabetiska tecknen i stickordet och högerkontexten. Beläggen av samma (normaliserade) graford förses med en löpnumrering. En sådan konkordans har ett mycket stort värde även bortsett från den här avsedda lemmatiseringen/homografsepareringen.

2. En ändelse-/paradigmlista upprättas i enlighet med Staffan Hellbergs alternativa metod. Något stamlexikon skall inte användas.

3. Lemmatiseringen (utan föregående homografseparering) försiggår on-line vid en textskärmsterminal. Detta förfarande minimerar omfattningen av felspridning vid lemmatiseringen. De sublemman (graford) som programmet anser skall bilda det aktuella lemmat visas på skärmen, exempelvis på det här sättet:



Lemmatiseringsalgoritmen bygger i huvudsak på Staffan Hellbergs alternativa modell. Eventuellt kan denna kompletteras för att ge ett 'intelligent' förslag till homografseparering automatiskt med hjälp av särskilt homograflexikon och viss kontextkontroll.

4. Med enkla kommandon skall det aktuella lemmat kunna modifieras:
   a. sublemman skall kunna tas bort ur lemmat (och behandlas senare),
   b. sublemman skall kunna flyttas till ett tidigare lemma (rubriken anges och lemmat ifråga visas på skärmen för kontroll),
   c. sublemman skall kunna separeras (vissa belägg (vars beläggställen hämtas ur konkordansen) bildar en ny sublemmaform, som sedan eventuellt flyttas ur lemmat),
   d. sublemman skall kunna märkas för intern homografi och ev. polysemi,
   e. tidigare behandlade lemman skall när som helst kunna inspekteras, modifieras och skrivas ut.

5. Då lemmats sublemmauppsättning godkänts föreslår programmet en lemmarubrik med ordklass- och lemmabeteckning.

6. Efter kontroll och eventuell ändring av lemmarubriken presenteras nästa lemma på skärmen (enligt punkt 3).

7. Helst bör möjlighet finnas att dynamiskt uppdatera paradigm/ändelseregistret under körningens gång.

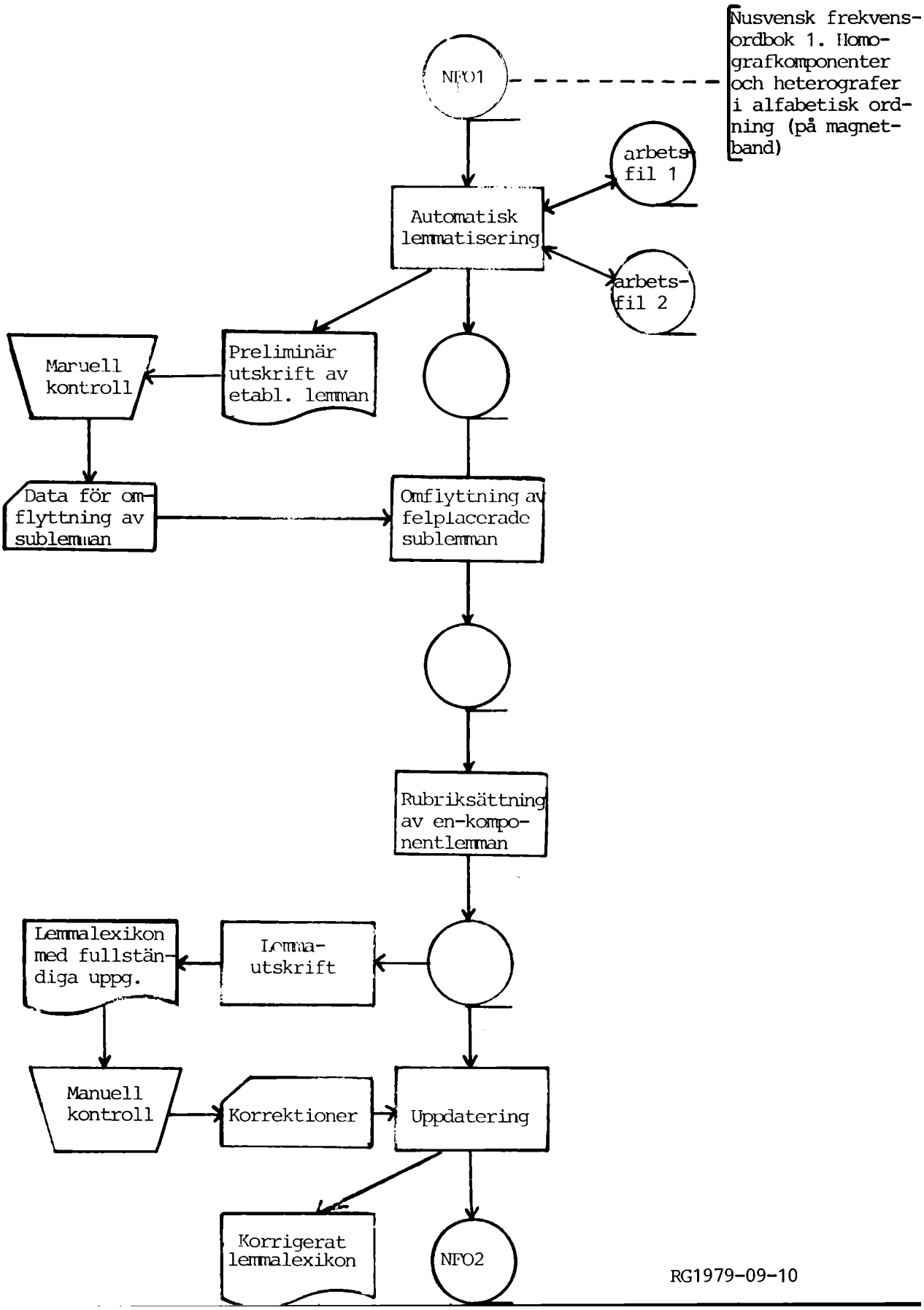
Det skisserade förslaget är utformat med tanke på krav på språkvetenskaplig noggrannhet. Helautomatisk lemmatisering med lägre krav på korrekthet kan förstås också i vissa fall vara önskvärd.

Att etablera grundformen till enstaka ordformer ställer andra krav på modellen. För att få en god säkerhet i detta fall behövs förmodligen ett morfem- eller stamlexikon, exempelvis av den typ som användes inom projektet Algoritmisk textanalys, se vidare Staffan Hellbergs bok The Morphology of Present-Day Swedish (Data linguistica 13, Almqvist & Wiksell International, Stockholm 1978).

51530	FOTOGENLAMPA-	NN -N			
1	FOTOGENLAMPA			0	
2	FOTOGENLAMPAN			0	
51540	FOTOGRAF-	NN -EN			
1	FOTOGRAF			0	*
2	FOTOGRAFEN			0	EN
3	FOTOGRAFENS			0	EN
4	FOTOGRAFER			0	
5	FOTOGRAFERNA			0	
6	FOTOGRAFERNAS			0	
51550	FOTOGRAFER-A	VB -AD			
1	FOTOGRAFERA	VB -AD	INP		
2	FOTOGRAFERA	VB -AD	INF		
3	FOTOGRAFERAD			A	
4	FOTOGRAFERADE	VB -AD	PRT		
5	FOTOGRAFERADE	VB -AD	PTP		
6	FOTOGRAFERADES	VB -AD	PRT		
7	FOTOGRAFERANDE	VB -AD			
8	FOTOGRAFERAR			A	
9	FOTOGRAFERAS	VB -AD	INF		
10	FOTOGRAFERAT	VB -AD	SUM		
51560	FOTOGRAFERING-	NN -EN			
1	FOTOGRAFERING			0	*
2	FOTOGRAFERINGEN			0	EN
51570	**********-				
1	FOTOGRAFERINGSFÖRBUDET				
51580	FOTOGRAFI-	NN -ET			
1	FOTOGRAFI			0	
2	FOTOGRAFIER			ER 0	UM
3	FOTOGRAFIERNA			0	UM
4	FOTOGRAFIET			0	UM
51590	FOTOGRAFISK-	AV -T			
1	FOTOGRAFISK				
2	FOTOGRAFISKA	AV -T	NEU	0	
3	FOTOGRAFISKA	AV -T	PLU		
51600	**********-	AB			
1	FOTOGRAFISKT	AB			
51610	**********-				
1	FOTOHISTORIA				

NUSVENSK FREKVENSORDBOK 2

Lemmatiseringsfasen (1969/70)



RG1979-09-10

Computerized Lemmatization without the Use of a Dictionary: A Case Study from Swedish Lexicology

Staffan Hellberg

Lemmatization, i.e., the bringing together of the inflectional forms (and variant forms) of a word under one heading, is one of the problems when making a frequency dictionary out of a large text corpus with the aid of a computer. Attempts have generally gone in the direction of confronting the material with an ordinary dictionary, presupposing that this dictionary would have an entry for practically every form in the corpus. This may be true for some texts, e.g., the classical ones, but it is definitely not true for a newspaper text corpus in a language like Swedish, which not only shows brand-new loan-words but is abundant in compounds of the more or less casual sort that will never appear in ordinary dictionaries. So the task we undertook in 1969-70 at the Research Group for Modern Swedish, University of Göteborg, was to lemmatize automatically about 112,000 different word forms without direct access to any existing dictionary. Homographs had been previously separated with the aid of a KWIC-index (the original number of different graphic words was about 103,000), a fact which meant that about one-third of the forms had been assigned grammatical information (word class and, roughly, gender or conjugation).[1]

As Swedish contains no inflectional prefixes, the procedure can operate with an alphabetically sorted version of the material. The computer passes through that version, successively grouping the forms into lemmas and printing them out, so that the whole lemmatization can be checked manually afterwards.

The program tests the first form in the projected lemma against the ones following alphabetically, one at a time, providing they have not already been included in a previously finished lemma. As soon as a form appears which is not identical with the first form as far as the stem of the latter goes, the testing is stopped and the lemma finished. The form that heads the remainder of forms alphabetically is then chosen as the first form of the next lemma, and the procedure is repeated.

[1] Allén, *Nusvensk frekvensordbok/Frequency dictionary of present-day Swedish 1*, Stockholm: Almqvist and Wiksell, 1970, presents the material on the level where homographs had been separated but no lemmatization done. The lemmatized version was published in 1971 as Allén, *Nusvensk frekvensordbok/Frequency dictionary of present-day Swedish* 2. A more detailed report of the lemmatization has been made (in Swedish) in Staffan Hellberg, *Automatisk lemmatisering*, 1971 (mimeographed). A general survey of the work at the Research Group will be found in Sture Allén, "Vocabulary data processing," *Proceedings of the International Conference of Nordic and General Linguistics, Reykjavik, 1969*, ed. Hreinn Benediktsson, Reykjavik 1970.

Staffan Hellberg is a member of the Research Group for Modern Swedish at the University of Göteborg.

The "stem" of a lemma thus had to be defined as the part of the word that was identical in all its inflectional forms. The remnants of the forms were called "endings." Obviously, these definitions don't altogether coincide with the usual linguistic ones: the word *titel,* 'title,' plural *titlar,* for instance, got the stem *tit-* and the endings *-el, -lar,* etc., though linguistically, the stem should rather be *titl-* and the plural ending *-ar.* An index was set up of those graphic sequences that might be endings in regular paradigms.[2] Lexical regularity proved not to be the same thing as grammatical regularity; for instance, an irregular noun occurring as the latter element in many compounds had to be taken account of. The word *man* 'man,' plural *män,* is as irregular as in English, but it appeared in over 150 compounds in the corpus, e.g., *adelsman,* 'nobleman,' plural *adelsmän,* and so a paradigm *-an, -än,* etc. was established. In all, 53 different paradigms were made the basis of the index, which contained 98 different endings. The figures give a somewhat exaggerated idea of the complexity of Swedish morphology, as one linguistic paradigm often had to be split into two or more paradigms here: compare *titel, titlar* (above) with the endings *-el, -lar,* etc. to *stol,* 'chair,' plural *stolar* with the endings *-ø, -ar,* etc.

For two forms to be brought into the same lemma, they were required to have an identical stem and compatible endings, i.e., such as could belong to the same paradigm. Whether the identity actually covered the whole stem was decided by checking whether the remnants of the forms were possible endings. So the index here served two purposes: to identify the latter parts of the forms as endings, and to give access to what was called the alpha-list, where for each ending the endings compatible with it were stored. But for the former procedure to function properly, it was necessary that every graph or graphic sequence Y which could not itself be an ending but which had a counterpart XY that was a possible ending appear in the index, where it was stored as a pseudo-ending with an empty alpha-row. An example is the final *-l* which didn't occur in any paradigm, while the sequence *-el* did (see *titel, titlar* above). In all, 14 pseudo-endings were required.

If the alpha test gave a negative result, it was repeated with the rightmost graph (roughly: letter) of the stem brought over to the endings, provided, of course, that these new endings were to be found in the index at all. But once a shorter stem had been recognized by a successful test of that kind, it was not allowed to be lengthened again as a result of a comparison with yet another form, because that would mean an obvious mixing of two paradigms.

The index served its third purpose when giving entry to the so-called beta-list, where the possible grammatical labels were given for each ending. The beta-list was consulted when one of the tested forms, or both, was a homograph and thus "marked" for grammatical category, and so a number of wrong lemmatizations could be prevented through the demand for grammatical compatibility. The beta-list was also used in the subprogram of automatic attributing of head forms and grammatical labels to all the lemmas, which will not be reported here.

The main course of the procedure is shown in the flow chart. Several improvements were suggested by our programmer, Rolf Gavare, who wrote the program in DATASAAB/ALGOL–GENIUS and DAC.

Some measures were taken to compress the lists. One of these made use of the structure of the Swedish inflectional system, where the ending *-s* plays a unique role. It always occupies the last position in the form, and it can be added to practically every form of nouns, adjectives, and verbs, having either a genitive or a passive function. If all those *s*-variants of the endings had been accounted for in the normal way, it would have meant nearly a doubling of the index and a considerable enlarging of the alpha-list.

[2] The bulk of the paradigms were taken from Björn Hammarberg, "Maskinell generering av böjningsformer och identifikation av ordklass," *Förhandlingar vid sammankomst för att dryfta frågor rörande svenskans beskrivning 3, Göteborg, 1965,* ed. Sture Allén, Göteborg, 1966 (mimeographed).

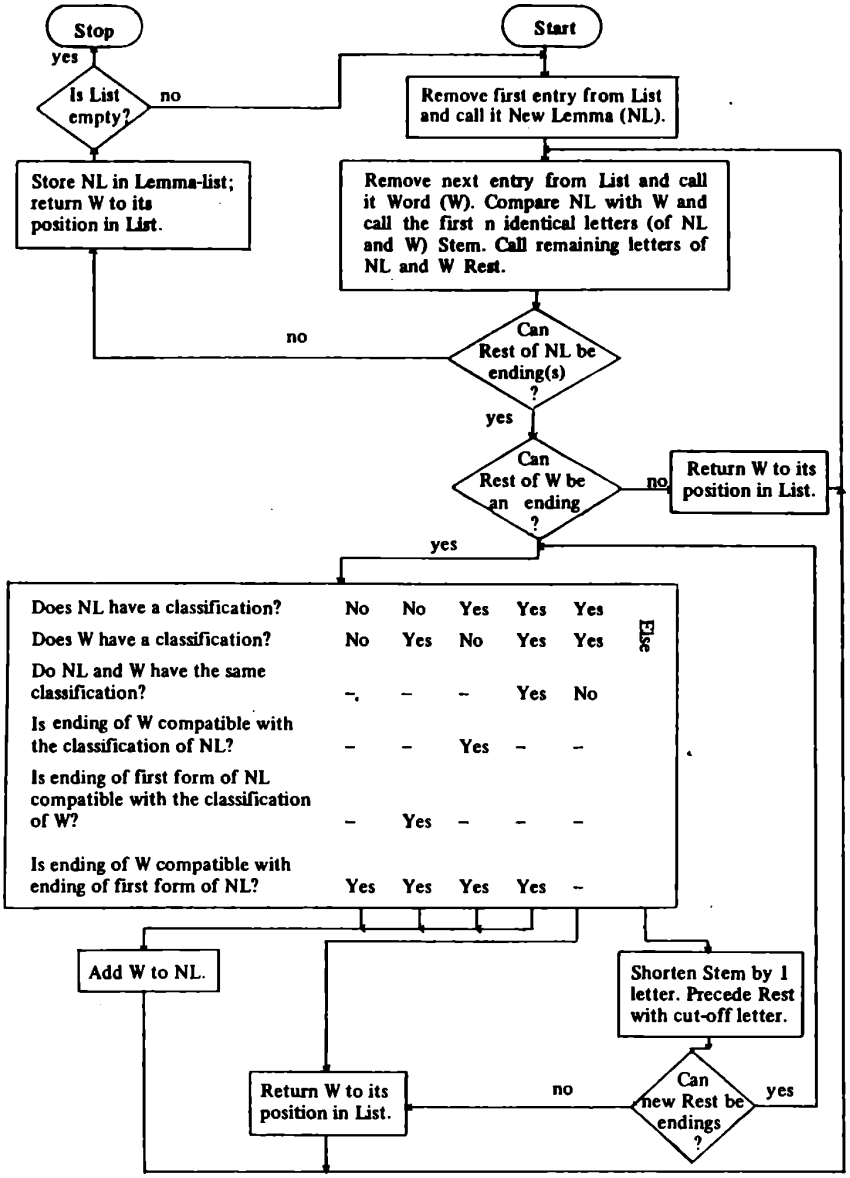


Figure 1

Instead, all forms ending with an *-s* were treated as if the *-s* wasn't there, except those where the *-s* belonged to the stem and which could be readily sorted out, as they were homographs internally with their own genitives and thus had a special "marking."

