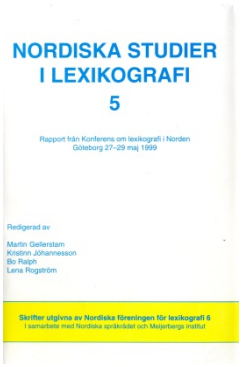
NORDISKE STUDIER I LEKSIKOGRAFI

Titel: Morfologisk analys och disambiguering som stöd i skapandet av frekvensordlistor

Forfatter: Fredrik Westerlund, Sjur Nørstebø Moshagen, Eva Grava og Juhani Birn

Kilde: Nordiska Studier i Lexikografi 5, 2001, s. 407-424
Rapport från Konferens om lexikografi i Norden, Göteborg 27.-29. maj 1999

URL: http://ojs.statsbiblioteket.dk/index.php/nsil/issue/archive



Søgbarhed
Artiklerne i de ældre Nordiske studier i leksikografi (1-5) er skannet og OCR-behandlet. OCR står for ’optical character recognition’ og kan ved tegngenkendelse konvertere et billede til tekst. Dermed kan man søge i teksten. Imidlertid kan der opstå fejl i tegngenkendelsen, og når man søger på fx navne, skal man være forberedt på at søgningen ikke er 100 % pålidelig.

Fredrik Westerlund
Sjur Nørstebø Moshagen
Eva Orava
Juhani Birn

Morfologisk analys och disambiguering som stöd i skapandet av frekvensordlistor

When applying morphological analysis (TWOL) and disambiguation (Constraint Grammar) to a text material, the linguist gets an opportunity to search for contextually disambiguated words on the basis of their base forms. A search on a specific base form gives a result that also includes all the inflected forms and some common derivations of the word. This way of assembling inflected words can offer a lot of help in text lemmatization and also in various types of frequency listings. This paper contains and discusses some frequency lists, which have been generated out of a morphologically analysed and disambiguated annual volume of the newspaper Göteborgs-Posten.

De frekvensordböcker som har kommit ut i Norden de senaste decennierna skiljer sig märkbart från varandra såväl till omfånget som till sin inre uppbyggnad. Materialet i de nordiska frekvensordböckerna är till stora delar baserat på tidningstext som har bearbetats och sammanställts i frekvenslistningar av olika slag: i form av listor med de mest frekventa orden i initial- och finalalfabetisk ordning, de mest frekventa orden fördelade på olika ordklasser, eller i form av listor på de ord som bara förekommer en gång i materialet, för att nämna några sorteringssätt i ordböckerna.

De kvantitativa uppgifterna som ges för varje ord varierar likaså allt efter på vilket sätt författarna presenterar sitt material och hur djupgående analysen av orden är. Vissa frekvensordböcker listar orden enbart enligt deras grafiska form, dvs. i den form de uppträder i textmaterialet. Till dessa ordböcker hör *Norsk frekvensordbok – De 10 000 vanligste ord fra norske aviser* (Heggstad 1982), *Dansk Frekvensordbog* (Bergenholtz 1992)

och *Hyppige Ord i Danske Aviser, Ugeblade og Fagblade* (Maegaard & Ruus 1986).

Nusvensk frekvensordbok baserad på tidningstext (del 1–3 Allén 1970–75, del 4 Allén et al. 1980) är den största frekvensordboken på den nordiska marknaden. I den finns materialet även lemmatiserat, dvs. alla böjningsformer av ett visst ord har förts upp under ifrågavarande ords grundform, såsom även textmaterialet i *Nynorsk frekvensordbok* (Vestbøstad 1989) och *Íslensk orðtíðnibók* (Pind et al. 1991). Också i frekvensordboken *Tiotusen i topp* (Allén 1972), som huvudsakligen bygger på de två första volymerna av *Nusvensk frekvensordbok*, har böjningsformerna av ett ord sammanförts under något som i boken benämns "lexikaliska ord".

I denna artikel redogör vi för hur morfologisk analys bland annat vid lemmatisering kan vara till stor hjälp i framställandet av material till frekvensordböcker och vid frekvenslistningar överlag.

1. Morfologisk analys och disambiguering

TWOL (Koskenniemi 1983) är en morfologisk analysmodell som Lingsoft Ab har tillämpat på flera språk. En TWOL-analysator förser varje ord i en text med en eller flera morfologiska analyser, beroende på om ordet är lexikalt entydigt eller flertydigt. En morfologisk analys består av en grundform och ett antal koder som upplyser om ordets ordklass samt, beroende på ordklass, bl.a. om genus, species, numerus, kasus, diates och tempus. Analysatorn markerar också sammansättningsgränserna i ordet. Modellens tillämpning på svenska, SWETWOL (Karlsson 1992), har vidareutvecklats på Lingsoft.

