Anna-Lena Sågvall Hein

CHARTANALYS OCH MORFOLOGI

Med automatisk grammatisk analys menas vanligen en automatiserad process, varigenom en grammatisk interpretation tillordnas till en språklig enhet. Beroende på om den språkliga enheten utgörs av en isolerad ordform eller ett längre uttryck brukar man traditionellt skilja mellan automatisk morfologisk resp syntaktisk analys.

Skiljelinjen mellan morfologi och syntax tenderar inom området datalingvistik att få en extra, artificiell markering av det faktum att existerande dataprogram som regel är inriktade antingen på morfologi eller syntax. Det rör sig om programsystem med helt skild uppbyggnad och komplexitet.

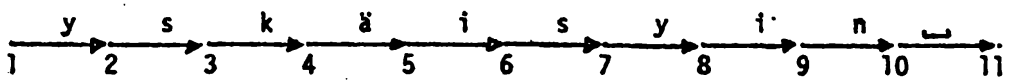
Existerande program för morfologisk analys är vanligen specifika till sin uppbyggnad i två avseenden; dels avgränsar man sig mot syntaxen genom vissa allmänna restriktioner som t ex att man utgår från att analysenheten utgörs av en teckensträng, avgränsad i löpande text genom mellanslag eller skiljetecken, dels anpassar man sig vanligen till de språkspecifika dragen i det språk man vill analysera. En sådan anpassning är givetvis eftersträvansvärd; problemen ligger i att dessa begränsningar vanligen byggs in i själva programlogiken. Man kan därför inte utveckla ett sådant system utöver de i förväg uppställda ramarna, t ex till att handskas med ett språk med större morfologisk variation eller till att klara av syntaktiska problem. Sålunda kan man t ex inte i ett traditionellt system för morfologisk analys känna igen analytiskt resp syntetiskt bildad komparativform som varianter.

Inom förefintliga system för syntaktisk analys, vilka vanligen skrivits för analys av engelska, har å andra sidan morfologin en rudimentär behandling. Man vill koncentrera sig på de 'intressanta' problemen och strävar efter att komma förbi snarare än igenom morfologin. Den morfologiska komponenten i sådana system är därför sällan utvecklingsbar.

Möjligheterna att inom ramen för ett och samma system kunna utföra såväl morfologisk som syntaktisk analys är en av grundtankarna bakom M Kay's 'Chartanalys' (1). Den kommer till uttryck bl a däri att analysenheten antingen den utgörs av en eller flera ordformer representeras med hjälp av samma struktur, en chart.

Fig 1 visar en grafisk representation av charten för den finska ordformen 'yskäisyin' (='hostning' i instr sg el pl) sådan den ter sig innan själva bearbetningen påbörjats.

fig 1



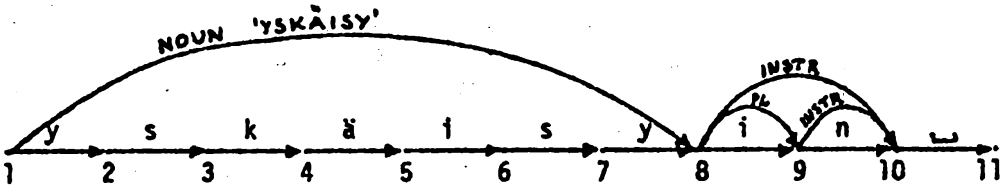
Anm. '␣' markerar mellanslag

Chartanalys och morfologi
Anna Sågvall Hein
Proceedings of NODALIDA 1977, pages 87-93

Charten består av numrerade vertices (1 till 11), sammanbundna av riktade, etiketterade edgar (1 - 2 'y', 2 - 3 's', etc). Den intar en central plats i det av Kay föreslagna systemet för grammatisk analys. Den används inte bara för att representera analysenheten sådan den ser ut innan analysen påbörjats utan även för att lagra såväl delresultat under pågående analys som slutresultat efter avslutad bearbetning. Väsentlig blir härvid möjligheten att inom charten representera alternativa analyser.

Fig 2 visar hur charten för 'yskäisyn' skulle kunna se ut efter avslutad morfologisk analys.