There were also quite a few *ad hoc* measures taken to obtain a better result, as several minor defects could be foreseen during the construction of the lists and by scrutinizing the result of test computations. Some of the measures simply meant omitting

an item from one of the lists, thereby replacing a number of wrong lemmatizations by a smaller number of missing correct ones. A measure of a different kind worth mentioning was the rearrangement of the alphabetically ordered material so that out of two homographs, one noun and one verb, the verb was placed before the noun. That saved a fair number of lemmas from going wrong.

The lemmatization yielded about 71,000 lemmas in all. The figure reveals that a large number of lemmas appeared in only one form. These lemmas did not cause any special troubles to the program, as a projected lemma could often be finished after its first form had been compared to—and shown too little similarity to—its nearest neighbor in the alphabetical order. A different subprogram had to be designed, though, for the attribution of head forms and grammatical labels (see above), as the beta-list gave no information in this case, where no boundary between stem and ending had been definitely established.

Though the whole corpus was treated in the manner now described, not all lemmas could, of course, be made to come out correctly from the computer. The program would have been hopelessly slow and complex if it had had to account for strong verbs, regular though they might be. There were also very rare paradigms that would have done more harm than good if they had been brought into the lists. In fact, the accomplished wrong lemmatizations are more notable than the missing correct ones. Not all clashes could be prevented by the above-mentioned *ad hoc* measures. And as the material also contained foreign words occurring in the newspaper corpus, there appeared a number of ridiculous lemmas, such as the one consisting of (English) *fair* and (French) *faire.*

The manual check of the computer output showed that 3.5 percent of the forms were in the wrong place and had to be moved. As this check was done with relative ease, the lemmatization program may well be said to have saved us from a considerable amount of dull routine work. Still, it could be asked whether the automatic procedure has actually been optimized. The number of wrong lemmatizations indicates that the alpha-list didn't have a sufficient discriminating function. This is actually natural for Swedish, where some sequences are very common as endings in different functions: the ending *-er,* for instance, occurs in 12 paradigms and is compatible with 29 other endings.

In closing, I will give a brief account of an alternative solution that I outlined after the computing of our material had been accomplished. In this solution, the ideas of alphabetical procedure and of an index of possible endings are taken over from the system used. But the alpha- and beta-lists are replaced by what could be called the gamma-list. That is, for each ending information is now given about which paradigms this ending can occur in, the paradigms having numbers from 1 to 53. For two forms to be brought together, it is now required that they have an identical stem and at least one paradigm number in common. If the common number or numbers are stored, a third form can be tested against them, and that means that any new tentative form will be tested against all the previously accepted forms in the lemma, which wasn't possible in the system used.

The beta-list is made superfluous by the grammatical labels being brought into the index and assigned paradigm numbers. So when two forms are tested, one of which is a homograph, it is required that at least one number occur three times in the gamma-list: with the two endings and with the grammatical label.

Most of the measures taken to improve the system used can be kept, as for instance the special treatment of forms ending in inflectional *-s.* Though the alternative solution hasn't been tested on the material, it seems fairly clear that it would have surpassed the one we chose. Of 12 different kinds of clashes that had been registered before the new system was developed, seven would have been avoided. What this would mean in figures is harder to guess. A reduction of the number of wrong lemmatizations by one-half is perhaps a somewhat too optimistic estimation.

NAVF'S EDB-SENTER FOR HUMANISTISK FORSKNING
VILLAVEI 10
POSTBOKS 53
5014 BERGEN - UNIVERSITET

Knut Hofland:

PRODUKSJON AV EN LEMMATISERT KONKORDANS TIL IBSENS SAMLEDE VERKER.

Ved NAVF's EDB-senter for humanistisk forskning er det i gang et prosjekt der målet er en lemmatisert konkordans til Ibsens skuespill og dikt. Prosjektet startet 1. april 1978 og skal gå over 3 år. Professor Harald Noreng er ansatt som faglig leder i hel stilling og konsulent Knut Hofland har ansvar for EDB-delen av prosjektet. Norges almenvitenskapelige forskningsråd finansierer tiltaket. Prosjektet disponerer ½ stilling til terminalarbeid og midler til timelønnet personale for vitenskapelig assistanse og til korrekturlesing og kontrollarbeid. All systemutvikling og programmering utføres av den faste staben ved NAVF's EDB-senter.

Tekstgrunnlaget er hundreårsutgaven av Ibsens verk og består av 28 skuespill derav 4 i to utgaver og et bind med dikt. Dette utgjør ca 3/4 million løpende ord.

Publiseringsmåten for konkordansen vil bli på mikrokort og det vil også bli aktuelt med en trykket utgave av en del av materialet, et sitatleksikon.

Ved tilrettelegging av teksten er det ført inn opplysninger om hvem en replikk er rettet til, type av scenehenvisning og markering av enderim etter type, der skuespill er på verseform.

Ibsen skifter rettskriving i 1870. Ved lemmatiseringen bruker en skrivemåten etter 1870 som grunnlag for oppslagsformen. Som en hjelp til brukeren av konkordansen (særlig i utlandet) innføres det ett sett med henvisninger. Dette gjelder fra moderne norsk rettskriving til Ibsens, fra Ibsen før 1870 til etter 1870 og ved sterk bøying fra ordform til oppslagsform.

Det er lagt opp til at hvert belegg kan få en nøye tilpasset kontekst. Deler av konteksten kan utelates og det kan også tilføyes opplysninger som f.eks. henvisning til pronomene. I tillegg til referansen vil det for hvert belegg være gitt opplysninger om hvem som har replikken og til hvem den er rettet.

Lemmatiseringen.

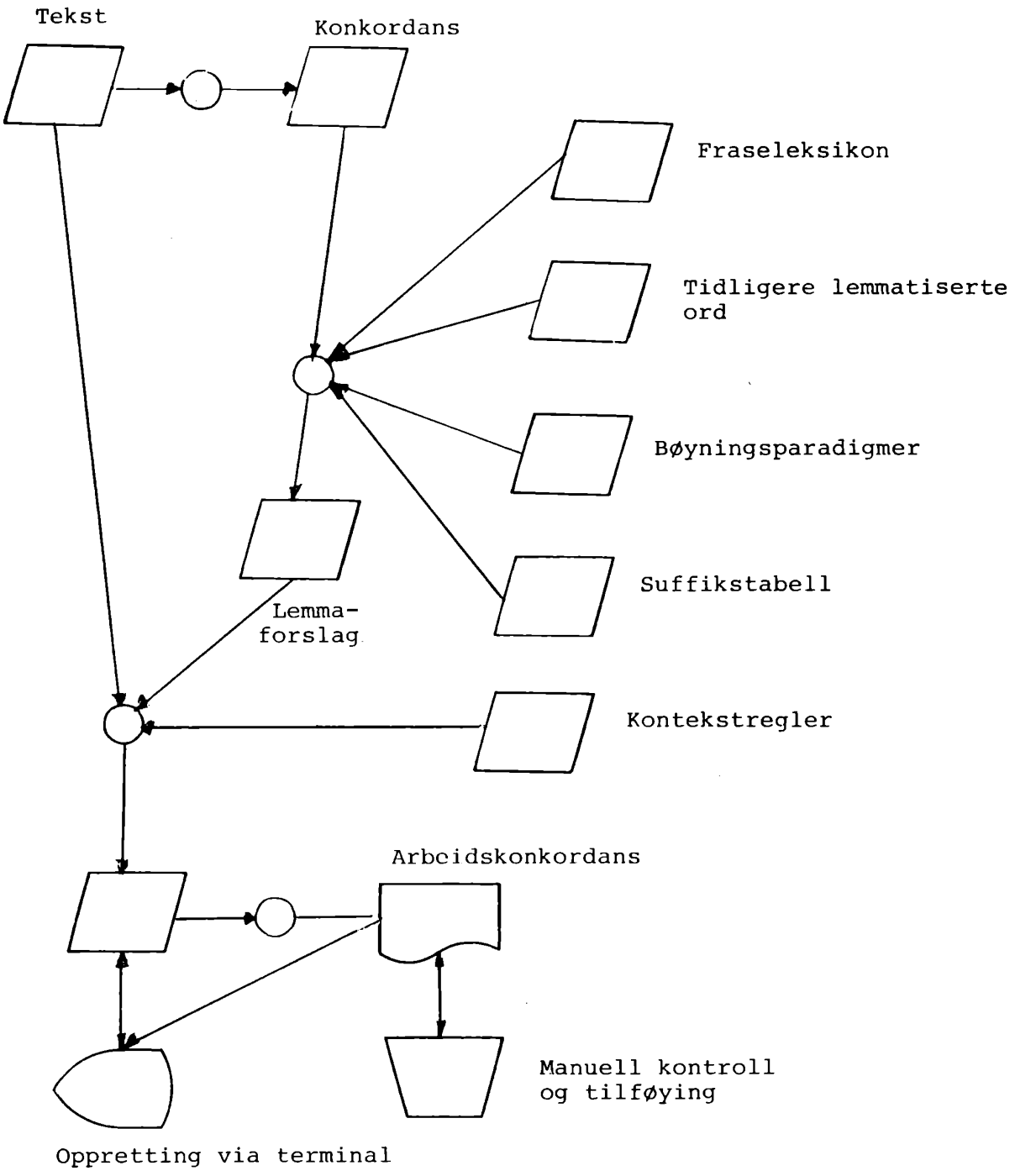
Tekstene lemmatiseres verk for verk. Den første teksten ble lemmatisert manuelt ut i fra en konkordans med 3 linjers kontekst.Det var her forslag til kontekst avgrensing basert på skilletegn. Ved de to neste verkene ble opplysningene fra det/de foregående verk stilt opp som forslag (se eksempel).

```
399 ********************************************************************************** 0 umeldt **********
400 L= 02685, O= 01, A= 001, S= 0061, FRA= 15, TIL= 1, SH-NR=
401    bitter Klage fra Din Læbe - / |EN GAMMEL SOLDAT (træder ind og nærmer sig Catilina) <1>
402  = Tilgiv mig, Herre! at jeg træder ind /  _umeldt_  ved Aftenstid her i din
403 Bolig; - /  =  Tilgiv - - - |CATILINA. <15>  - - Sig frem! hvad vil Du her? /
404 ********************************************************************************** 0 umild **********
405 L= 01286, O= 04, A= 002, S= 0077, FRA= 6, TIL= 7,8,9 & 10, SH-NR=
406       hører først, hvad jeg / vil Eder sige;  = - vi har os besluttet / med væbnet Haand
407      at ville vinde, hvad / en haard, en  _umild_  Skjæbne har os nægtet; / =  nu
408 undertrykkes vi, - vi ville herske; / vi lide Mangel, - Rigdom er vort Maal! / |PLEBE.
409 ********************************************************************************** 0 umuligt **********
409 #1  O: umuligt  GF: umulig  OK: adj.  BØ: n. sg., ubest.  BR:  MO:
410 #2  O: umuligt  GF: umulig  OK: adj.  BØ: n. sg., ubest.  BR: adv.  MO:
412 L= 02034, O= 10, A= 003, S= 0097, FRA= 1, TIL= 4, SH-NR=
413    - / |CURIUS (fortvivlet) <1> Nei, nei, Du er forraadt, din Plan er røbet. / |CATILINA.
414   <4> Ha, raser Du?  = Nei, nei, det er  _umuligt!_  / =  |CURIUS. <1> Jo det er
415 saa, men fly i denne Stund, / end kan maaskee ved hurtig Flugt Du reddes. / |CATILINA.
416 ---------------------------------------------------------------------------------- 0 umuligt
417 L= 02037, O= 13, A= 003, S= 0097, FRA= 1, TIL= 4, SH-NR=
418      men fly i denne Stund, / end kan maaskee ved hurtig Flugt Du reddes. / |CATILINA. <4>
419    = O, skulde det - - - men nei, det er  _umuligt!_  - / =  |CURIUS (river sin Dolk
420 ud og rækker den hen til Catilina.) <1> Der, Catilina! tag, igjennemboer mit Bryst - / -
421 ********************************************************************************** 0 under **********
421 #1  O: under  GF: under  OK: prep.  BØ:  BR:  MO:
422 #2  O: under  GF: under  OK: adv.  BØ:  BR:  MO:
423 #3  O: under  GF: unde  OK: verb  BØ: pr.  BR:  MO: unne
425 L= 00267, O= 02, A= 001, S= 0049, FRA= 1, TIL= 4, SH-NR=
426  det er vel Spøg? / |CATILINA. <4> En Spøg? ja vist! som al min Kjerlighed, / men Alvor er
427   dog, hvad jeg nys Dig sagde; - /  = da  _under_  sidste Fest igjennem Romas Gader /
428 i Procession Præstinderne jeg saae, / tilfældigviis paa een af dem jeg fæsted' / et  =
```

a) ny klassifikasjon
b) forslag til klassifikasjon rangert etter hyppighet i tidligere verk
c) ny kontekstavgrensing

Utsnitt av arbeidskonkordans

Oversikt over lemmatiseringen.



Ved det videre arbeid vil det i økende grad bli tatt i bruk maskinell lemmatisering. Metoden som skal brukes er tradisjonell og bygger bl.a. på arbeid som har vært gjort andre steder (1), (2), (3).

En oversikt over metoden er gjort i figuren. Metoden er basert på at forskeren skal kunne arbeide med konkordanser utskrevet på papir. På bakgrunn av markeringer i konkordansen gjøres rettinger/ tilføyelser via terminal.

På grunnlag av de tekster som er lemmatisert blir det laget maskinelt et sett med tabeller som brukes ved den automatiske lemmatiseringen. Den nye teksten som skal lemmatiseres kjøres først ut i en konkordans med begrenset kontekst. Denne konkordansen kjøres mot et fraseleksikon som er en serie med ordkombinasjoner som opptrer samlet med en viss frekvens. Til hvert ord i denne kombinasjonen er det grammatiske opplysninger. Ordene i konkordansen flettes sammen med ordene i den lemmatiserte ordliste. Ved likhet brukes opplysninger fra ordlisten. I tillegg blir to og to ord sammenlignet tegn for tegn. Der hvor likheten opphører starter endelsen. Disse endelsene sjekkes mot ett sett med bøyningsparadigmer for Ibsens språk. Dersom de to endelsene kan høre til samme paradigme antas det at de to ordene er bøyningsformer av samme ord.

De ordene som ennå ikke har fått grammatisk klassifikasjon blir testet mot en suffikstabell, som er en liste med (grafiske) endelser og mulige grammatiske klassifikasjoner.

Ordene i konkordansen blir så sortert tilbake til sin opprinnelige rekkefølge. Det brukes da et sett med kontekstregler for å skille homografer. Disse kontekstreglene er en sekvens av grammatiske klassifikasjoner som kan opptre samlet, eller kombinasjoner som ikke kan forekomme.

Til slutt skrives det ut en arbeidskonkordans med 3 linjers kontekst.

