Innhold

Dokumentasjon av den norske stavekontrollen

1 Generell oversikt

1.1 Kort om stavekontrollen

Den norske stavekontrollen er et fri programvare-prosjekt for ordlister og stavekontroll for bokmål og nynorsk. Prosjektet er en fortsettelse av arbeidet til Rune Kleveland, og er kilden til alle fritt tilgjengelige stavekontroller for disse språkene. Prosjektet støtter ispell, aspell, myspell og hunspell.

Vedlikeholdet av orddatabasen gjøres ved hjelp av systemet speling.org. Vi bruker et e-postdrevet system der alle kan hjelpe til med å kvalitetsjekke og legge inn ord. Prosjektet inneholder

- en orddatabase med ordklasser og bøyningsinformasjon
- en synonymordbok
- orddelingsregler (i TeX-format, automatisk overført til OpenOffice.org-format).
- affix-regler (i ispell-format, automatisk overført til aspell- og myspell-format).
- ordfrekvensinformasjon i samarbeid med «An Gramadoir»-prosjektet og andre.

Systemfilene vedlikeholdes på alioth og speling.org. Alioth og Tyge blir sikkerhetskopiert.

1.2 Om dette dokumentet

Dette dokumentet skal hjelpe bidragsytere med å komme i gang eller gå dypere inn i stavekontrollsystemet, slik at generell og spesiell informasjon skal være lettere å finne og mer samlet. Dokumentasjonen skal også inneholde vyer, ønsker og endringsforslag slik at man har et utgangspunkt for å gå videre, og ikke er avhengig av de personene som er med for at systemet skal kunne drives også i framtiden, slik tilfellet var da de nåværende personene overtok etter flere års stillstand.

1.3 Nærmere oversikt

Hvorfor ikke bare hunspell?

Dog er det ganske irrelevant hva en kunne tenke seg å kaste så lenge det varierer hva de ulike programmene støtter av stavekontroll. Så vidt jeg vet støtter emacs kun ispell (har den lært aspell?), mens f.eks. KDE forstår aspell, og OpenOffice.org brukte før myspell, men nå hunspell. Det betyr at hvis en ønsker at stavekontrollen skal fungere i alle programmer må en støtte flere systemer inntil noen implementerer støtte for hunspell i alle programmer.

Plassering

Ordene ligger nå på to steder:

- www.speling.org, (TYGE) en database der vi mater inn ord ved hjelp av et e-postsystem.
- www.no.speling.org der alle ordene ligger i en fil som heter norsk.words sammen med en god del skript for å legge inn synonymer, orddeling, lage ispell-, myspell- og hunspell-versjoner o.a.

Håvard er kontaktperson mot www.speling.org.

Innhold og funksjon

På www.speling.org ligger diverse skript og dokumentasjon av speling.org-databasen.

De opprinnelige ordene fra norsk.words (som bygger på Rune Klevelands ordliste) er alle matet inn i databasen på www.speling.org. Ordbankens ord er derimot bare matet inn i databasen og ikke i ordlistefila (norsk.words).

Begge ordlistene (både i databasen og i norsk.words) har mange sammensatte ord for å omgå dårlig gjenkjenning av sammensatte ord.

Både vår ordliste og den lista vi har fra ordbanken er fullformsordlister. Derfor er også arbeidet med e-postinnmeldingene av nye ord viktige (se punkt 2.2.2).

Ordbankens liste er komprimert ned fra 1,2 mill ord med dobbeltoppføringer for tvetydige ord til det halve (600.000 ord) med bare enkeltoppføringer uten hensyn til betydning. Det er denne komprimerte versjonen som ligger i databasen. Dessverre ligger også enkeltord og tegn og annet der som må fjernes, men vi vet foreløpig ikke hvordan.

Egennavn legges inn i orddatabasen og vaskes ut ved hjelp av frekvensinformasjon, slik at selve stavekontrollen kun inneholder mye brukte navn, mens vi har notert ordklasse og korrekt stavemåte i databasen til framtidig bruk.

Angående rettigheter

Katalogvern

Kilde: http://www.lovdata.no/all/tl-19610512-002-041.html#43

§ 43. Den som frembringer et formular, en katalog, en tabell, et program, en database eller lignende arbeid som sammenstiller et større antall opplysninger, eller som er resultatet av en vesentlig investering, har enerett til å råde over hele eller vesentlige deler av arbeidets innhold ved å fremstille eksemplar av det og ved å gjøre det tilgjengelig for allmennheten.

