Alex Speed Kjeldsen
Notater til Menotas lemmatiseringsmøde
Oslo, 29–30 maj 2006

1. Lidt om projektet og om hvorfor jeg ikke benyttede MLA

Jeg er ved at afslutte lemmatiseringen af det islandske kongesagahåndskrift Morkinskinna der indeholder knap 100.000 ord. Godt 20.000 af disse (hånd B's del) havde jeg allerede lemmatiseret i mit konferensspeciale. Som en del af mit ph.d.-projekt har jeg nu næsten afsluttet den resterende del af hs. (hånd A). Jeg mangler i første omgang at kigge nærmere på visse afsnit som i første omgang blev oversprunget pga. deres sværtlæselighed, og derefter skal jeg rette en del inkonsekvenser og fejl. Alt i alt er lemmatiseringen imidlertid kommet ganske langt.

Jeg kan ikke sige så meget om MLA da jeg kun har meget begrænsede erfaringer med den (jeg har blot prøvet den nogle få gange kort inde i den første tekstfase). Jeg kan derimod sige en smule om min egen fremgangsmåde, og hvorfor jeg ikke valgte at bruge MLA på trods af de mange gode elementer i denne. Dette kan forhåbentlig være med til at belyse hvilke elementer jeg mener man med fordel kunne overveje at integrere i MLA.

Den absolut væsentligste grund til at jeg ikke benyttede MLA var at min transskription af hånd A kun forelå på facs-niveau. Jeg ville i videst mulig udstrækning slå genereringen af dipl-niveau, norm-niveau, lemmatisering og morfosyntaktisk opmærkning sammen i én fase, hvilket ikke umiddelbart var muligt i MLA.

For at mindske det manuelle lemmatiseringsarbejde mente jeg også at der i en eller anden udstrækning måtte integreres en form for disambiguering i den halvautomatiske lemmatiseringsprocess -- i hvert fald på et basalt niveau. Så vidt jeg forstod var -- og er -- dette ikke understøttet af MLA.

2. Fremgangsmåden

Udgangspunktet for den halvautomatiske lemmatisering af MorkA var de 20% af håndskriftet som MorkB havde skrevet, og som jeg allerede havde lemmatiseret for et par år siden. Jeg lavede en "lemmatiseringsbase" der indeholdt alle ordmanifestationer i MorkB (ved homografer blev de mindre frekvente poster sorteret fra). Denne base forøgedes ved at variantformer genereredes (fx 'or'/'oƨ', 'ſſ'/'s', 'ɑ'/'o'/'ǫ'). Den viden om MorkA's ortografi som jeg havde opnået i forbindelse med transskriberingen trak jeg naturligvis på ved genereringen af variantformer

Den udvide lemmatiseringsbase blev så brugt som inputfil for lemmatiseringen af MorkA. Jeg skrev et simpelt script som sammenlignede facs-formerne i MorkA med facs-formerne i lemmatiseringsbasen. Hvis der var overensstemmelse på facs-niveau, overførte scriptet alle andre oplysninger fra lemmatiseringsbasen (dipl-niveau, lemma og POS. Efter at scriptet havde kørt igennem hele filen havde det kommet med et bud på dipl- og norm-niveau, lemma og pos i godt 70% af alle ordene.

Jeg gik herefter i gang med at kontrollere computerens bud og rette en hel del af disse (jeg havde lavet en mængde makroer der gjorde at manglende grammatiske oplysninger hurtigt kunne indføres uden for megen indtastning). Når nye lemmata og POS-oplysninger blev indtastet for en given ordform, sørgede et script for at dette blev gennemført for samtlige identiske ordformer i resten af filen.

For at lette korrekturarbejdet lavede jeg vha en lang række grep-baserede søg&erstat en simpel form for kontekstafhængig disambiguering. Nogle eksempler på dette kunne være (de kunne sagtens finpudses):

(NB: Jeg har i min opmærkning anvendt lfg. struktur:
<f>ƿyrſta</f><d>ƿyrſta</d><n>fyrsta</n><w l="fyrstr" pos="xTO gN nS cA sD"></w>
Ved en simpel grep-baseret søg&erstat ændres dette senere så det er i overensstemmelse med Menotas retningslinjer).

