

Ny termbasestruktur

0.1: Fyrste versjon, diskusjonsgrunnlag.

by Sjur N. Moshagen

Table of contents

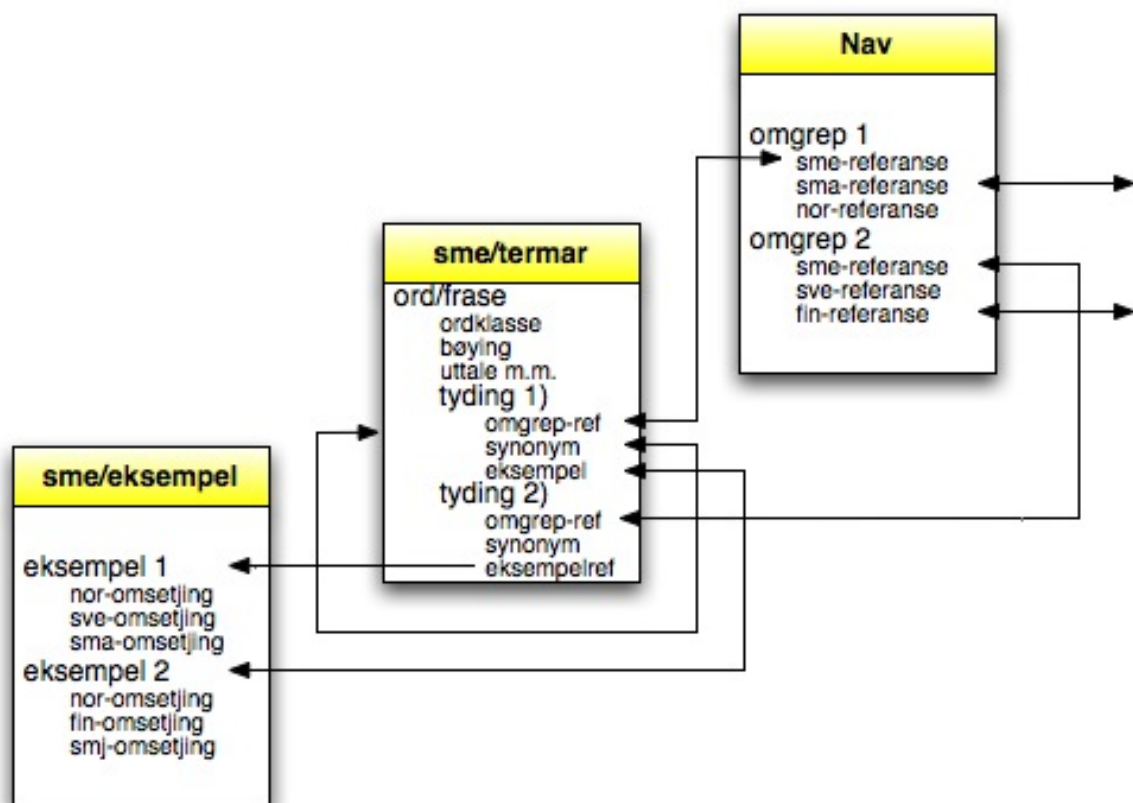
1 Grunnleggjande mål og arkitektur.....	2
2 Detaljert struktur.....	3
2.1 Navet.....	3
2.2 Formen på ID-ar.....	5
2.3 Termpostar (ordbokspostar).....	6
2.4 Eksempel.....	9

1. Grunnleggjande mål og arkitektur

Eg har hatt nokre mål i hovudet medan eg har arbeidd med å byggja opp ein ny struktur for termbasen. Dei er:

- omgrepsorientert fleirspråkleg nav
- minimalt med duplisering av info (men nokre stader gjev det meining å kopiera heller enn å referera)
- enkelt å søkja (frå ein teknisk synsstad)
- enkelt å byggja opp termpostar for vising
- effektiv prosessering

Ut i frå desse krava føreslår eg ein struktur som kan skisserast som i figuren nedanfor:



Ny struktur

Uttrykt punktvis:

- navet er eit samlepunkt for kvart omgrep, med peikarar til termar/termpostar på kvart språk
- peikarane peiker til eit separat XML-dokument, som er i strukturen lik ei einspråkleg ordbok:

- eitt oppslagsord (eller frase)
- alle grammatisk (og andre) opplysningar som heng i hop med oppslagsordet (og ikkje med tydingane) er samla rett etter oppslagsordet
- dei ulike tydingane svarar til ulike omgrep i navet, og har peikarar dit
- synonyma er eigne oppslagsord med komplette ordboksartiklar; gjer det mogleg at same oppslagsordet kan vera synonym til ein term for eitt fagfelt (=tyding), og hovudterm for eit anna fagfelt; synonyma vert refererte til frå hovudtermen
- alle eksempel er lagra i ei separat XML-fil, og er refererte til frå term-posten:
 - same eksempelet kan brukast mange stader utan at ein treng halda ved like meir enn eitt eksemplar
 - eksempla er "retningsorienterte", dvs. dei har eitt kjeldespråk og ei eller fleire omsetjingar - det er oftast slik at dei omsette eksempla ikkje sjølve er gode eksempel for målspråket, dei er berre gode eksempel på korleis ein kan omsetja uttrykket i kjeldespråket til målspråket i ein gjeve kontekst
 - det er mogleg gjenbruksverdien er så liten at det kunne vera like enkelt å ha eksempla direkte i "ordboksartikkelen", då slepp ein samtidig den ekstra prosesseringa for å fylja referansen og inkorporera eksemplet
 - eksempla bør truleg vera domenespesifikke, dvs gjelda for eit visst fagfelt
- når det gjeld å unngå duplisering, vil denne organiseringa hindra duplisering av oppslagsord, av synonym og av eksempel

I høve til søking vil ein slik arkitektur gjera det enkelt å søkja på termar og generera treffliste: alle søk vil gjerast på dei einspråklege ordboksliknande filene. Sett med treff kan enkelt filtrerast til å innehalda oppslagsorda + emnekategori med ei underliggjande lenke til term-posten i navet. Når ein klikkar på denne lenka, vil ein generera ein fleirspråkleg term-post frå navet, ettersom navet inneheld peikarar til alle oppslagsord med definisjonar som gjeld for eit visst omgrep, på alle språk som termen er definert for. Dersom desse peikarane er uttrykt med XInclude, burde det vera mogleg å få heile term-posten med eitt søk.

2. Detaljert struktur

I teksten nedanfor har eg brukt nokre konvensjonar:

@tekst

eit attributt med namnet **tekst**

tekst@attributt="verdi"

elementet **tekst** med verdien til attributtet **attributt** sett til **verdi**

2.1. Navet

Navet skal innehalda term-postar ordna etter omgrep, identifiserte med ein ID, og med referansar til det eigentlege innhaldet i term-posten som XInclude-uttrykk som peiker til den språkspesifikke delen av ein term-post. Til slutt kjem ei liste med endringshistoria, frå

konverteringa frå SQL og framover. Lista blir laga automatisk, og listar berre opp kva for element (=språk) som har vorte endra eller lagt til. Her fylgjer DTD (dokumentstruktur) for navet:

```
<!-- The root element. Contains attributes for id, reference to
      a classification system, and reference to a separate front
      document containing the name of the terminology collection,
      authors, copyright etc. -->
<!ELEMENT termCenter entry+>
<!ATTLIST termCenter id          ID
                  ClassSystem CDATA
                  xmlns:xi      CDATA          #FIXED
                  "http://www.w3.org/2001/XInclude"
                  front         CDATA >

<!ELEMENT entry (topicClass, entryref+, changes)>
<!ATTLIST entry id ID>

<!ELEMENT topicClass EMPTY >
<!ATTLIST topicClass top      CDATA
                  mid        CDATA
                  botm      CDATA >

<!-- We need to encapsulate the XInclude tag in another tag, to be
      able to identify the language of the resulting section (the
      XInclude tag is replaced with the content it points to, thus
      having the @xml:lang in it won't work - it will disappear). -->
<!ELEMENT entryref (xi:include)>
<!ATTLIST entryref xml:lang (sme | smj | sma | nor | sve | fin)
#REQUIRED >

<!ELEMENT xi:include EMPTY >
<!ATTLIST xi:include href          CDATA          #REQUIRED >

<!ELEMENT changes (change+) >

<!ELEMENT change EMPTY >
<!ATTLIST change when CDATA
                  what CDATA
                  who  CDATA >
```