Eneretten etter første ledd gjelder tilsvarende ved gjentatt og systematisk eksemplarfremstilling eller tilgjengeliggjøring for allmennheten av uvesentlige deler av arbeid som nevnt, dersom dette utgjør handlinger som skader den normale utnyttelse av arbeidet eller urimelig tilsidesetter frembringerens

legitime interesser.

Eneretten til et arbeid som nevnt i første ledd varer i 15 år etter utløpet av det år arbeidet ble fremstilt. Dersom arbeidet i løpet av dette tidsrom offentliggjøres varer vernet i 15 år etter utløpet av det år arbeidet første gang ble offentliggjort.

Er arbeid som nevnt foran, helt eller for en del gjenstand for opphavsrett, kan også denne gjøres gjeldende.

Bestemmelsene i §§ 2 andre og tredje ledd,[1] 6 til 8, 11a til 22, 25, 27, 28, 30 til 38b og 39h fjerde og femte ledd gjelder tilsvarende.

Avtale som utvider frembringerens rett etter første ledd til et offentliggjort arbeid kan ikke gjøres gjeldende.

1: Skulle være andre til fjerde ledd.

1.4 Historikk (usortert)

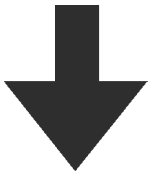
16.000 ord ble lagt inn av Tom Grydeland automatisk ved å opprette fullformer ut fra grunnformene av mange ord (et eksempel: finger).

Det mystiske formatet som rådatafilen i stavekontrollen var laget på, er « .sq». Det er et arkaisk komprimeringssystem kalt squeeze. Filen ble pakket ut med «unsq». Etter at vi fant ut det gikk vi over til å bruke gzip.

1.5 Flytdiagram

Oversikt over den veien ordene går fra noen melder dem inn til en ordliste til de kommer ut, som et flytdiagram.

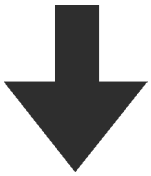
1 **Brukeren som vil melde inn ord.**



Metoder

- e-postinnmelding
- ord fra ordbanken og fra Rune Klevelands ordliste
- innmeldinger vha. skript direkte i databasen.
- ord kan legges inn i «missing.nb/nn» i cvs-en (se over)

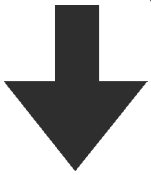
2 Tyges database



Overføringsmetode: Overføring vha. skriptet «speling2words», laget bl.a. av Petter Reinholdtsen. Skriptet ligger i no.speling.orgs cvs, i mappa src/spell-norwegian/scripts/
I dette skriptet filtreres noen av ordbankens ord bort, f.eks. ordendelser og tall (disse endelsene har alle bindestreker foran, og bør være enkle å sortere ut).

3 Fila «norsk.words» på no.speling.org

Databasen på www.speling.org (egentlig en ren tekstfil og ikke en ekte database).



Fila «norsk.words» på no.speling.org – Hovedkilden for alle de norske ordlistene
Overføring vha. Rune Klevelands skript «Makefile». Ordlistene, som så legges manuelt ut på alioth.org.

4 Ordlistefiler

Ferdige ordlistefiler for ispell/aspell/myspell/hunspell som kan offentliggjøres og tas i bruk av f.eks. OOo, Thunderbird, Firefox, Koffice o.a.

2 De ulike delene av systemet

2.1 Ordbankens liste

Ordbankens liste inneholder omtrent 1,6 millioner ord, som en fullformordliste. Lista er lagt ut til fri benyttelse under GPL på http://www.edd.uio.no/prosjekt/ordbanken. Sida krever at du registrerer deg.

Ordbankens ord mangler sammensatte ord og genitivsformer, derfor legger vi det inn i vår egen ordliste.

Vår ordliste er en fullformsordliste, det samme gjelder ordbankens liste. Sistnevnte vil i

komprimert form gå fra over en million ord til ca. 600 000 ord (doble oppføringer med ulik betydning er da fjernet).

Før ordbankens ord ble lagt inn, var omtrent 1/3 av ordene felles for både ordbankens liste og den frie stavekontrollen, 1/3 var bare i ordbankens liste og den siste tredjedelen bare i den frie stavekontrollen.

