Systemisk ordning i informationsstruktur

- Laborationsrapport -

Eva Forsbom
evafo@stp.ling.uu.se
Uppsala universitet
Språkteknologiprogrammet
Datamaskinell diskurs
Lärare: Jennifer Spenader

16 juni 2000

Sammanfattning

Informationsstrukturen i en text hjälper till att visa vilken information som är i fokus för tillfället. Enligt vissa forskare finns det en språkspecifik systemisk ordning för fyllnads- och tilläggsled som motsvarar en omarkerad informationsstruktur. Om ett led flyttas till en annan plats så förändras dess kommunikativa dynamik. Jag har gjort ett program som beräknar den kommunikativa dynamiken, och testat det på några svenska meningar.

1 Bakgrund

Med hjälp av informationsstrukturen i en text kan författare signalera vilken information som för tillfället är i fokus. I språk med relativt fri ordföljd kan det t.ex. ske genom förändringar i ordföljden, medan man i språk som engelska, danska och svenska har mindre möjligheter till manipulering av ordföljden. Men det finns vissa tänkbara variationer.

Hajičová et al. [2] beskriver i en artikel om automatisk identifiering av tema-fokus[1] en ordning för fyllnads- och tilläggsled i engelskan som motsvarar en omarkerad informationsstruktur vad gäller kommunikativ dynamik (KD)[2], en s.k. systemisk ordning (SO). Teorin bygger på att ett led normalt har högre informationsvikt ju längre till höger det står i en mening, men att det finns en naturlig ordning för leden som i viss mån åsidosätter detta. I SO för engelskan kommer exempelvis ett objekt efter ett indirekt objekt:

1. [She/S] gave [several children/IO] [a few apples/O].
2. [She/S] gave [a few apples/O] to [several children/IO].

Det innebär att det första exemplet, som följer SO, har en omarkerad informationsstruktur där fokus ligger på verbet och båda komplementen, medan det andra exemplet, där objektet kommer före det indirekta objektet,

[1] Det motsvarar ungefär tema-rema.

[2] En skala som beskriver hur dynamiken utvecklas från det egentliga temat till det egentliga fokus.

har en markerad informationsstruktur där fokus enbart ligger på det indirekta objektet. Eftersom objektet flyttats åt vänster, där den informationsmässiga vikten är svagare, minskar dess KD så att det kommer att ingå i temat, medan det indirekta objektet, som flyttats åt höger, ökar sin KD och finns kvar i fokus.

Paggio [6] har tagit fram motsvarande SO för danska, och eftersom svenska och danska är såpass lika borde det gå att använda den även för svenska, möjligen med några justeringar.

1. TIME / VIEWPOINT / ATTITUDE
2. SUBJECT
3. INDIRECT OBJECT
4. OBJECT
5. SUBJECT COMPLEMENT / OBJECT COMPLEMENT
6. ORIGIN
7. DIRECTIONAL FROM
8. DIRECTIONAL TO
9. LOCATIVE
10. MANNER / MEANS
11. INSTRUMENT
12. TIME

Som synes finns TIME med två gånger, och det beror på att det verkar ge en lika omärkt informationsstruktur oavsett om man placerar ett tidsuttryck först eller sist i en mening.

Paggio anmärker att Hajičová et al. använder namn på semantiska roller för de olika leden, men att de i praktiken speglar mer ytnära grammatiska funktioner så att en ACTOR t.ex. alltid motsvarar det grammatiska subjektet[3]. Hon har därför valt att använda en mer traditionell grammatisk terminologi, och eftersom jag utgår från hennes beskrivning gör jag också det.

I denna rapport beskriver jag en implementering som beräknar den kommunikativa dynamiken för några givna, manipulerade, svenska meningar, enligt en föreslagen, grovt förenklad, algoritm.

2 Metod

De givna meningarna som skulle användas som indata till programmet var inte segmenterade och uppmärkta för olika fyllnads- och tilläggsled så det måste göras först, och det var inte helt problemfritt, vilket beskrivs närmare senare.