Det skulle gje eit dokument som ser slik ut:

```
<termCenter id="some-id" ClassSystem="system-ref" front="front-ref">
  ...
  <entry id="3122">
    <topicClass toplevel="E" midlevel="E0000" bottomlevel="EN0010" />
    <entryref xml:lang="sme">
      <xi:include
href="terms-sme.xml#xpointer(//entry[@id='aktoriekti\s'])" />
    </entryref >
    <entryref xml:lang="nor">
      <xi:include
href="terms-nor.xml#xpointer(//entry[@id='enerett\s'])" />
    </entryref >
    ...
    <changes>
      <change when="2005-03-01T09:53:16.256+02:00" what="Converted from
SQL" who="admin" />
  </entry>
</termCenter>
```

```
</changes>  
</entry>  
...  
</termCenter>
```

Heile ordboksartikkelen er referert til på kvart språk. Det betyr at ein må filtrera vekk dei tydingane som ikkje gjeld dette omgrepet. Det gjer ein ved å kontrollera fagfeltet i dei ulike tydingane mot fagfeltattributta i oppslaget i navet. Berre der det er treff, går tydinga igjennom, i normale tilfelle vil ein alltid og berre få eitt treff pr artikkel. Det er mogleg at ein bør eller vil omforma term-posten enno meir før presentasjon, men det bør i så fall skje med eit separat stilark i ein eigen prosess.

Synonym er ikkje refererte frå navet, berre hovudtermen. Synonym og andre sekundære former er refererte i term-posten for hovudtermen for kvart språk, og er referert til med eit ID-attributt. ID-en bør få ei form som gjer det mogleg å transformera han til ordet/frasen det gjeld, slik at ein slepp å gjera fleire søk for å kunna visa synonyma. Meir om forma på ID-en nedanfor.

2.2. Formen på ID-ar

For å effektivera prosesseringa og forenkla visinga av term-postar, føreslår eg at alle ID-ar har ein klår struktur som samtidig inneheld oppslagsordet eller -frasen han identifiserer. Modellen er som fylgjer:

oppslagsord

ID = oppslagsord\ordklasse (d.e. oppslagsordet + '\' + ordklassa i forkorta form)

ord i oppslagsfrase

ID = ord_i_oppslagsfrase

Føresetnaden er at innanfor ein termbase/språk finst det ikkje to oppslagsord som er like innanfor same ordklasse, og for oppslagsfraser vil det aldri finnast to like fraser som kvart sitt oppslag. Ein skal hugsa at alle oppslagsord og -fraser er organiserte som i ei einspråkleg ordbok, med dei ulike definisjonane, synonyma m.m. ordna i ulike tydingsgrupper, der kvar tydingsgruppe svarar til eitt omgrep i det sentrale navet. Termar som har fleire bruksområde innanfor ulike fagfelt vil difor ikkje få fleire oppslag, men fleire ulike tydingsgrupper, uansett om termen er laga av eitt eller fleire ord. Føresetnaden bør difor halda. For å gjera systemet konsistent kunne ein tenkja seg at fraser òg fekk ei ordklassemarkering etter kjerna i frasen.

Med ein slik struktur kan ID-en genererast automatisk så snart oppslagsordet/-frasen og ordklassen har vorte skrivne inn, og ID-en er samtidig heilt forståeleg for alle som ser han. Men det viktigaste argumentet for ein slik struktur er at det er svært enkelt å konvertera ID-en til klårtekst som kan visast direkte til brukaren. Dette forenkler prosesseringa når ein skal skapa hypertekstlenker og visingstekst av til dømes synonymreferansar for ein hovudterm. For eit døme, sjå eksempla i neste punkt, samt dømet på ein term-post i navfila over.