2.1.1 Overføring

Overføringa ble gjort ved at Petter hentet ut alle ordene for henholdsvis bokmål og vha. cut -fN (han husker ikke hvilke felt det var), kopierte filen over til Tyge og kjørte et skript for å generere fila som speling.org trenger, og kopierte den fila over til riktig sted på Tyge. **I neste runde bør dette automatiseres.**

Fra Petter:

> **Jeg fikk endelig tid til å sette meg ned for å mate fullformsordlisten** fra ordbanken inn i no.speling.org. Det er 628382 unike fullformsord for bokmål, og 469284 for nynorsk.
>
> For å gjøre dette måtte jeg først redigere fullform_bm.txt og fullform_nn.txt for å fikse en formateringsfeil i toppteksten. Dernest hentet jeg ut alle unike ord sortert uten å ta hensyn til store og små bokstaver (for å forenkle sammenligning med words.nb og words.nn) Til slutt kopierte jeg filene over til stavekontrolltjeneren tyge.sslug.dk og logget inn der for å gjøre kommandoene som la ordene inn i stavekontrollsystemet.
>
> Her er kommandoene brukt for bokmål. Tilsvarende ble gjort for nynorsk.
>
> ```
> vi fullform_bm.txt # Korriger topptekst
> grep -v '^*' fullform_bm.txt |cut -f3 | sort -uf > ordbok-nb-ord
> vi ordbok-nb-ord # Fjern oppføring for \r
> ```
>
> på tyge.sslug.dk:
>
> ```
> /opt/speling.org/bin/words_to_ds \
>   --authority 'Norsk ordbank ordbank_bm.zip 2007-07-09' \
>   --editor 'Norsk ordbank <ordbanken@iln.uio.no>' \
>   --status + \
>  < ordbank-nb-ord > /var/speling.org/nb/incoming.ds/start
> ```
>
> Fullformslista inneholder endel merkelige oppføringer, som #, $, A, a-, -abelt, etc. Jeg beholdt dem for å ha en komplett liste som vi heller får fjerne ved å gi ordene negative stemmer. Jeg er usikker på om dette er den beste måten å gjøre dette på.
>
> Neste steg blir så å mate ordene fra no.speling.org inn i stavekontrollpakken. Håper noen har

tid til å se på dette igjen snart.

2.2 http://tyge.sslug.dk

2.2.1 Om «databasen»

Dette er ingen ordentlig database, men en tekstfil. Det ligger en kopi av Tyges «database» på http://tyge.sslug.dk/~korsvoll/nb.speling.org/htdocs/status. Originalfila heter collected.ds, og ligger på tyges tjener (der man kan logge seg inn om man har en ssh-konto), en kopi av fila «/var/speling.org/nb/complete.ds» (på Tyge), som igjen lages utfra alle filene i katalogen «/var/speling.org/nb/data/» (og kanskje også den eksisterende utgaven av «/var/speling.org/nb/complete.ds»).

Når «update_dictionary» kjører, leser den «... /collected.ds« og alle filene i «... /incoming.ds/». Filene i «... /data/» er bare en sikkerhetskopi av de innkomne dataene. Det burde altså være nok å rette i «... /collected.ds».

Ord fjernes ved å stemme dem ut, det vil si at et flertall sier at ordet er feil. Det finnes dessuten et eget valg som heter «AUTHORITHY» som overstyrer andres innmelding.

Ordene legges hovedsaklig inn i databasen ved at man sender ordinformasjon per e-post, men man kan også legge inn ord direkte ved hjelp av ulike skript.

For å redigere kildefilene manuelt må en logge inn på tyge med ssh.

Ordene i ordlista blir oppført i ASCII-betisk rekkefølge med feltene ordnet i denne rekkefølgen:

ds_to_sds WORD CATEGORY CLASS ROOT DATE EDITOR

(Linja er hentet fra programmet «update_dictionary» iflg. Jacob Sparre Andersen.)

2.2.2 Legge inn nye ord i ordlista

For at nye ord skal havne i ordlista, kan man gjøre følgende:

- Send e-post til i18n-no@ med [ordliste] som emne med beskjed om hvilke ord som skulle vært akseptert av stavekontrollen.
- Noen av leserne på listen legger disse ordene inn i missing.{nb,nn}, alt etter hvilket av bokmål og nynorsk det gjelder.
- Magi skjer
- Ordlista blir oppdatert slik at ordene blir akseptert i neste utgave av stavekontrollpakkene.

Informasjon om dette ligger her: http://spell-norwegian.alioth.debian.org/dokumentasjon.html

2.2.3 E-post-rettesystemet på Tyge

For å korrigere ord bruker vi et e-postinnmeldingssystem. Disse havner databasen på Tyge, som igjen havner i ordlistene etterhvert. Via dette systemet kan de som melder seg på, sjekke ordene i ordlista ved å svare på regelmessige e-poster som sendes ut automatisk. Hvilke ord som sendes ut kan styres, men utvalget gjøres av et program. Som deltager kan man velge hvor mange ord per dag man vil ha. Også de bøyde ordene skal godkjennes.