Programmet skulle sedan läsa in mening för mening, analysera dem och skriva ut resultatet av analysen. Analysen innebar att alla led först registrerades med det ordningsnummer de hade i den aktuella meningen. Därefter jämfördes varje led i nummerordning med de efterföljande ledens SO-nummer. Om det aktuella ledets SO-nummer var lägre än eller lika med alla efterföljande leds SO-nummer hade ledet en omärkt position. Var det däremot högre än något led så hade det aktuella ledet flyttats till vänster i meningen (eller det andra till höger) och deras kommunikativa nivåer uppdaterades nedåt respektive uppåt (jfr figur 1 på s. 3).

Eftersom leden jämfördes i den ordning de förekom i meningen skulle det kunna hända att ett led som förekom långt till höger i en välmatad mening

[3] I den algoritm som beskrivs av Hajičová et al. analyseras meningarna först enligt en dependensgrammatik och ges en underliggande ordföljd. De tar även hänsyn till om leden är kontextuellt bundna (t.ex. markerade med bestämd form) eller obundna, samt en systematisk semantisk klassifikation av verbet i föregående mening. I de exempelmeningar de tar upp i artikeln motsvarar de grammatiska rollerna de semantiska, så jag har ingen möjlighet att verifiera Paggios omdöme. Hon refererar dock till andra verk av författarna, där det säkert finns en mer detaljerad beskrivning.

Indata: En ledmärkt mening LM
Utdata: En mening KLM märkt med varje leds kommunikativa dynamik

1. Registrera varje led i den ordning de förekommer i LM
2. Gör följande för varje led L_i^{n-1}:
   (a) Gör följande för varje led $L_{k=i+1}^n$:
      i. Om $\text{so}(L_i) > \text{so}(L_k)$ så minska $\text{kd}(L_i)$ och öka $\text{kd}(L_k)$
3. Ange $\text{kd}(L)$ i KML för varje led L i LM

Figur 1: Algoritmen för beräkning av kommunikativ dynamik

fick sin nivå höjd flera gånger (och eventuellt sänkt en gång), men det som egentligen var intressant var om nivån var oförändrad (omärkt) eller hade förändrats uppåt eller nedåt (märkt). Därför hade skalan bara tre steg: 0, + eller -.

När varje leds kommunikativa nivå var uträknad skulle ledet märkas med den och meningen skrivas ut.

3 Implementering

Programmet är skrivet i Perl. Det tar en textfil med annoterade meningar, extraherar fyllnads- och tilläggsleden med hjälp av reguljära uttryck, jämför deras SO-nummer enligt en inbyggd SO (för ett mer generellt program skulle man kunna ha även SO som indata) och skriver ut texten igen med tillagda markeringar för KD.

3.1 Indata

Eftersom de givna meningarna[4] inte var analyserade måste det göras först (se figur 2 på s. 8). Det visade sig vara något komplicerat, vilket berodde på att det var autentiska meningar (och manipulationer av autentiska meningar) och därigenom oftast inte lika strömlinjeformade som de konstruerade meningar som Hajičová et al. och Paggio använde som illustration. I meningarna förekommer det också syntaktiska led som är svåra att kategorisera och där olika forskare har olika åsikter (om ens någon) om vilken kategori leden ska ha.

Det gäller exempelvis konstruktionen med ett formellt och ett egentligt subjekt (jfr mening 1a), som diskuterats av bl.a. Malmgren [5]. Varken Hajičová et al. eller Paggio har med någon sådan distinktion, så jag fick ingen ledning därifrån. Eftersom båda de refererade systemiska ordningarna bygger på grammatiska roller, och inte semantiska som vore naturligare, valde jag att märka båda som subjekt, vilket gör att deras KD inte påverkas alls enligt algoritmen. Men det egentliga/semantiska subjektet har ju egentligen flyttats åt höger för att få mer fokus (på objektsplats enligt Paul Diderichsens positionsgrammatik, beskrivet i Holm & Larsson [3]) och borde få högre KD, så det är ett ganska tveksamt val.