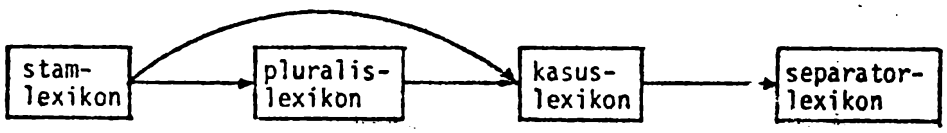
fig 2



Om vi jämför med fig 1, så finner vi att 4 nya edgar har införts, nämligen från 1 till 8, från 8 till 10, från 8 till 9 och från 9 till 10. De svarar mot två alternativa läsningar, d v s två alternativa segmenteringar, nämligen
1. 'yskäisy' - 'in' (subst 'yskäisy' i instruktiv) och
2. 'yskäisy' - 'i' - 'n' (subst 'yskäisy' i instruktiv pl).[1)]
Dessa analyser förutsätter att systemet konsulterat ett stamlexikon (huvudlexikon enl Kay's terminologi), där stammen 'yskäisy' med informationen 'substantiv' återfinns, ett suffixlexikon, som upptar suffixet 'i' med informationen 'pluralis', ett suffixlexikon, som upptar suffixen 'n' och 'in', båda med informationen 'instruktiv' samt slutligen ett 'separatorlexikon', som upptar ' '.[2)]

Det finns i det generella systemet som sådant inga begränsningar på hur många morfemsegment som kan följa på varandra i en ordform eller någon uppgift om vilka de är eller om den inbördes ordningen mellan dem. Som språkspecifik information måste man därför, förutom ett huvudlexikon, även tillföra systemet ett antal lexikon (Kay's terminologi) som upptar de faktiska realisationerna av de grammatiska morfemen, jämte de morfotaktiska reglerna. Morfotaxen avspeglas i en hänvändelse vid de olika segmenten i huvudlexikon resp suffixlexikon till vilket lexikon skall konsulteras för igenkänning av följande segment. Kopplingen mellan de olika lexikonen, som ligger bakom analysen i fig 2, dvs den bakomliggande morfotaxen illustreras i fig 3.

fig 3



1) Tolkningen av morfemsegmenten ligger som synes i etiketterna för motsvarande edgar.
2) Konsultationen i separatorlexikonet har till uppgift att kontrollera att ordformen i fråga är slut, d v s i det anförda exemplet att analysenheten verkligen är uttömd i och med återfinnandet av kasussegmentet.

Införandet av ett särskilt separatorlexikon bidrar till systemets generalitet och konsistens. Det innebär, att kontroll av huruvida en ordform är slut eller inte kan ske helt i analogi med hur en jämförelse mellan en bokstav i analysenheten - en edge i charten - och en bokstav i något av lexikonen, går till. Det allmänna begreppet för en sådan aktion är 'task' (uppgift). En uppgift kan t ex också svara mot tillämpningen av en grammatisk regel på en eller flera edgar. Under bearbetningens gång lagras uppgifterna i en (eller flera) agenda (agendor), varifrån de aktiveras. Hela analysprocessen kan beskrivas som en följd av uppgifter. [1]

Att låta de morfotaktiska reglerna komma till uttryck enbart via kopplingen mellan de olika lexikonen är emellertid inte tillräckligt för att garantera korrekta analysresultat. Vi kan illustrera problematiken med ett exempel. Antag att våra lexikon (för den finska morfologin) innehåller bl a följande uppslagsord, fig 4.

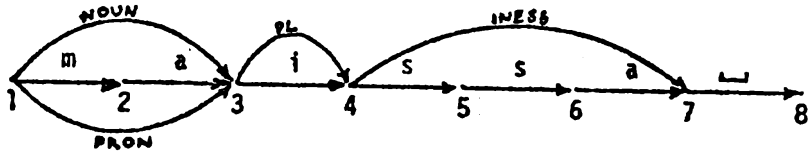
fig 4

```
*MAIN:

        MA      1. MORPHCAT=WORD, LEXEM=MA, SYNCAT=PRON
                   continue in dictionary SEP

                2. MORPHCAT=STEM, LEXEM=MAA, SYNCAT=NOUN
                   continue in dictionary NUM

+ + + + + + + + + + + + + + + + + + + + + + + + + + + + + + + + + +

NUM:

        I       1. NUM=PL
                   continue in dictionary CASE

+ + + + + + + + + + + + + + + + + + + + + + + + + + + + + + + + + +

CASE:

        SSA     1. CASE=INESS
                   continue in dictionary SEP

+ + + + + + + + + + + + + + + + + + + + + + + + + + + + + + + + + +

SEP:

        ␣       1. continue in dictionary *MAIN
```

Då skulle den morfologiska analysen av ordformen 'maissa' (inessiv plur av 'land') ge följande chartstruktur, fig 5.