De feltene vi foreløpig bruker er bare noen få av alle de som er dokumentert (se: http://no.speling.org/filformat.html), nemlig:
NB: Orddelinger i tyges database kan ikke endres automatisk i etterkant, bare manuelt. Det samme gjelder alle andre felter enn «STATUS» og «WORD».

Angående rekkefølgen på ordene i databasen: Hvis samme ord meldes inn to ganger samme dag vil ikke nyeste havne sist, da det bare er datoen som blir registrert. Ordene oppføres ellers i ASCII-alfabetisk rekkefølge med feltene ordnet i denne rekkefølgen:

ds_to_sds WORD CATEGORY CLASS ROOT DATE EDITOR
(linje fra programmet «update_dictionary»).

2.2.4 Dokumentasjon

Jacob har vel sendt en lenke til en presentasjon han holdt om hvordan speling.org-systemet fungerer (finn den og legg den her!). Tror det kan være en grei start for å lære hvordan det fungerer. Ellers er kildekoden tilgjengelig for nedlasting.

2.2.5 Struktur

Eksperimenter med Perl viser at det ikke er spesielt vanskelig å hente ut relevant informasjon, da tekstfila er godt strukturert, iallfall innad i hver enkelt ordblokk av typen:

WORD: ord
STATUS: +
EDITOR: meg
DATE: i dag

Det finnes også andre felt, men de har vi foreløpig ingen metode til å utnytte, derfor brukes de ikke i innmeldingene per e-post.

Reglene for de ulike feltene finnes oppsummert her: http://no.speling.org/filformat.html

2.2.6 Felter som er i bruk

Vi hentet ut statistikk over hvilke felt som er i bruk og hvor ofte hver av dem er blitt brukt. De fire første er brukt hver gang (og er obligatoriske), mens de andre er dels ganske sjeldne.

DATE	1 206 265
EDITOR	1 206 265
STATUS	1 206 265
WORD	1 206 265
AUTHORITY	628 388
COMPOSITE-WORD	24 968
HYPHENATION	19 657
ROOT	18 094
CONJUGATION	17 645
CLASS	17 735
CONJUGATION-RULE	17 048
CORRECTION	6 039
COMMENT	3 566
SYNONYM	398
EXAMPLE	76
CATEGORY	13
DESCRIPTION	7
SOURCE	3
ANTONYM	1

Noen av disse vil bare ligge lagret enn så lenge, da vi ikke har noe å bruke dem til, men f.eks. informasjon om CLASS og CONJUGATION-RULE kan brukes til å fastslå hvilke ord som kan settes sammen om noen tar dette i bruk. Men det vil på den annen side også kreve at dette er angitt for mer enn bare noen få ord hvis det skal være nyttig.

De som har fått lov til det av redaksjonen, kan også bruke »AUTHORITY«-feltet til å understreke viktigheten av at ordet blir strøket fra listen – eller er korrekt.

Hvordan kan vi stemme ut ord som er registrert med AUTHORITY-feltet? Ved å late som om du selv er en like stor autoritet. (Eller ved å gå inn og slette de relevante »AUTHORITY«-feltene manuelt i de relevante filene.)

2.3 no.speling.org

2.3.1 Generelt

Prosjektet på no.speling.org er fra Klevelands tid, og inneholder byggesystemet for ordlistene, mens Tyge-databasen inneholder ordene slik de er kommet fra ordbanken og e-postrettesystemet.

2.3.2 Å bygge ordlista

I hovedsak skal denne fila ta seg av å bygge ordlistefiler for aspell, myspell, ispell og hunspell ut fra bl.a. fila «norsk.words». Men koden er altfor kryptisk, da det er uklart hva hver kodesnutt gjør. Vi bør ha som siktemål å «løse hva den gjør» og dokumentere den, eller starte fra bunnen av med et nytt skript som vi kan dokumentere og forstå.
En kort veileding til å bruke skriptet og byggavhengigheter finner du her:
http://no.speling.org/lagNyeOrdlister.html
En noe kortere veiledning på engelsk finnes her:
http://no.speling.org/developer.html

2.3.3 Ekstrakommandoer fra «Makefile» (Klevelands skript)

Bruk kommandoen:

make speling-new.nb speling-new.nn

for å sammenligne stavekontrollen med no.speling.ord-dataene

Kommandoen

make update-from-spelingorg speling-new.nb speling-new.nn

vil endre norsk.words, og legge inn bokmål- og nynorsk-ordene fra no.speling.org som mangler i norsk.words. Deretter vil den lage litt statistikk over differansen.

