STEFÁN BRIEM

Automatisk morfologisk analyse af islandsk tekst

Abstract

Automatic Morphological Analysis of Icelandic Text

One of the projects worked on at the Institute of Lexicography at the University of Iceland is a frequency analysis of Icelandic vocabulary and grammar. The most time-consuming part of the work consists in morphological analysis of text samples containing in all more than half a million running words. For every single word the analysis results in registration of the word class, the flexion form and the lemma to which the text word belongs.

If manually performed, this kind of analysis would be enormously monotonic work requiring high precision. A method has been developed to perform the analysis to a great extend automatically using a computer. However the manual work can not be eliminated, but it has already been reduced significantly, and at the same time the character of the manual work is altered to be mainly a matter of correcting activity.

The method of automatic analysis is based on a corpus of tags and word forms originating in a previous manually performed analysis of more than 54,000 text words and on a set of rules for possible relations between words of the same sentence. On the basis of frequencies compiled by the previous analysis and of points given by the rules when fulfilled, a computer program automatically selects the probably 'best sentence' among a (usually great) number of homograph sentences. Furthermore, in case of words not found in the collection of word forms, the program makes use of a collection of more than 5,000 back parts of word forms in order to make an intelligent guess.

At the current stage the result of the automatic analysis is completely correct for about 70% of the text words and partly correct for about 15%. Our experience shows that the manual effort is reduced by about 2/3. By extension of the word form collection and by improvement of the relation rules and points giving, a significant improvement of the automatic analysis is expected in the near future, e.g. leading to 85%–90% of text words being correctly analysed.

Proceedings of NODALIDA 1989, pages 3-13

1 Indledning

Automatisk morfologisk analyse af islandsk tekst hører til en af Leksikografisk Instituts opgaver. Det drejer sig om en fase af et større projekt, som er udarbejdelse af en islandsk frekvensordbog. Bogen skal give oplysninger om brugen af islandsk nutidssprog, det skriftlige sprog, i form af forskellige slags oversigter over hyppigheden af ord, ordformer, bøjningsformer, ordklasser o.s.v.

Det kan give en idé om projektets omfang at det vil omfatte ca. en halv million tekstord. Langt den største del af arbejdet ligger i den morfologiske analyse, som bliver særlig omhyggeligt udført. I den sidste ende vil analysen blive manuelt udført eller i hvert fald manuelt kontrolleret. Men det har allerede vist sig at arbejdet reduceres i betydelig grad ved at man i første omgang udfører analysen automatisk ved hjælp af en datamat.

2 Formål

Formålet med den morfologiske analyse uanset om den bliver udført manuelt eller maskinstøttet er følgende.

For det første skal analysen for hvert enkelt tekstords vedkommende føre til registrering af ordklasse og bøjningsform.

Desuden registrerer man hvilket kasus verber og præpositioner styrer. De fleste ord, som man plejer at klassificere som præpositioner, bruges i nogen tilfælde som adverbier, og omvendt, mange ord, som traditionelt klassificeres som adverbier, bruges også som præpositioner. I dette projekt har man derfor valgt at behandle præpositioner og adverbier under ét, hvilket medfører, at man også registrerer adverbiers kasusstyrelse, når den forekommer.

Endvidere registrerer man for hvert tekstord det tilhørende leksikonsord, også kaldt lemma.

Som et eksempel tager vi følgende korte tekst og det tilstræbte resultat af dens analyse:

TEKST:

```
Það er þriðjudagur í dag. Magnús kemur á
morgun. Hann dvaldist ásamt dr. Jósteini
Samúelssyni alllengi í Danmörku. Félagi
hans hefur orðið eftir.
```