Vid tillämpning av restriktionsgrammatiken Constraint Grammar, CG, disambigueras de olika morfologiska analysmöjligheterna av ett ord så att i idealfallet endast en analys av ordet kvarstår. (För en fullständig presentation av CG-formalismen se Karlsson et al. 1995.) I detta sammanhang är det SWECG, disambigueraren för svensk text, som är aktuell, se Birn 1998.

Om man har tillämpat den morfologiska analysen på ett textmaterial är det möjligt att söka ord på deras grundformer, och med en enda sökning samtidigt få fram belägg på alla böjningsformer av ett ord. Lemmatiseringen av ett textmaterial förlöper på detta sätt betydligt mer smidigt än om man vore tvungen att söka varje böjningsform separat, och man slipper också den redundanta information som sökning med trunkerade ord ger.

För att med exempel kunna belysa hur morfologisk analys och disambiguering av en text kan vara till stöd vid frekvenslistningar har vi tillämpat morfologianalysatorn och restriktionsgrammatiken på årgång 1997 av Göteborgs-Posten, totalt drygt 20 miljoner ord fördelade på knappt 94 000 artiklar. Tidningsmaterialet finns följaktligen i en form som tillåter en mångsidig frekvensanalys med möjligheter till olika slag av frekvenslistningar. De små frekvenslistor som ingår i denna artikel är av utrymmesskäl begränsade till de tio ord som befinner sig på toppen, dvs. de tio mest frekventa orden i respektive exempelgrupp. Det kan i vissa fall anses vara i minsta laget, men med tanke på materialets omfång i förhållande till det tillgängliga utrymmet är det motiverat att begränsa listan. Om flera ord i slutet av listan har samma frekvens tas alla ord med ifrågavarande frekvenstal med. Varje exempelord anförs i den form som den morfologiska analysen ger det, dvs. som en grundform, vilken åtföljs av koder som berättar i vilken form ordet uppträder i texten.

De ord som den morfologiska analysatorn inte känt igen har uteslutits ur frekvensberäkningarna. Det rör sig till en stor del om felstavningar och ord på främmande språk. Siffror ingår inte heller i undersökningen. En automatisk analys av en text innebär alltid att det i resultatet finns en liten felmarginal som endast en manuell genomgång kunde förhindra. Förskjutningar i frekvensvärden beror dels på att disambigueraren i vissa fall lämnar kvar flera än en analys, dels på att den inte helt kan undvika att den korrekta analysen slopas. I en undersökning av en text på 10 600 ord klarade disambigueraren av att till 96,59 % ge orden en entydig analys, medan en andel på 0,43 % inte hade kvar den korrekta analysen (Birn 1998).

2. Frekvensen för ett visst lemma

Som ett första exempel kan vi ta fasta på frekvensen för ett visst ord och alla böjningsformer av ordet. Den morfologiska analysen är speciellt välkommen när det exempelvis gäller sökning av verb som genomgår vokalväxling vid konjugation, eftersom det vanligen skulle krävas flera sökningar för att man skall kunna komma åt alla böjningsformer av ordet. Valet av exempelord är i detta fall slumpartat såtillvida att vi gjort två uppslag i tolfte upplagan av *Svenska Akademiens ordlista* och på respektive uppslag letat oss fram till det första verbet med vokalväxling i böjningen. Metoden gav verben *snyta* och *gråta*. Med en sökning på vartdera ordet fås frekvensen på alla böjningsformer samt vissa avledningar av de båda verben i Göteborgs-Posten 1997.

TABELL 1. *Lemmat* snyta.

9	"snyta" V ACT PRES
7	"snyta" V ACT PAST
5	"snyta" V ACT INF
1	"snyta" V PASS PRES
1	"snyta" V ACT SUPINE
1	"snyta" V ACT IMP
1	"snyta" DER/-ning N UTR INDEF SG NOM
1	"snyta" DER/-ning N UTR INDEF PL NOM

TABELL 2. *Lemmat* gråta.

196	"gråta" V ACT PAST
174	"gråta" V ACT INF
147	"gråta" V ACT PRES
55	"gråta" <V/DER> <PCP1> A UTR/NEU DEF/INDEF SG/PL NOM
29	"gråta" DER/-nde N NEU INDEF SG NOM
26	"gråta" V ACT SUPINE
6	"gråta" <V/DER> <PCP2> A UTR INDEF SG NOM
4	"gråta" V ACT IMP
2	"gråta" DER/-nde N NEU DEF SG NOM
1	"gråta" V PASS PAST

Eftersom analysatorn i hög grad kan skilja på homografer är det också möjligt att med en sökning få fram t.ex. alla former av ord som har samma grafiska grundform, men som hänför sig till

skilda kategorier. Till denna grupp räknas t.ex. ordet *resa*, som ju både är ett verb och ett substantiv.