Än mer komplicerat var den emfatiska omskrivningen i mening 4a, där det semantiska objektet framhävts genom utbrytning och blivit till subjektskomplement enligt de refererade kategorierna. Som sådant får det i detta fall ingen förändring i KD, eftersom det inte kan förekomma något objekt före det (det var ju det semantiska ob-

[4] Ur *Pettson tältar* av Sven Nordqvist.

jektet som framhävdes). Men i andra fall kan ett på detta sätt framhävt led komma att få lägre KD, t.ex. vid ett utbrutet DIRECTIONAL TO:

(1) [Det/S] var [till Åmål/Dto] som hon åkte.

Det gör att KD-beräkningen av sådana meningar blir ganska missvisande. Ytterligare en svårighet med konstruktionen är att kategorisera det efterföljande ledet *som Pettson hade uppfunnit*[5]. Det är ju inte ett attribut till *honom*. Jörgensen & Svensson [4] menar att hela *en fiskebåge som Pettson hade uppfunnit* kan betraktas som "ett satsled", men att det inte utgör vare sig någon fras eller satsdel i egentlig bemärkelse". De har dock inget förslag på hur man ska kategorisera det. Thorell [9] menar att *en fiskebåge* är "ett starkt betonat predikativ"(motsvarande subjektskomplement i de refererade kategorierna) och att *som*-satsen ansluter sig till den i "en något oklar relation".

Båda tolkningarna anger att fokus på det utbrutna ledet ökar och antyder att fokus på det efterföljande ledet minskar. I vilket fall som helst kommer en sådan fokusförändring inte att komma fram i någon KD-beräkning utifrån SO. Jag valde ändå att märka *en fiskebåge* som subjektkomplement, men lät helt enkelt bli att märka det efterföljande ledet eftersom det är så oklart vad det ska kategoriseras som.

Det kan också vara intressant att ta i beaktande Prince's [7] iakttagelse att det finns en tendens att använda syntaktiska konstruktioner som gör att en nominalfras som har ett lågt värde på hennes bekanthetsskala, dvs. är relativt obekant, inte hamnar i subjektsposition.

Det kan också förklara varför konstruktionen med passiv i exempel 10a, där både *Sulitelma* och *vi* har ett högre bekanthetsvärde än *en flock vita vargar*, känns betydligt mer autentisk än exempel 10b. I det exemplet finns det ingen anledning att använda passivformen, eftersom den aktiva skulle ge samma ordning på de ingående leden.

Vid märkningen av den meningen saknade jag en kategori för agenten, så jag lät helt enkelt bli att märka den.

Ett annat problem var att meningarna innehöll flera satser och varken Hajičová et al. eller Paggio nämner något om hur man ska analysera sådana meningar[6]. Det går väl i och för sig att tänka sig att underordnade satser analyseras som ett led i den överordnade meningen, men hur gör man med samordnade satser och i synnerhet samordnade satser med elliptiska led?

Jag valde att analysera underordnade satser som ett enda led, av den anledningen att de oftast bakas in i en annan mening för att de inte har så hög informationsvikt, och då borde inte ett enskilt led i en sådan sats få högre eller lägre fokus, utan den kommunikativa nivån borde gälla hela satsen.

Samordnade satser analyserade jag var för sig. I annat fall hade den givna algoritmen gjort att exempelvis det andra subjektet fick den kommunikativa nivån för alla led till vänster om det att minska, vilket förmodligen inte var meningen. Det skulle i och för sig kunna vara så att man samordnar satser för att ge fokus åt den sista satsen, men det verkar snarare så att det finns två fokus, där det senare kanske är lite starkare.

[5] Hajičová et al. tar upp ett sådant exempel som illustration på att det efterföljande ledet hamnar i temat, men analyserar det inte vidare än så.

[6] Hajičová et al. skriver också att deras algoritm bara klarar av enkla meningar och bara verbkomplement.

3.2 Analys

De uppmärkta indata kördes sedan genom ett beräkningsprogram som först segmenterade texten och alla kategorier, lagrade dem i två synkroniserade listor som motsvarade deras ordning i texten, samt initierade ledens KD till 0. Beräkningen utfördes enligt algoritmen i figur 1 på sidan 3. (Koden finns i bilaga A.)