ANALYSE:

f p h e n	það	það
s f g 3 e n	er	vera
n k e n	þriðjudagur	þriðjudagur
a o	í	í
n k e o	dag	dagur
n k e n s	Magnús	Magnús
s f g 3 e n	kemur	koma
a o	á	á
n k e o	morgun	morgunn
f p k e n	hann	hann
s f m 3 e þ	dvaldist	dvelja
a þ	ásamt	ásamt
n k e þ	dr.	dr.
n k e þ s	Jósteini	Jósteinn
n k e þ s	Samúelssyni	Samúelsson
a a	alllengi	alllengi
a þ	í	í
n v e þ s	Danmörku	Danmörk
n k e n	félagi	félagi
f p k e e	hans	hann
s f g 3 e n	hefur	hafa
s s g	orðið	verða
a a	eftir	eftir

I den midterste kolonne har vi tekstordene, ét i hver linie, og til højre de tilhørende leksikonsord. Til venstre har vi så tegn for de grammatiske oplysninger som analysen har ført til. Det første tegn står for ordklasse. For hvert tekstord er der højst 7 grammatiske tegn i en bestemt rækkefølge. Den kan man kalde en grammatisk streng. En blank linie betegner begyndelsen af en ny sætning.

3 Den automatiske analyse

Den metode, som man her benytter til automatisk morfologisk analyse, er baseret på en tidligere manuelt udført analyse af godt 54.000 tekstord (Friðrik Magnússon 1988) og fungerer ved hjælp af et sæt af morfologiske regler og ved betragtning og vurdering af sandsynligheder.

Hovedanalysen udføres med én sætning ad gangen. Det må derfor være helt klart, hvor hver sætning begynder. For at opnå det udføres der en foranalyse som har til formål at registrere startfeltet for hver sætning af den tekst, som skal analyseres.

3.1 Foranalyse

Foranalysen udføres halvautomatisk, d.v.s. i samarbejde mellem menneske og maskin.

TEKST:

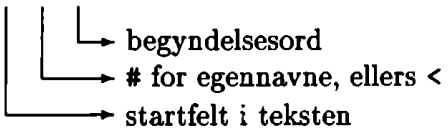
```
Það er þriðjudagur í dag. Magnús kemur á
morgun. Hann dvaldist ásamt dr. Jósteini
Samúelssyni alllengi í Danmörku. Félagi
hans hefur orðið eftir.
```

BEGYNDELSESORD:

```
< Það
< Hann
< Í
< Ég
< Og
< En
< Þá
< Hún
< Þegar
< Við
< Á
< Þar
< Þetta
< Nú
< Um
< Þú
< Að
< Ekki
< Hér
< Af
......
......
```