TABELL 3. *Substantivet respektive verbet* resa.

```
1174  "resa" N UTR INDEF SG NOM
889   "resa" V ACT INF
```

Det kan även vara svårt, eller åtminstone tidsödande, att i en text få tag i samtliga former av ord där efterleden utgörs av ett visst lemma men där förleden varierar. En morfologiskt analyserad text är betydligt mer lättbearbetad. En sökning ger vid handen att t.ex. ord med former av verbet *bryta* som sista led oftast uppträder i form av ordet *utbröt* (271 ggr), följt av olika former av verbet *avbryta*. Först på femtonde plats i frekvenslistan kommer följande nya lemma i form av perfekt particip-formen *uppbruten*. Participen (PCP) klassificeras som (verbavledda) adjektiv (A) på basis av böjning och syntaktisk funktion.

TABELL 4. *Lemman på* -bryta.

```
271  "utbryta" V ACT PAST
201  "avbryta" V ACT INF
146  "avbryta" V PASS PAST
123  "avbryta" V ACT PAST
83   "avbryta" V ACT PRES
65   "avbryta" V PASS INF
59   "avbryta" <V/DER> <PCP2> A UTR INDEF SG NOM
49   "utbryta" V ACT PRES
49   "avbryta" V PASS PRES
42   "avbryta" <V/DER> <PCP2> A UTR/NEU DEF/INDEF PL NOM
     /---/
20   "uppbryta" <V/DER> <PCP2> A UTR INDEF SG NOM
```

3. Ord i sammansättningar

Frekvensen för ett visst enskilt ord i en text kan intressera en språkvetare, men även frekvensen för det aktuella ordet i eventuella sammansättningar, samt vilken position det har i det sammansatta ordet, kan vara av intresse. Med hjälp av den morfolo-

giska analysen kan man söka efter ett ord och få veta vilken ställning det har, dvs. hur ofta det förekommer initialt, medialt respektive finalt.

Ordet *medium* är i detta sammanhang ett intressant exempelord, med tanke på att det ofta uppträder i varierande former. Som förled i substantivsammansättningar förekommer ordet 467 ggr i den språkriktiga formen *medie-* och 250 ggr som *media-*. I medial ställning förekommer ordet 84 ggr i formen *-medie-*, och i final ställning flest gånger i form av *-media* (542 ggr) och endast hälften så många gånger i form av efterleden *-medium* eller någon böjd form av ordet.

Då man t.ex. tar en titt på de frekventa orden *medie-* och *mediaföretag* märker man att den form som rekommenderas av språkvården, *medieföretag*, förekommer mer än dubbelt fler gånger jämfört med den andra formen. De språkriktiga formerna (den första kategorin i figur 5) har allmänt taget också ett stort försprång, trots att variationen mellan de båda formerna i materialet är bred.

TABELL 5. *Ordet* medium *som förled i sammansatta ord.*

22 "medie_värld" N UTR DEF SG NOM
18 "medie_företag" N NEU INDEF SG/PL NOM
13 "medie_koncern" N UTR DEF SG NOM
12 "medie_våld" N NEU DEF SG NOM
12 "medie_program" N NEU DEF SG NOM
11 "medie_våld" N NEU INDEF SG NOM
11 "medie_marknad" N UTR DEF SG NOM
11 "medie_företag" N NEU DEF SG NOM
9 "medie_område" N NEU DEF SG NOM
9 "medie_bransch" N UTR DEF SG NOM
/--/

9 "media_program" N NEU DEF SG NOM
8 "media_företag" N NEU INDEF SG/PL NOM
8 "media_dag" <TIME> <DAY> N UTR DEF PL NOM
7 "media_bild" N UTR DEF SG NOM
6 "media_våld" N NEU INDEF SG NOM
6 "media_värld" N UTR DEF SG NOM
6 "media_magasin" N NEU DEF SG NOM
5 "media_uppbåd" N NEU DEF SG NOM
5 "media_teknik" N UTR DEF SG NOM
5 "media_program" N NEU INDEF SG/PL NOM

5 "media_område" N NEU DEF SG NOM
5 "media_mogul" N UTR INDEF SG NOM
5 "media_bransch" N UTR DEF SG NOM
5 "media_bolag" N NEU INDEF SG/PL NOM
/---/