(1) Staffan Hellberg: Automatisk lemmatisering, Språkdata, Gøtegorg 1971

(2) Green & Rubin: Automatic Grammatical Tagging of English, Brown University, Providence, 1971

(3) Gerd Willé: LEMMA - ein programsystem zur automatischen lemmatiserung deutscher wortformen, IKP, Bonn, 1978

LEMMATISERING AF ISLANDSK:

Sidste sommer ville man forsøge at experimentere med automatisk lemmatisering af islandsk. Til rådighed havde man 1-2 måneders tid (juli/august), en PDP 11/60 (Fortran og Basic) og tre medarbejdere Baldur Jónsson, Höskuldur Þráinsson og Björn Ellertsson. Vi diskuterede grundigt de forskellige muligheder vi havde for at kunne komme frem til nogen resultater og besluttede til sidst at bruge et sytem med ordbogskonsultation. For også at have en mulighed for at få et system implementeret, blev det Basic-programm, som Martin Kay fremviste i Bergen 1973, benyttet. Det var i forvejen klart at det ikke kan analysere sammensatte ord (medmindre disse findes i ordbogen), men vi håbede på, at med mindre ændringer og tilføjelser skulle det kunne klare størstedelen af den islandske morfologi. Oprindelig kunne det klare genkendelse og ændringer af præfikser og suffikser, udføre ordbogssøgning, addere oplysninger til en identifikationsstreng om analyseordet, benytte subrutiner og udføre analysen på en ikke-deterministisk måde. Det var altså klart i forvejen,at uændret kunne programmet ikke klare den overordentlig hyppige u-omlyd (a til ö, barn - börn), og ikke uden besvær synkope (bortfald af a,i,u i endelser), f.eks. akur - ökrum, samt andre fenomener af lignende art (får - færri; assimilation af stam og endelse (muligvis med stamforandring) osv.)
Med tilføjelse af instruktioner for at genkende og ændre infikser (indlysende kompletering til præfikser og suffikser), en instruktion, som med omvendte transformationer gør u-omlyd tilbagegængelig, samt mulighed for ordbogsopsøgning af ord med synkope er systemet blevet temmelig komplet. Det er blevet afprøvet på størstedelen af substantivernes deklination og adjektivernes deklination og komparation. Også verbernes konjugation synes det at kunne klare (ganske vist med op til 6 stammer for stærke verber og op til 2 stammer for svage verber). Vi syntes,at resultatet var tilfredsstillende, systemet er enkelt at benytte og instruktionerne lette at indlære sig. På den negative side står, at programmet selv interpreterer hver enkelt grammatik-instruktion hver gang den skal udføres (uden f.eks. at indsætte run-time-adresser), hvad der kan accepteres (med fordel?) på et experimentelt system, men ellers er uacceptablet. Det samme gælder iøvrigt for benyttelse af Basic, medmindre man ikke overfører sytemet til en egen home-computer.
Sammenfattende kan man altså sige, at det drejer sig om et beskedent system, som også kan klare kompliceret flexion, dens benyttelse er enkel og nem at tilegne sig (ideel for undervisning), det kan let udvides, men til produktion skulle det omskrives til et kompileret sprog (Fortran eller lign.) og koden skulle også optimeres, hvis det ikke drejer sig om grammatikker som stadig bliver ændret.

Reykjavik, den 17.sept. 1979

For gruppen

Björn Ellertsson

Björn Ellertsson.

Hanne Ruus
Institut for nordisk filologi
Københavns Universitet
Njalsgade 80
DK-2300 København S.

SEMANTIK I AUTOMATISK LEMMATISERING.

De fleste, der beskæftiger sig med lemmatisering, har som data autentiske tekster på naturlige sprog og som nært mål at kunne svare på spørgsmål som: Hvor mange og hvilke ord er almindelige i aviser, i en forfatters værker, i børnebøger osv. ?

I denne situation er der flere grunde til at automatisere lemmatiseringsprocessen. For det første den rent praktiske, at en så godt som automatisk lemmatisering gør det muligt at komme igennem et stort materiale på overskuelig tid. For det andet den vigtigere grund, at kun ved at beskrive tilordningen af bøjningsformer til lemmaer algoritmisk kan man sikre sig en helt konsekvent behandling af materialet. For det tredie kan en automatiseret lemmatiseringsprocedure anvendes igen og igen på mange slags tekst til forskellige formål: frekvensundersøgelser, automatisk syntaktisk analyse, maskinoversættelse.

En automatisering fordrer en entydig definition af lemma: man må kræve udtryksforskel i mindst én form i to serier af ordformer for at opstille to lemmaer (jf. Nusvensk Frekvensordboks og DANwORDs[1] lemmadefinition). Selv med denne formelle lemmadefinition kommer man ikke uden om at anvende semantiske oplysninger, når man vil nærme sig en fuldautomatisk lemmatisering. For at få en enkel og billig lemmatiseringsprocedure må man imidlertid vride så meget information som muligt ud af udtryksforskelle og begrænse sig til et minimum af semantik. Derfor er jeg på udkig efter semantiske træk, der kan bruges ved mange forskellige ord, og helst sådanne, som kobles med udtryksforskelle.

Heterografer er ord, der i alle bøjningsformer staves anderledes end alle andre ord. De kan altså lemmatiseres automatisk alene ud fra deres udseende - uden semantik. Men vejen til den automatiske analyse af naturlige sprog er brolagt med entydiggjorte homografer. Man kan komme et godt stykke i entydiggørelsen ved hjælp af formelle træk i konteksten, men derefter er man henvist til semantikken.

Overvejelserne om entydiggørelse nedenfor bygger på excerpering af ordformerne i fig. 1 og 2 i DANwORDs[1] prøver fra fiktionstekster for voksne (godt 250 000 løbende ord), hvad der gav godt 200 belæg på rejse I, II og III og godt 50 belæg på øre I og II. Eksemplerne er valgt, så de viser homografi inden for samme ordklasse: øre I og II er substantiver, rejse II og III er verber.

øre I "legemsdel" sb. -t, pl. -r el. -n

øre	
øret	
ører	
(øren)	
ørerne	
ørene	

øre II "betalingsmiddel" sb. -n, pl. dsm. el -r

øre	
(øren)	
ører	
ørerne	
ørene	

fig. 1.

I figuren er entydige former streget under. Da de to substantiver har hvert sit genus, kan en del forekomster af øre i singularis entydiggøres ud fra kongruensbøjede former foran ordet: dit øre, det indre øre til øre I og en øre til øre II. øre brugt som pluralis af øre II kan også bestemmes ud fra konteksten: det har altid et talord som adled tolv øre, halvtreds øre. Men ved de øvrige homografe pluralisformer er man henvist til semantiske hjælpemidler.

rejse I sb. pl. -r

rejse
rejser

rejse II "travel" vb. -te	rejse III "raise" vb. -te, -ning
rejse	rejse
rejser	rejser
rejste	rejste
rejst	rejst

fig. 2.

I figuren er fra substantivet rejse kun anført de former, der er homografe med verbalformer. Disse substantivformer vil i de fleste tilfælde kunne udskilles ved hjælp af kongruerende adled, f.eks. en besværlig rejse, hele denne rejse, den oplevelsesrige rejse. Begge verber er transitive, men kun rejse III kan have refleksivt objekt. Denne syntaktisk-semantiske oplysning er kvantitativt set vigtig, da knap 2/3 af belæggene på rejse er former af rejse sig. Til resten af verbaleksemplerne må man finde mere forfinede, semantiske deskriptorer.

Når man ser sig om i den datalingvistiske forskning for at finde ideer til semantiske oplysninger og anvendelsen af dem, viser det sig, at de projekter, der arbejder med automatisk behandling af semantiske oplysninger, befinder sig inden for området simuleret intelligens, hvor man interesserer sig mere for metoderne end for resultaterne og følgelig udvikler disse metoder på ret begrænsede tekstmængder. Her arbejder man på at afbilde så meget betydning, at maskinen kan simulere en form for forståelse af den indlæste tekst. Man er derfor mindre interesseret i udtrykket og bruger til gengæld komplicerede sæt af semantiske primitiver og slutningsregler, som det i deres nuværende form ville være uoverkommeligt at anvende på større tekstmængder. Men måske kan man låne et lille sæt anvendelige kategorier?

øre I	øre II	
HEAD: PART	HEAD: THING	B. HEAD: SIGN
QUAL: ANI	QUAL: "metal"	

fig. 3.

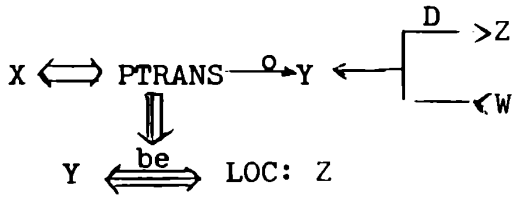
Figuren viser de vigtigste dele af de semantiske udtryk for øre I og II formuleret i Yorick Wilks'[2] semantiske formler. øre II har to betydninger, en der henviser til selve mønten (THING "metal") og en anden til værdien (SIGN), mens øre I er beskrevet som del af menneske eller dyr (PART ANI)

Ved brug af semantiske oplysninger kan man nøjes med at udvælge sådanne, der entydigt kan bestemme forekomster af den ene af to tolkningsmuligheder, hvis man kan finde semantiske oplysninger, der gør dette tilstrækkeligt sikkert.

Da øre II har flere betydninger i Wilks' system, er det oplagt at forsøge at finde træk, der entydigt kan udskille øre I-eksemplerne. Af excerpterne fremgår det, at øre I ofte forekommer med possessiver som adled, jf. de possessive neutrumformer, der blev brugt til at bestemme singularisformerne af øre I. Hvis man tilføjer en regel om, at possessiv ofte står foran "legemsdel", kan man entydiggøre ved hjælp af en delvis formelt afgrænselig klasse af ordformer, de possessive pronominer og personnavne i genitiv, Denne regel vil klare over 3/4 af de resterende flertydige belæg og virker bedre end en ren semantisk regel, der bygger på, at ord for legemsdele ofte forekommer sammen, og derfor forudsætter, at alle ord for legemsdele er mærket som sådanne uanset entydighed. Den semantiske regel ville kun klare halvt så mange eksempler som possessiv-reglen og yderligere kræve, at ansigt, mund, øjne, kæbeben og rottehaler er mærket som legemsdele.

Til entydiggørelse af former af rejse II og III kan man hente hjælp i Roger Schanks Conceptual Analysis[3]. Beskrevet i Schanks diagramform ser rejse II og III således ud:

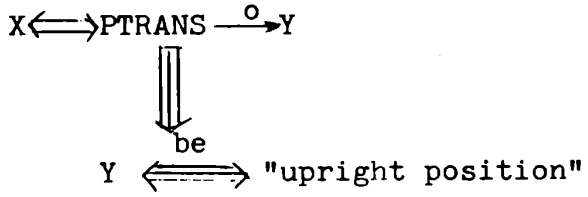
rejse II "travel"



Betingelser til de størrelser, der indgår på pladserne:

X: human; Y = X; Z, W: place

rejse III "raise"



Betingelser til de størrelser, der indgår på pladserne:

X: animate; Y: physobj

fig. 4.

Diagrammet over rejse II kan parafraseres: at X fysisk flytter Y fra W til Z, forårsager, at Y er på lokaliteten Z.

Diagrammet over rejse III kan parafraseres: at X fysisk flytter Y, forårsager, at Y befinder sig i oprejst stilling.

rejse III har altid "ægte" objekt og burde derfor kunne udskilles af den syntaktiske analyse. Denne kan imidlertid ikke laves uden visse semantiske oplysninger: mulige kerner i alle de tids- og målsadverbialer, der kan have form af et nominalhypotagme, må særmærkes. Ellers ville man få forkert analyse af sætninger som Søren rejste en del omkring. Det er altså ikke så lige til at udskille rejse-III-eksemplerne. Man kan derimod bestemme ganske mange af rejse-II-belæggene ved at anvende den semantiske rolle retning, Schanks D-case (Directive). Den manifesteres nemlig, så den er nogenlunde let genkendelig f.eks. ved retningsadverbier og præpositionssyntagmer indledt med til og fra (Hanne Ruus: Sproglig betydningsanalyse, i Nydanske Studier 10-11, 1979, s.186). En regel om at søge efter en retningsangivelse i konteksten vil bestemme godt halvdelen af de resterende flertydige belæg rigtigt.

Entydiggørelsen af den sidste restmængde vil formentlig kræve adgang til ganske fyldige semantiske beskrivelser af mange ord f.eks. kan rejse III have så semantisk forskellige objekter som hoved, sigtelse, spørgsmål, galge og hær.

Denne undersøgelse af 2 sæt homografer har vist, at man forholdsvis let kan opstille syntaktisk-semantiske regler, der ved hjælp af, til dels, formelt afgrænsede klasser, possessiver og retningsadverbier, entydiggør de fleste af de belæg, der ikke kan klares uden semantik.

Som bemærket ovenfor er det vigtigt, at de semantiske træk, man vælger ud til brug i entydiggørelsesprocedurer, kan anvendes ved mange ord og ikke medfører krav om semantiske oplysninger ved alle ord. De træk, jeg har skitseret brugen af her, opfylder begge dette krav: den semantiske rolle "retning", som blev foreslået ved rejse-eksemplerne, vil kunne anvendes ved flertydige bevægelsesverber som fare, føre, lede, mens klassen af possessiver, som blev indført ved øre, vil være nyttig ved flertydige ord for mere eller mindre umistelige legems/ejen-dele som arm, tunge og stol.

Ved kvantitative opgørelser over lemmaer behøver man næppe bekymre sig om den sidste rest af flertydigheder, men til en fuldt automatisk analyse kan man forudse, at der kræves både en ret gennemarbejdet syntaktisk analyse og et omhyggeligt udvalgt, større sæt af semantiske oplysninger.

Noter.

(1) DANwORD, Hyppighedsundersøgelser i moderne dansk, ved Bente Maegaard og Hanne Ruus, se f.eks. SAML III, 4, 5.
(2) se. f.eks. Yorick Wilks: The Stanford Machine Translation Project, i Natural Language Processing ed. by Randall Rustin, New York 1973.
(3) se f.eks. Roger Schank: Identification of Conceptualizations Underlying Natural Language, i Computer Models of Thought and Language ed. by Roger C. Schank and Kenneth Mark Colby, San Francisco 1973.

Foredrag ved Nordiske Datalingvistdage på IAML, Københavns Universitet, 9.-10. oktober 1979.

Henrik Holmboe:

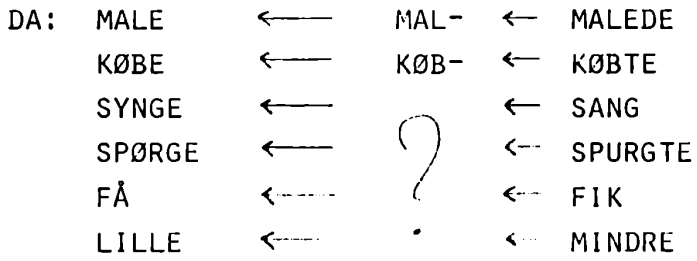
Lemmatisering - hvilke af de ideelle krav til lemmatisering er opfyldelige eller opfyldte?

Lemmatisering er en term, der er kurant i snævre kredse, hvor termen anvendes på en måde, der vel ikke er entydig, men dog har et centralt betydningsområde, som alle er enige om hører med til termen. Hvis man konsulterer en række gængse lingvistiske terminologiske ordbøger eller oversigtsværker, konstaterer man imidlertid, at termen ikke er optaget og defineret i disse værker.

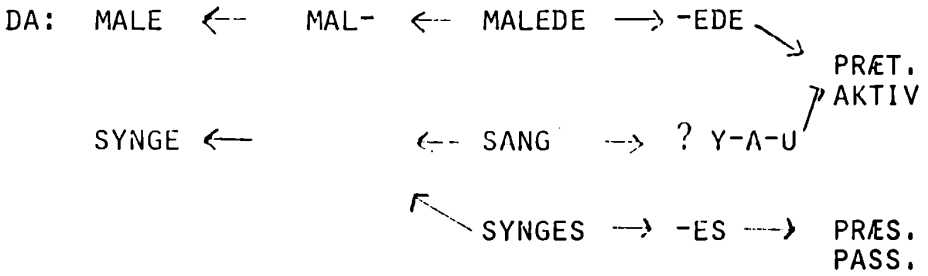
At lemmatisere betyder minimalt at henføre et ord fra en tekst til en bestemt type eller kategori, som det i teksten aktuelt forekommende ord kan påstås at være en bøjet form af. Dette forudsætter en analyse af ordet og eventuelt dets omgivelser i teksten, men behøver ikke at forudsætte informationer, der ligger uden for ordet og teksten selv.