3.3 Utdata

Resultaten (se figur 3 s. 9) är vad de förväntas vara efter den tidigare diskussionen om svårigheterna att märka upp indata. Ett undantag är att inledande tidsled får sänkt KD, vilket kan bero på att jag valt en hashtabell som datastruktur för SO. Eftersom det finns två SO-nummer för tidskategorin vet programmet inte vilket det ska välja, utan väljer alltid det högsta värdet så att ledet får neutral KD om det står sist i meningen. Det skulle naturligtvis vara möjligt att ge programmet mer information att gå på (specialregler), eller bestämma sig för att bara använda ett värde. Jag är nämligen inte så säker på att Paggios tolkning att tidsledet har två KD-neutrala positioner är helt korrekt, utan att det bara är den meningsfinala positionen som är neutral.

4 Slutord

Eftersom den systemiska ordningen används tillsammans med annan information vid identifiering av tema-fokus och det program jag implementerat bara beräknar den kommunikativa dynamiken utifrån den systemiska ordningen är det svårt att säga något alls om resultatet. Det är helt enkelt för många bitar som saknas. Dessutom är det oklart hur meningarna ska segmenteras och leden kategoriseras, vilket gör det så gott som meningslöst att göra analysen. Det finns andra metoder och teorier som förklarar informationsstrukturfenomen och fokus bättre, t.ex. positionsscheman à la Diderichsen och ny-given informationà la Prince. För vissa enkla meningar och ordföljder är metoden dock tillräcklig.

Referenser

[1] Peter Cole, red. *Radical Pragmatics*. Academic Press, London, 1981.

[2] Eva Hajičová, Hana Skoumalová och Petr Sgall. An Automatic Procedure for Topic-Focus Identification. *Computational Linguistics*, 21(1):81–94, Mars 1995.

[3] Lars Holm och Kent Larsson. *Svenska meningar*. Ord och stil 7, Språkvårdssamfundets skrifter. Studentlitteratur, Lund, 2a uppl., 1980.

[4] Nils Jörgensen och Jan Svensson. *Nusvensk grammatik*. Liber, Malmö, 1987.

[5] Sven-Göran Malmgren. Är det egentliga subjektet egentligen subjekt? I Teleman [8], s. 56–62.

[6] Patrizia Paggio. *The Treatment of Information Structure in Machine Translation*. Doktorsavhandling, Center for Sprogteknologi, Njalsgade 80, DK-2300 Copenhagen S, Denmark, Oktober 1997. (http://www.cst.ku.dk/patrizia/PhD/-thesis.ps.gz).

[7] Ellen F. Prince. Toward a Taxonomy og Given–New Information. I Cole [1], s. 223–255.

[8] Ulf Teleman, red. *Grammatik på villovägar*. Skrifter utgivna av Svenska språknämnden 73. Esselte Studium, Solna, 1987.

[9] Olof Thorell. *Svensk grammatik*. Esselte Studium, Solna, 2a uppl., 1977.