POINTERREGISTER:

```
   0 < Það
  26 # Magnús
  49 < Hann
 115 < Félagi
```



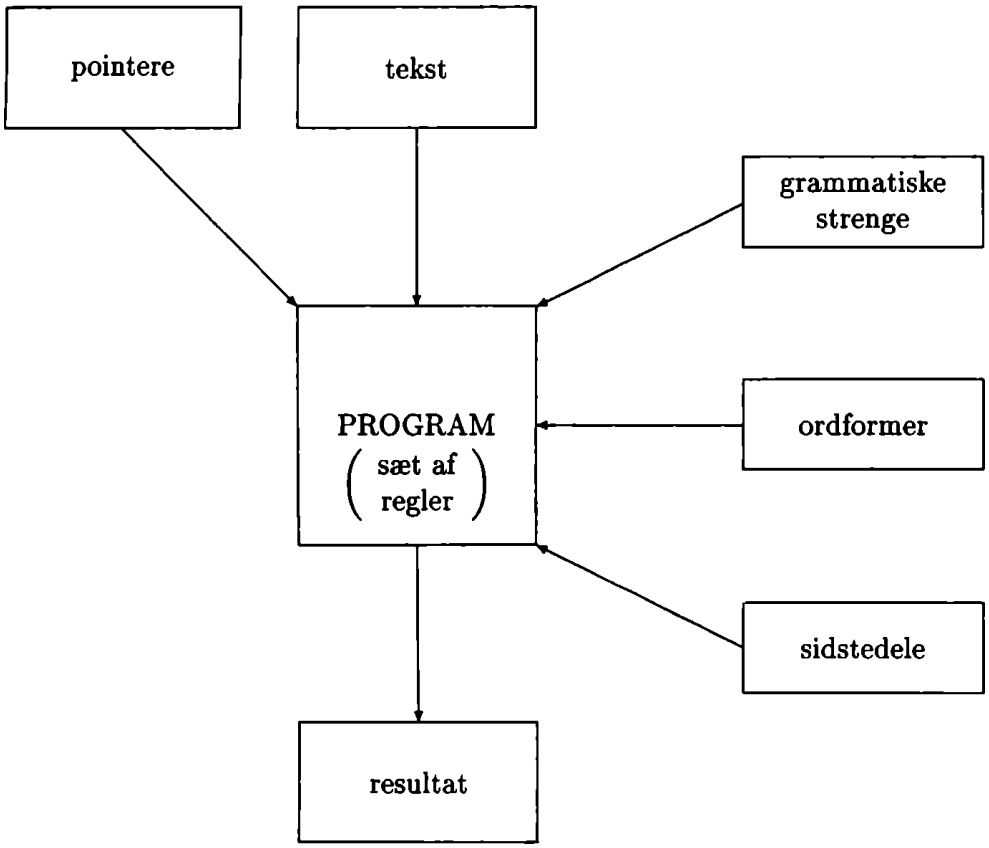
Her har vi den samme tekst igen. Til højre er vist de første 20 ord i en samling af 649 begyndelsesord, ordnet efter hyppighed. Disse 20 ord er altså de hyppigste i begyndelsen af islandske sætninger, i hvert fald i den tekst som samlingen er baseret på. Det program, som udfører foranalysen, bruger denne samling til automatisk at finde ud af de fleste sætningsbegyndelser i en tekst. Når der er tvivl stopper programmet op, inviterer mennesket til en afgørelse og fortsætter når en beslutning er truffet.

I dette eksempel må et menneske træde til for at afgøre, at Magnús er et proprium og at Jósteinn ikke er begyndelsen af en ny sætning.

Resultatet, et pointerregister, bruges så sammen med teksten under hovedanalysen, den egentlige automatiske morfologiske analyse, som vi nu går over til at betragte.

3.2 Den automatiske hovedanalyse

SYSTEMETS BESTANDDELE



Dette billede viser systemets bestanddele. I centrum er et program, hvori der er indbygget et sæt af 75 regler, især morfologiske regler. Programmet gør brug af en samling af ordformer og en samling af sidstedele af ord. Det er programmets opgave for hvert enkelt ord i en given tekst at finde frem til den mest sandsynlige streng blandt de 566 forskellige grammatiske strenge, samt at finde det tilhørende leksikonsord.

Vi vil nu betragte de enkelte bestanddele nærmere.

3.2.1 Støtteregistre

Lad os begynde med de 3 støtteregistre som programmet bruger uændrede fra tekst til tekst.

Det første støtteregister indeholder 566 grammatiske strenge i alfabetisk rækkefølge. Tallene er hyppighedstal.

GRAMMATISKE STRENGE:

```
  434 a a
   27 a a e
   34 a a m
   17 a e
   40 a o
   13 a u
.....
.....
   25 n v f þ g
    1 n v f þ g s
   24 s b g 2 e n
    1 s b g 2 f n
   57 s f g 1 e n
.....
.....
```

Det andet støtteregister indeholder knap 18.000 ordformer i alfabetisk rækkefølge sammen med det tilhørende leksikonsord og længst til venstre nummeret på en grammatisk streng. Hver ordform kan optræde mange gange p.g.a. forskelle i leksikonsord og forskellige bøjningsformer. Tallene i anden kolonne er hyppighedstal fra den tidligere analyse.

ORDFORMER:

```
.....
.....
  421    1  vélstjórum   vélstjóri
  453    5  vélum        vél
  455    3  vélunum      vél
  116    4  vér          ég
  116  171  við          ég
    0   52  við          við
    4  498  við          við
    6   27  við          við
  484    1  viða         viða
  215    1  viðamesta    viðamikill
  208    1  viðamikið    viðamikill
.....
.....
```

Og det tredje støtteregister indeholder godt 5.000 sidstedele af ordformer i baglæns alfabetisk rækkefølge, hver sammen med den tilsvarende sidstedel af et leksikonsord og nummeret på en grammatisk streng. Tallene i anden kolonne er sandsynlighedstal som man kan knytte til hver sidstedel for at styre programmets anvendelse af dette register.

