

Konverteringsprosessen

Table of contents

1 Hovudstega.....	2
2 Pseudokode.....	2
3 Ferdig kode (Perl).....	9

1. Hovudstega

1. Reinsa kjeldefilene:
Fjerna testpostar og andre oppslag som ikkje er «ekte» ved å filtrera dei fem hovudfilene (Oversettelse, Ordbeskrivelse*2, Synonym*2) mot ei liste med Oversettelses-ID-ar som ikkje skal med i den vidare prosessen
2. Trekkja ut ei liste med alle brukte teikn (utgangspunkt for konverteringa til UTF-8 - den endelege spesifiseringa av omkodinga er eit manuelt steg).
I den fyrste versjonen koda eg om all tekst likt, men det er mogleg eg må skilja mellom omkoding av samisk og av norsk tekst for å få alt rett - det ser ut som om same teiknet i kjeldematerialet av og til eigentleg er to ulike teikn, avhengig av språk.
3. Les inn alle fem hovudtabellane i kvar sin hash-struktur (ID er nykkel)
4. Konvertera dei gamle tabellane til tre nye hash-strukturar: navet, samiske oppslag og norske oppslag. Dei nye hash-strukturane får nye ID-ar (sjå [spec for ny termstruktur](#)) som fungerer som nykkel i hashen.

Som ein del av konverteringa prøver eg å parsa Kommentar-feltet i Ordbeskrivelse, og gjera den implisitte strukturen eksplisitt. Eg må vera forsiktig, og gje meg med ein gong det er noko som ikkje stemmer, og i så fall kopiera alt ut i eit samla element i den nye strukturen.
5. Skriwa ut dei nye hash-strukturane som XML-kode

I tillegg må emnehierarkiet konverterast, og ordklasser og bøyingsklasser.

2. Pseudokode

```
Strukturar:
- %Oversettelse:
  - KEY: ID
  - VALUE: hash med faste nyklar:
    MACRO
    MICRO
    NANO
    POS
    S-ORD
    N-ORD
    LATIN
- %OrdbeskS og %OrdbeskN:
  - KEY: ID
  - VALUE: hash med faste nyklar:
    WORD
    POS
    MAININFL
    SUBINFL
    INFLEXMPL
    LATIN
    COMMENT
- %SynoS og %SynoN:
  - KEY: Ord-ID
  - VALUE: liste med synonym
```

```
- NAV-hash: her er all informasjon, denne hashen er
utgangspunktet
  for XML-navet:
  KEY: ID (gamal ID)
  VALUE: Hash med faste felt, felte er:
        MACRO
        MICRO
        NANO
        Sref (reff til ordartikkel)
        Nref (reff til ordartikkel)
        Lref (reff til ordartikkel)
- MainID: hash med alle og berre dei postane som blir
hovudpostar
  i den ferdige termbasen, inneheld ei liste med alle synonyme
postar.
  NB! Den fyrste ID-en i arrayet er ID-en til posten sjølv! Eg
kom
  ikkje på nokon annan måte å unngå ei feilmelding på. Så hugs å
hoppa
  over posisjon 0 (null) i arrayet.
  KEY: ID
  VALUE: ARRAY: ID-ar for alle synonyme postar
- duplID: duplikat med same emnekode får ID-en sin her som
nykkel,
  ID-en til duplikatet er verdi; skal brukast for å kunna gå frå
ordbeskrivelsen til eit duplikat til datastrukturen til samle-
posten (for å fjerna alle doble skildringar)
  KEY: duplikat-ID
  VALUE: NAV-ID
- %SOppslagsID/%NOppslagsID: inneheld unike ID-ar (sjå KEY-def)
for
  å kunna identifisera om eit oppslagsord allereie finst eller
ikkje;
  to versjonar, ein for kvart språk.
  KEY: string of type: word#POS#class
  VAL: ID of the entry
- OppslagsSyn: som oppslagsordhashen, men inneheld berre
oppslags-
  ord der det andre språket har vore identisk med ein
eksisterande
  post - oppslagsordet for det fyrste språket blir altså eit
synonym
  for eit eksisterande oppslagsord
  KEY: oppslagsord#micro
  VALUE: NAV-ID
- OverRusk: hash med postar som ikkje skal vera med i
konverteringa;
  ID-ane er nykkel, verdi '1' (vi skal berre raskt kunna sjekka
om ein ID finst i hashen)
  KEY: NAV-ID
  VALUE: "1"
- samisk-hash/norsk-hash/latin-hash: inneheld oppslagsorda med
definisjonar og synonym
  KEY: oppslagsord#ordklasse
  VALUE: Reff til key/value hash:
        Word
        POS
        PL (indicates whether a word is pluralis tantum)
```

```

InflMain
InflSub
InflEx
Senses (Ref)-> KEY: GmlID (%nav)
                VALUE: Reff til Hash:
                    Macro
                    Micro
                    Nano
                    Definisjon (frå Comment)
                    Status ("main", "syn")
                    Main (mainref, only if
status=syn)
                    Synref -> Array:
                                Syn1-reff
                                Syn2-reff
                                Syn3-reff
                                ...

Les inn dei tre rusklistene:
    - legg alle ID-ane som nyklar i hashen.

Opne Oversettelse.txt
Les inn line for line:
    - kløyv lina i enkeltdelar
    - sjekk om ID-en er med i rusklista:
      - om ja, les neste line
    - finst det emnekode? (vi vil berre ha "termar" = postar med
emnekode)
      JA:
        - reins emnekodane (mellomrom føre, midt i og etter;
          nano: for få teikn)
      NEI:
        - gå til neste line
    - sjekk at dei to ID-ane er like (det skal dei vera):
      - om ikkje, skriv ut ei åtvaring m. ID-ane, og gå til
neste line
    - reins oppslagsorda (mellomrom etc. i starten/slutten)
    - lagra alle delane av oppslaget i ein hash med faste nyklar:
      - emnekode*3, ordklasse, og uferdige reffar til 3 språk:
        - se, no, la
    - lagra lista i Oversettelse-hashen, med ID som nykkel
    - gå til neste line

Opne OrdbeskrivelseS.txt
Les inn line for line:
    - kløyv lina i enkeltdelar
    - sjekk om ID-en er med i rusklista:
      - om ja, les neste line
    - finst ID-en i Oversettelse-hashen?
      - NEI: gå til neste line
    - prøv å parsa kommentar-feltet:
      - trekk ut bøyingsformer
      - sjå etter latinsk ekvivalent
    - reins ordklassane
    - legg ordklassa inn i Oversettelse-hashen (same ID)
    - dersom latinsk ekvivalent:
      - legg til latin i Oversettelse-hashen
    - lagra alle delane i Ordbesks
    - neste line