TABELL 6. *Ordet* medium *som mellersta led i sammansatta ord.*

6 "mass_medie_koncentration" N UTR INDEF SG NOM
6 "mass_medie_forskning" <RETAIN> N UTR INDEF SG NOM
4 "mass_medie_uppbåd" N NEU DEF SG NOM
4 "mass_medie_ekonomi" N UTR INDEF SG NOM
3 "mass_medie_uppbåd" N NEU INDEF SG/PL NOM
3 "mass_medie_forskare" N UTR DEF SG NOM
2 "multi_medie_teknik" N UTR INDEF SG NOM
2 "multi_medie_projekt" N NEU INDEF SG/PL NOM
2 "multi_medie_företag" N NEU INDEF SG/PL NOM
2 "mass_medie_intresse" N NEU DEF SG NOM
2 "eter_medie_område" N NEU DEF SG NOM
/---/

TABELL 7. *Ordet* medium *som efterled i sammansatta ord.*

351 "massmedia" N NEU DEF/INDEF PL NOM
102 "multi_media" N UTR/NEU DEF/INDEF PL NOM
100 "multi_media" N UTR INDEF SG NOM
95 "multi_media" N NEU DEF/INDEF PL NOM
29 "massmedia" N NEU DEF/INDEF PL GEN
18 "info_media" N UTR/NEU DEF/INDEF PL NOM
18 "info_media" N UTR INDEF SG NOM
18 "info_media" N NEU DEF/INDEF PL NOM
16 "eter_media" N UTR/NEU DEF/INDEF PL NOM
16 "eter_media" N UTR INDEF SG NOM
16 "eter_media" N NEU DEF/INDEF PL NOM
/---/

84 "mass_medium" N NEU DEF PL NOM
62 "mass_medium" N NEU INDEF PL NOM
13 "mass_medium" N NEU DEF PL GEN
12 "nyhets_medium" N NEU INDEF PL NOM
7 "nyhets_medium" N NEU DEF PL NOM
7 "film_medium" N NEU DEF SG GEN
6 "mass_medium" N NEU INDEF SG NOM
5 "eter_medium" N NEU INDEF PL NOM
4 "orts_medium" N NEU INDEF PL NOM
4 "köld_medium" N NEU INDEF PL NOM
4 "film_medium" N NEU DEF SG NOM
/---/

Ordet *medium* förekommer som väntat som efterled flest gånger i former av *massmedia* och *multimedia*. Analysatorn anger grundformen *-media* för de ord som i texten står i former som slutar på *-media* (den första kategorin i tabell 7 nedan). De ord som däremot står i en böjd form, t.ex. *massmediernas* och *filmmediet*, får slutleden *-medium* som grundform (den andra kategorin i tabell 7). Som efterled tycks *-media* vinna överlägset över *-medium* vad frekvensen beträffar.

Tack vare att ordens sammansättningsgränser markeras kan man också undersöka hur ofta ord som har samma form, sånär som på en bokstav, förekommer t.ex. som mellersta led i ett sammansatt ord. Nedan listas de sammansatta ord vars mellersta led utgörs av en konsonant plus bokstavskombinationen *and*.

TABELL 8. *Ord på konsonant* + and *som mellersta led i sammansatta ord.*

```
57 "folk_tand_vård" N UTR DEF SG NOM
23 "privat_tand_läkare" N UTR INDEF SG/PL NOM
19 "dam_hand_boll" N UTR INDEF SG NOM
14 "video_band_spelare" N UTR INDEF SG/PL NOM
 8 "studie_hand_ledning" N UTR INDEF SG NOM
 8 "klubb_hand_boll" N UTR INDEF SG NOM
 7 "över_tand_läkare" N UTR INDEF SG/PL NOM
 7 "privat_tand_läkare" N UTR DEF PL NOM
 7 "kassett_band_spelare" N UTR INDEF SG/PL NOM
 7 "folk_tand_vård" N UTR INDEF SG NOM
 7 "dörr_hand_tag" N NEU DEF SG NOM
   /--/
```

4. Frekvensen för en viss böjningsform

Eftersom orden i det analyserade materialet är kodade enligt den form de står i är det även möjligt att ta fram frekvensen för en viss böjningsform. Nedan följer en titt på verben, substantiven och adjektiven i tidningstexten.

4.1. Verb

För verbens del får man med hjälp av den morfologiska analysatorn uppgifter om att Göteborgs-Posten år 1997 innehöll flest verbformer i presens, hela 1 576 546 st. (46,6 %), därefter i infinitiv, 625 840 st. (18,5 %). Preteritumformerna kommer på tredje plats, 558 453 st. (16,5 %), och supinumformerna på fjärde plats, 248 091 (7,3 %). Ovanstående uppgifter gäller verb i aktiv form. Verbet *vara* toppar inte oväntat alla böjningskategorier. Olika hjälpverb figurerar också flitigt bland de mest frekventa verben.