1) Erstatte verbalformern "er" med relativpartiklen, når der efter denne følger et adverbium OG et finit verbum:
(<w l=")vera(" pos="x)VB fF tPS mIN p3 nS vA iST5("></
w>.*?\r.*?xAV.*?\r.*?xVB fF.*?)
-> \1er\2RP\3

2) Erstatte verbalformern "er" med relativpartiklen, når der efter denne følger en præp. OG et finit verbum og samtidig ændre præpositionen til et adverbium:
(<w l=")vera(" pos="x)VB fF tPS mIN p3 nS vA iST5("></
w>.*?\r.*?)xAP(.*?\r.*?xVB fF.*?)
-> \1er\2RP\3xAV rP\4

3) Erstatte verbalformen "er" med relativpartiklen når der følger et propr., et appell. og et finit verbal:
<w l="vera" pos="xVB fF tPS mIN p3 nS vA iST5"></

w>(.*?\r.*?xNP.*?\r.*?xNC.*?\r.*?xVB fF)
-> <n>er</n><w l="er" pos="xRP"></w>\1

4) Erstatte verbalformen "er" med relativpartiklen, når et finit verbal følger, og der ikke er noget interpunktionstegn efter "er":
(<w l=")vera(" pos="x)VB fF tPS mIN p3 nS vA iST5("></w>.*?\r.*?xVB fF)
-> \1er\2RP\3

5) Erstatte finit form med infinitiv umiddelbart efter en finit form:
(xVB fF.*?\r.*?xVB)fF tPS mIN p3 nP
-> \1fI tPS

6) Erstatte infinitiv med finit form mellem "ok" og et pluralt sb. i N/A:
(<w l="ok" pos="xCC"></w>.*?\r.*?xVB f)I tPS (v.*?\r.*?xNC.*?nP c(N|A))
-> \1F tPS mIN p3 nP\2

7) Erstatte infinitiv med finit form mellem "ok" og "þeir"
(<w l="ok" pos="xCC"></w>.*?\r.*?xVB f)I tPS (v.*?\r.*?<n>þeir</n><w l="sá" pos="xPE gM nP cN"></w>)
-> \1F tPS mIN p3 nP\2

8) Erstatte infinitiv med finit form når et "þeir" følger umiddelbart efter:
fI tPS(.*?\r.*?<n>þeir</n><w l="sá" pos="xPE gM nP cN"></w>)
-> fF tPS mIN p3 nP\1

9) Erstatte akkusativ med nominativ af þat når et 3pers.sg. verb. følger:
(<w l="sá" pos="xPD gN nS c)A("></w>\r.*?xVB.*?p3 nS)
-> \1N\2

10) Erstatte nom. med gen. (af "Haraldr konungr") efter et substantiv:
(xNC.*?\r<f>har̄</f>____<d>har<e>all)ðr(</e></d>____<n>Harald)r(</n><w l="Haraldr" pos="xNP gM nS c)N(sI"></w>.*?\r<f>k̄</f>____<d>k<e>onong)r(</e></d>____<n>konung)r(</n><w l="konungr" pos="xNC gM nS c)N
-> \1z\2s\3G\4f\5s\6G

11) Erstatte 3.pers.-tag med 1.pers.-tag efter "ek":
(<w l="ek" pos="xPE p1 nS cN"></w>.*?\r.*?mIN)p3(nS)
-> \1p1\2

12) Erstatte AP med AV rP når et finit verbum følger:

```
xAP("></w>.*?\r.*?xVB fF)
      -> xAV rP\1
```

13) Erstatte dat.pl. med dat.sg. af poss.pron. efter mask.sg.dat.:

```
(nS cD sI"></w>.*?\r.*?xPP gM n)P( cD"></w>)
      -> \1S\2
```