2.3.4 Frekvensinformasjon

Om et ord skal legges inn ordlista avgjøres av frekvensinformasjonen knyttet til de enkelte ordene. Sjeldne ord blir ikke tatt med.

Frekvenskategoriene for hvert ord er regnet ut fra absolutt frekvens. I koden

if (s<=5) {t=s} else {t=-9+15*log(1+log(s))}

er «s» antall ganger order har forekommet totalt.

Frekvensordlista til Kleveland var generert fra ca 4 mill norske nettartikler med lengde mer enn 1000 tegn og «få» spesielle ord. Ut fra dette laget han automatisk en liste med 100 000 forslag til nye rotord delt med gamle orddelingsmønstre.

Vi har tilgang til endel frekvensinformasjon for norske ord fra http://helmer.aksis.uib.no/nta. Der er blant annet en frekvensliste med 465.000 ord. Den bør kunne brukes til å oppdatere frekvenstallene i norsk.words, men etter å ha tittet på dette en stund så stopper det hele opp. Hva betyr egentlig frekvenstallet i norsk.words? Hvordan oversetter jeg fra frekvensinformasjonen tilgjengelig i f.eks. http://torvald.aksis.uib.no/nta/ord10k.txt, der forekomsten er oppgitt i promille og over til tallet som brukes i norsk.words? Det ser ut til å være et tall i området 0–31.

2.3.5 Manglende ord

Filene «missing.nb» og «missing.nn» inneholder ord som ennå ikke er lagt inn. Disse må manuelt legges inn i fila «norsk.words», ingen automatikk der.

Alle disse filene ligger i CVS-en til spell-norwegian-prosjektet på alioth:
http://cvs.alioth.debian.org/cgi-bin/cvsweb.cgi/src/spell-norwegian/?cvsroot=spell-norwegian

2.3.6 Annet

Skriptet «bokmaal» kan slå opp ord på http://www.dokpro.uio.no fra kommandolinja.

2.4 «speling2words»

2.4.1 Plassering

[SVN]/Speling.org/src/spell-norwegian/scripts

2.4.2 Kort overblikk

I hovedsak legger dette skriptet ord fra Tyges «database» inn i fila «norsk.words». Samtidig utelater den alle ord vi foreløpig ikke kan behandle pga. begrensninger ved byggesystemet (se neste punkt).

Meningen med skriptet er omtrent følgende: Hent inn norsk.words, del i ord, ordtypemarkør og statistikktall. Hent inn Tyge-ordene også og sammenlign dem med norsk.words. Legg så inn de ordene fra Tyge som mangler i norsk.words, utelatt alle sære ord vi foreløpig ikke kan håndtere (f.eks. ord med bindestrek eller trippelkonsonanter).

2.4.3 Bruk

Oppskriften på: https://lister.ping.uio.no/pipermail/i18n-no/2007-November/005610.html
sier: «make update-from-spelingorg speling-new.nb speling-new.nn»
Skriptet kjøres via en makefile.

2.4.4 Nærmere om skriptet

- Noen byggeavhengigheter måtte løses før skriptet kunne brukes, se: http://no.speling.org/lagNyeOrdlister.html
- Skriptet trenger minst 700 MB, helst 1 GB, for å kjøres innen rimelig tid (ikke bruk vekselminne), fordi det er såpass mange ord som skal hentes inn i minnet og behandles.
- Kommandoen «make update-from-spelingorg speling-new.nb speling-new.nn» eller bare «make update-from-spelingorg» skal virke, gjør ikke den feilen å bruke makefila i undermappa scripts, det virker ikke.
- «speling-good.nb/nn», som er nevnt i skriptet opprettes av skriptet underveis.
- Alle skriptfilene er kodet i Latin1, da det er det eneste ispell forstår. Dette påvirker også aspell, myspell og hunspell.
- Linje 114 er en lengre streng som skal filtrere ut uønskede ord, som antatt. Meningen med linja:
  return unless m/^[a-cçd-eéêèëf-oóôòp-uüv-zæäøöåA-CÇD-EÉÊÈËF-OÓÔÒP-UÜV-

ZÆÄØÖÅ]+$/;
er å bare bruke de ordene som inneholder gyldige norske tegn, og ingen andre. Dette fordi bl.a. en god del tall og uttrykk (setninger, altså mer enn ett ord) o.a også er med i ordbankens liste. Disse behøver vi ikke og/eller kan vi ikke bruke i stavekontrollen, så de filtrerer vi ut.