DA: MAL- ——— MALEDE
KØB- ——— KØBTE

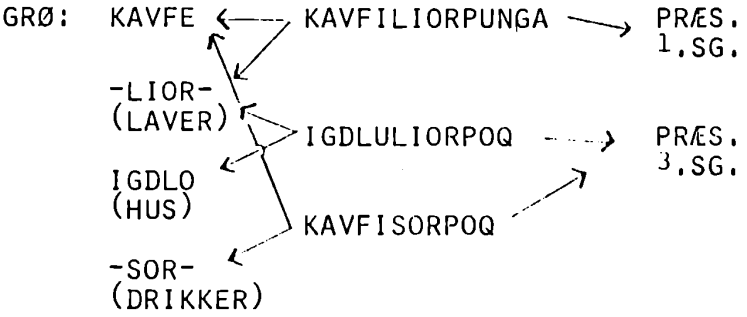
Ud over dette minimale krav vil man eventuelt også forlange, at lemmatiseringen skal resultere i en henvisning til den leksikalske indgang, som ordet skal søges under i gængse ordbøger, hvad enten dette er en abstrakt form eller en bestemt forekommende bøjningsform af ordet. Dette vil implicere viden, der ikke nødvendigvis er til stede i ordet eller teksten på stedet.



Endvidere vil man eventuelt forlange, at lemmatiseringen skal henføre ikke blot den del af ordet, der rummer dets centrale betydning, til en leksikalsk type eller kategori, med alle ordets dele til leksikalske eller morfologiske typer eller kategorier. Dette vil implicere en endnu større viden.

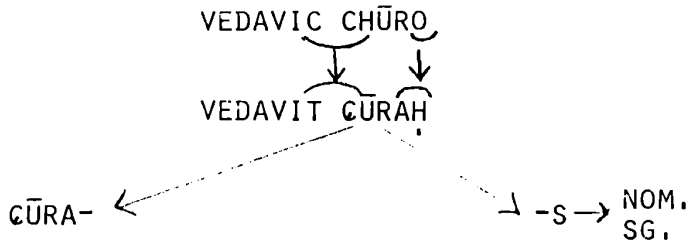


Men dette krav er vigtigt, hvis man vil opnå en mere generel definition af ordet lemmatisering, således at termen kan anvendes også i forbindelse med sprog, der f.eks. tillader mere end én orddel med det, vi vil kalde central betydning, inden for ét og samme ord:

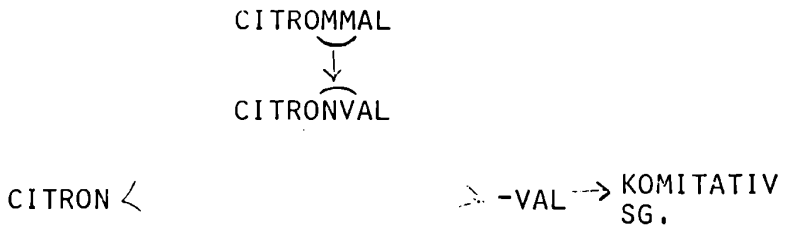


Lemmatiseringen skal også kunne opløse sandhi-fænomener internt og eksternt af forskellig kompleksitet (assimilation, fusion):

ASSIMILATION:
SANSKRIT:

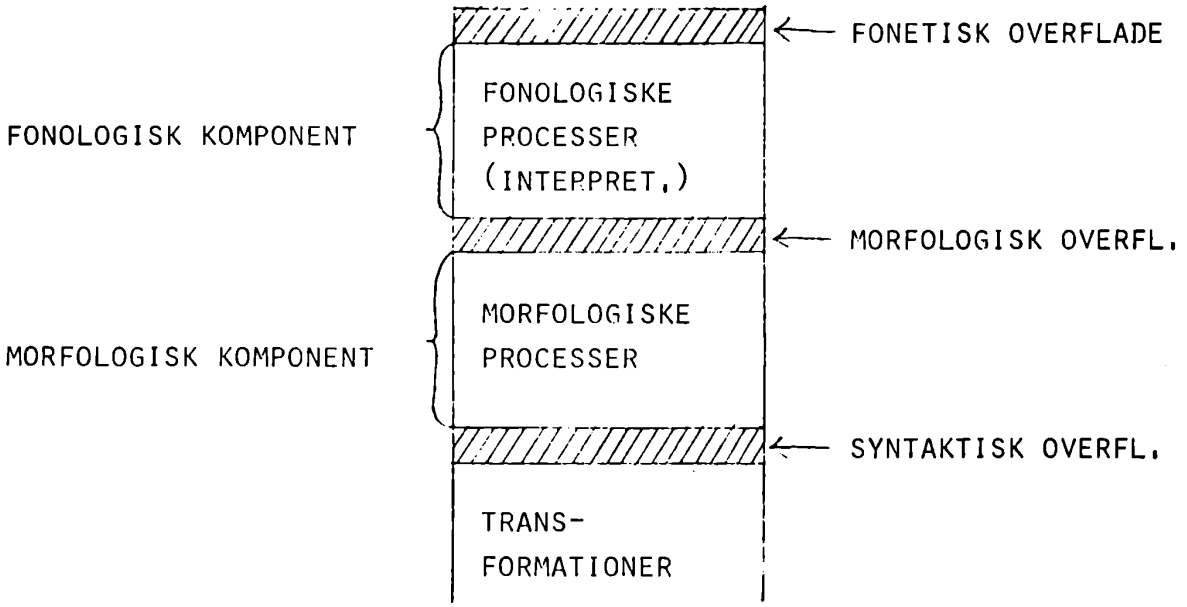


FUSION:
UNG.:

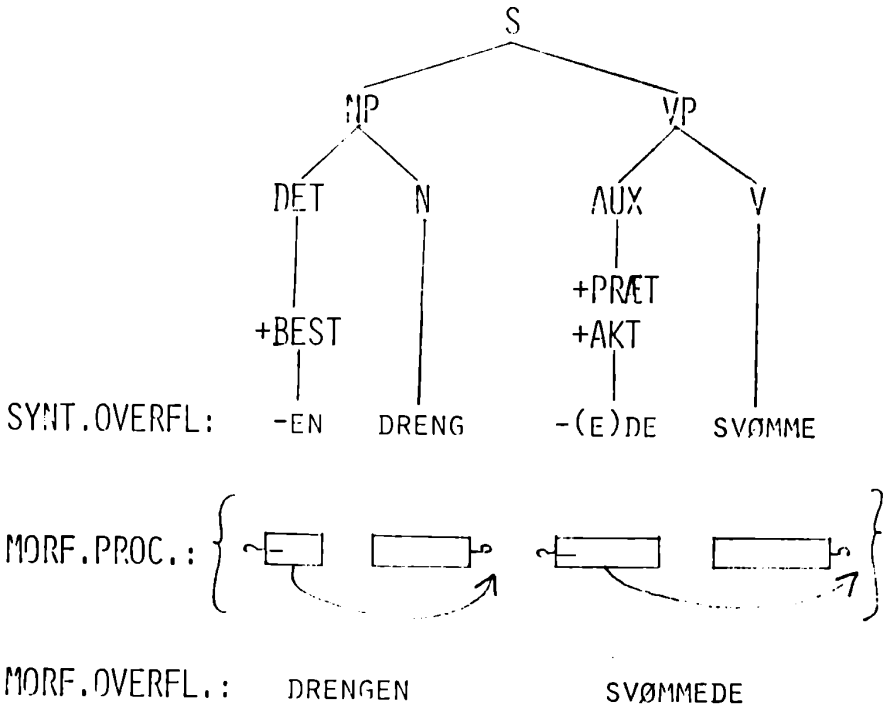


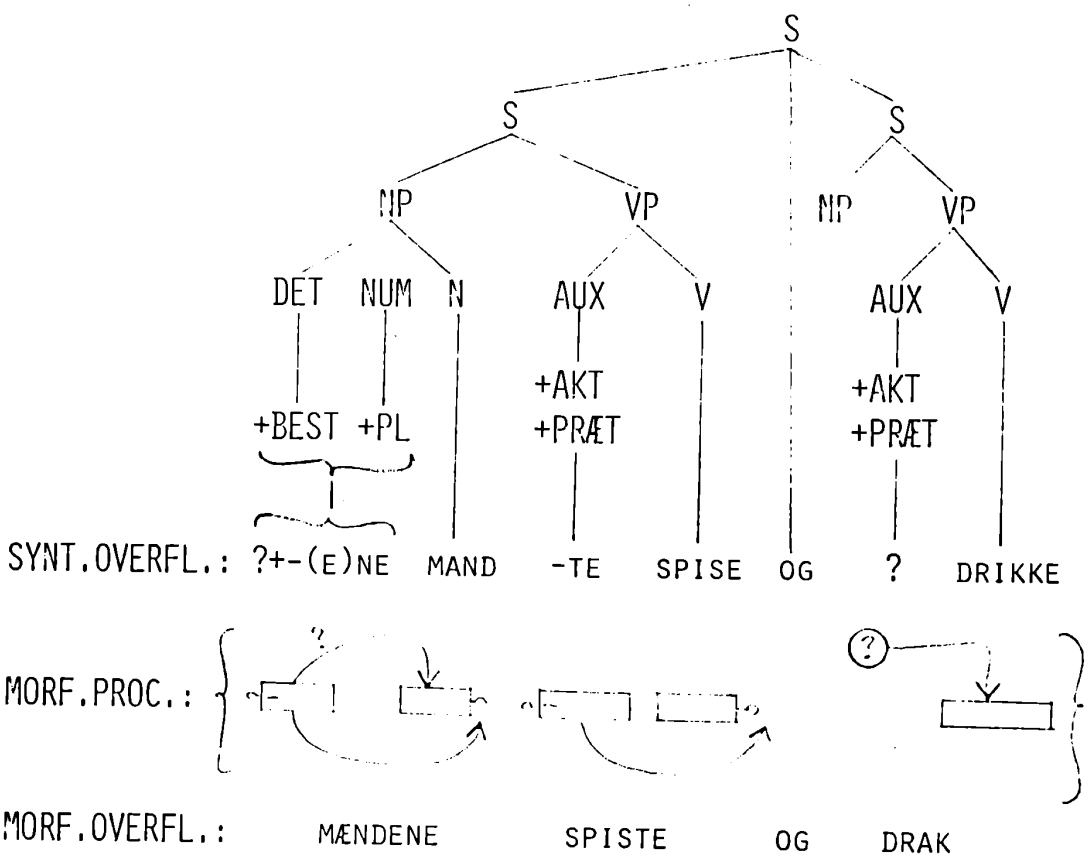
Når termen lemmatisering ikke kan siges at være en gængs lingvistisk term i almindelighed uden for leksikografers og datalingvisters kreds, skyldes det ikke, at lingvistikken ikke har beskæftiget sig med det problemkompleks, som lemmatisering omfatter, men at beskrivelsen af disse problemer skal søges under disciplinerne morfologi og fonologi og evt. morfofonologi. Morfologisk analyse og lemmatisering skal altså kunne opvise en række fælles resultater. Ser vi på en transformationsgrammatisk model, møder vi straks den vanskelighed i f.eks. Aspects-modellen, at TG beskæftiger sig meget lidt med morfologi. Den går næsten direkte fra transformationer af hovedsagelig syntaktisk natur til en fonologisk komponent, der producerer den fonetiske

repræsentation. Ind imellem skyder man de såkaldte readjustment-rules, hvis status ikke er ganske klar. En lidt ændret transformationsgrammatisk model kunne være følgende, idet kun de hér relevante dele er medtaget:



f.eks.:





Anskuet på denne måde er det lemmatiseringens opgave at bevæge sig fra den morfologiske overflade til den syntaktiske overflade.

Jeg vil ikke hævde, at en lemmatisering skal forløbe som et baglæns gennemløb af den morfologiske komponent, men resultaterne af de to skulle gerne være sammenfaldende: inputtet til den morfologiske komponent (eller den del af grammatikken, der rummer de såkaldte readjustment-rules) skal være det samme som outputtet fra en lemmatiseringsproces. Dvs. en lemmatisering "skal kunne" det samme som en grammatiks morfologiske komponent (hvad enten denne anskues som analyserende eller genererende), men ikke nødvendigvis på samme måde. Dette må være det ideelle krav til lemmatisering.

I lingvistikken har man beskrevet indholdet af den morfologiske komponent på lidt - men ikke meget- forskellig måde. Bl.a. J.H. Greenberg og P.H. Matthews har elaboreret

og forfinet den beskrivelse, vi finder hos Sapir i Language fra 1921, men for ikke at fortabe mig i detaljer vil jeg i store træk holde mig til Sapir, der omtaler følgende morfologiske eller grammatiske processer:

A: AFFIXERING:

	FUSING	JUXTAPOSING
PRÆ- IN- SUB-	+ SANDHI	- SANDHI

B: INTERN MODIFIKATION:

VOKALHARMONI, OMLYD, AFLYD, STADIEVEKSLING } FJERNASSIMILATION { PROGRESSIV, REGRESSIV

A: REDUPLIKATION:

PRÆFIGERING AF EN "DUBLETDEL".

Vokalharmoni og i visse tilfælde omlyd kan anskues som henholdsvis progressiv og regressiv fjernassimilation og er altså forudsigelige eller redundante. Dette gælder ikke aflyd som vi kender den fra f.eks. germanske stærke verber eller intern flexion i arabisk.

Dette er altså fundamentalt to typer af processer:

A. Den ene vil jeg kalde <u>forøgelse</u>, dvs. noget forøges med noget andet. Herunder hører affixering og reduplikation. Lemmatiseringsopgaven er her at identificere det, der er blevet forøget, og forøgelsen.

B. Den anden vil jeg kalde <u>mønsterændring</u>, dvs. ét mønster erstattes af et andet; f.eks.

B: ARAB.: KITĀB ⟵ KUTUB (PL)
RASŪL ⟵ RUSUL (PL)
BAJT ⟵ BUJUT (PL)
JAUM ⟵ AJJĀM (PL)
AJUĀM (*PL)

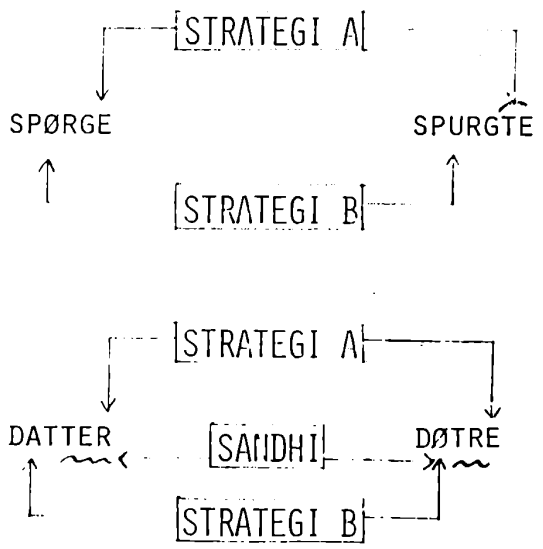
DA.: SPRINGE ⟵ SPRANG (PRÆT)
GÅS ⟵ GÆS (PL)

Herunder hører intern modifikation.

Lemmatiseringsopgaven er hér at skelne mønsteret fra baggrunden og derpå identificere disse to. I begge tilfælde kan intern og extern sandhi sløre billedet. Der vil så vidt jeg kan se være tale om to helt forskellige lemmatiseringsstrategier alt efter, om man skal identificere en forøgelse eller et mønster.

Resultaterne og erfaringerne viser, at man lettest kan automatisere lemmatisering af agglutinerende strukturer uden sandhi. Sandhi-fænomener er brydsomme, men ikke uovervindelige. I systemer, der mestrer problemer af denne type, vil strategien over for mønsterproblemerne være at henvise disse til undtagelseslister. Denne strategi er langt fra ideel, men praktisk og anvendelig, så længe man beskæftiger sig med sprog, hvori de agglutinerende fænomener er de hyppigste og de flekterende befinder sig i relativt små, lukkede klasser, men strategien ville være uanvendelig, hvis forholdet var det omvendte.

Af de ideelle krav til lemmatisering mangler man at opfylde dem, der vedrører mønstergenkendelse og formodentlig også metoder til at styre, hvornår den ene og hvornår den anden strategi skal bringes i anvendelse.



Den almindelige opfattelse af, hvad der er svært og let, svarer nøje til, hvad vore maskinelle metoder i dag kan klare: tyrkiskens morfologi, der er agglutinerende næsten uden sandhi, er lettere end f.eks. ungarskens, der er agglutinerende med sandhi. Noget sværere er f.eks. sanskrit med sin blanding af agglutinerende og flekterende morfologi med udstrakt sandhi og sværest er klassisk arabisk med sin internt flekterende morfologi med en del sandhi.