TABELL 9. *Verb: aktiv presens.*

315546	"vara" <COP> V ACT PRES
201493	"ha" <AUX> <SUPINE> V ACT PRES
70389	"kunna" <AUX> V ACT PRES
58889	"säga" <VCOM> V ACT PRES
43353	"ska" <AUX> V ACT PRES
40452	"komma" <ATT-INF> V ACT PRES
37279	"få" V ACT PRES
34335	"skulle" <AUX> V ACT PRES
33264	"bli" <COP> V ACT PRES
29895	"vilja" <AUX> V ACT PRES
	/---/

TABELL 10. *Verb: aktiv infinitiv.*

33386	"vara" <COP> V ACT INF
28080	"få" V ACT INF
27983	"ha" <AUX> <SUPINE> V ACT INF
21153	"bli" <COP> V ACT INF
18873	"ta" V ACT INF
16142	"göra" V ACT INF
12599	"kunna" <AUX> V ACT INF
12363	"gå" V ACT INF
11289	"se" <VCOG> V ACT INF
8832	"komma" <ATT-INF> V ACT INF
	/---/

TABELL 11. *Verb: aktiv preteritum.*

94097	"vara" <COP> V ACT PAST
35778	"ha" <AUX> <SUPINE> V ACT PAST
25439	"få" <AUX> V ACT PAST
23528	"bli" <COP> V ACT PAST
15702	"komma" <ATT-INF> V ACT PAST
11946	"kunna" <AUX> V ACT PAST
11251	"gå" V ACT PAST
10075	"göra" V ACT PAST
9575	"ta" V ACT PAST
9201	"säga" <VCOM> V ACT PAST
	/---/

TABELL 12. *Verb: aktiv supinum.*

19108	"vara" <COP> V ACT SUPINE
14793	"få" V ACT SUPINE
9275	"bliva" <COP> V ACT SUPINE
8361	"göra" V ACT SUPINE
6899	"ha" <AUX> <SUPINE> V ACT SUPINE
5879	"gå" V ACT SUPINE
5050	"ta" V ACT SUPINE
4515	"komma" <ATT-INF> V ACT SUPINE
3618	"se" <VCOG> V ACT SUPINE
2697	"kunna" <AUX> V ACT SUPINE
	/---/

Även bland verben i passiv är det presensformerna som är allmännast, 111 769 st. (3,3 %). Materialets preteritumformer i passiv är 66 344 till antalet (2 %), dvs. något fler än passiv infinitivformerna, 64 355 st. (1,9 %). Antalet verb böjda i passiv supinum är 36 712 till antalet (1,1 %).

TABELL 13. *Verb: passiv presens.*

2603	"behöva" <INF> V PASS PRES
2445	"kräva" V PASS PRES
1724	"vänta" V PASS PRES
1627	"använda" V PASS PRES
1578	"tvinga" V PASS PRES
1496	"ta" V PASS PRES
1496	"kalla" V PASS PRES
1366	"beräkna" V PASS PRES
1345	"anse" V PASS PRES
1306	"göra" V PASS PRES
	/---/

TABELL 14. *Verb: passiv preteritum.*

1366	"tvinga" V PASS PAST
1194	"gripa" V PASS PAST
1172	"skada" V PASS PAST
1152	"göra" V PASS PAST
1114	"ta" V PASS PAST
985	"döma" V PASS PAST
924	"föra" V PASS PAST
822	"hitta" V PASS PAST
705	"bilda" V PASS PAST
670	"drabba" V PASS PAST
	/---/

TABELL 15. *Verb: passiv infinitiv.*

1852	"använda" V PASS INF
1421	"göra" V PASS INF
1325	"ta" V PASS INF
911	"bygga" V PASS INF
869	"se" <VCOG> V PASS INF
731	"tvinga" V PASS INF
658	"utveckla" V PASS INF
657	"lägga" V PASS INF
625	"hålla" V PASS INF
575	"genom_föra" V PASS INF
	/---/

TABELL 16. *Verb: passiv supinum.*

1375	"göra" V PASS SUPINE
882	"drabba" V PASS SUPINE
695	"ta" V PASS SUPINE
659	"tvinga" V PASS SUPINE
476	"tilldela" V PASS SUPINE
471	"använda" V PASS SUPINE
441	"utveckla" V PASS SUPINE
421	"döma" V PASS SUPINE
403	"utsätta" V PASS SUPINE
396	"utse" V PASS SUPINE
	/---/

Förutom verben i aktiv och passiv förekommer det 70 141 deponensverb i materialet. I spetsen står presensformen *finns* med frekvenstalet 37 136.