SIDSTEDELE:

```
  377   0  trjáa      tré
  287   0  ædda       ældur
  311   0  ædda       ældur
   41   0  alda       öld
  287   1  falda      faldur
  311   1  falda      faldur
  377   0  halda      hald
  377   0  elda       eldi
.............
.............
  377   0  efla       efli
  429   0  regla      regla
  377   0  bila       bil
  373   0  heimila    heimili
  377   0  heimila    heimili
  287   0  mikla      mikill
  311   0  mikla      mikill
  409   0  jökla      jökull
  417   0  jökla      jökull
  377   0  falla      fall
  377   0  fjalla     fjall
  409   0  valla      völlur
.............
.............
```

3.2.2 Sæt af regler

Lad os nu betragte det regelsæt som er direkte indbygget i programmet. Til hver regel er der knyttet et antal points, som gives hver gang reglen er opfyldt.

Programmet arbejder med én sætning ad gangen. Det slår tekstordene op i samlingen af ordformer. I de fleste tilfælde findes der nogle muligheder for hvert tekstord. Det kan føre til et stort antal af homografe sætninger, der bliver kandidater til stillingen 'den bedste sætning'. Hvis tekstordene betragtes hver for sig uanset deres stilling i sætningen, er det umuligt at afgøre hvilken af de homografe sætninger er den rigtige eller den bedste.

Men her træder reglerne til. For hver enkelt af de homografe sætninger kalkulerer programmet det totale antal points som reglerne giver hver gang de er opfyldt. Den sætning, som får de fleste points, anses for at være den bedste og bliver valgt som analysens resultat.

I praksis kan antallet af homografe sætninger blive så enormt, at det ville tage datamaten måneder eller endog år at betragte dem alle. Men da antallet er kendt tidligt under processen, løses dette problem nemt ved optimalisering af kun en del af sætningen ad gangen. Det medfører en nedsættelse af analysens korrekthed på kun ca. 1%.

De fleste af reglerne er meget enkle og kan derfor nemt omskrives til et programmeringssprog. Her er vist nogen få af reglerne i dansk oversættelse. Tallene er de tilhørende points.

- Hvis der eftir et faldstyrende adverbium følger et ikke faldstyrende adverbium og derpå følger et faldbøjet ord, så vil det faldbøjede ord stå i det fald som det første adverbium styrer. **400**

- Det er sandsynligt at *að* er en konjunktion, hvis et ukendt ord følger efter. **500**

- Hvis der efter et adjektiv følger et substantiv, så har de næsten altid samme køn, tal og fald. **1000**

- Hvis der efter et adjektiv følger et substantiv i bekendt form, så er der større sandsynlighed for at adjektivet har bestemt form end ubestemt. **10**

- Et verbum i perfektparticipium er sandsynligt, hvis det følgende ord er verbet *vera* eller hvis verbet *vera* er et af de to foranstående ord. **200**

Kongruensbøjning er en af de stærkeste støttepiller for den automatiske analyse. Et eksempel på kongruensbøjning har vi her i tredje regel i tilfælde af substantiv og tilhørende adjektiv. Andre analoge regler for kongruensbøjning omfatter også pronominer og talord.

3.2.3 Resultat

Korrektheden af den automatiske analyse fremgår af sammenligning mellem den på næste side anførte og den tidligere viste korrekte analyse af samme tekst. I dette eksempel har programmet bl.a. taget fejl af tekstordet **félagi**. Ifølge den automatiske analyse skulle det være singularis dativ af substantivet *félag* som betyder *forening*, men i virkeligheden drejer det sig om singularis nominativ af substantivet *félagi* som betyder *kammerat*. Lidt senere får vi flere eksempler på homografi.