- Kommandoen «make distcheck» sjekker at alt lar seg bygge (men ikke om innholdet er i orden).
- Filene «nn.phonet.dat» og «nb.phonet.dat» er symbolske lenker til samme fil, nemlig «aspell-phonet.dat».
- Ord på færre enn 3 bokstaver (altså ett eller 2 tegn) fjernes av «speling2words», dette da de fleste (eller alle) disse ordene er lagt inn allerede, dessuten finnes en god del enkelttegn i ordbankens liste som vi slik filtrerer ut. Men med funksjonen på linje 114 (se over) er ikke dette siste lenger viktig.
- Som før nevnt er rekkefølgen: Ordbankens ordliste(1) –> Tyges database(2) –> Norsk.words(3). Fra (1) må endel ord filtreres bort (se nedenfor). Til (2) retter vi ved å sende inn rettede og sjekkede ord via e-post til en egen e-postliste. Alt på (3) ligger også på Tyge.
- Følgende ord filtreres bort (se kildekoden til «speling2words» der dette er dokumentert nærmere):
  - Enkelttall, enkelttegn
  - Ord med punktum, hermetegn, apostrof, skråstrek og mellomrom, da ordlistene ikke klarer å håndtere dette (dette gjelder uttrykk som «hoppe over bord» o.a.). – Som nevnt tas bare de ordene med som inneholder de tegnene som er listet opp ovenfor. (Hvis noen mener noen er utelatt, kan lista forlenges)
  - Ord med bindestrek (som ikke skriptet vårt håndterer).
  - Vi velger å utelate ord som inneholder orddelingsregler, da vi fant mange feil der og ikke har tid nå til å sjekke alle. Dette er uansett nye ord, så vi får bare færre nye ord på denne måten (det vil si færre av ordbankens ord og færre av de som er lagt inn siden sist via e-post-innmeldingssystemet). Petter bygger. (?)
- Da vi kjørte skriptet for å lage en ny ordliste i mars, ble f.eks. 18.947 ord avvist (se ovenfor for hva slags ord dette er).
- Om «sub load_norsk_words»: Henter inn Klevelands ord og legger frekvenstallene inn i en liste. «Headers» = alle kommentarer.
- Det må være minst to som har stemt på en orddeling for at den skal foretrekkes, jo flere stemmer jo bedre.

3 Ønsker, gjøremål og spørsmål

Her kommer en liste med både langsiktige og kortsiktige oppgaver som vi etterhvert som de besvares og løses fletter inn i de andre delene av dette dokumentet.

1. Hunspell

1. Finne ut hvordan hunspell virker (formatet m.m.).

2. Hva gjør hunspell, har den en fullformsordliste, eller ekspanderer den selv en komprimert liste?
3. Hva må gjøres rent konkret med ordbankens liste for at hunspell skal kunne bruke dem?
4. Gaute: Hvilket format behøver OOos ordlister? (har han lenker?)

2. Klevelands skript og tilhørende filer

1. Det skriptet vi i dag bruker for norsk.words er for dårlig dokumentert og mange steder må man gjette og prøve seg fram for å finne ut hvordan det hele virker. Rune Kleveland selv husker heller ikke hvordan alt ble laget, da dette var for 10 år siden.
2. Skriptet er til dels obskurt og ikke fullstendig dokumentert. Hvis noen vil bygge dette om, f.eks. i Python, så trenger vi informasjon om hva ispell/aspell/myspell/hunspell har av formatkrav. Vi har prøvd å søke litt, men fant foreløpig ingen nøyaktig nok dokumentasjon på nettet.
3. Kleveland husker ikke selv alt om hvordan skriptet som lager ordlistene virker. (Repetisjon av punkt 1 over)
4. Hva brukes ordklasseinformasjonen til? (avis, normal osv). Brukes noe av informasjonen til å gjette på sammensatte ord? Hvorfra får vi informasjon om orddeling?
5. Hva betyr \J osv. i nb-no.dic-fila?
6. Hvilke problemer som ennå må løses, f.eks. bindestrek i ord. Bindestreksproblematikken kan kanskje løses ved å lage en erstatning for Makefile. Foreløpige forsøk på å endre koden der så det virker (fra Petters side) har ikke virket.
7. Bruken av «ssed» i skriptet «speling2words» kan forbedres til å sjekke om den genererte fila er i orden (ikke tom) før den erstatter den opprinnelige.
8. I makefila kan ssed-kommandoene forbedres slik at den sjekker om de filene som er opprettet er i orden før den legger inn de foreløpige filene som nye filer.
9. Fila «nb_no.aff» ser ut til å være en liste over ordendelser. Hvor kommer denne informasjonen fra? Fra Norsk.words? I så fall hvordan og hva brukes den til?
   /j er ikke meta-info. Det er ordendelsesinfo. Les nb.aff.in for å se hva /j betyr. Let etter 'flag *j:'. Tenk på det som komprimering, der ordene Volvo, Volvoene, Volvoen, Volvoer, Volvos og Volvoens slås sammen til Volvo/AEGJ. Du kan bruke ispell til å ekspandere en slik komprimert versjon:
   % echo Volvo/AEGJ | ispell -d nb -e
   Volvo Volvoene Volvoen Volvoer Volvos Volvoens %
10. Til side 6: Det eneste Petter fant for å forklare hva dette tallet representerer er følgende kommentar i toppen av fila: «Each word is hyphenated at compound points, and has a frequency indicator essentially of log log type.»
    Hva betyr «log log type» her, og hvordan oversetter man fra frekvens i prosent eller promille og til denne «log log type»?
11. **Egennavn**: Skal vi endre statistikkfunksjonen til en mye enklere og mer gjennomsiktig variant? Et eksempel: For egennavn så har vi det slik at vi gir dem en egen tagg: Finn **Det** forholdstallet vi har i dag som antallet egennavn i de ferdigbygde ordlistene delt på antallet