4.2. Substantiv

Med hjälp av analysatorn får man för substantivens del veta att 1 545 363 substantiv (25,8 %) står i grundform, dvs. som indefinita nominativer i singular, medan motsvarande siffra för substantiven i plural är 653 417 (10,9 %). SWETWOL ger den underspecificerade analysen SG/PL för de ord vilka kan kombineras med såväl singulara som plurala framförställda bestämningar (t.ex. *hus* som i *ett/några hus*). Frekvensen för de orden presenteras i en separat lista (se tabell 19).

Av den första gruppen substantiv nedan kan det nämnas att 8 002 st. (av drygt 1,5 miljoner) står som första led i uttryck av typen *lager- och kontorsfastighet.*

TABELL 17. *Substantiv: indefinit singular nominativ.*

15051	"del" N UTR INDEF SG NOM
12072	"plats" N UTR INDEF SG NOM
11895	"tid" N UTR INDEF SG NOM
9189	"gång" N UTR INDEF SG NOM
7719	"man" N UTR INDEF SG NOM
6796	"dag" <TIME> <DAY> N UTR INDEF SG NOM
6431	"publik" N UTR INDEF SG NOM
5978	"kommun" N UTR INDEF SG NOM
5817	"väg" N UTR INDEF SG NOM
5633	"hand" N UTR INDEF SG NOM
	/--/

TABELL 18. *Substantiv: indefinit plural nominativ.*

32591	"krona" N UTR INDEF PL NOM
19782	"miljon" <NUM> N UTR INDEF PL NOM
9917	"människa" N UTR INDEF PL NOM
9410	"peng" N UTR INDEF PL NOM
8272	"person" N UTR INDEF PL NOM
5718	"kvinna" <FEM> N UTR INDEF PL NOM
5405	"miljard" <NUM> N UTR INDEF PL NOM
4991	"dag" <TIME> <DAY> N UTR INDEF PL NOM
4858	"gång" N UTR INDEF PL NOM
4356	"man" <MASC> N UTR INDEF PL NOM
	/--/

TABELL 19. *Substantiv: indefinit singular/plural nominativ.*

59045	"år" <TIME> N NEU INDEF SG/PL NOM
20635	"procent" <RETAIN> N UTR INDEF SG/PL NOM
11295	"barn" <FamRel> N NEU INDEF SG/PL NOM
7829	"exempel" N NEU INDEF SG/PL NOM
6670	"sätt" N NEU INDEF SG/PL NOM
6499	"problem" N NEU INDEF SG/PL NOM
6491	"par" N NEU INDEF SG/PL NOM
6228	"ordförande" N UTR INDEF SG/PL NOM
6172	"mål" N NEU INDEF SG/PL NOM
5997	"jobb" N NEU INDEF SG/PL NOM
	/---/

Då man undersöker samma böjningskategorier bland de substantiv som är böjda i genitiv erhåller man följande frekvensuppgifter: 45 458 substantiv (0,76 %) står i indefinit form genitiv singular (2 241 är första led i uttryck av typen *tävlings- och äventyrsprogram*) och 8 865 substantiv (0,15 %) står i indefinit form genitiv plural.

Som jämförelsematerial till de nyss presenterade substantiven i obestämd form kan man ställa upp tidningstextens substantiv i bestämd form. I singular är de definita substantiven i nominativ 1 265 375 (21,1 %) till antalet, i plural 324 604 (5,4 %).

TABELL 20. *Substantiv: definit singular nominativ.*

10363	"år" <TIME> N NEU DEF SG NOM
9070	"polis" N UTR DEF SG NOM
8521	"regering" N UTR DEF SG NOM
7472	"kommun" N UTR DEF SG NOM
7299	"tid" N UTR DEF SG NOM
7094	"fråga" N UTR DEF SG NOM
6583	"man" <MASC> N UTR DEF SG NOM
5675	"land" N NEU DEF SG NOM
5443	"ställe" N NEU DEF SG NOM
5308	"sida" N UTR DEF SG NOM
	/---/

TABELL 21. *Substantiv: definit plural nominativ.*

5634	"år" <TIME> N NEU DEF PL NOM
4164	"barn" <FamRel> N NEU DEF PL NOM
3331	"peng" N UTR DEF PL NOM
2487	"politiker" N UTR DEF PL NOM
2345	"elev" N UTR DEF PL NOM
2275	"social_demokrat" N UTR DEF PL NOM
2084	"man" <MASC> N UTR DEF PL NOM
2078	"kommun" N UTR DEF PL NOM
1706	"moderat" N UTR DEF PL NOM
1661	"företag" N NEU DEF PL NOM
	/---/

