Tekniska spörsmål

Facs-nivå

I stort sett all text som transkriberats inom Vadstena-projektet återfinns enbart på facs-nivå. Därför efterfrågar vi möjlighet att lägga in text i lemmatisatorn på facs-nivå.

För Barlaamfilen har vi gjort en ad-hoc lösning genom att ändra <facs> till <dipl>. Detta skapar en del problem med framför allt struken text (text placerad inom <del></del>). Om skrivaren t.ex. har struket delar av ordet får vi <del> innanför <w>. Men lemmatisatorn ser inte ut till att ta hänsyn till sådana <del>-taggar, utan för upp de strukna tecknen som del av ordformen. Och vi får ordformer som t.ex. (blocknr inom parentes):

thænnoa (9): <w><facs>thænn<del hand=”scribe” type=”overstrike”>o</del><add hand=”scribe” place=”supralinear”>a</add></facs></w>

førd (463): <w><facs>før<seg type=”multiple del”><del hand=”scribe” type=”subpunction”><del hand=”scribe” type=”overstrike”>d</del></del></seg></facs></w>

thættane (1512): <w><facs>thæ<del hand=”scribe” type=”overstrike”>tta</del><add hand=”scribe” place=”supralinear”>ne</add></facs></w>

Vi önskar att det som är struket inte kommer med i de ordformer som skall lemmatiseras.

Ett annat önskemål är att all struken text markeras i lemmatisatorn. För tillfället är det enbart längre textpartier som markeras, medan enskilda tecken och ord står utskrivna som en del av texten.

Redigering

Vi önskar att kunna redigera på flera nivåer. Dels önskar vi att kunna redigera i vår xml-fil när vi upptäcker fel i transkriberingen och kodningen. Därför efterlyser vi procedurer för nedlastning och upplastning av filen igen i lemmatisatorn. Vidare önskar vi att kunna göra ändringar i lemmatisatorn. Vi är i den situationen att vi bygger upp lexikon, och vi vill gärna

kunna redigera detta efterhand. Till exempel önskar vi att kunna ta bort lemma som felaktigt lagts in.

I nuläget har vi möjlighet att trycka ”Slå undertrykking på/av” och vi undrar vad sker med den formen som markerats med ’ut’, försvinner den ur databasen, eller?

Dessutom undrar vi vad är det för skillnad mellan ”Last dokumentet ned” och ”Last det taggede dokumentet ned”?

Funktioner i lemmatisatorn

Vi skulle dessutom vilja ha möjlighet till att:

- kontrollera vilka lemman som har lagts in
- få fram en lemmalista
- göra en KWIC konkordans på det som har lemmatiserats
- se hur stor del av texten som har lemmatiserats och hur mycket som är kvar
- kunna söka i lemmatisatorn efter t.ex sådant som inte blivit inlagt eller lagrat
- göra övergripande ändringar i basen, t.ex kunna ändra ortografin i en ordform

Flera av de här funktionerna finns kanske redan i lemmatisatorn. I så fall skulle vi vilja lära oss hur dessa fungerar.

Vi har även ett önskemål om en fast ordklasslista att välja från samt en kontroll i programmet på att det man registrerar faktiskt återfinns i listan. Detta vill medföra mindre fel i basen, för tillfället godkänner t.ex. lemmatisatorn ’n’ som en ordklass (däremot inte ’v’).

Filologiska spörsmål

Samman- och särskrivningar

Vissa ord förekommer i både sammanskriven och särskriven form, t.ex *berghahulu*/*bergha hulu*, *öghnabliki*/*öghna bliki*, *utkasta*/*ut kasta*, *bortköra*/*bort köra*, *atenast*/*at enast*, *thäntidh*/*thän tidh* osv. Dessa har transkriberats i enlighet med hur de är nedtecknade i

handskriften, vilket innebär att de har placerats inom en word-tagg när man tydligt kunnat se att de är sammanbundna till ett ord: <w><facs>berghahulu</facs></w> och inom två word-taggar när man tydligt kunnat urskilja två helt åtskiljda ord: <w><facs>bergha</facs></w><w><facs>hulu</facs></w>. När osäkerhet förelegat huruvida ordet skall betraktas som samman- eller särskrivet har det placerats inom en word-tagg med ett mellanrum mellan de två orden: <w><facs>bergha hulu</facs></w>. I samtliga fall handlar det om ett och samma ord som skall kopplas till ett lemma. Men i lemmatisatorn blir detta problematiskt eftersom det sammansatta ordet blir länkat till ett lemma medan de båda särskrivna till två separata lemman. *Utkasta* blir lemmatiserat som verb och *ut kasta* som adverb respektive verb. Sammanfattningsvis: Det förekommer en växling mellan en och två <w>-taggar för samma sammansättning. Vi har kodat olika något som vi vill skall behandlas lika i lemmatisatorn. Hur kan detta lösas på bästa sätt?

Nästa spörsmål i sammanhanget blir till vilket lemma en sådan här sammansättning ska knytas. Skall t.ex. *renlivis man* utgöra ett eget lemma *renlivis madher* eller placeras under *madher* eller *renlive*? Söderwall placerar *renlivis madher* i fet stil under *renlive*.

Participer

Hur ska man förhålla sig till participer? Om man går efter formen bör participer placeras under verb, men om man går efter funktionen under adjektiv, ibland som substantiv och adverbial.

En form – en ordklass – olika betydelser – olika former

Ett och samma ord kan ha flera betydelser. Bör dessa markeras på olika sätt? I nuläget placeras alla former under ett lemma. Ta t.ex. verbet *löna* som har två olika huvudbetydelser: 1) löna, vedergälla och 2) dölja. Eller verbet *vita* som förekommer i ett antal former beroende på betydelse.

Romerska tal

Lemmatisering av romerska tal

Person- och ortnamn

Normalisering av personnamn/ortnamn