```

```
Steng OrdbeskrivelseS.txt
Gjenta for OrdbeskrivelseN.txt -> i hash OrdbeskN

Opne SynonymS.txt
Les inn line for line:
- kløyv lina i enkeltdelar - NB! det kan finnast meir enn eitt
  synonym pr. post (kommaseparert)
- finst syn-ID-en i rusklista for SynS?
  JA: neste line
  NEI: fortsett
- finst ord-ID-en i rusklista for hovudoppslag?
  JA: neste line
- finst ord-ID-en i Oversettelse-hashen?
  JA: fortsett
  NEI: neste line
- lagra synonyma i hashen SynoS:
  - KEY: Ord-ID
  - VALUE: liste med synonym
  - neste line
Lat att SynonymS.txt

Gjenta for SynonymN.txt

Slå i hop synonyme postar:
- Ta omsyn til oppslagsord, emnekode og ordklasse når ein
  slår i hop postar som verkar synonyme
- definer hovudoppslag:
  - fyrste unike post
  - duplikat er pr. def synonyme
- list opp synonyme oppslag for seg:
  - lagra gml ID for hovudoppslag, med peikar til alle
    synonyme postar (med gamal ID)
  - lagra ID for synonyme postar med peikar til
hovudoppslag

Konverter hovudoppslaga til ny struktur, ein gong pr språk

Gå over alle synonyma (både eksplisitte synonym frå gamal base, og
deriverte synonyme oppslagsord), og lag ordboksartikar/tydingstillegg
for dei.

Gå gjennom dei fem kjelde-hashane:
- Oversettelse
- OrdbeskS
- OrdbeskN
- SynoS
- SynoN
og restrukturer dei til fire hashar (navet etter omgrep, dei andre
etter språk, og deretter oppslagsord):
- NAV-hash
- samisk-hash
- norsk-hash
- latin-hash

Meir detaljert:
- gå gjennom Oversettelse-hashen, ID for ID, tre gonger:
  - 1. gong: samisk perspektiv
  - 2. gong: norsk perspektiv
```

```

- 3. gong: latinsk perspektiv
- kvar gong:
  - lag komplette reffar til oppslaget i målspråket
  - lag fulle oppslag av oppslagsorda
  - lag fulle oppslag av alle synonyma

Skriv ut dei fire hashane som XML-kode:
- NAV-hash
- samisk-hash
- norsk-hash
- latin-hash

Klypp frå tidlegare forsøk (skal gjenbrukast, men med modifikasjonar):

Opne Oversettelse.txt
Les inn line for line:
- kløyv lina i enkeltdelar
- sjekk om ID-en er med i rusklista:
  - om ja, les neste line
- finst det emnekode? (vi vil berre ha "termar" = postar med
emnekode)
  JA:
  - reins emnekode (mellomrom føre, midt i og etter;
  nano: for få teikn)
  NEI:
  - gå til neste line
- sjekk at dei to ID-ane er like (det skal dei vera):
  - om ikkje, skriv ut ei åtvaring m. ID-ane, og gå til
neste line
  - reins oppslagsorda (mellomrom etc. i starten/slutten)
  - slå opp i oppslagsordhashen (oppsl.+emnek.) for å sjå om
  oppslagsorda finst frå før:
  JA:
  - om baa finst:
    - lagra ID som nykkel i ID-hashen, med ID-en til
oppslaget
      som finst frå før som verdi
      - gå til neste line
  - om språk A finst:
    - sjekk om oppslagsordet+emnek finst i
synonymhashen:
  NEI:
  - lagra oppslagsordet som ikkje er likt
i
  OppslagsSyn (oppslagsord+emnek.), med
verdi =
  ID til Nav-posten som ordet er synonym
til
  - legg til synonymet i SynonymS/-N under
ID-en til
  navet, med eigen ID som nykkel og
synonymet som
  verdi i ein anon. hash inni
SynonymS/-N
  JA: fortsett
  - lagra ID-en som nykkel i ID-hashen, med ID-en
til den
  eksisterande posten som verdi
  - gå til neste line