Henrik Holmboe

Institut for Lingvistik
Aarhus Universitet

Bente Maegaard og Hanne Ruus
Københavns Universitet, Amager
Njalsgade 80
DK 2300 København S.
DANMARK

STRUKTURERING AF LINGVISTISKE DATA TIL BRUG VED MASKINOVERSÆTTELSE.

EUROTRAs overførselsstruktur.

Siden feb. 1978 har forskere fra videnscentre inden for datamatisk lingvistik og maskinoversættelse i EF-landene arbejdet på at formulere et projekt til et europæisk maskinoversættelsessystem, EUROTRA.

Det er efterhånden blevet et meget stort problem for EF, at mængder af dokumenter skal foreligge på 6 sprog. Oversættelsesafdelingerne vokser og alligevel kan de knap klare efterspørgslen.

Til afhjælpning af oversættelsesbehovet har Kommissionen anskaffet et amerikansk system, SYSTRAN, der kan oversætte mellem visse par af EF-sprog (engelsk-fransk, fransk-engelsk). Dette systems oversættelser er imidlertid ikke tilfredsstillende og Kommissionen har derfor taget initiativ til at få udarbejdet en plan for et bedre system i Europa. Det europæiske system skal være flersproget, skal kunne oversætte mellem de 6 EF-sprog og skal kunne udvides til at omfatte nye sprog. I arbejdet med projektformuleringen har vi deltaget fra Danmark (siden sept. 1978).

I arbejdsgruppen deltager i øvrigt forskere fra følgende universiteter:

Leuwen

Manchester/Essex

Grenoble

Pisa

Saarbrücken

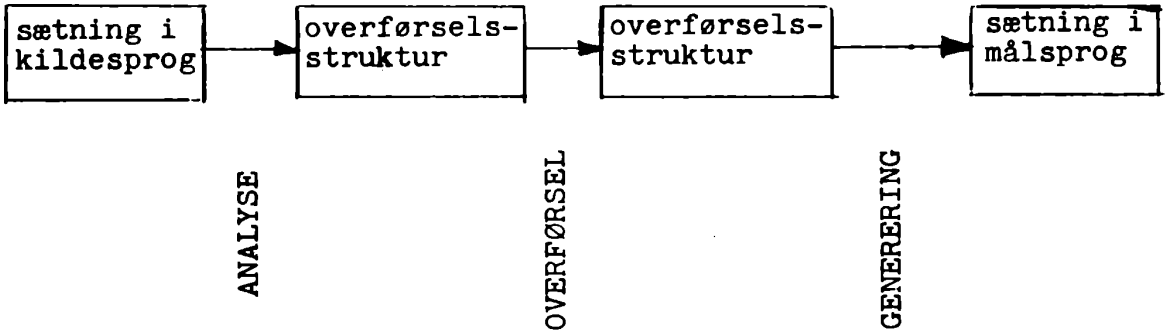
Arbejdsgruppens formand er englænderen Margaret King.

Det overordnede princip ved udformningen af projektet er, at det skal udvikles decentralt: Så meget som muligt skal udvikles separat og lokalt for det enkelte sprog af en arbejdsgruppe, der har det pågældende sprog som modersmål. Oversættelsesprocessen er derfor delt i tre dele: analyse, overførsel og generering, af hvilke de to, analyse og generering, er ensprogede og udvikles helt af

den enkelte arbejdsgruppe. Overførsel mellem sprogene skal derimod udarbejdes af grupperne i fællesskab, sprogpar for sprogpar.

Det er nok klart, at deltagerne i arbejdsgrupperne, der skal udvikle analyse og generering for et sprog og samarbejde med andre grupper om overførsel, må have det pågældende sprog som modersmål; men dette kunne også opnås ved et centralt system. Når man lægger så stor vægt på, at systemet skal udvikles decentralt, er det især fordi man herved styrker datamatisk lingvistiske miljøer i alle EF-landene.

Oversættelsesprocessen i EUROTRA kan fremstilles således:

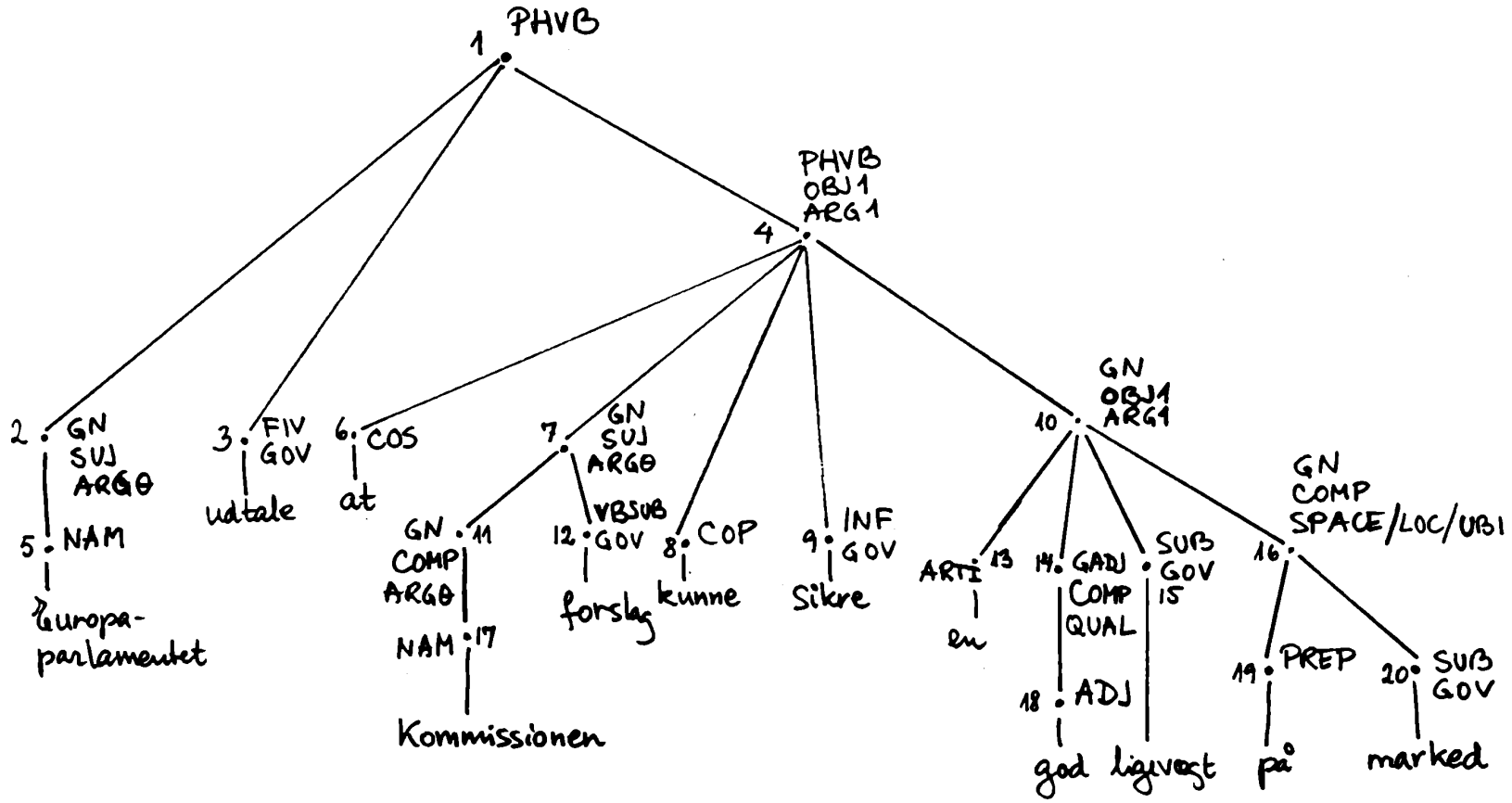


Hvis man forestiller sig, at en engelsk tekst skal oversættes til dansk, vil det analysemodul, der anvendes, være udarbejdet af den engelske gruppe, overførselsmodulet i samarbejde mellem den engelske og den danske gruppe, og generering vil være udarbejdet af den danske gruppe. Analysemodul og genereringsmodul er uafhængige af henholdsvis målsprog og kildesprog; det samme engelske analysemodul benyttes altså ligegyldigt, hvilket sprog der skal oversættes til, hvorimod der udarbejdes to overførselsmoduler for hvert sprogpar. Det er derfor vigtigt, at overførslen begrænses til det allernødvendigste: det, som kræver, at man har adgang til informationer om begge sprog på én gang.

Inden for rammerne af det tilbudte, fælles programmel kan den enkelte arbejdsgruppe selv bestemme sin strategi, når blot resultater, der skal være input til næste modul, opfylder helt bestemte, veldefinerede krav, som fastlægges i projektbeskrivelsen. Resultatet af analysen, såvel som resultatet af overførslen, udmøntes i en overførselsstruktur: tekststykker og sætninger beskrives ved træstrukturer.

En overførselsstruktur består af et eller flere dependenstræer, hvor hver knude er forsynet med etiketter på forskellige sproglige analyseniveauer: morfosyntaktisk, syntaktisk og logisk-semantisk. Det er herved muligt at bevare ordstillingen fra kildeteksten i bunden af træstrukturen samtidig med, at man har oplysninger om sætningens dybdestruktur i træets etiketter.

fig. 1.



Europaparlamentet udtalte at Kommissionens forslag kunne sikre en bedre ligevægt på markedet.

I fig. 1 er vist en overførselsstruktur for den danske sætning Europaparlamentet udtalte, at Kommissionens forslag kunne sikre en bedre ligevægt på markedet.

For hver moderknude (knuderne 1,4,7,10,16) gælder det, at en af dens døtre er kerne i forhold til sine søstre. Kernen mærkes med etiketten GOV på det syntaktiske niveau. Søstrene til kernen står i relation til denne gennem moderknuden. F.eks. står knude 2 og 4 i relation til kerneknuden 3. En kerneknude dominerer altid direkte et blad på træet, der indeholder henvisning til den leksikalske enhed, som indeholder det ord eller udtryk i inputsætningen, som indgår på den pågældende plads i træet. Etiketter på knuderne anføres i rækkefølgen morfosyntaktisk funktion, syntaktisk funktion, logisk-semantisk funktion: f. eks. er knude 2 en nominalgruppe (GN), der syntaktisk er subjekt i sætningen (SUJ). Logisk-semantisk er den dybdesubjekt (ARG0) i forhold til kerneknuden 3, som har den morfosyntaktiske etikette finit verbal (FIV) og syntaktisk er kerne (GOV). Træet under knude 4 er morfosyntaktisk en sætning (PHVB), syntaktisk direkte objekt (OBJ1) og logisk-semantisk dybdeobjekt (ARG1). Knude 2 dominerer knude 5, der indeholder oplysning om den morfologiske klasse, proprium (NAM) for Europaparlamentet.

Træets blade er leksikalske enheder og oplysninger om tekstens bøjningsformer findes i den knude, der dominerer den leksikalske enhed, f. eks. indeholder knude 18 oplysninger om, at god i teksten optræder i komparativ.

Knude 11 viser et tilfælde, hvor det logisk-semantiske subjekt ikke falder sammen med det syntaktiske subjekt. Man ser, at Kommissionen er ARG0 for forslag, idet det er Kommissionen, der foreslår noget. I denne model kan verbalsubstantiver altså have subjekter og objekter ligesom de tilsvarende verber, og det logisk-semantiske niveau benyttes til at beskrive dette forhold. Andre anvendelser af logisk-semantiske etiketter ser man i knude 16, hvor etiketten SPACE/LOC/UBI er brugt på et adverbial, der angiver stedet,hvor noget sker. Der findes en række etiketter, der benyttes ved stedsangivelser, og en tilsvarende række for tidsangivelser. Endelig kan man bemærke etiketten QUAL (knude 14), der anvendes på adjektiver i attributiv stilling, på relativsætninger mv. Disse logisk-semantiske etiketter svarer til semantiske roller (deep cases) i Fillmores forstand.

I knude 8 ser man, at kunne er markeret med den morfosyntaktiske etikette COP, dvs. som tilhørende ordklassen hjælpeverber. Dette skal ikke tages som udtryk for en bestemt mening om den mest hensigtsmæssige beskrivelse af modalverber, idet modalitet og hjælpeverber stadig er under overvejelse i arbejdsgruppen.

En vigtig generel egenskab ved træet er, at det er fladt: antallet af niveauer er begrænset, og herved spares både lagerplads og søgetider. Man kan f. eks. sammenligne dette træs grundstruktur med den tilsvarende i sædvanlig IC-analyse (Immediate Constituents) eller en tilsvarende kontekstfri grammatik:

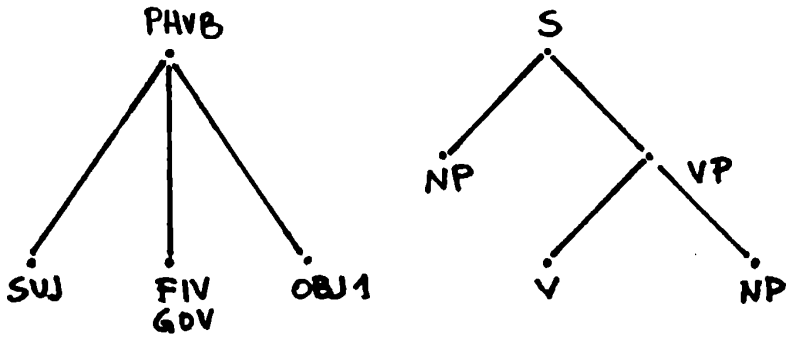


fig. 2.

Dependenstræet har blot tre grene gående ud fra roden, af hvilke en er markeret som GOV, og det giver den samme information som det andet træ (eller endda mere). Også ved beskrivelse af nominalgrupper er det flade træ enklere: se f. eks. en bedre ligevægt (det deltræ, der er domineret af knude 10).

En overførselsstruktur skal indeholde de oplysninger, der er nødvendige for at overføre tekstens indhold fra et sprog til et andet, men heller ikke flere. At de lingvistiske oplysninger på knuderne i overførselstræerne er nødvendige, ser man bedst gennem en beskrivelse af den vej, en overførselsstruktur tilbagelægger fra et sprog til et andet. Vi skal derfor vise i hovedtrækkene, hvordan en fremmedsproget overførselsstruktur ændres til dansk under udnyttelse af de informationer, der findes på træets knuder.

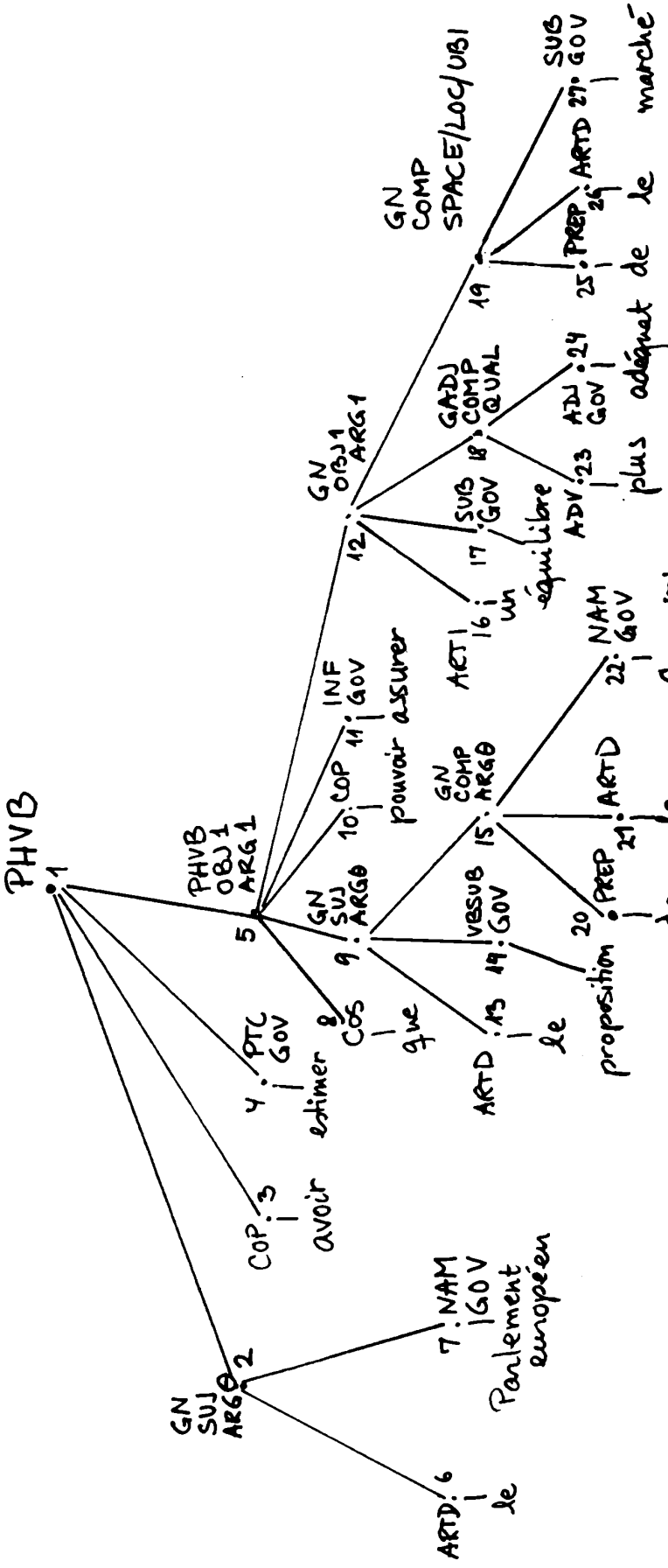
Fig. 3 viser en overførselsstruktur for den franske sætning Le Parlement européen a estimé que les propositions de la Commission pouvaient assurer un équilibre plus adéquat du marché.