1) Se vidare om systemets generella uppbyggnad i (1) och (2).

fig 5



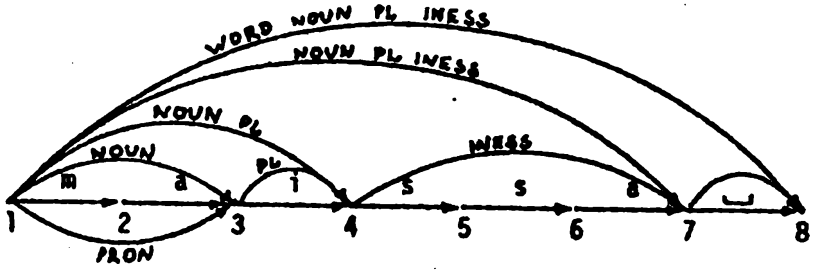
Denna chartstruktur medger två läsningar, nämligen 1. substantivet MAA i inessiv plur och 2. pronominet MA i iness plur, varav endast den första är korrekt. Orsaken till att vi i fall av homografi, t ex stamhomografi som ovan, får en felaktig alternativ analys är tvåfaldig; dels beror det givetvis på själva chartstrukturen, dels på det faktum att lexikonsökningen och segmenteringen sker helt kontextoberoende, d v s att en ny edge läggs in i charten så snart man finner överensstämmelse mellan ett segment (uppslagsord) i något av lexikonen och en följd av edgar i charten. Med bibehållande av chartstrukturen kan vi nalkas problematiken på två sätt, antingen genom att göra segmenteringsprocessen kontextsensitiv eller genom att formulera fristående kompatibilitetsregler, som appliceras på den kontextfritt genererade chartstrukturen, och vars tillämpning leder till att övergripande edgar införs för att markera kompatibilitet mellan segment.

I min egen implementering av chartanalysen har jag valt att göra segmenteringsprocessen kontextkänslig. Här arbetar jag på så sätt, att ingen ny edge introduceras i charten förrän systemet verifierat, att det segment, som representeras av edgen, är kompatibelt med angränsande segment. När man arbetar med en ordform som analysenhet innebär det i praktiken, att införandet av nya (segment-)edgar måste undertryckas, till dess man undersökt hela ordformen, inkl separatorn. I exemplet ovan (fig 5) skulle alternativet 'pron' aldrig ha införts i charten, enligt min implementering, då systemet ej återfunnit den förväntade separatorn. Fördelen med den här metoden är dels den, att man får ett väsentligt mindre belastat chart, eftersom endast de korrekta alternativen återspeglas i charten, dels den att analysen kommer att ske i färre etapper, då man i samma bearbetningssteg både segmenterar och verifierar kompatibilitet.[1] Detta tillvägagångssätt har å andra sidan den nackdelen, att det är svårt att vara lika generell som vid en kontextfri segmenteringsprocess. Språkspecifik information smyger sig lätt in i programmen, varför programsystemet måste ha en modulär uppbyggnad, som gör de språkspecifika programmen lätt utbytbara.

Låt oss diskutera det andra alternativet, där man bibehåller en kontextoberoende segmenteringsprocess och låter kompatibiliteten mellan segmenten kontrolleras i påföljande bearbetningssteg. Fig 5 illustrerar då endast en etapp i analysen av 'maissa'. En möjlig slutgiltig chartstruktur visas i fig 6.

1) För en illustration av denna metod, se (3).

fig 6



Vid läsningen av chartstrukturen tillämpar man den konventionen, att det slutgiltiga analysresultatet finns i etiketten (etiketterna) till den (eller de) edge (edgar) som går från första till sista vertex. Om den resulterande charten inte uppvisar någon sådan edge har analysen inte lyckats. Felaktiga delanalyser, som i exemplet tolkningen av 'ma' som ett pronomen, stör därigenom inte slutresultatet. Edgen från vertex 1 till vertex 7 representerar enbart en delanalys, då man ännu inte har något belägg för att ordet de facto slutar efter kasussegmentet. Återfinnandet av separatorn tillför här igen ny information utan har endast kontrollfunktion.