```

```
NEI:
- lagra alle delane av oppslaget i ei liste:
  - emnekode, og uferdige reffar til 3 språk:
    - se, no, la
- lagra lista i Nav-hashen, med ID som nykkel
- lagra kvart oppslagsord som nykkel i
oppslagsord-hashen, med ID
  som verdi
  - gå til neste line

Opne OrdbeskrivelseS.txt
Les inn line for line:
- kløyv lina i enkeltdelar
- sjekk om ID-en er med i rusklista:
  - om ja, les neste line
- finst ID-en i NAV-hashen eller DuplID-hashen?
  - NEI: gå til neste line
- prøv å parsa kommentar-feltet:
  - trekk ut bøyingsformer
  - sjå etter latinsk ekvivalent
- hent oppslagsord
  - er det eit hovudoppslag/duplikat/synonym?
DERSOM HOVUDOPPSLAG:
- hent emnekode frå Nav-hashen (bruk ID)
- lag ny ID (oppslagsord + ordklasse)
- oppdater NAV-hashen med betre reff for SE (ID frå førre lina)
  - men berre dersom det ikkje er eit synonym/duplikat
- finst oppslaget frå før i Samisk-hash (ny ID er nykkel):
  JA:
  - sjekk at bøyingskodane er dei same:
    JA: ok, fortsett
    NEI:
    - skriv ut ei feilmelding, og fortsett
  - legg til gml ID (ref til omgrepstext i NAV-hash) i
Sense-
  lista inne i lista for oppslagsordet
  - legg til emnekode i Sense-lista
  - legg til definisjon i Sense-lista (frå Comment-feltet)
  - legg til status for oppslaget i denne tydinga i Sense-
    feltet ("main" om det er hovudoppslag, "syn" om det er
    eit synonym)
  - sjå om det finst synonym i synonym-hashen med gml ID:
    JA:
    - for kvart synonym:
      - bruk ordklassa til å laga ein ekte
reff
  - legg inn den nye reffen i Sense-lista
  - lag eit nytt oppslag med synonymet:
    - sjekk om det finst frå før:
      JA:
      - legg inn gml ID,
main-reff og emne-
      kode, og status i
Sense-lista til ordet
      NEI:
      - legg inn synonymet med
ny ID/reff i
      oppslagsord-hashen
      - legg inn gml ID,
```

```

main-reff og emne-
ordet
                                kode i Sense-lista til
                                NEI: - fortsett
                                NEI:
                                - legg oppslagsord, ordklasse og bøyning inn i ei liste
                                for oppslagsordet
                                - legg gml ID (ref til omgrepstext i NAV-hash) inn i
Sense-
                                liste inne i lista for oppslagsordet
                                - sjå om det finst synonym i synonym-hashen med gml ID:
                                JA:
                                - for kvart synonym:
                                - bruk ordklassa til å laga ein ekte
reff
                                - legg inn den nye reffen i Sense-lista
                                - lag eit nytt oppslag med synonymet:
                                - sjekk om det finst frå før:
                                JA:
                                - legg inn gml ID,
main-reff og emne-
ordet
                                kode i Sense-lista til
                                NEI:
                                - legg inn synonymet med
ny ID/reff i
                                oppslagsord-hashen
                                - legg inn gml ID,
main-reff og emne-
ordet
                                kode i Sense-lista til
                                NEI: - fortsett
                                - lagra oppslagsordlista i oppslagsord-hashen, med ny ID
                                som nyckel
DERSOM SYNONYM:
- stort sett som for main, men Sense-behandlinga skal gje:
- status = syn
- reff til main
DERSOM DUPLIKAT:
- jamfør bøyingskodar:
- likt: fortsett
- ulikt: skriv ut melding, fortsett
- parsa kommentar
- legg til bøyingsdøme dersom det ikkje finst
- legg til restkommentar dersom han er ulik eksisterande
- legg ID (KEY) og nav-ID til main (VALUE) i %SamDup (gjev
enklare
prosessering av synonyma lenger ned)
FOR ALLE:
- finst det latinsk ekvivalent?
JA:
- bruk ordklassa til oppslagsordet til å laga ny reff/ID
- legg til reffen i NAV-hashen for hovudoppslaget (OBS!
Syn.!)
- finst ordet frå før i latin-hashen?
JA:
- legg til gml ID (ref til omgrepstext i
NAV-hash) i
                                Sense-lista inne i lista for oppslagsordet

```