För genitivens del ger en sökning i det morfologiskt analyserade materialet följande siffror: 113 257 substantiv (1,9 %) står i definit form genitiv singular och 23 872 substantiv (0,4 %) i definit form genitiv plural. Så här ter sig de tio mest frekventa i respektive grupp:

TABELL 22. *Substantiv: definit singular genitiv.*

4778	"dag" <TIME> <DAY> N UTR DEF SG GEN
4776	"år" <TIME> N NEU DEF SG GEN
2959	"land" N NEU DEF SG GEN
2781	"värld" N UTR DEF SG GEN
2423	"kommun" N UTR DEF SG GEN
2053	"regering" N UTR DEF SG GEN
1792	"gårdag" N UTR DEF SG GEN
1355	"stat" N UTR DEF SG GEN
1071	"stad" N UTR DEF SG GEN
1028	"polis" N UTR DEF SG GEN
	/---/

TABELL 23. *Substantiv: definit plural genitiv.*

784	"barn" <FamRel> N NEU DEF PL GEN
506	"social_demokrat" N UTR DEF PL GEN
469	"moderat" N UTR DEF PL GEN
467	"år" <TIME> N NEU DEF PL GEN
320	"kommun" N UTR DEF PL GEN
264	"företag" N NEU DEF PL GEN
257	"elev" N UTR DEF PL GEN
230	"politiker" N UTR DEF PL GEN
228	"förälder" <FamRel> N UTR DEF PL GEN
224	"dam" N UTR DEF PL GEN
	/---/

I de ovanstående frekvensuppställningarna är inte personnamn inkluderade, eftersom morfologianalysatorn inte klassificerar dem som vare sig definita eller indefinita. Det framgår dock på annat vis att 5,4 % av alla substantiv i Göteborgs-Posten 1997 utgörs av förnamn i nominativ och 0,12 % av förnamn i genitiv. De förnamn vars grafiska form i nominativ är identisk med den i genitiv (t.ex. *Niklas* och *Ann-Lis*) utgör 1 % av alla substantiv i tidningstexten. Tio i topp-listan på förnamnsfronten innehåller enbart mansnamn. Det första kvinnliga förnamnet, *Anna*, kommer först på sjuttonde plats.

TABELL 24. *Förnamn.*

```
8867  "*peter" <*c> <FIRST_NAME> <MASC> N ? SG NOM
8682  "*anders" <*c> <SWE> <MASC> <FIRST_NAME> <HUMAN>
      N ? SG NOM/GEN
7013  "*lars" <*c> <SWE> <MASC> <FIRST_NAME> <HUMAN> N ?
      SG NOM/GEN
5842  "*magnus" <*c> <SWE> <MASC> <FIRST_NAME>
      <HUMAN> N ? SG NOM/GEN
5829  "*jan" <*c> <SWE> <FIRST_NAME> <MASC> N ? SG NOM
5669  "*göran" <*c> <SWE> <FIRST_NAME> <MASC> N ? SG NOM
5541  "*stefan" <*c> <SWE> <FIRST_NAME> <MASC> N ? SG
      NOM
5189  "*thomas" <*c> <SWE> <MASC> <FIRST_NAME> <HUMAN>
      N ? SG NOM/GEN
4995  "*johan" <*c> <SWE> <FIRST_NAME> <MASC> N ? SG NOM
4383  "*hans" <*c> <MASC> <FIRST_NAME> <HUMAN> N ? SG
      NOM/GEN
      /---/
3624  "*anna" <*c> <FIRST_NAME> <FEM> N ? SG NOM
      /---/
```

På samma vis som den morfologiska analysen kan underlätta sökandet efter frekvensen på en viss böjd form i ett material kan den också få fram antalet förekomster av ord som har en viss avledd form.

4.3. Adjektiv

Det finns totalt drygt 1,77 miljoner adjektiv i tidningstexten. Den största gruppen (45,8 %) utgörs av adjektiv som står i indefinit singular nominativ. Adjektiven i plural nominativ, som har iden-

tisk definit och indefinit form, utgör 20,4 %. Den tredje största gruppen är de definita adjektiven i nominativ singular med 16,4 %. Adjektiven i genitivform står för en marginell del av alla adjektiv. Det kan också nämnas att 5,2 % av adjektiven står i komparativ och 3,6 % i superlativ.