Jeg lavede en del beslægtede (og ubeslægtede) søg&erstat (desværre var jeg så dum kun at nedskrive en mindre del af dem:-(). Som det fremgår, var der tale om meget simple mønstre, men det viste sig alligevel forbløffende effektivt. Efter at jeg havde arbejdet ca. 14 dage med filen (hvoraf tiden primært var gået med korrekturlæsning, men herunder altså også en del grep-baserede søg&erstat), lavede jeg for morskabs skyld en lille statistisk undersøgelse af 500 ord (prosa) for at se hvor mange rigtige gæt der var. Resultatet blev det følgende:

Helt korrekte	369	= 74%
Fejlagtige:	65	= 13%
Uudfyldte:	66	= 13%
Ord i alt:	500	= 100%

Dette vil jeg betragte som ganske opløftende, ikke mindst når man tager i betragtning at teksten kun forelå på facs-niveau.

3. Forslag til MLA

Disse bemærkninger skal naturligvis tages med et gran salt eftersom jeg kun har snuset til MLA i den indledende testfase. Hvis jeg skulle benytte mig af MLA i et fremtidigt projekt kunne jeg tænke mig, at der bl.a. var taget hensyn til flg.:

- Mulighed for at arbejde direkte på facs-niveau (helt centralt)
- Mulighed for lettere og frem for alt hurtigere at kunne tilføje nye former (vigtigt)
- Mulighed for at arbejde helt uden mus. Det lyder måske ikke så vigtigt, men jeg betragter det faktisk som noget meget centralt. Dette mener jeg ikke kun fordi det ergonomisk set er et frygteligt værktøj, men også fordi musen er et langsomt og hjælpeværktøj, der efter min mening ikke burde bruges i forbindelse med tektredigering.
- Mulighed for at redigere i den rå xml-fil (centralt), herunder grep-baseret (eller lignende) søg&erstat og meget gerne understøttelse af konkordanslignende funktioner
- Understøttelse af en simpel form for kontekstbaseret disambiguering

En del af begrænsningerne i MLA skyldes vel at den er webbaseret. Har det været overvejet om man skulle udvikle et kraftigere værktøj som kunne nedlastes som et egentligt program? (er klar over at der er en række problemer af forskellig art forbundet med dette).

4. Diverse lemmatiseringsproblemer
Her kommer der lidt af en blandet landhandel (både emne- og præsentationsmæssigt!!!)

Kompositum vs. simplicia
Som både IS (Haraldur og Jóhannes) og SE (Vadstenaprojektet) nævner, er det ofte problematisk at tage stilling til om der i et givet tilfælde er tale om to simplicia eller ét kompositum. Kan forstå at Vadstenaprojektet har valgt en meget "grafisk" løsning ved at tage udgangspunkt i den enkelte skrivers praksis. Dette gør naturligvis transskriptionsfasen lettere og mindre analytisk/mere objektiv). I et lemmatiseringsmæssigt perspektiv er det vel imidlertid et spørgsmål om hvor meget man skal tage hensyn til den grafiske manifestation (sat på spidsen kunne man sige at det ville føre til at alle præp. der er sammenskrevet med styrelsen, ville blive en del af denne!). Jeg tror ikke at man kommer uden om en mere eller mindre subjektiv bedømmelse af hvornår der er tale om hhv. kompositum og simplicia. Flere spm. melder sig imidlertid.

- I hvor høj grad skal man lægge sig op ad ONPs lemmaliste (vel vidende at der på dette punkt sikkert kommer til at ske ændringer. (I Widdings tid var der -- så vidt jeg husker -- vist en tendens til at opføre flere komposita end der medtages i den endelige ordbog)?
- Hvis man lægger sig op ad ONP, kan man stadig have et problem i poesi, der dels behandles meget dårligere af ONP (selvom meget er blevet skrevet ind i ordlisten) og dels måske ikke kan underlægges de samme regler.