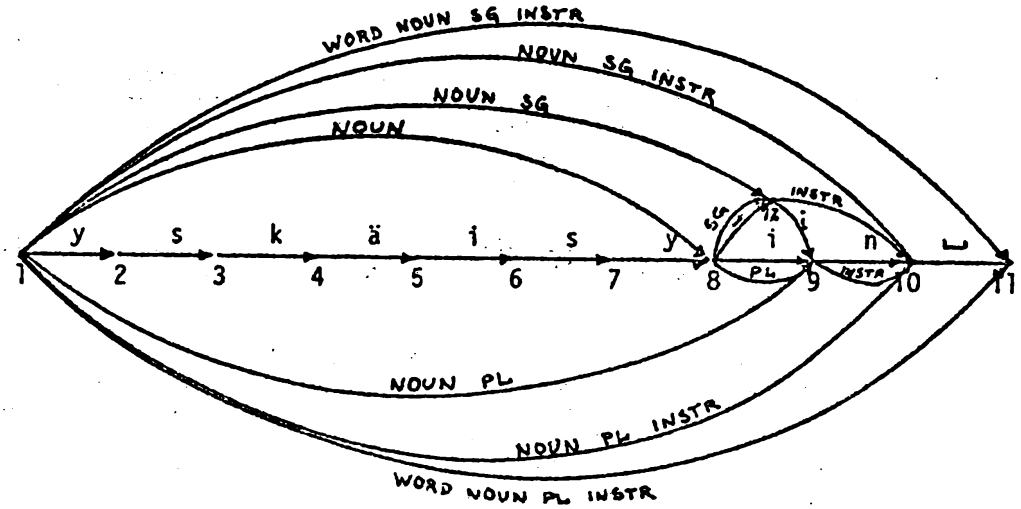
Vilka är då problemen inom chartanalysens ram, då det gäller att komma fram till en sådan analys som i fig 6? Successivt måste man verifiera kompatibiliteten mellan segmenten, t ex 'noun - pl', 'pron - pl'. Om det råder kompatibilitet, skall en övergripande edge införas, som i sin etikett bär den relevanta informationen från de båda edgarna. Detta skall ske genom att en regel appliceras på de båda edgarna. Tillämpningen av en regel på en eller flera edgar formuleras som alla andra aktioner inom chartanalysteorin som en uppgift. När skall nu den uppgiften genereras och varifrån skall den hämta information om vilket test, som skall utföras, samt om vad, som skall ingå i etiketten till den övergripande edgen? Jag har valt följande strategi:

1. Kompatibilitetstestet jämte informationen om etikettens innehåll och uppbyggnad sammanförs i en regel.
2. Reglerna skall formuleras enligt samma notation, i vilken de syntaktiska reglerna är uttryckta samt kunna utnyttja samma grammatiska operatorer, se (4).
3. Reglerna skall sammanföras i en morfologisk grammatik, helt i analogi med den syntaktiska grammatiken.
4. Namn på aktuell regel skall ingå i lexikoninformationen till resp suffix.
5. Reglerna skall initieras, d v s leda till generering av aktuell uppgift, i samband med återfinnandet av suffix-segmenten.
6. Reglerna skall verka från höger till vänster, d v s på aktuellt suffix jämte föregående segment.

Strategin fungerar, och jag har formulerat 3 morfologiska regler, som garanterar kompatibilitet mellan stamsegment, numerussegment, kasussegment och separator hos finska nomina. Kopplingen till de syntaktiska faciliteterna, d v s möjlighet att utnyttja de väldefinierade grammatiska operatorerna, har också gjort det möjligt att utan ytterligare utveckling av Kay-systemet kunna utföra kompatibilitetskontroll, som motiveras av förekomst av såväl stam- som suffixallomorfer.