Overførslen af det franske træ består af en leksikalsk overførsel, dvs. at de franske leksikalske enheder på træets blade skiftes ud med tilsvarende danske, mens strukturelle ændringer i træet henvises til genereringsfasen.

Det skaber ingen særlige problemer at udveksle Parlement européen med Europaparlamentet, her er den morfologiske etikette proprium (NAM) en ekstrasikring. Tilsvarende veksles équilibre til ligevægt.

Når det franske ord har flere oversættelser til dansk, bliver overførslen imidlertid mere kompliceret og man får brug for etiketterne på de forskellige niveauer. For eksempel kan estimer (knude 4) betyde agte, vurdere, mene og assurer (knude 11) forsikre og sikre. Ved overførsel af assurer skal i dette tilfælde vælges det danske sikre. Dette valg træffes ved at udnytte informationer i de knuder, som assurer er kerne for: når assurer optræder med et ARG0, hvis kerne er non-human, vælges sikre. Ved overførslen af estimer

fig. 3.



Le Parlement européen a estimé que les propositions de la Commission pouvaient assurer un équilibre plus adéquat du marché.

vælges mene, fordi det franske verbums objekt (OBJ1, knude 5) har form af en sætning (PHVB). I den ene valgsituation udnyttes altså semantiske informationer - dels den logisk-semantiske etikette i en knude (ARG0), dels et semantisk træk, der er anført ved en leksikalsk enhed i ordbogen (non-human) - i den anden har vi udnyttet syntaktisk (OBJ1) og morfosyntaktisk information (PHVB) fra etiketterne.

Ved udgangen fra den leksikalske overførsel er den franske overførselsstrukturs leksikalske enheder altså erstattet med de tilsvarende danske. De fleste funktionsord som artikler, konjunktioner og præpositioner udskiftes dog ikke, da deres udfyldning i målsproget først kan bestemmes ud fra syntaktiske og logisk-semantiske etiketter på de knuder, der dominerer dem.

Den overførselsstruktur, der er output fra overførselsmodulet, fungerer som input til genereringsmodulet. Ved generering skal konstituenterne i sætningen og i de enkelte led anbringes i korrekt orden. F. eks. skal knuderne 16,17,18 i det franske træ ændre rækkefølge, således at man fra un équilibre plus adéquat når til en mere passende ligevægt.

Ved genereringen skal også vælges verbaltider og til brug for dette valg må de nødvendige oplysninger findes i overførselsstrukturens etiketter. Det franske træ skal således indeholde information, der gør det muligt at vælge dansk præteritum, mente, som oversættelse for den sammensatte franske verbaltid i knude 3 og 4, a estimé.

Til bestemmelse af funktionsord, der ikke er oversat ved overførslen, benyttes den lingvistiske analyse af konteksten, som er udmøntet i etiketterne, og oplysninger fra den danske ordbog, - f. eks. om hvilke præpositioner et verbum konstrueres med. I den her betragtede overførselsstruktur for den franske sætning skal que i knude 8 udskiftes med at, fordi det indleder en sætning, der er objekt (OBJ1) for mene.

Præpositionen de i nominalgruppen domineret af knude 19, du marché, skal oversættes ved på. Denne afgørelse træffes ved hjælp af den logisk-semantiske etikette SPACE/LOC/UBI sammenholdt med ordbogsoplysninger for den leksikalske enhed marked.

Som et sidste eksempel på, hvad genereringsfasen må omfatte, skal nævnes anvendelse af morfologiske regler: substantiver, adjektiver mm. skal bøjes, der skal være kongruens visse steder osv. Som et lidt mere indviklet tilfælde kan man betragte de knuder, der i det franske træ er domineret af knude 15, de la Commission. De skal på dansk blive til Kommissionens. Den regel, der sørger for det, vil gå ud på, at en nominalgruppe, der er ARG0 for et VBSUB, på dansk sættes i genitiv. De logisk-semantiske etiketter bruges altså ikke blot ved overførsel, men også i genereringsfasen.

Med denne gennemgang af en overførselsstruktur for en forholdsvis enkel, dansk sætning og den skitsemæssige overførsel af en tilsvarende fransk overførselsstruktur fra fransk til dansk har vi vist, hvilken slags informationer

der skal bruges i det planlagte oversættelsessystem; vi har derimod kun antydet, hvor disse informationer skal komme fra.

Det vil afhænge af den enkelte gruppes valg af analysestrategi, hvornår de vil fremtage og bruge oplysningerne på de forskellige lingvistiske niveauer i deres arbejde hen mod overførselsstrukturer: én gruppe kan støtte sig mest på semantiske oplysninger, en anden på syntaktiske oplysninger i analysefasen. Det er dog fastlagt, at alle lingvistiske oplysninger skal hentes fra grammatikker og ordbøger, der er adskilt fra og uafhængige af det anvendte programmel.

Vi vil således kunne udbygge og anvende den danske morfologiske analyse, som vi bruger i DANwORD (se f. eks. SAML III, 4, 5), ligesom en del af de frekvensoplysninger, vi fremtager i DANwORD, vil kunne indgå i arbejdet med at formulere regler for valg af den rette oversættelse ved flertydige ord og udtryk. Selv med den forholdsvis udførlige lingvistiske analyse, som man sigter mod i EUROTRA, kan man nemlig ikke vente, at man altid har informationer nok til at vælge mellem to oversættelser ud fra træk i de lingvistiske omgivelser, ligesom man heller ikke kan forvente, at analysen af alle perioder vil resultere i én overførselsstruktur. I sådanne tilfælde vil oplysninger om hyppige og sjældne ord og konstruktioner være en god støtte til at træffe det rigtige valg.

GÖTEBORGS UNIVERSITET
SPRÅKDATA
Bo Ralph

November 1979

LEXIKOLOGI SOM DATALINGVISTIK

Alla teoretiska språkmodeller brukar ha en komponent med språkliga byggstenar, ett lexikon, och en komponent med regler för byggstenarnas sammanfogning till satser och texter, en grammatik. Vid lemmatisering – angett som ett huvudtema vid datalingvistikdagarna – aktualiseras ett par grundläggande frågor som rör den lexikaliska komponenten. Med lemmatisering syftar man i första hand till att samklassificera formella och funktionella varianter till abstraktare lexikaliska enheter och upprätta kanoniska grundformer. Jag vill med det här bidraget lägga några generella synpunkter på lexikaliska enheters egenskaper och inbördes relationer, dvs. på den lexikaliska komponentens allmänna struktur.

Det lexikaliska modellbyggandet är i och för sig inte någon specifikt datalingvistisk angelägenhet. Det finns emellertid flera skäl att diskutera lexikologi i ett datalingvistiskt sammanhang. För det första har inte allmänlingvisterna ägnat sig i särskilt hög grad åt lexikaliska frågor. Huvudvikten har alltid legat vid andra grammatikkomponenter, såsom den syntaktiska, den semantiska, den fonologiska och i någon mån den morfologiska, olika starkt betonade under olika perioder. Eftersom den teoretiska modellen ändå förutsätter en lexikalisk komponent har denna tenderat att bli något slags "garbage component", som man hänskjutit problemen till då man inte velat ta itu med dem i det sammanhang som just varit aktuellt. Som kontrast har datalingvister ofta haft att lösa lexikaliska problem inom ramen för praktisk verksamhet. Datalingvisten har ofta nog fått bli sin egen lexikolog. Som en följd har rätt mycket av den teoretiskt inriktade lexikologi som över huvud taget bedrivits på senare tid presterats av datalingvister.

Det finns emellertid skäl även för andra lexikologer att närma sig datalingvistiken. Själva det lexikaliska materialets art gör det nödvändigt att lexikologi bedrivs i nära samråd med data-

lingvister. Välkänd är Dr. Johnson's karakteristik av lexikografen som "a harmless drudge that busies himself in tracing the original and detailing the signification of words". Även lexikologi är i långa stycken ett träljobb; i högre grad än inom de flesta andra lingvistiska forskningsgrenar förutsätts det att stora datamängder undersöks. Datorn är den idealiske trälen. Metodologiskt ställs lexikologin i ett helt nytt perspektiv genom möjligheterna till interaktiv bearbetning. Struktureringen accentueras om fasta format används. Modellen vinner stadga genom möjligheterna till smidig kategorivis genomgång. Men framför allt är själva den experimentella inriktning som präglar datalingvistiken en tillgång. Trycket på många datalingvister att komma till praktiska resultat ger upphov till idéer som bör prövas inom en allmänteoretisk ram. De systematiseringar av datamängder som utförs av informationsbehandlare och de informationsstrukturer som därigenom byggs upp kan likaledes skänka inspiration åt lexikologin. En verksamhetsgren som i synnerhet måste bevakas av lexikologer är databastekniken, som säkerligen kan ge impulser till lexikologiskt nytänkande.

När lexikon skapas i datalingvistiska sammanhang sker det kanske vanligen med siktet inställt på att klara av en given uppgift. Det är naturligt och legitimt att enheterna i diverse maskinlexikon kan ha högst varierande karaktär beroende på syftet. Avvägningen mellan lexikonets och grammatikens respektive kraftfullhet, fördelningen mellan rent automatisk och interaktiv bearbetning etc. blir praktiskt baserade optimeringsproblem. Avser man att approximera människans eget sätt att fungera språkligt – och det kan man naturligtvis göra även om det primärt är ett praktiskt problem som skall lösas – ställer sig saken annorlunda. Då kommer t.ex. resonemanget om psykologisk relevans in i bilden. Hela synen på den lexikaliska enheten påverkas om hänsyn tas till den psykologiska verkligheten.

Lingvister tycks ofta föreställa sig det mänskliga lexikonet som något slags idealiserad ordbok av relativt traditionell utformning. Det betyder att man övertar åtskilliga av de problem som förknippas med ordböckers utformande. Det är exempelvis en grundläggande fråga vad som skall konstituera en lexikalisk enhet. Problemet har sin rot i förhållandet mellan form och betydelse.

Så länge relationen är någorlunda ett-till-ett, som i *distikon*, *hata*, *vit*, är läget okomplicerat, men hur många lexikaliska enheter bildar det berömda *krona*? Det kan betyda 'huvudprydnad för furste', 'konungamakten el. staten som institution', 'den övre delen på träd', 'den greniga delen på hjortdjurs horn', 'utstyrd taklampa' och 'myntenhet i Norden m.fl. områden' osv. Sture Allén har för svenskans del diskuterat dessa problem i flera skrifter (t.ex. i inledn. till NFO 1 och i Allén 1967).

I den livliga lingvistiska debatten under sextiotalet tilldrog sig knappast den lexikaliska komponenten något huvudintresse, men några utkast till lexikonstruktur som då presenterades är typiska för tiden. Katz och Fodor (1963) tänkte sig att en lexikalisk enhet kunde ha underordnade betydelsevarianter (*bachelor*), ungefär som i traditionell lexikografi. McCawley (1968) uppfattade varje ny kombination av ett uttryck och ett innehåll som en egen lexikalisk enhet. Dessa båda infallsvinklar ger som resultat höggradig polysemi respektive homonymi. I övrigt har diskussionen kring den lexikaliska enheten mera gällt vilken information som skall knytas till denna än hur själva enheten skall avgränsas.

I traditionell lexikografi företas uppdelningen som bekant på etymologins grund. Därigenom blir *tunga* 'slags plattfisk' och *tunga* 'rörlig muskelkropp i munhålan' betraktade som e t t uppslagsord med polysemkomponenter, medan *axel* 'skuldra' är en självständig lexikalisk enhet helt skild från *axel* 'tvärstång el. tänkt centrum kring vilken/vilket något roterar' (jfr fsv. *axl* resp. *axul*). Från strikt synkronisk synpunkt är det inte mer naturligt att associera fisken till muskelkroppen än att associera skuldran till andra bärande axlar. Betecknande nog har *axel* 'tvärstång etc.' åtminstone fakultativt fått akut accent trots ursprunglig tvåstavighet, säkerligen under inverkan från *axel* 'skuldra'. Allén (1978) ger fler exempel av samma slag.

I ordböcker måste praktiska hänsyn (konsekvens, lättillgänglighet och klarhet, utrymme m.m.) vägas mot de teoretiska kraven. En etymologisk uppdelningsprincip kan faktiskt i långa stycken ge ett gott s y n k r o n i s k t r e s u l t a t, nämligen i den mån etymologisk härledning sammanfaller med produktiv morfologi. Principiellt är förstås varje historisk indelning psykologiskt otillfreds

ställande, även då det endast gäller att upprätta ordboksenheter. Ett bättre alternativ är enligt vår mening den lemma-lexem-modell som vi tillämpar i projektet Lexikalisk databas. Modellen finns beskriven i Allén (1978), mera som utkast i Ralph, Järborg och Allén (1977). Enligt modellen etableras den lexikaliska enheten först med hänsyn till form och funktion, vilket ger lemmat, därefter med hänsyn till betydelse, vilket ger lexemet. Gemensamma formella och funktionella uppgifter redovisas i lemmadelen, specifika betydelseuppgifter i lexemdelen. Enheten utgörs av en lemmadel j ä m t e en lexemdel.

Lemmat, såsom det definieras av Allén (1967 m.fl. ställen, särsk. i inledn. till NFO 1), utnyttjades ursprungligen som operativ enhet i en kvantitativ korpusundersökning på datamaskinell basis (NFO 1-4). När det nu läggs till grund för en definitionsordbok är det delvis för att det tillsammans med lexemet möjliggör en klar och konsekvent presentationsform. Men vi anser också att ords uttryckssida hör till det som är psykologiskt relevant. Något tillspetsat kan vi påstå att uteslutande form-funktion är minst lika psykologiskt relevant som etymologi. Ordet *balja* kan i svenskan ha betydelserna 'kar' respektive 'fröskida'. I ISO (1977) ger olika historiskt ursprung upphov till två olika enheter. Lemma-lexem-modellen ger en enhet med två skilda betydelser, eftersom båda substantiven *balja* har samma formella och funktionella egenskaper (se äv. Allén 1978). Språkpsykologiskt är det lika rimligt att uppfatta de två enheterna *balja* som varianter av någon grundbetydelse 'behållare' som att hålla isär dem. Den etymologiska information som hänför sig till respektive betydelse är för övrigt mycket specifik (Hellquist 1966); den trotsar all normal intuition och kräver språkhistoriska specialkunskaper. Nu är det kanske inte meningsfullt att betona den språkpsykologiska aspekten alltför starkt så länge det bara handlar om ordböckers presentationssystem. Det är helt klart att även lemman torde kunna associeras till varandra, och lika uppenbart knyter sig vissa formella egenskaper snarast till enskilda lexem. Av en idealisk lexikonmodell, som alltså inte nödvändigtvis behöver svara mot en ordboks krav, bör även sådana relationer som de just antydda framgå.

När lexikonet fokuserats i den nyare teoretiska lingvistiken, har det vanligen varit Chomskys (ursprungligen Bloomfields) lexikonbegrepp man arbetat med: lexikonet uppfattas som en minimal lista över oregelbundenheterna, det unika i språket, dvs. det som inte kan genereras med regler. I den generativa grammatiken sätts de lexikaliska enheterna in i redan genererade strukturer. De lexikaliska enheterna blir med ett sådant synsätt ungefär liktydiga med morfem. Nu är det uppenbart att människan inte lagrar enbart morfem i sitt lexikaliska minne. Särskilt tydligt framgår detta av egennamnen. Skulle vi "komma ihåg" former som {etiop}, {hondur}, {jav} och {ungr} för *Etiopien-etiopier-etiopisk*, *Honduras-honduran-honduransk*, *Java-javanes-javanesisk*, *Ungern-ungrare-ungersk*? Särskilt nationsnamnens former är här helt oförutsägbara. Troligen har människan en inbyggd morfologisk kompetens, men den fungerar säkert på ett helt annat plan.