Til almindelig "moro" anfører jeg lige nogle enkelte eksempler der kan illustrere visse af vanskelighederne ved analysen og opmærkningen af komposita/ simplicia. Betragt fx en sætning som den følgende (ONP har kompositummet "danalið").

> beþı ðana lıð oc norþm ̋(34v.23)

Hvis man vitterligt betragter det som et kompositum i dette tilfælde, så har man problemet med Norðmanna (men det er selvfølgelig ikke anderledes end at

lemmatisere mod. dansk "kaffe- og testue", men stadig et problem!)

Tilsvarende problemer finder man også i de følgende to eksempler:

> Sıgurþr k¯var opſtopa m̃mıkıll oc veırþar v¯alla lvtı (36r.6-7)

> h¯bıo norþr í nor¯a̋þıgr m̃oc ranglatr capſ ꝓvllr oc veırþar (36r.37-8)

Særlig i forbindelse med poesi kan man støde på det problem at et kompositum er opsplittet (tmesis). Jf. flg. eksempler med appellativet jastostr m. og proprierne Þjólarnes, Ǫngulsey og Hlésey:

Problemet omkring indbyrdes afhængighed mellem lemma -og pos-oplysninger
Dette problem forekommer mig ganske centralt. Det drejer sig om situationer hvor flere lemmata og pos-analyser er mulige, men hvor der er indbyrdes afhængighed mellem ét eller flere lemmata og én eller flere grammatiske analyser. Jf. eksempelvis

> þ¯mvn þa ſumra ᴍ̊mal at þu takız mıcıt ı ꝓang ꝓyrſta ſın̄¯flıc ſcallð ſem oꝓ mıc haꝓa oƨt (16r.2-3)

```
<f>ꝓyrſta</f><d>ꝓyrſta</d><n>fyrsta</n><w l="fyrstr" pos="xTO gN nS cA sD | xTO gN nS cD sD"></w>
<f>ſın̄¯</f><d>ſin<e>n</e>i</d><n>sinni</n><w l="sinn | sinni" pos="xNC gN nS cD sI | xNC gN nS cA
sI"></w>
```

Hvis lemma opfattes som "sinn", kan der kun være tale om cD, hvis det derimod opfattes som "sinni" kan der både være tale om cD og cA. Der er altså ikke et simpelt 1:1 forhold mellem lemma og pos.

Et lignende eksempel er:

> vıl ec v̊a tıonaðar m̃yþrara [sic!] ꝓıolſcẏllða (25r.19-20)

```
<f>ꝓıolſcẏllða</f><d>ꝓiolſcẏllða</d><n>fjǫlskylda</n><w l="fjǫlskylda | fjǫlskyld | fjǫlskyldi" pos="xNC gF nP cG
sI | xNC gN nP cG sI"></w>
```

For at løse dette problem er man vel i princippet nødt til at foretage en eksplicit sammenkobling af lemma og pos. Hvorledes kan man tage højde for dette uden at det får for vidtgående konsekvenser og forårsager praksisændring i alle andre, utvetydige, eksempler? Kunne man gøre det på flg. måde (sikkert ikke, men det kan i det mindste danne basis for en diskussion)

<n>sinni</n><w l="sinn" pos="xNC gN nS cD sI" | l="sinni" pos="xNC gN nS cD sI | xNC gN nS cA sI"></w>

Dette problem optræder naturligvis ikke i hver anden linje i et hs., men det forekommer dog ikke helt sjældent, og der må findes en løsning på det.

Angående enklise

Som også Haraldur og Jóhannes er inde på, synes det nødvendigt at udvide brugen af eE (ikke mindst i poetisk kontekst). Dette medfører (i hvert fald) følgende udvidelser af brugen:

1. Ud over det enklitiske þú må man regne med
   A. enklitisk ek (k), fx barðak for "barða ek"
   B. enklitisk er/es (s/z), jf. forbindelser som þannz, þars, þeims
   C. enklitisk a/at med nægtende betydning (fx vara for "var eigi" og fekkat for "fekk eigi").
   D. enklitisk eru ((r)ó), 3.ps.pl.præt. af vera (jf. flg. eksempler fra Morkinskinna: "þat lið er callaþ ro þınga meŋ̄", "Oc eıtt ſın er þ̲ baþ̄ o ſam̄ og "þ ̇at ħ ro m̓kı mın")

   Bemærk at fx A og B kan kombineres i former som "skalka" (="skal ek eigi"). I sådanne tifælde har jeg forsynet hvert af de tre "ord" med eE-tag.