Slike ID-ar skal brukast overalt kor ein vil referera til ein term/oppslagsord. Derimot er det ikkje mogleg å bruka slike ID-ar som ID for postane i navet: i og med at desse postane er omgrepspostar utan noka eiga uttrykksside (uttrykkssida er dei ulike termene på kvart språk som omgrepsposten refererer til), kan ein ikkje leggja eit uttrykk (ord, frase) til grunn for ID-en. Samtidig trengst slike ID-ar for at ein skal kunna referera til omgrepa frå kvart oppslagsord (rettare: frå tydingsgruppene i oppslaga) for kvart språk. Det enklaste er å bruka numeriske id-ar (ev med ein bokstav framfor: i XML kan ikkje attributt av typen ID byrja på tal, men ein kan sjølvsagt velja ein annan type for attributtet). I og med at handteringa av desse id-ane i all hovudsak er automatisk, burde det ikkje vera noko problem å bruka numeriske id-ar for omgrepa.

2.3. Termpostar (ordbokspostar)

Sidan navet ikkje eigentleg inneheld noko som helst, men berre samlar alle postar som høyrer til same omgrepet, ligg all eigentleg informasjon i det eg kallar termpostar. Det er eigentleg eit missvisande namn, sidan tradisjonelle termpostar er strengt ordna etter omgrep. Difor vil ein i tradisjonelle terminologisamlingar finna same termen definert fleire gonger, dersom han er i bruk i ulike tydingar innanfor forskjellige fagfelt.

For å forenkla vedlikehaldet og unngå problem med duplisering av informasjon, har eg i staden organisert all språkleg informasjon i ulike dokument, eitt for kvart språk. Kvant slikt dokument er ordna som ei einspråkleg ordbok, der kvart oppslagsord (=term) er ført éin gong. All informasjon (grammatisk, ortografisk m.m.) som gjeld sjølve oppslagsordet står dermed òg berre ein gong, medan definisjon, synonym og anna som gjeld oppslagsordet som term i ein bestemt samanheng er ordna i ulike tydingsgrupper. Det er berre eitt nivå med tydingsgrupper.

Kva som er eit oppslagsord har ein streng teknisk definisjon i XML-fila: eit oppslagsord er ein unik tekststreng sett saman av eitt eller fleire ord og ordklassa for ordet eller kjerna i frasen. Dette tyder at ortografiske variantar av «same ordet» er ulike oppslag (med peikarar til kvarandre), likeeins at forkortingar er egne oppslag, med tilvisingar til den utskrivne frasen (som òg er eit eige oppslag). Det kan kanskje verka som mykje skrik for lite ull, men det forenkla strukturen til oppslaga ein heil del om ein ikkje treng ta opp i seg alle moglege variantar av ein ordboksartikkel, og det forenkla dermed òg prosesseringa av oppslaga.

Dette tyder òg at alle tydingar av eit ord blir plassert under same oppslagsordet, så lenge ordklassen er den same. Det er mogleg ein må ta med bøyning som eit kriterium i tillegg til ordklasse, slik at eit ord som har ulik bøyning i ulike tydingar vil bli ført som to oppslag. Dette er t.d. ikkje uvanleg i norsk (jf *plan*, både mask. og nøy., eller *rett*, som har ulik (mask.) bøyning avhengig av om det gjeld mat eller juss), men eg veit ikkje om det gjeld for samisk.

Det er uproblematisk å presentera alle variantar som eitt oppslag, slik at det for brukaren ikkje blir forskjellig frå det ein finn i ei vanleg ordbok. Det same gjeld (meir eller mindre)

for redaktørane.