AUTOMATISK ANALYSE:

f p h e n	það	það
s f g 3 e n	er	vera
n k e n #	þriðjudagur	þriðjudagur
a o	í	í
n k e o	dag	dagur
n k e n s	Magnús	Magnús
s f g 3 e n	kemur	koma
a o	á	á
n k e o	morgun	morgunn
f p k e n	hann	hann
	dvaldist	dvaldist
a þ	ásamt	ásamt
	dr.	dr.
n k e þ s #	Jósteini	Jósteinn
n k e þ s #	Samúelssyni	Samúelssonur
a a	alllengi	alllengi
a þ	í	í
n v e þ s	Danmörku	Danmörk
n h e þ	félagi	félag
f p k e e	hans	hann
s f g 3 e n	hefur	hafa
s s g	orðið	verða
a þ	eftir	eftir

betyder at analysen er baseret på ordets sidstedel.

3.3 Præstation og kvalitet

Resultatet af analysen af en prøvetekst på 5.000 ord blev:

Præstation:

Maskin/menneske:	Foranalyse	2– 5	min.
Maskin:	Hovedanalyse	15–20	min.
Menneske:	Korrektur	20	timer

Kvalitet:

70%	tekstord korrekt analyseret
15%	tekstord ukorrekt analyseret
15%	tekstord slet ikke analyseret

Foranalysen tager næsten ingen tid. Hovedanalysen tager heller ikke lang tid og det er jo datamatens tid. Størstedelen af arbejdstiden er stadigvæk den menneskelige arbejdstid som kræves til korrekturlæsning af den automatiske analyses resultat.

Kvaliteten af den automatiske analyse på det nuværende stadium er vist her i procenter.

3.4 Problemer

De største problemer hidrører fra følgende tre faktorer:

1. Mange ukendte tekstord
2. Uregelmæssig interpunktion
3. Homografer blandt hyppige ordformer

Ukendte ord medfører at det bliver svært at analysere de nærmest liggende ord korrekt og præcist.

På grund af uregelmæssig brug af interpunktion i islandsk har man i den automatiske analyse helt set bort fra interpunktionen, selv om den selvfølgelig i mange tilfælde kunne give værdifulde oplysninger.

Jeg vil nu give et par eksempler på homografer blandt hyppige ordformer, som tit bliver fejlagtigt analyseret under den automatiske analyse.

ordform	*leksikonsord*		*dansk*
við	*við*	pron. pers. 1. p. pl. nom.	*vi*
	við	præp./adv.	*ved*
	(*viður*	sb. masc. sg. akk.	*ved*)

Ordformen **við** har tre helt forskellige meninger. De mest almindelige er personligt pronomen og præposition. Den tredje er her sat i parenteser, fordi den er ikke nær så hyppig som de andre; og den er faktisk slet ikke med i samlingen af ordformer.

ordform	*leksikonsord*		*dansk*
orðið	*verða*	vb. p.p./sup.	*blevet*
	orð	sb. neutr. sg. nom./akk. bek.	*ordet*
orðin	*verða*	vb. p.p.	*blevet*
	orð	sb. neutr. pl. nom./akk. bek.	*ordene*

De to andre ordformer, **orðið** og **orðin** har hver for sig to helt forskellige meninger.

Disse bestemte homografer udgør måske ikke ret store problemer. Men der kræves i hvert fald mere præcise regler end de hidtidige til at skelne i mellem dem.

3.5 Forbedringer

Til slut skal vi betragte de muligheder der gives for at forbedre den automatiske analyse.

- Større korpus, d.v.s. flere ordformer og flere sidstedele
- Flere og mere præcise regler
- Præcisering af pointsgivning
- Udnyttelse af interpunktion

Den forbedring som man venter at opnå uden større besvær skulle føre til 85%–90% korrekt analyse.

Litteratur

Magnússon, Friðrik. 1988. Hvað er títt? Jón Hilmar Jónsson [ed.]. I *Orð og tunga 1*: 1–49. Orðabók Háskólans. Reykjavík.