2. Pga. 1.D. er det ikke kun verbalformer som enklitiske former knytter an til. Spørgsmålet er om man i alle tilfælde skal forsynes begge de i enklisen deltagende ordformer med eE (som tilfældet er i anbefalingerne omkring det enklitiske þú), eller om man i øvrige tilfælde (hvor der ikke kan regnes med samme fonetiske "bivirkninger" skal udelade eE i den ordform som det enklitiske element knytter an til. Jeg har valgt i alle tilfælde at indføre eE som i tilfældene med þú".

Et interessant problem, der kan opstå ved enklitiske former har man i "hnıǧt" i flg. strofe:

Der er jo tale om samme problem som det af Haraldur og Jóhannes nævnte 'þaz' for "þat +s". Hvordan hulen løser man dette?

Diverse spredte bemærkninger
Ville det være ønskværdigt at kunne skelne mellem den almindelige konjunktion 'en' "men" og sammenligningskonjunktion 'en'/'an' "end"?

Som bemærket af SE kan det være særdeles problematisk at lemmatisere numeralier, når de er skrevet som symboler (hvad enten det nu måtte være araber eller romertal". Problemet opstår når tallet består af flere led, hvad enten det er et eks. som '.ccc.' eller 'xl'. Hvis man vælger at betragte det som flere led ('þrjú hundruð' og 'fjórir tøgir', så må det vel deles op i to lemmata, men hvordan gør man det? En lettere løsning er måske derfor i stedet at betragte det som ét lemma og så regne med bøjning af hvert led (i princippet adskiller det sig jo ikke fra et pron. som "hvárrtveggi"). Et problem i denne henseende er ikke blot at man får et uendeligt antal numeralier (en praksis man jo ellers ikke følger i leksikografisk arbejde), men også at man må opmærke på inkonsekvens vis, når talordet opdeles ved indskud af ord mellem de flere af leddene (hvad gør man egentlig når man lemmatiserer moderne tekster?)

Har det været diskuteret hvilken politik man har over for sammensatte konjunktioner, adverbier og præpositioner? Jf. ikke mindst tilfælde som þótt/þó at, þars/þar er og þvít(t)/því at.

Angående bestemthedsspørgsmålet ved proprier taler det imod Haraldurs og Jóhannes' forslag om at udelade det, når man kan finde variantformer som Vík/Víkin. Visse stednavne har man øjensynligt både kunnet bruge i bestemt og ubestemt form (selvom der naturligvis ikke er tale om en traditionel skelnen mellem noget bestemt og ubestemt).

Mener ikke at det umiddelbart er en god ide at slå de forskellige analyser af 'einn' sammen i én fælles kategori. Det er korrekt at det ofte er meget svært eller umuligt at skelne mellem de forskellige muligheder, men i en lang række tilfælde er det ikke desto mindre muligt (og det vil også være tilfældet i ONP). I de tvetydige tilfælde kan man jo fortsat anvende "|" til at markere dette. Det er jo

ikke som eksempelvis GP og DP af pronominer/adjektiver, hvor det ekstra besvær så at sige ikke fører noget med sig.

Hvorfor skal man i følge Menotas anbefalinger markere en tve- eller flertydighed i genus som gMF, gMN ...? Det synes for det første inkonsekvent at markere det således i en vis kategori (substantiver), men ikke i andre, og for det andet bryder det med den generelle ide om at anvende "|" til markering af alternativer. Det forekommer mig at være betydelig mere konsistent udelukkende at anvende "|" og helt droppe tags som gMF.