Her fylgjer eit utkast til dokumentstruktur for ei einspråkleg terminologiordbok:

```
<!-- The root element. Contains attributes for id and language -->
<!ELEMENT terminology entry+>
<!ATTLIST terminology id ID
                        xml:lang NMTOKEN >

<!-- The senses element is usually required, but is disallowed
      when the entry is an orthographical variant of another head word
-->
<!ELEMENT entry (common, senses?, changes)>
<!ATTLIST entry id NMTOKEN >

<!ELEMENT common (head, spoken?, infl, orth, qa) >

<!ELEMENT head EMPTY >
<!ATTLIST head pos ( s | a | adv | v | pron | pp | num | phrase )
              head CDATA > <!-- contains the word-count position
                                of the head of a phrase, used
                                when infl the head -->

<!ELEMENT spoken (#PCDATA) >

<!ELEMENT infl (#PCDATA) >
<!ATTLIST infl major CDATA
            minor CDATA>

<!-- The orthographical status of the word; it is indicated with
      the @status. Using the 'false' value, one can also include
      spellings of a term that fall outside accepted orthography.
      The @mainref should point to the main entry of the word
      when the entry is not the main entry; if so, don't use the
      senses element - it should only be used on the main entry. -->
<!ELEMENT orth (variant*) >
<!ATTLIST orth status ( main | false | abbr | variant ) "main"
            mainref CDATA >

<!ELEMENT variant EMPTY >
<!ATTLIST variant status ( false | abbr | variant )
            variantref CDATA >

<!-- QA field of the headword -->
<!ELEMENT qa EMPTY >
<!ATTLIST qa checked ( true | false ) "false"
            when CDATA
            who CDATA >

<!ELEMENT senses sense+ >

<!ELEMENT sense (topicClass, def?, examples?, synonyms?) >
<!ATTLIST sense idref CDATA
              status ( main | synonym | depr ) "main"
              mainref CDATA >

<!ELEMENT topicClass EMPTY >
<!ATTLIST topicClass toplevel CDATA
                  midlevel CDATA
```

```

        bottomlevel CDATA>
<!ELEMENT def    (#PCDATA) >
<!ELEMENT examples (example+) >
<!ELEMENT example EMPTY >
<!ATTLIST example xmplref CDATA >
<!ELEMENT synonyms (synonym+) >
<!ELEMENT synonym EMPTY >
<!ATTLIST synonym synref CDATA >
<!ELEMENT changes (change+) >
<!ELEMENT change EMPTY >
<!ATTLIST change when CDATA
                who CDATA
                what CDATA >

```

Det skulle gje eit dokument som ser omtrent slik ut:

```

<terminology id="SD-terms" last-update="" xml:lang="sme">
  ...
  <entry id="aktoriekti\S">
    <common>
      <head pos="s">aktoriekti</head>
      <spoken>IPA-transkripsjon av uttale?</spoken>
      <infl major="I" minor="o">riekti - riektái - rivttiide</infl>
      <orth status="main" >
        <variant status="false" variantref="akturiekti\S"/>
      </orth>
      <qa checked="true" when="DATO" who="SIGNATUR" />
    </common>
    <senses>
      <sense idref="3122" status="main">
        <topicClass toplevel="E" midlevel="E0000" bottomlevel="EN0010"
        />
      />
      <def>(ein definisjon)</def>
      <examples>
        <example
xmplref="file:///examples.xml#xpointer(//entry[@id='example-id'])"/>
        ..
      </examples>
      <synonyms>
        <synonym synref="aktovuoigatvuohta\S"/>
        <synonym synref="oktovuoigatvuohta\S" />
      </synonyms>
    </sense>
    <sense idref="1888" status="syn" mainref="aktogávperiekti\S" >
      <topicClass top="E" mid="E0000" botm="EN0010" />
    </sense>
  </senses>
  <changes>
    <change who="SIGNATUR" when="DATO" what="La til ny bøyingsinfo"/>
    ...
  </changes>
</entry>
  ...

```

</terminology>

Oppsummering: ein termartikkel inneheld ein felles del med oppslagsord, uttale, bøyning, ortografisk status og eit felt for kvalitetssikring. Etter denne felles delen kjem ei eller fleire tydingar, der kvar av tydingane inneheld ein referanse til omgrepsposten i navet, fagfelt, samt eventuelt ein definisjon, eitt eller fleire døme og eitt eller fleire synonym. Etter dei ulike tydingane kjem det til slutt ei liste over arbeidet med denne term-posten.

Om ein fører opp ortografiske variantar skal `orth`-elementet innehalda ein referanse til hovudoppslaget for ordet. I slike tilfelle treng ikkje posten innehalda eit `senses`-element. Dette skal skje automatisk, slik at korkje brukarar eller redaktørar treng tenkja på dei tekniske sidene ved det. Motsett vil ein i hovudoppslaget finna referansar til dei ulike variantane, med ein indikasjon om kva slags variant det gjeld. Slik er det mogleg å gå frå variant til hovudoppslag, og frå hovudoppslaget kan ein finna alle variantane. Dersom `orth`-elementet inneheld variantar, må `orth@status="main"` - det er berre hovudoppslaget som kan innehalda peikarar til variantar.

