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

BETA-systemet som verktyg för lingvistiska experiment - morfologisk analys enligt ytkriterier
Benny Brodda
Proceedings of NODALIDA 1979, pages 103-121

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: $X \left\{ \begin{array}{l} \hat{} \ V \ \hat{} \ Mc \\ -S \end{array} \right\} \hat{} \ \underline{E} \ \#$

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

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

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.

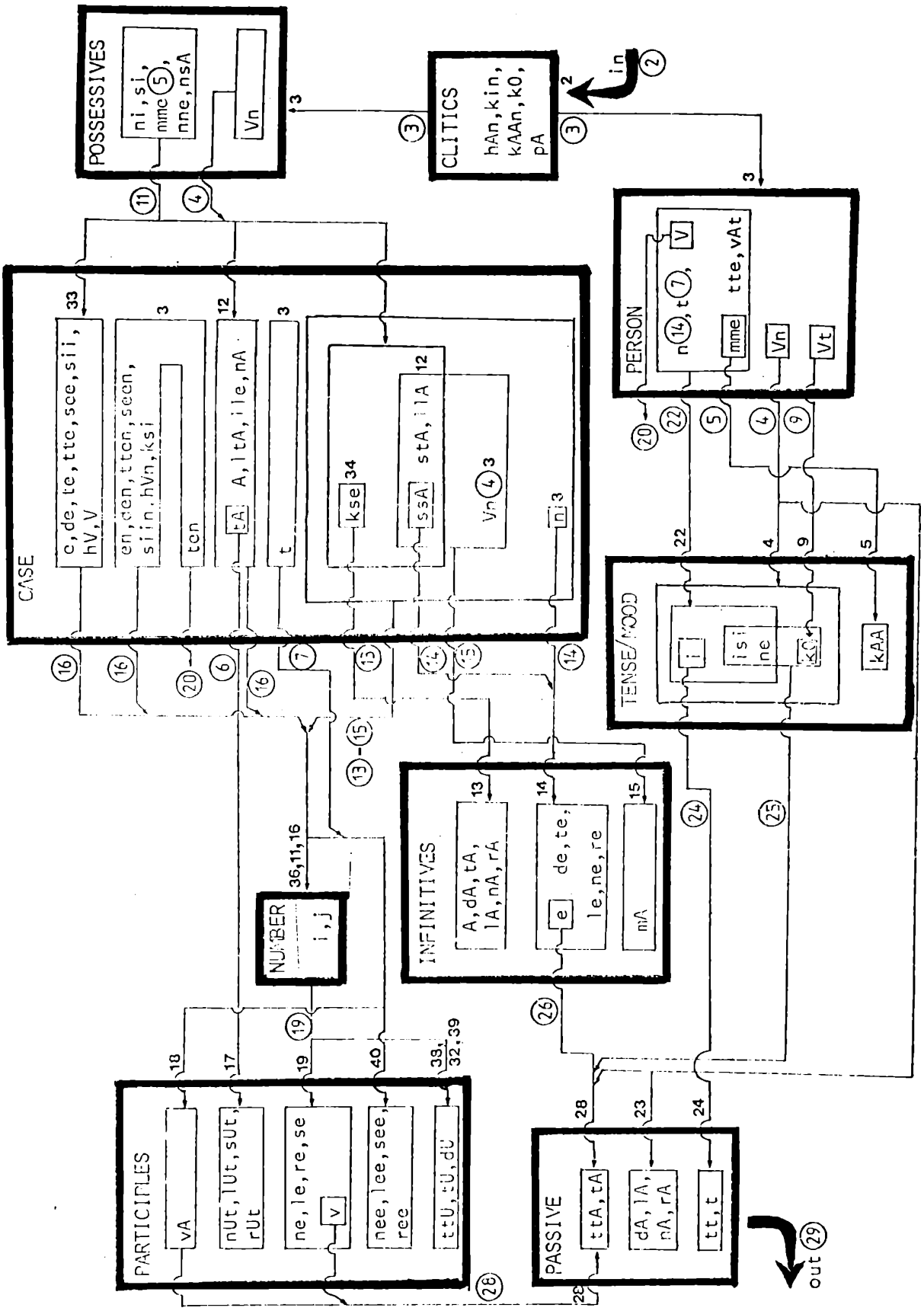

Fig. 2 Analysschema för finskan.

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-RMEAR=AN
1  ORMEAR=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!

III. 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
P|   P|     13   0   3   6   5
```

TXTLST: LATIN.X12 DATUM :07-OKT-79 PAGE: 3

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