TABELL 25. *Adjektiv: indefinit singular nominativ.*

```
11899  "stor" A UTR INDEF SG NOM
10971  "ny" A UTR INDEF SG NOM
9075   "samtidig" A NEU INDEF SG NOM
7374   "svår" A NEU INDEF SG NOM
7052   "egen" A UTR INDEF SG NOM
6314   "klar" A NEU INDEF SG NOM
6131   "stor" A NEU INDEF SG NOM
5864   "lång" A NEU INDEF SG NOM
5675   "svensk" A UTR INDEF SG NOM
5509   "ny" A NEU INDEF SG NOM
       /---/
```

TABELL 26. *Adjektiv: definit singular nominativ.*

```
12523  "ny" A UTR/NEU DEF SG NOM
11631  "svensk" A UTR/NEU DEF SG NOM
6869   "stor" A UTR/NEU DEF SG NOM
4858   "god" <SUP> A UTR/NEU DEF SG NOM
4743   "stor" <SUP> A UTR/NEU DEF SG NOM
3950   "liten" A UTR/NEU DEF SG NOM
2251   "politisk" A UTR/NEU DEF SG NOM
2108   "hög" <SUP> A UTR/NEU DEF SG NOM
2072   "internationell" A UTR/NEU DEF SG NOM
2053   "närmast" <SUP> A UTR-MASC DEF SG NOM
       /---/
```

TABELL 27. *Adjektiv: plural nominativ.*

```
13439  "ny" A UTR/NEU DEF/INDEF PL NOM
11711  "stor" A UTR/NEU DEF/INDEF PL NOM
11461  "olik" A UTR/NEU DEF/INDEF PL NOM
8211   "svensk" A UTR/NEU DEF/INDEF PL NOM
4601   "anställa" <V/DER> <PCP2> A UTR/NEU DEF/INDEF PL NOM
4314   "små" A UTR/NEU DEF/INDEF PL NOM
4308   "viss" A UTR/NEU DEF/INDEF PL NOM
4112   "egen" A UTR/NEU DEF/INDEF PL NOM
3810   "ung" A UTR/NEU DEF/INDEF PL NOM
3226   "god" <SUP> A UTR/NEU DEF/INDEF PL NOM
       /---/
```

5. Slutord

Sätten på vilka man kan frekvenslista ett textmaterial i elektronisk form är många, och den morfologiska analysen möjliggör ytterligare flera sorteringssätt. Med årgång 1997 av Göteborgs-Posten som grund utföll frekvensnedslagen på detta vis – om det är representativt för tidningstext eller dagens svenska i skrift överlag är svårt att avgöra, men vissa språkligt allmängiltiga tendenser kan säkert urskiljas.

Litteratur

A. ORDBÖCKER

Dansk Frekvensordbog. Baseret på danske romaner, ugeblade og aviser 1987–1990. 1992. Red. Henning Bergenholtz. København.

Hyppige Ord i Danske Aviser, Ugeblade og Fagblade. 1986. Red. B. Maegaard & H. Ruus. København.

Íslensk orðtíðnibók. 1991. Red. Jörgen Pind et al. Orðabók Háskólans.

Norsk frekvensordbok. De 10 000 vanligste ord fra norske aviser. 1982. Red. Kolbjørn Heggstad. Universitetsforlaget.

Nusvensk frekvensordbok baserad på tidningstext. Del 1–3. 1970-1975. Red. Sture Allén. Stockholm.

Nusvensk frekvensordbok baserad på tidningstext. Del 4. 1980. Red. Sture Allén et al. Stockholm.

Nynorsk frekvensordbok. Dei vanlegaste orda i skriftleg nynorsk. Hente frå aviser, sakprosa og romanar, 1978–84. 1989. Red. Per Vestbøstad. Bergen.

Svenska Akademiens ordlista över svenska språket. Tolfte upplagan. 1998. Svenska Akademien.

Tiotusen i topp. Ordfrekvenser i tidningstext. 1972. Red. Sture Allén. Stockholm.

B. Annan litteratur

Birn, Juhani. 1998. "Swedish Constraint Grammar: A Short Presentation." Tillgänglig på http://www.lingsoft.fi/doc/swecg/.

Karlsson, Fred. 1992. "SWETWOL: A Comprehensive Morphological Analyser for Swedish." *Nordic Journal of Linguistics* 15:1–45.

Karlsson, Fred et al. 1995. *Constraint Grammar. A Language–Independent System for Parsing Unrestricted text.* Berlin and New York: Mouton de Gruyter.

Koskenniemi, Kimmo. 1983. *Two-level Morphology. A General Computational Model for Word-form Recognition and Production.* Department of General Linguistics, University of Helsinki, Publication No. 11.