Synonym er i denne strukturen komplette oppslag, som berre er synonym i høve til eit bestemt omgrep, og då samtidig i høve til eit hovudoppslag. Dersom eit oppslagsord er synonym til eit anna, skal det i eit `sense`-element merkjast med `@status="synonym"`, og med `@mainref` som peikar til hovudoppslaget. `Sense`-elementet skal innehalda `topicClass`, og ingenting anna. Tilsvarande skal hovudoppslaget ha `@status="main"` i det relevante `sense`-elementet.

2.4. Eksempel

I arkitekturskissa i starten har eg plassert eksempla direkte under kvar einspråkleg fil, og at eksempla har berre éi retning - frå kjeldespråket til målspråka. Som nemnt i starten føreslår eg ein slik struktur fordi det kan vera problematisk å bruka omsette eksempl som om dei var originale. Om vi tenkjer oss ein term som finst på nordsamisk, norsk, svensk og sørsamisk, og originaleksemplet er skriva på nordsamisk, med omsetjingar til norsk, svensk og sørsamisk, er det sannsynleg at den sørsamiske eksemplsetninga *ikkje* er eit godt døme på korleis den sørsamiske termen skal brukast, og tilsvarande for norsk og svensk.

Med ein del omtanke og arbeid kan ein tenkja seg at omsetjingane blir så bra at dei faktisk kan brukast som eksempel på bruken av ein term på kvart språk. Eg er difor slett ikkje sikker på at den modellen eg skisserer over er den beste. I det noverande forslaget vil ein kunna kopiera og bruka eksemplsetningar om att i kvart språk, men det vil vera som ein kopi, og ikkje som ein referanse. Eller for å summera opp:

1. **Anten:** eksempla er språkspesifikke, og tilpassa kvart kjeldespråk. **Fordel:** potensielt høgare kvalitet på eksempla. **Ulempe:** mykje kopiering og duplisering av eksempla, tek mykje plass
2. **Eller:** alle eksempla samla i ei fil, eksempla er i prinsippet likeverdige på alle språk (ingen skilnad på kjeldespråk og målspråk når det gjeld eksempla). **Fordel:** tek mykje mindre plass, lettare gjenbruk. **Ulempe:** vil lett gje dårlegare eksempl i mange tilfelle

der eksemplet er omsett og ikkje skrive originalt på språket det gjeld.

Warning:

Dette valet må gjerast ganske snart. Ei mogleg løysing er å velja 1. for reine ordbøker (for slike er det eigentleg ingen diskusjon, det må gjerast som i 1.), og velja 2. for terminologi - i det minste i denne omgang. Då vil vi ha ein modell for både alternativa i framtida, og kan vurdere fordelane og ulempene betre seinare.

Uansett kva vi vel, bør eksempelsetningar og -fraser plasserast i ei eiga fil for lettare gjenbruk. Ein mogleg struktur for ei slik eksempelfil kan vera:

```
<!-- The root element. Contains attributes for id, and reference to
the classification system used. -->
<!ELEMENT termExmpl entry+>
<!ATTLIST termExmpl id ID
ClassSystem CDATA >

<!ELEMENT entry (topicClass, example+)>
<!ATTLIST entry id ID
origlang (sme | smj | sma | nor | sve | fin) >

<!ELEMENT topicClass EMPTY >
<!ATTLIST topicClass toplevel CDATA
midlevel CDATA
bottomlevel CDATA>

<!ELEMENT example (#PCDATA) >
<!ATTLIST example xml:lang (sme | smj | sma | nor | sve | fin) >
```

Med denne strukturen vil ei eksempelfil kunna sjå slik ut:

```
<termExmpl id="some-id" ClassSystem="system-ref">
...
<entry id="xyz" origlang="sme">
  <topicClass toplevel="R" midlevel="R8100" bottomlevel="RN8120" />
  <example xml:lang="sme">nordsamisk eksempel</example>
  <example xml:lang="nor">norsk eksempel</example>
  ...
</entry>
...
</termExmpl>
```

Oppsummering: alle omsetjingane eller versjonane av same eksemplet er samla i éin post, og originalspråket er markert som eit attributt på denne posten. Kvar slik post/eksempelsamling er spesifikk for eitt fagfelt, som er merkt ut med ein eigen tægg. Deretter kjem kvar (språk)versjon av eksemplet lista opp etter kvarandre. Meir burde det ikkje vera.