SCARRIE leksikalsk ressurs, LMF-kompatibel versjon

Den originale SCARRIE-databasen ble laget for å bli brukt til automatisk korrekturlesing av norsk bokmål (nob). Denne LMF-kompatible versjonen er laget av Koenraad De Smedt ved Universitetet i Bergen med utgangspunkt i det opprinnelige leksikonet. Dette foregikk innenfor rammen av META-NORD-prosjektet, og målet var å gjøre ressursen enklere å dele og gjenbruke.

Norsk bokmål har et stort antall alternative (sidestilte) former. SCARRIE-databasen er så vidt vi vet den første og så langt eneste leksikalske ressursen på norsk som inneholder informasjon om hvilket stilnivå eller "subnorm" hver enkelt ordform tilhører. Databasen er laget som støtte til et korrekturlesingsprogram, slik at dette kan finne ordformen som best passer med det overordnede stilnivået i en tekst. SCARRIE er ikke oppdatert på offisielle normeringer som er foretatt etter 1998, og materialet er ikke fullstendig manuelt kontrollert.

Lisens

Forfattere og institusjoner som skal krediteres i forbindelse med SCARRIE er: Victoria Rosén og Koenraad De Smedt ved Universitetet i Bergen, og Torbjørn Nordgård ved Norges teknisk-naturvitenskapelige universitet.

Format

SCARRIE-leksikonet er et fullformsleksikon der ordformer som tilhører samme lemma er gruppert sammen, mens selve lemmaet ikke inneholder noen informasjon.

Denne versjonen er konvertert fra de originale IDF-filene til en fil som er kompatibelt med LMF-formatet (Lexical Markup Framework, se http://www.lexicalmarkupframework.org/). Den er validert mot DTD_LMF_REV_16.dtd. Alt ligger i en fil med en enkelt *LexicalResource*, som inneholder sju *Leksikon* elementer med følgende navn:

- *prefixes* - prefikser
- *suffixes* - suffikser
- *gramwords* - grammatiske ord
- *gramwords2x* - flere grammatiske ord
- *abbrevwords* - ord som forekommer i forkortelser
- *idiomwords* - ord i flerordsuttrykk
- *main* - vanlige ord fra de åpne ordklassene

- Antall ordformer i *prefixes* 327
- Antall lemmaer i *prefixes* 327
- Antall ordformer i *suffixes* 562
- Antall lemmaer i *suffixes* 125
- Antall ordformer i *gramwords* 707
- Antall lemmaer i *gramwords* 539
- Antall ordformer i *gramwords2x* 39
- Antall lemmaer i *gramwords2x* 8
- Antall ordformer i *abbrevwords* 23
- Antall lemmaer i *abbrevwords* 23
- Antall ordformer i *idiomwords* 811
- Antall lemmaer i *idiomwords* 811
- Antall ordformer i *main* 359. 684
- Antall lemmaer i *main* 72.617
- Antall ordformer totalt 362.153
- Antall lemmaer totalt 74.450

Hver leksikalske enhet (*LexicalEntry*) har et tomt lemma (*Lemma*) og minst en ordform (*WordForm*). Vanlige *WordForm*-elementer har trekk (*Feat*) med følgende attributter (*att*):

- *writtenForm* - ordform
- *corrStyle* - stiltype/stilverdi (til bruk i korrektur)
- *featureList* - grammatiske trekk for sammensetningsanalyse
- *replacement* - erstatningsord i en gitt stil
- *synCat* - grammatiske trekk for setningsanalyse

Noen oppføringer, særlig prefikser og ordformer som forekommer i forkortelser eller idiomer har verken *replacement*- eller *synCat*-attributtet. Bruken av attributtene er forklart i *SCARRIE deliverable 3.3.1 Tagset*.

Om SCARRIE-prosjektet

SCARRIE (Scandinavian Proofreading Tools) var et forsknings- og utviklingsprosjekt (LE3-4239) innenfor språkteknologi under EUs *Telematics*-program. Prosjektet startet 1. desember 1996 og ble avsluttet 28. februar 1999. Koordinator for prosjektet var WordFinder Software AB (Växjö, Sverige). De andre hovedpartnerne i prosjektet var Universitetet i Bergen, Institutionen för lingvistik ved Uppsala Universitetet, Center for Sprogteknologi (København) og Svenska Dagbladet (Stockholm).

Prosjektets mål var å lage korrekturlesingsverktøy for dansk, norsk og svensk. For å nå dette målet ble det innenfor rammene av prosjektet forsket på effektive feildetekterings- og korreksjonsmekanismer for de skandinaviske språkene. Ressurser for disse språkene har blitt integrert i CORRie-plattformen, som opprinnelig ble laget for nederlandsk av Cognitech. Prototypen til korrekturlesingssystemet inneholder flere lingvistisk baserte feildetekterings- og feilkorreksjonsmekanismer både på ord- og setningsnivå.

Arbeidet med norsk ble koordinert ved Universitetet i Bergen. Victoria Rosén var forskningsleder, mens professor Koenraad De Smedt sto for den vitenskapelige koordineringen.

Den norske delen av SCARRIE-prosjektet fokuserte på avansert stavekontroll for bokmål. Systemet bruker fullformsordlister kombinert med spesielle mekanismer for å håndtere flerordsuttrykk og for å gjenkjenne nye sammensetninger, egennavn og andre ord som ikke finnes i ordlistene. I samarbeid med NTNU har det blitt laget en tilpasset fullformsordliste. Ordformene i denne listen er merket med informasjon om lemma (grunnform), standard, stilnivå eller skriftnorm, morfosyntaktiske trekk og mulige erstatningsord. Forutsigbare feilstavinger er supplert med anbefalte korrekturord.

Nye sammensetninger gjenkjennes ved hjelp av en analyse basert på regler som er laget ved Universitetet i Oslo. Ord som ikke finnes i ordboken og som er sannsynlige feilstavinger prosesseres av en korrekturmekanisme som blant annet baserer seg på lydlig (fonologisk) likhet. Dessuten ble det utviklet en robust grammatikk for å finne og korrigere bestemte typer feil som ikke kan håndteres på ordnivå, for eksempel feil i samsvarsbøying. Systemet er innrettet slik at forslag til korrektur velges så det passer inn i den skriftlige normen som det aktuelle dokumentet er skrevet på (fra konservativt til radikalt bokmål).

Mer informasjon om prosjektet er arkivert under
http://ling.b.uib.no/projects/scarrie/