Man kan märka en tendens till att avfärda egennamn från den lexikaliska diskussionen som triviala och inom språket extrema fall. Då skall man minnas att språkstatistiskt är namnen långt ifrån marginella. I det material på drygt en miljon löpande ord som beskrivits ur olika synvinklar i NFO 1-4 förekommer nästan 11 000 namn på knappt 45 000 ställen. Således är ca 4.5 % av orden i den löpande texten egennamn. Lexikaliskt stiger proportionen till 15 %; totalt har ca 71 000 lemman framanalyserats. Om en så stor del av lexikonet måste lagras som hela ord finns det egentligen ingen anledning att anta något annat än att resten av det lexikaliska materialet lagras i ordform på samma sätt.

I ordböcker lagras onekligen lexikoninventariet normalt i ordform. Ordböcker förlorar emellertid i psykologisk relevans genom de inneboende presentations- och organisationsproblemen: linearite-ten, den orättmätiga betoningen av alfabetisk ordning, själva den grafiska bildens dominans etc. Vilken uppslagsform man väljer dikteras också till stor del av traditionen. Såväl det lemma som skapas vid lemmatisering som den kanoniska form som väljs som uppslagsform kan representera högst olika entiteter, och förhållandet mellan enhetens realiseringsformer kan variera betydligt. Hur förhåller sig exempelvis en defekt flexionsserie till en fullt utbyggd? Till en uppslagsform *fort* anmäler sig komparationsformerna *fortare-fortast* på ett naturligt sätt, men uppfattas

serien *hellre-helst*, som saknar positivform, som en homogen enhet på samma sätt som den kompletta serien *fort-fortare-fortast*? Kan någondera formen *hellre* eller *helst* anses vara primär framför den andra, såsom kanske *fort* är det jämfört med *fortare-fortast*?

Över huvud taget är paradigmen utfyllda i mycket olika utsträckning för olika ord. En principiell skillnad föreligger mellan sådana paradigm som har stabila luckor och sådana som åtminstone potentiellt kan uppvisa kompletta system. Det finns åtskilliga fall av ofullständiga komparationsserier som visar partiell överensstämmelse med fullständiga, såsom

fort	fortare	fortast	(adverb)
tung	tyngre	tyngst	(adjektiv)
—	hellre	helst	(adverb)
—	sämre	sämst	(adjektiv el. adverb).

Man kan här inte märka någon som helst tendens till nybildning av "analogiska" positivformer som **hell*, **säm*. I NFO 2, som bygger på ett material från 1965, saknar ordet *resurser* på ett liknande sätt singularformer. I detta fall uppfattas emellertid *resurs* spontant som en möjlig singularform. Intuitionen stöds av ett likartat men tio år yngre material, Press 76, där singularformen är belagd:

NFO 2 (material från 1965)		Press 76	
RESURS	121	RESURS	140
s:a sin	*0*	resurs	5
resurser	104	resursen	1
resurserna	17	*s:a sin*	*6*
s:a plu	*121*	resurser	109
		resurserna	25
		s:a plu	*134*

Ordens inherenta betydelse bestämmer naturligtvis i mycket hög grad de olika böjningsformernas inbördes styrkeförhållanden. Så är delvis fallet med *resurs-resurser*, där den plurala eller kollektiva betydelsen ligger nära till hands. På motsatt vis förekommer ord som *värld* och *liv* oftast i singularis, såsom också framgår av siffrorna i NFO 2:

VÄRLD	570	LIV	560
värld	136	liv *sin*	278
världen	310	livet	205
världens	115	livets	39
s:a sin	*561*	livs *sin*	18
världar	9	*s:a sin*	*540*
s:a plu	*9*	liv *plu*	19
		livs *plu*	1
		s:a plu	*20*

Mellan dessa extremer, vars fördelningsförhållanden tycks bero mest på ordens betydelse, återfinns rader av mindre självklara fall. I NFO 2 har *problem* och *väg* totalfrekvenser (559 resp. 552) som är jämförbara med såväl varandra som dem för *värld* och *liv*. Medan *problem* visar relativt jämn fördelning mellan singularis och pluralis (273-286) förekommer *väg* betydligt oftare i singularis än i pluralis (442-103). Även i fråga om andra grammatiska kategorier kan man urskilja individuella profiler för orden. Genitiv är en på det hela taget sparsamt företrädd kategori i svenskan. För enstaka ord bryts dock det vanliga mönstret, som då genitiven är den nästan dominerande realiseringsformen för lemmat *slag*. Man kan jämföra *sak* som har samma frekvens som *slag* men frekvensmässigt negligibla genitivformer.

SAK	453	SLAG	453
sak	183	slag *plu*	69
saken	149	slag *sin*	133
saker	104	slagen	1
sakerna	6	slaget	50
s:a grundform	*442*	*s:a grundform*	*253*
sakens	8	slags	13
sakernas	3	slags	187
s:a genitiv	*11*	*s:a genitiv*	*200*

Fraseologiska konstruktioner kan bidra till att frekvensen höjs för enskilda böjningsformer. Singularformen *väg* understöds av mer eller mindre metaforiska uttryck som *på väg*, *ta vägen*, *vara i vägen* m.m., och genitiven *slags* får sin höga frekvens delvis genom uttryck som *ett slags*, *något slags*. Fraseologi kan slå helt olika inom en klass. Följande adverb, som i NFO 2 har ungefär samma frekvenser (från 297 till 346), ingår enligt NFO 3 i mycket varierande utsträckning i fraser (från 2 till 239).

	frekvens	varav i fraser
verkligen	297	2
däremot	296	8
dessutom	346	13
knappast	307	39
ytterligare	291	99
snart	288	107
alltför	288	144
bort	298	179
ner	286	183
helst	303	215
tillbaka	333	239

Klasser med starkt inskränkt flexion, såsom adverben, kan således vara mycket heterogena med hänsyn till vilken benägenhet de ingående orden visar till att förekomma i fraser. Vi har tidigare sett att enstaka böjningsformer kan framhävas inom paradigmet genom fraseologins inverkan. Härigenom ökas också asymmetrin mellan olika medlemmar av en ordklass. Ovan jämfördes adverben *fort*, inklusive komparationsformerna *fortare-fortast*, och *hellre-helst*. Förutom att den ena serien saknar positivform markeras diskrepansen mellan de båda serierna av den roll fraseologin spelar.

	frekvens	varav i fraser
fort	42	28
fortare	7	3
fortast	2	2
hellre	40	8
helst	303	215

Fort förekommer uteslutande i relativt triviala fraser såsom *för fort*, *lika fort*, *så fort* m.m., *fortare än*, *fortast möjligt*, där *fort* är syntaktiskt utbytbart mot snart sagt vilket adverb som helst. Beträffande *helst* däremot förekommer det hela 209 gånger i den stelnade kombinationen *som helst*, som i uttryck som *vad som helst*, *vem som helst* etc. markerar indefinit betydelse.

Tendens till frasbundenhet är avgjort en viktig lexikalisk uppgift. Det är också tydligt att denna uppgift hänför sig till enskilda ordformer snarare än till den abstrakta enhet som ett ords samtliga böjningsformer bildar tillsammans. I själva verket finns det en del tecken som tyder på att det är ordens böjningsformer man lagrar. Barn är ofta mer direkt medvetna om specifika böjningsformer än om relationerna mellan dem. Det tycks vara en mognadsprocess att vi går från syntagmatiska associationer till paradigmatiska (Salus och Salus 1978). Om människan bygger upp sitt ordförråd genom att först observera enskilda böjningsformer i kontext som tills vidare lagras undan för att först senare associeras till varandra, bör den lexikaliska lagringsformen bära spår av denna process. Frekvenser, som här i viss mån tagits till utgångspunkt för resonemanget men som i språkteoretiska sammanhang ofta nedklassas som ett "performansfenomen", blir därmed ett viktigt begrepp, eftersom frekvenser sedan länge har ansetts vara av grundläggande betydelse för inlärningen. Robinson

nämner exempelvis frekvens som den viktigaste grunden för association (Robinson 1932).

Ordböcker och lexikonmodeller brukar båda vara underkastade stränga ekonomiprinciper. För ordböcker finns det materiella skäl, men lexikonets ekonomi får ses som ett resultat av språkteorins allmänna ekonomitänkande. Kravet på ekonomi och generaliseringar är ofta ett effektivt hjälpmedel i den vetenskapliga analysen, men det är egentligen inget som säger att den mänskliga hjärnan skulle domineras av någon övergripande ekonomiprincip. Tvärtom verkar åtminstone lexikonet vara ganska redundant i sin organisation. Man kan bara tänka på idiomen, som har en perfekt syntaktisk struktur med tolkbara ingående delar men som dessutom fungerar semantiskt på ett fullkomligt arbiträrt sätt. Sådan dubbelhet svär mot varje ekonomiprincip.

Det är troligt att såväl produktionsinriktade som perceptionsinriktade fenomen sätter sina spår i det mänskliga lexikonet. Kanske betonas den semantiska sidan vid perception och upplagring i minnet, medan formen aktualiseras mer vid produktion och framtagning ur minnet. Det mesta av den information man mottar tycks lagras undan rent semantiskt så att den exakta ordalydelsen i meddelandet endast kan rekonstrueras med möda, om alls. Vid produktion, å andra sidan, är det exakta uttrycket viktigt. Det ovan nämnda *slag* kan i de flesta kontexter inte ersättas av det i stort sett likbetydande *art*.

Den enda realistiska modellen av det mänskliga lexikonet torde vara ett flerdimensionellt nätverk som innehåller många sorters information, om betydelse, om form, om funktion osv., med länkar som sammanhåller böjningsformer, synonymer, antonymer, som förbinder fraser med de i dem ingående orden såväl som med synonyma uttryck med annan form osv. Det säger sig självt att en mängd tidskrävande lingvistiska undersökningar behövs innan en sådan modell kan ta form. Lika uppenbart är det att de programmerings- och implementeringstekniska problemen är svårlösta om modellen skall kodas in på datamaskin. I all anspråkslöshet är det ändå principiellt en sådan modell vi har för ögonen när vi bygger upp den lexikaliska databasen i vårt projekt. Informationsmängden kan utökas successivt och måste kanske delvis omstruktureras,

länkningen får sofistikeras efterhand. Det viktiga som skall slås fast här är att en psykologiskt rimlig lexikonmodell i mycket överensstämmer med de databasstrukturer i form av länkade nätverk som nu utvecklas på flera håll. Därigenom understryks ytterligare vikten av att datalingvister engagerar sig i det lexikologiska arbetet.

Referenser

Allén, S. 1967. Förhållandet mellan tal och skrift. I: Allén et al., Språk, språkvård och kommunikation. Lund 1967.

Allén, S. 1978. Lexical entry, linguistic sign, and lexical data base. Föredrag vid 7th International Conference on Computational linguistics. Bergen 1978. Utkommer i förhandlingar från konferensen.

Hellquist, E. Svensk etymologisk ordbok. 3 uppl. 4y tr. Lund 1966.

ISO = Illustrerad svensk ordbok. 3 uppl., 3 tr. Stockholm 1977.

Katz, J.J. och J.A. Fodor. 1963. The structure of a semantic theory. I: Language 39. Omtryckt i: The structure of language, utg. av J.A. Fodor och J.J. Katz. Englewood Cliffs 1964.

McCawley, J.D. 1968. The role of semantics in a grammar. I: Universals in linguistic theory, utg. av E. Bach och R.T. Harms. New York etc. 1968.

NFO = Nusvensk frekvensordbok. 1-3 (1970-75), 4 (utkommer). Stockholm.

Press 76. Konkordans över ett tidningsmaterial från 1976. Språkdata, Göteborg. [Radskrivarutskrift.]

Ralph, B, J. Järborg och S. Allén. 1977. Svensk ordbok och Lexikalisk databas. Förstudierapport. Språkdata, Göteborg. [Stenc.]

Robinson, E.S. Association theory today. New York 1932.

Salus, M.W. och P.H. Salus. 1978. The acquisition of opposites and the structure of the universe. Papers in language use and language function 1. Scarborough College, University of Toronto. [Stenc.]

Denna artikel har delvis kunnat utarbetas inom ramen för projektet Lexikalisk databas.

Sture Allén, som var forhindret i at deltage i mødet, har sendt denne artikel, som stort set dækker, hvad han ville have sagt i sit foredrag.

STURE ALLÉN, som är född i Göteborg 1928,
avlade fil kand-examen 1954, fil mag-examen 1956 och fil lic-examen 1961 samt disputerade för fil doktorsgrad på avhandlingen "Grafematisk analys som grundval för textedering med särskild hänsyn till Johan Ekeblads brev till brodern Claes Ekeblad 1639–1655" (Nordistica Gothoburgensia 1–2) 1965, allt i Göteborg. Förordnad till docent i nordiska språk vid Göteborgs universitet 1965. Tjänstledig för att leda projektet "Datamaskinell undersökning av tidningsprosa" med anslag från Riksbankens jubileumsfond från 1966. Förordnad till t f professor i nordiska språk vid Göteborgs universitet periodvis 1967–69. Förordnad till innehavare av en forskartjänst i språklig datateknik vid Statens humanistiska forskningsråd 1970. Utnämnd till professor i språklig databehandling vid Statens humanistiska forskningsråd 1972. Tilldelades Henrik Ahrenbergs pris av Göteborgs universitet 1966 och erhöll Svenska Akademiens språkvårdspris 1979. Invaldes i Kungliga Vetenskaps- och Vitterhets-Samhället i Göteborg 1976.

Utnämnd att fr o m den 1 juli 1979 vara professor i språkvetenskaplig databehandling vid universitetet i Göteborg.

Språkvetenskaplig databehandling

Frågorna från olika håll blir många, efterhand som ett nytt ämne växer fram. De kan ha sin utgångspunkt i att det hela verkar obegripligt eller fascinerande eller överflödigt eller nyttigt eller något annat. I allmänhet bottnar de i ett intresse att få veta mer. Undringarna tar här gestalt i en *kursiv* samtalspartner. Det utspinner sig ett belärande samtal eller som det en gång hette en dialogus.

Hur började det?

I vårt fall har ämnet spirat ur forskning i nordiska språk med intresse för frågor av allmänt språkvetenskaplig natur. Konkret gällde det i början att utveckla metodik för att snabbt få fram information om vilket element som helst i en text, t ex beläggen på ett ord eller en bokstav. Behovet kunde tillgodoses av en datamaskinell konkordans. Konkordansen till Johan Ekeblads brev togs fram 1961–62.

Sedan dess har en rad mer raffinerade konkordanstekniker skapats. Konkordanser av växlande slag är fortfarande det mest efterfrågade forskningshjälpmed-

let både utom och inom institutionen. Här är några rader ur en konkordans över ordledet *log* med de fulla orden som kontext. Markeringarna till höger har att göra med klassifikation och beläggställe.

bi o	log	isk	AV -T	7306
bok kata	log		NN -EN	7944
chef s id eo	log		NN -EN	10224
Chicago soci o	log	i	NN -N	10262
dia	log		NN -EN	11807
dia	log	is er a nde	AV -=	11811
dia	log	scen	NN -EN	11815
djup psyk o	log	i	NN -N	12211
djur fys io	log	i	NN -N	12231
eko	log		NN -EN	13592
ent myt o	log	is er a	VB -AD	14182
epi	log		NN -EN	14245

Det finns också andra utgångspunkter för ämnet. På 1950-talet hade datamaskinen bland annat använts på olika håll för försök med språköversättning och informationssökning och för beräkningar i samband med textattribution eller författarbestämning.

Frekvensundersökningar nämns ofta. Vilken roll spelar de?

De är inte så dominerande som många tror. Jag menar då frekvensundersökningar som går utöver den rena bokföringen av teckensträngar mellan blanka typrum i en text, i den här meningen alltså "Jag", "menar", "då" osv. Mer avancerade frekvensundersökningar kräver utveckling av välspecificerade beskrivningsmodeller, något som leder rakt in i centrala lingvistiska problem. Märk väl att frekvensstudier inte alls nödvändigtvis gäller ord – som redan det är ett svårfångat begrepp – utan också ordförbindelser, ordled, grundbetydelser eller kärnbetydelser, syntaktiska konstruktioner och åtskilligt annat.