i databasen totalt. La oss ta utgangspunkt i de n% vi da får, og i fremtiden bare ta med de n% mest populære navnene. Så kan en eventuelt heller justere denne prosenten i fremtiden for å få med færre eller flere navn.

12. Apropos **egennavn**, kan det være en idé å ta kontakt med Statistisk sentralbyrå og be om en liste over navn brukt i Norge, med frekvensinformasjon, og så mate de mest brukte inn i stavekontrollen med og uten genitivs-s?
13. Sjeldne ord bør med såfremt de er kontrollert og godkjent.
14. Hvordan beregnes dagens statistikk, og ut fra hvilke kilder?
15. Fra byggeskriptet (Klevelands skript) på no.speling.org skal vi fjerne:
    1. Funksjonen for å velge ord utfra statistikk (vanskelig forståelig og ikke så veldig nyttig)
    2. Dobbeltoppføringer med k-markøren (konservativt bokmål: foreløpig følgende: taksten, torv, torva, torvene, torvenes = duplikater som nå fjernes) o-markøren (oljebransjen, foreløpig bare: asfalt, et duplikat som fjernes) og M-markøren (matematisk, eneste ord som nå fjernes: polynom, et duplikat som fjernes).

3. Struktur

1. Sikkerhetskopiering: Med en felles plassering gjøres dette kanskje enklere? Mer sårbart kanskje også …
2. Vi har mange plasseringer av stavekontrollen. Vi har nettstedene:
    1. synonymer.merg.net
    2. no.speling.org
    3. speling.org
    4. tyge (og noe på Håvards konto på tyge)
    5. og kanskje flere jeg har glemt

    Kunne flere av disse slås sammen? Kanskje tyge-filene kunne legges inn på no.speling.org og også de ferdige ordlistene kunne ligge der. Er det noe som taler imot dette?

4. Ordbanken

1. Har ordbankens ordliste en ekspansjonsfunksjon innebygget? Hvis ja, må dette sammenlignes med hva hunspell bruker.
2. Kanskje ordbankens metainformasjon kan brukes bedre?

5. E-postinnmeldingssystemet på Tyge

1. Dokumenter Tyges vedlikeholdsmetode (eller finn en lenke til informasjon).
2. Forslag: Status=rettelse som en ny verdi for dette feltet.
3. Legg inn nærmere informasjon om et skript på Tyge som lar en legge inn mange ord på en gang, skal være postet på i18n-lista.
4. Databasen på www.speling.org (egentlig en ren tekstfil og ikke en ekte database) har en

del feil og mangler, kanskje noen av dem kan rettes på om man leser dokumentasjonen. Eks.: Hvordan fjerne ord? Hva gjør vi med dobbeltoppføringer?