A Kod för beräkning av KD

```
#!/usr/local/bin/perl -w
##################################################
# so:          An annotator for systemic ordering
# usage:       so < <file>
# created:     2000-06-09 evafo@stpling.uu.se
###############################################

# PROGRAM
# limit the usage of variables
use strict;

# store the systemic ordering in a hash table
# there can be alternative complements or adjuncts in one position,
# signalled by the same value in the table
# T = TIME
# V = VIEWPOINT
# A = ATTITUDE
# S = SUBJECT
# IO = INDIRECT OBJECT
# Ob = OBJECT
# SC = SUBJECT_COMPLEMENT
# OC = OBJECT_COMPLEMENT
# Or = ORIGIN
# Dfrom = DIRECTIONAL FROM
# Dto = DIRECTIONAL TO
# L = LOCATIVE
# Ma = MANNER
# Me = MEANS
# I = INSTRUMENT
my %so = ("T" => 0, "V" => 0, "A" => 0, "S" => 1,
          "IO" => 2, "Ob" => 3, "SC" => 4, "OC" => 4,
          "Or" => 5, "Dfrom" => 6, "Dto" => 7, "L" => 8,
          "Ma" => 9,"Me" => 9, "I" => 9, "T" => 10);

# while a line from the input file can be read
while (<>) {

    # store actual text (synchronized with @order and @cd)
    my @text = ();

    # store actual ordering
    my @order = ();

    # store communicative dynamism
    my @cd = ();

    do {

        # for all tags, extract the text: $1, and the tag: $2
        # store them in two synchronized lists

        $_ =~ s/([^\/]+\/)([^\]]+)(\].*)/$1$2$3/;
        if ($2) {
            push @text, $1;
            push @order, $2;
            push @cd, '0';
            $_ = $3;
        }

        # store whatever is left last in the text list

        else {
            push @text, $_;
            $_ = "";
        }
    }
    while ($_ ne "");

    # for each tag, compare its systemic ordering to
    # the systemic ordering of all tags to its right

    my $i = 0;
```

```
    my $j = 0;
    for ( ; $i < (scalar @order) - 1; $i++) {
        for ($j = $i + 1; $j < scalar @order; $j++) {
            if ($so{$order[$i]} > $so{$order[$j]}) {
                if ($cd[$i] eq '+') {
                    $cd[$i] = '0';
                }
                else {
                    $cd[$i] = '-';
                }
                if ($cd[$j] eq '-') {
                    $cd[$j] = '0';
                }
                else {
                    $cd[$j] = '+';
                }
            }
        }
        print "$text[$i]$order[$i].$cd[$i]";
    }

    if (scalar @text > 1) {
        print "$text[$i]$order[$i].$cd[$i]";
        print $text[(scalar @text) -1];
    }
#    elsif (scalar @text < 1) {
#        print "\n";
#    }
    else {
        print $text[0];
    }
}
```

1. (a) [Det/S] ligger [ett hoprullat tält/S] [inne i påsen/L].
   (b) [Inne i påsen/L] ligger [ett hoprullat tält/S].
2. (a) [En dag/T] var [gubben Pettson/S] [uppe på sin vind/L] och letade efter [en påse med flöten/Ob].
   (b) [Uppe på sin vind/L] var [gubben Pettson/S] [en dag/T] och letade efter [en påse med flöten/Ob].
   (c) [En påse med flöten/Ob] letade [gubben Pettson/S] efter [uppe på sin vind/L] [en dag/T].
3. (a) [Pettson/S] gick [till hönsgården/Dto] och ropade på [dem/Ob].
   (b) [Till hönsgården/Dto] gick [Pettson/S] och ropade på [dem/Ob].
4. (a) [Det/S] var [en fiskebåge/SC] som Pettson hade uppfunnit.
   (b) [Pettson/S] hade uppfunnit [en fiskebåge/Ob].
5. (a) [Kroken och flötet/S] satt fast [på en pil/L].
   (b) [På en pil/L] satt [kroken och flötet/S] fast.
6. (a) [Med pilbågen/I] kunde [han/S] skjuta iväg [pilen med kroken/Ob] [långt ut i vattnet/Dto].
   (b) [Långt ut i vattnet/Dto] kunde [han/S] skjuta iväg [pilen med kroken/Ob] [med pilbågen/I].
   (c) [Han/S] kunde skjuta iväg [långt ut i vattnet/Dto] [med pilbågen/I] [pilen med kroken/Ob].
7. (a) [När de kom tillbaka/T] hade [hönorna/S] [mycket riktigt/A] tröttnat på [att vara i tältet/Ob].
   (b) [Hönorna/S] hade [mycket riktigt/A] tröttnat på [att vara i tältet/Ob] [när de kom tillbaka/T].
   (c) [Att vara i tältet/Ob] hade [hönorna/S] [mycket riktigt/A] tröttnat på [när de kom tillbaka/T].
8. (a) [Findus/S] kröp ner [i sovsäcken/Dto] och [Pettson/S] satte [sig/Ob] [bredvid/L].
   (b) [I sovsäcken/Dto] kröp [Findus/S] ner och [bredvid/L] satte [sig/Ob] [Pettson/S].
9. (a) Men [det/S] var [ett litet tält/S] och [han/S] satt inte [särskilt bekvämt/Ma].
   (b) Men [det/S] var [ett litet tält/S] och [särskilt bekvämt/Ma] satt [han/S] inte.
10. (a) [Vid Sulitelma/L] blev [vi/S] [jagade/SC] av en flock vita vargar, så [vi/S] gick [vilse/Ma].
    (b) En flock vita vargar blev [vi/S] [jagade/SC] av [vid Sulitelma/L], så [vi/S] gick [vilse/Ma].