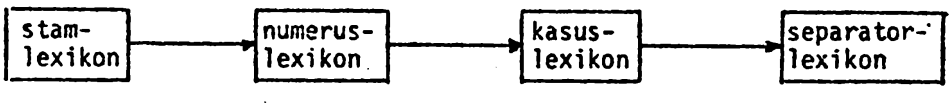
Låt oss gå tillbaka till fig 2. Här betraktas distinktionen mellan singularis och pluralis som en distinktion markerad/omarkerad, där pluralis är den markerade parten, d v s om inget uttryck för pluralis återfinns, så tolkas formen i fråga som singulär. Alternativt kan man vilja betrakta uttryck för numerus som obligatoriskt, där det singulära segmentet realiseras som ett noll-segment. Systemet erbjuder denna möjlighet. Då får man som alternativ till segmenteringen i fig 2 den segmentering som presenteras i fig 7.

fig 7



1. 'yskäisy' - '␣' - 'in'
2. 'yskäisy' - 'i' - 'n'

Som synes sker en omskrivning i ursprungscharten på så sätt att ett noll-segment explicit läggs in i charten som en edge (8 - 12 '␣'). Denna omskrivning sker selektivt, nämligen vid morfemgräns vars andra morfem kan realiseras som ett noll-segment. Motsvarande suffixlexikon upptar noll-segmentet med tillhörande tolkning. Med denna strategi kommer också kopplingen mellan de olika lexikonen att se något annorlunda ut, jfr fig 8 och fig 3.



Behovet av att kunna göra omskrivning ('rewriting') av ursprungscharten aktualiseras inte bara i samband med noll-segment utan även vad det gäller hela det problemkomplex som rör hantering av morfofonematiska växlingar, såvida man ej väljer att i sina lexikon explicit uttrycka alla varianter. Det senare alternativet tvingar oss att avstå från att

fånga upp intressanta generaliseringar i det språkliga materialet och därigenom skapa ett mindre insiktsfullt system, förutom att det leder till en väsentlig belastning på de olika lexikonen. För finskans del gäller det såväl fonologiskt som grammatiskt betingad morfofonematisk växling, nämligen volkalharmoni, vokalstrykning, kvantitativ resp kvalitativ stadieväxling, vokalförändring före suffix på -i och vokalassimilation. Hela detta problemkomplex kan spaltas upp i ett antal mindre frågor. Hur skall omskrivningsreglerna formuleras? Vilken alternant skall väljas som lexikonrepresentant? Hur skall reglerna integreras i den övriga bearbetningen? Kan och bör skillnaderna mellan de fonologiskt och de grammatiskt betingade morfofenomatiska växlingarna reflekteras genom principiellt olika behandling under analysen?

Av de språkliga fenomenen ovan har jag hittills endast bearbetat vokalharmonin. Den metod jag utarbetat ger möjlighet att behandla den finska vokalharmonin i icke sammansatta ord. Som arkisymboler för harmonivokalerna har jag valt /a,o,u/. Sålunda representeras t ex inessivmorfemet, med allomorferna 'ssa', 'ssä', av segmentet 'ssa' i kasuslexikonet.
Omskrivningen av 'ä' till 'a', 'ö' till 'o' samt 'u' till 'y' sker helt kontextfritt i samband med uppbyggandet av ursprungscharten, varvid ett speciellt 'bokstavslexikon' konsulteras. Erforderlig kompatibilitetskontroll sker med hjälp av den ovan skisserade metodiken. Stammar, som enbart innehåller de neutrala vokalerna 'e' och 'i', måste markeras som främre i lexikon. Övriga stammar innehåller ingen lexikoninformation om huruvida de är främre eller bakre, utan den informationen härleds under analysens gång.

Anm. För närmare information om status på projektet 'Chartanalys och finsk morfologi', vilket jag bedriver i samarbete med Erling Wande från Finsk-ugriska institutionen i Uppsala hänvisas till (5).

Referenser

1. Kay, M, Morfological and syntactic analysis, Lecture notes from the 3rd International Summer School of Computational and Mathematical Linguistics, Pisa 1974

2. Kay M, Syntactic Processing and Functional Sentence Perspective, Handbook from 1977 Nordic Summer School in Computational Linguistics

3. Sågvall Hein, A-L, An Approach to the Construction of a Text Comprehension System for X-ray Reports, Proc of the IFIP Working Conference on Computational Linguistics in Medicine, eds W Schneider, A-L Sågvall Hein, North-Holland Publ Comp 1977, pp. 91-99

4. Kay, M, Reversible Grammar, A Summary of the Formalism, Handbook from 1977 Nordic Summer School in Computational Linguistics

5. Analysresultat, lexikon, grammatiska regler och funktionsdefinitioner från projektet 'Chartanalys och finsk morfologi', Datalistor, UDAC, Uppsala