Det är fullt klart att en frekvensundersökning av modernt svenskt material, bedriven på det antydda sättet, har varit av grundläggande betydelse för ämnets framväxt här. De teoretiska och metodologiska övervägandena och tillämpningen av de utarbetade modellerna har kommit konturerna till ämnesområdet att framträda. Naturligtvis har internationella kontakter genom konferenser, arrangemang av forskarkurser och arbetsseminarier, samarbetsprojekt, publikationer och annat också givit betydelsefulla impulser.

Det projekt som har resulterat i Nusvensk frekvensordbok baserad på tidningstext skisserades 1964. Materialet samlades in 1965. Det bestod av färdigstansade hålremsor från tidningssätterier. Bearbetningarna började 1966 och utfördes inom den då bildade forskningsgruppen för modern svenska. Det är alltså i hög grad fråga om ett lagarbete. Den fjärde och sista ordboksdelen, som just gäller ordled och kärnbetydelser, produktionskörs hösten 1979. Formellt upplöstes forskningsgruppen i samband med att institutionen bildades 1977.

Hur kan man då karakterisera ämnet?

Språkvetenskaplig databehandling är som namnet anger ett språkvetenskapligt ämne. Det allmänna målet för ett sådant är att vinna insikt i naturligt språk och det vetenskapliga studiet av naturligt språk. Inom vårt speciella område är det grundläggande att betrakta språket som en process. Allmänt sett är processer mer komplicerade än strukturer genom att man måste ta hänsyn till ordningsföljd. Exempel på frågor som aktualiseras är hur man skall finna en satsdels gränser, hur man skall komma åt en diskontinuerlig förbindelses delar – ett exempel är *på* ett på förhand givet *sätt;* hur man skall identifiera och sammanföra böjningsformer och variantformer, hur man skall känna igen en lexikalisk huvudbetydelse – ***kursiv*** adjektiv 1 lutande (om stil) 2 mindre noggrann, oförberedd (om läsning och översättning) 3 fortlöpande, obegränsat pågående (om verbhandling); i vilken ordning man skall generera och presentera de olika konstituenterna i en struktur osv.

Denna dynamiska inriktning kräver lexikaliska och grammatiska komponenter som är långt mer detaljerade, täckande och välstrukturerade än vad som nu står till buds. De måste med ett ord vara explicita. Processynen leder enligt min tanke till uppställande av ett långsiktigt mål för forskningen inom ämnet som kan sammanfattas i formuleringen datorn som språkbrukare.

Datorn som språkbrukare?

Låt oss försöka undvika missförstånd. Den oförlikneliga språkbrukaren är människan. Det är vi överens om. Analogin med datamaskinen innebär inte att människan betraktas som en automat, inte heller att datamaskinen någonsin kan förväntas nå människans kommunikativa höjder. Genom att placera datorn på språkbrukarens plats ställer vi de lingvistiska frågorna på sin spets. Det är det som är den vetenskapliga poängen.

Det finns en viktig poäng till. Datorn får ju allt fler användningar i samhällslivet. Därigenom ökar behovet av kommunikation med den. Det är väsentligt att denna kommunikation sker på människans villkor.

Var ligger tyngdpunkten?

Låt oss först kasta en blick på modellen. Den ser ut ungefär så här. Språkliga yttringar kommer in till datamaskinen. Det gäller att avkoda dem – jag undviker ordet förstå. De skall i sin tur ge upphov till utdata, som kan vara av många slag, t ex en omskrivning, en konkret åtgärd, placering av språkgodset på rätt plats i en språkbank, ett svar på en fråga, en sammanfattning eller ett utkast till översättning.

Allt detta kräver naturligtvis kommunikativa resurser. Dem kan man ordna i fyra grupper. Med din tillåtelse betecknar jag dem med några fackterm er och kommer strax in på vad de står för: lingvistisk kapacitet, encyklopedisk information, dynamiska faciliteter och algoritmer. Den lingvistiska kapaciteten gäller det lexikaliska systemet, det grammatiska systemet och språkbruket. Observera att data om språkbruket, användningen av språksystemet, inte är mindre viktiga än data om systemet självt. Man kan varken analysera eller generera naturligt språk på ett tillfredsställande sätt utan en uppsjö av information om hur och i vilken utsträckning språksystemets olika delar används.

Den encyklopediska information som krävs är systematiserad kunskap om världen. Vad kan exempelvis befinna sig ovanpå något annat? Fanns det verkligen bussar på Karl XII:s tid? Uppgiften är lindrigt talat stor och långsiktig. Den inbegriper flera olika vetenskapsgrenar. Detsamma gäller de dynamiska faciliteterna, som innefattar sådant som modeller för uppbyggande av kunskap genom perception och slutsatsdragning och strategier för konversation. Den forskning som bedrivs inom området artificiell intelligens är av stor betydelse härvidlag. Algoritmer slutligen behövs naturligtvis över hela fältet. Deras roll är ju att i detalj ange de procedurer som programmeringen bygger på.

Svaret på frågan om var tyngdpunkten ligger blir efter detta, att utforskningen av den lingvistiska kapaciteten och de därmed oskiljaktigt förenade algoritmerna är det centrala.

Människan har ju också en psykologisk dimension. Vad blir det av den?

Analogin med språkbrukaren framhäver i själva verket den aspekten. Att en modell har vad som brukar kallas psykologisk relevans får allt större betydelse. Den psykolingvistiska forskningen är också livaktig och följs med uppmärksamhet. Resultaten är emellertid ännu ganska motsägelsefulla och därför inte utan vidare tillämpliga. Men om det kan göras troligt att den ena av två i övrigt likvärdiga modeller är psykologiskt mera träffande än den andra, är den naturligtvis att föredra.

I vårt arbete på att komma åt de operativa enheterna i det lexikaliska betydel-

sesystemet har vi försökt att fånga de kategorier som kan antas vara relevanta för en observant och intresserad språkbrukare med tillgång till några vanliga nuspråkliga ordböcker.

Har detta synsätt stått klart från början?

Det har vuxit fram efterhand. Några saker har stått klara från början. Dit hör betydelsen av att undersöka autentiskt språk, att göra det på basis av teoretiskt grundade modeller och att utveckla datamaskinell metodik för det.

Vad kan datamaskinen redan nu göra av det vi har varit inne på?

Jag kanske får precisera din fråga på en viktig punkt. Datamaskinen gör inget annat än exakt det som vi programvägen har gett den möjlighet att göra. Det är alltså vi som verkar genom den. När detta är sagt, vill jag svara att vi under överskådlig tid får räkna med samarbete mellan maskinens programsystem och människan genom interaktiv databehandling. Successivt kan behovet av mänskligt ingripande under körningarna av allt att döma bli mindre. I flera avseenden kommer det förmodligen att bestå, såsom vid lösning av besvärliga flertydigheter, hantering av avsiktliga ordlekar av typen "Niagara är ett gränsfall", analys av texter där olika sociala och religiösa system interfererar med varandra, översättning av många slags texter osv. Å andra sidan finns det redan nu programsystem som kan utföra ganska goda morfologiska, syntaktiska och lexikaliska analyser, svara på frågor inom vissa områden osv.

Sedan hör det till saken att datorn också är ett redskap i själva forskningsprocessen. Den används nämligen för att bygga upp de kunskapsmängder och utveckla de system som krävs enligt språkbrukarmodellen. När det gäller den lingvistiska kapaciteten kan datamaskinen bland annat användas för experiment, t ex testning av språkregler, och insamling och bearbetning av lexikaliskt material. Den kan också användas som en stor och flexibel informationsbevarare.

Vilken roll spelar datateknikens utveckling?

Den är viktig på det sättet att den successivt gör tidigare omöjliga ting möjliga och tidigare tidsödande ting snabbt avklarade. Och utvecklingen har gått raskt. En stordator från mitten av 1950-talet som kostade 3 000 000 kronor kunde utföra 2000 instruktioner per sekund och hade en vikt av 7 ton. En mikrodator från mitten av 1970-talet som kostade 3000 kronor kunde utföra 200 000 instruktioner per sekund och hade en vikt av 7 kilogram.

För oss är minidatorn inte minst viktig. Den första modellen av en sådan lanserades 1960. Man har jämfört utvecklingen av minidatorn med utvecklingen av

en bil som Volkswagen för att konkretisera förhållandena. Om bilen hade haft motsvarande utveckling av prestanda och pris, så skulle den nu ha haft en topphastighet av 100 000 kilometer i timmen och kostat 10 kronor. Den avgörande slutsatsen av detta är, att vi inte skall låta oss hindra av för tillfället rådande teknologiska begränsningar vid utvecklingen av forskningsmetodiken på området. Datamaskinen är trots allt bara ett tredjedels sekel gammal.

Vilken utrustning har institutionen?

Vi har vad man kan kalla ett stort minidatorsystem som är speciellt utformat för behandling av språkligt material. Det kräver bland annat stor minneskapacitet, god strängbehandling, obegränsad typuppsättning och terminaler för interaktiv bearbetning. Forskarnas närkontakt med datamaskinen via textskärmsterminaler är av mycket stor vikt. Vi utnyttjar också flitigt ett par terminaler som är kopplade till Göteborgs Datacentral, där vi får tunga bearbetningar utförda.

Vilken inriktning har institutionens forskning för närvarande?

Frågan förutsätter helt rimligt att vi inte försöker täcka hela fältet. Man kan säga att datalingvistiska undersökningar av det lexikaliska systemet och av språkbruket utgör huvudområdena. Efter frekvensundersökningen som nu avslutas är det största projektet det som syftar till att lägga upp en omfattande svensk lexikalisk databas på modern lingvistisk grund och med det levande språket som källa. Genom denna förs utvecklingen av lexikon för datamaskinell analys och syntes vidare. Från databasen skall vi också generera en ordbok över modern svenska för allmänt bruk. Andra lexikaliska arbeten gäller ordlistor för invandrare och en studie av alla svenskars namn. Namn är ett rätt stort inslag i språkets lexikon. Under hösten 1979 utkommer Förnamnsboken.

Vi strävar efter att också arbeta med projekt som rör system och procedurer på den mera renodlat grammatiska sidan inom ramen för algoritmisk textanalys. En studie i program för morfologisk och syntaktisk analys är just avslutad.

Av flera skäl är språkmaterialet oftare skrivet än talat språk. En del av undersökningarna tar emellertid hänsyn till talspråk. Ett projekt arbetar helt med det. Undersökningen gäller vissa kommunikativa drag i talat vardagsspråk i Göteborg och deras relation till sociala faktorer.

Vad får forskningen för konsekvenser för teoribildningen?

Tillkomsten för första gången i historien av en symbolbehandlande maskin är naturligtvis en utmaning för lingvistiken. Detta understryks av att minneskapacitet, hastighet och kompakthet nu närmar sig den mänskliga hjärnans. I stället för

att begränsa oss till spekulationer om språkbeskrivningens utformning kan vi använda datamaskinen som redskap. Det visar sig att också starkt uppmärksammade teoretiska inriktningar råkar i svårigheter. Det som skymtar är en lingvistisk teori som en teori om analys och syntes av naturligt språk med hjälp av en uppsättning kommunikativa resurser inklusive algoritmer.

Det talas ibland om mjukdata. Vad blir det av dem?

I vissa sammanhang talar man mycket riktigt om mjukdata, som då gärna ställs i motsats till hårddata. Detta väcker lite av samma föreställningar som mjukvara och hårdvara inom databehandlingen, dvs program respektive maskinutrustning. Intressant är då att gränsen mellan dessa håller på att mjukas upp. Med hjälp av mikroteknik kan program nu monteras in som maskinkomponenter. Vi får hård mjukvara.

Som mjukdata betecknas exempelvis uppgifter om människors attityd till olika språkliga uttryck, t ex yrkesbeteckningar, människors upplevelse av sjukdom eller åldrande osv. Som hårddata betecknas då observerade språkliga frekvenser, resultat av laboratorieprov osv. I båda fallen gäller emellertid att uppgifterna måste svara mot vetenskapliga krav på noggrannhet i dokumentationen. De kan då principiellt också behandlas med likartade metoder. Orden mjukdata och hårddata representerar därför snarare olika vetenskapliga forskningsinriktningar än fundamentalt olika slag av data.

Kan man då också säga att epistemiska, fenomenologiska, holistiska och ontologiska frågeställningar inom språkvetenskapen är förenliga med forskningen på området?

Naturligtvis.

Är det riktigt att beteckna ämnet som tvärvetenskapligt?

Ja och nej. Beteckningen tvärvetenskap är inte alldeles klar. Vid dess sida står bland annat mångvetenskap som benämning på forskningsföretag som bygger på samverkan mellan olika vetenskaper. Sann tvärvetenskap innebär en sammansmältning av flera ämnen till en ny helhet. Detta är den tvärvetenskapliga paradoxen. När ett nytt ämne har bildats på det sättet, kan det ju inte längre med rätta kallas tvärvetenskapligt annat än från historisk synpunkt.

Vårt ämne har emellertid livliga förbindelser med andra ämnen. Det är också hjälpvetenskap för ämnen inom och utom språkvetenskapen och utnyttjar i sin tur andra ämnen som hjälpvetenskaper.

Vad gäller för utbildningen?

Som den är upplagd nu riktar den sig enbart till doktorander och utgår från grundexamen med tre terminers studier i allmän språkvetenskap eller motsvarande kunskaper. Studier i informationsbehandling är inte ett förkunskapskrav. Den som saknar sådana kunskaper får tillägna sig dem under utbildningen genom kursläsning och laborationer. Utöver doktorandutbildningen ger vi ibland orienteringskurser för olika grupper. Den verksamheten borde nog vidgas.

Vart vänder man sig om man vill utnyttja institutionens språkmaterial?

Då tar man kontakt med institutionens servicecorgan Logoteket. Det är en dynamisk språkbank som inrättades 1975 som en permanentning av en länge bedriven serviceverksamhet. Logoteket har vissa nationella uppgifter, nämligen att samla in och bevara datamaskinellt läsbara texter, att tillhandahålla sådana texter och vissa bearbetningar av dem, att bygga upp en ordbank och att ge råd på sitt område, allt i mån av resurser.

Textsamlingarna omfattar nu omkring 30 miljoner ord ur bland annat skönlitteratur, tidningar och författningar. Två speciella textmängder är särskilt aktuella. Den ena är riksdagens snabbprotokoll från arbetsåret 1978-79, tillsammans omkring fyra miljoner ord i en språkform någonstans mellan talspråk och skriftspråk. Den andra är Strindbergs samlade skrifter som skall kodas in och bearbetas här som ett led i kulturrådets nyutgivning. Omfånget uppskattas till mellan sex och sju miljoner ord.

Ordbanken innehåller ett par hundra tusen ord försedda med varierande uppgifter. Huvudinslag är frekvensordbokens olika bearbetningar och Svenska Akademiens ordlista. De inkommande texternas ordförråd tillförs successivt.

Konkordanser på mikrokort framställs systematiskt. Belägg på olika typer av ord och konstruktioner efterfrågas också. Avnämarna finns inom språkforskning och samhällsforskning, språkvård och journalistik, informationsbehandling och informationssökning, olika grenar av grafisk industri osv.

Hör arbetet kring språkbanken till den vetenskapliga verksamheten?

På flera sätt. Urvalet av språkmaterial och utformningen av databaserna, innehållsligt och tekniskt, är exempel på forskningsproblem. Men det finns skäl att anlägga ett vidare perspektiv. Inom en vetenskapsgren, och då inte minst språkforskningen, står vad man kunde kalla vetenskapsvården i centrum. Vad det gäller är teoribildning, metodutveckling och kunskapsgenerering. Ett annat betydelsefullt område är utbildningen, som ju är en förutsättning för att den vetenskapliga genomlysningen skall föras vidare. Ett tredje område har jag kallat resultatvård.

Här hör servicen hemma tillsammans med praktiska tillämpningar inom olika samhällssektorer. Forskarnas del i ansvaret för resultatens användning kommer också in i bilden.

En sak till. Intuitionen, vart tar den vägen?

I grunden finns det ingen motsättning, menar jag, mellan ett datalingvistiskt betraktelsesätt och en intuitiv uppläggning. Ytterst bör det nämligen vara så, att intuition är grundad på skarp observation. Förhåller det sig på det sättet, och min intuition får mig att tro det, blir ämnesområdets framtid inte mindre intressant.

Efter avslutat samtal förfogar man sig på klassiskt maner till caldarium, där armhävning och hett bad väntar.

STURE ALLÉN

ALFABETISK FORFATTERREGISTER