Figur 2: Segmenterade och uppmärkta led i indatameningarna

1. (a) [Det/S.0] ligger [ett hoprullat tält/S.0] [inne i påsen/L.0].
   (b) [Inne i påsen/L.-] ligger [ett hoprullat tält/S.+].
2. (a) [En dag/T.-] var [gubben Pettson/S.+] [uppe på sin vind/L.+] och letade efter [en påse med flöten/Ob.0].
   (b) [Uppe på sin vind/L.-] var [gubben Pettson/S.+] [en dag/T.0] och letade efter [en påse med flöten/Ob.0].
   (c) [En påse med flöten/Ob.-] letade [gubben Pettson/S.+] efter [uppe på sin vind/L.0] [en dag/T.0].
3. (a) [Pettson/S.0] gick [till hönsgården/Dto.0] och ropade på [dem/Ob.0].
   (b) [Till hönsgården/Dto.-] gick [Pettson/S.+] och ropade på [dem/Ob.0].
4. (a) [Det/S.0] var [en fiskebåge/SC.0] som Pettson hade uppfunnit.
   (b) [Pettson/S.0] hade uppfunnit [en fiskebåge/Ob.0].
5. (a) [Kroken och flötet/S.0] satt fast [på en pil/L.0].
   (b) [På en pil/L.-] satt [kroken och flötet/S.+] fast.
6. (a) [Med pilbågen/I.-] kunde [han/S.+] skjuta iväg [pilen med kroken/Ob.+] [långt ut i vattnet/Dto.+].
   (b) [Långt ut i vattnet/Dto.-] kunde [han/S.+] skjuta iväg [pilen med kroken/Ob.+] [med pilbågen/I.0].
   (c) [Han/S.0] kunde skjuta iväg [långt ut i vattnet/Dto.-] [med pilbågen/I.-] [pilen med kroken/Ob.+].
7. (a) [När de kom tillbaka/T.-] hade [hönorna/S.0] [mycket riktigt/A.+] tröttnat på [att vara i tältet/Ob.+].
   (b) [Hönorna/S.-] hade [mycket riktigt/A.+] tröttnat på [att vara i tältet/Ob.0] [när de kom tillbaka/T.0].
   (c) [Att vara i tältet/Ob.-] hade [hönorna/S.0] [mycket riktigt/A.+] tröttnat på [när de kom tillbaka/T.0].
8. (a) [Findus/S.0] kröp ner [i sovsäcken/Dto.0] och [Pettson/S.0] satte [sig/Ob.0] [bredvid/L.0].
   (b) [I sovsäcken/Dto.-] kröp [Findus/S.+] ner och [bredvid/L.-] satte [sig/Ob.0] [Pettson/S.+].
9. (a) Men [det/S.0] var [ett litet tält/S.0] och [han/S.0] satt inte [särskilt bekvämt/Ma.0].
   (b) Men [det/S.0] var [ett litet tält/S.0] och [särskilt bekvämt/Ma.-] satt [han/S.+] inte.
10. (a) [Vid Sulitelma/L.-] blev [vi/S.+] [jagade/SC.+] av en flock vita vargar, så [vi/S.0] gick [vilse/Ma.0].
    (b) En flock vita vargar blev [vi/S.0] [jagade/SC.0] av [vid Sulitelma/L.0], så [vi/S.0] gick [vilse/Ma.0].

Figur 3: Segmenterade och uppmärkta led i indatameningarna