5. Hadde det ikke vært bedre om vi gjorde om «databasen» på Tyge til postgresql? Terje sa han kunne se på muligheten for det. Tekstfila som sådan er iallfall tilpasset en slik omgjøring, men det er sikkert ikke optimalt med:

   E-post --> tekstfil --> database

   bedre er nok

   E-post --> Database

   Men jeg regner med det krever en endring av Tyge. Prøvde å finne dokumentasjon på hvordan systemet på tyge virker, fra noe sendes per e-post (for det er vel den innmeldingsmetoden vi baserer oss på?) til den ligger i fila, men fant ingenting. Ligger dette noe sted jeg har oversett? Hvis denne overgangen er ukomplisert, kanskje den enkelt kunne bygges om til å legge alt i en database isteden? Muligens kan tekstdatabasen være grunnlaget (da den er enklere og sikrer oppetid), mens den ekte databasen brukes når vi bygger pakker?
6. Orddelinger i Tyges database kan ikke endres automatisk i etterkant, bare manuelt. Det samme gjelder alle andre felter enn «STATUS» og «WORD». Hva kan vi gjøre for å endre dette? Et problem med systemet er at hvis man sender samme ord to ganger samme dag, så trenger man en tilleggsregel. Om ordene legges inn med det sist innmeldte sist, så kan rekkefølgen være et tilleggskriterium. Denne regelen kan brukes til å rette opp feiloppføringer i andre felt enn WORD. Rettelser for dette feltet ser vi fortsatt ingen automatisk løsning på, men kanskje man kunne lage et system for å sende ut ord til korrektur der automatikken ikke strekker til. Nøyaktig hvilke ord dette er, må fastlegges. Reglene kunne vi tenke oss brukt til en automatisk bearbeidelse, for eksempel å hente ut ord fra databasen.
7. Hva betyr: EDITOR: i Rune Klevelands ordliste? Hører ikke den informasjonen til i feltet SOURCE?
8. Hva om CORRECTION er det samme som WORD?
9. Hva om en av oppføringene er mer omfattende (flere angitte felt, f.eks. informasjon om ordklasser i den ene av dobbeltoppføringene, men ikke de andre)
10. Hvilken oppføring skal ha forrang om vi har flere, kan dette gjøres automatisk?

    Løsningsforslag

    Regler:

    Ranger viktigheten etter authority-lista (det følgende er alle de oppføringene vi fant brukt)

    AUTHORITY: Norsk ordbank, ordbank_bm.zip 2007-07-09 (Likestilte, da dette er manuelt sjekket i en ordbok)

    AUTHORITY: Norske synonymer, blå ordbok av Dag Gundersen

    AUTHORITY: Norsk Ordbok

    AUTHORITY: Norsk ordbok

    AUTHORITY: Norsk Ordbok med 1000 illustrasjoner

    AUTHORITY: Norsk ordbok med 1000 illustrasjoner, annen utg.

    Alt annet

I en ordblokk: Hvis WORD og CORRECTION er like, dropp linja CORRECTION
Hvis EDITOR og WORD i en ordblokk er lik EDITOR og WORD i en annen ordblokk, la da nyeste ordblokk være autorativ (bruk DATE-feltet). Hvis EDITOR ikke er angitt prioriteres ordet lavere enn andre oppføringer av samme ord av andre EDITOR.

11. Vi som lager stavekontrollen bør ha en plass å notere slike ord som vi er enige om ikke skal være godkjente ord, og som har vist seg problematiske tidligere, slik at vi sikrer at de ikke kommer tilbake som godkjente ord. Det de gjør i DSDO er å ta en beslutning i redaksjonsgruppa og deretter sender inn en post om beslutningen med «AUTHORITY: DSDO-redaksjonen».
12. Dokumentasjonen vår mangler litt når det gjelder egennavn. Dette bør være en egen tagg, og sjeldne ord skilles ut ved hjelp av statistikkinformasjon.

6. Annet

1. Legge inn lenke til og siste versjon av ordbankens liste i no.speling.orgs cvs og de tilhørende nettsidene. Disse ligger her: http://www.edd.uio.no/prosjekt/ordbanken/
2. Foreløpig får alle ordlistene det samme (de samme ordene, samme format osv.). Dette kan eventuelt endres senere.
3. Hvordan overfører vi ordene fra speling.org til norsk.words?

4 Viktige lenker

https://alioth.debian.org/scm/?group_id=30577
http://no.speling.org
http://www.linux.com/articles/51675?tid=93
http://wiki.services.openoffice.org/wiki/Dictionaries
http://no.speling.org/lagNyeOrdlister.html
http://www.edd.uio.no/prosjekt/ordbanken
http://wiki.debian.org/SpellNorwegian/Møteplan
http://alioth.debian.org/frs/?group_id=30577
http://en.wikipedia.org/wiki/MySpell
http://lingucomponent.openoffice.org
http://hunspell.sourceforge.net