ANNA BRAASCH

Udnyttelse af maskinlæsbare ordbogsdata til maskinoversættelse

Abstract

The multilingual machine translation project EUROTRA works with a transfer-based system. For the running prototype, which is going to be produced in the last phase of the project, each language group has to reach the goal of approximately 20.000 dictionary entries.

In the Danish EUROTRA group, we have started a small-scale pilot project. The purpose is to investigate the possibilities for reuse of existing machine-readable dictionary data in the extension and improvement of the monolingual Danish dictionary.

We have two different data sets available: The machine readable version of the official Danish Spelling Dictionary (RO2) and a small subset of records from the Danish-French Dictionary Base (DFOB).

The two main points of the pilot project are:

1. To get an idea as to how the dictionary information should be systematized for it to be useable in different applications;
2. How we can exploit existing machine-readable dictionary information in the coding process of the general vocabulary.

The state of the art:

1. We have made a comparison between information types needed for the EUROTRA Translation System (ETS) and the information types represented in the two data sets.
2. We carried out some tests with a subset of available dictionary data to estimate how much of the information needed by the ETS can be deduced and/or converted without too extensive programming effort.

Proceedings of NODALIDA 1989, pages 127-137

1 Indledning

Kort skitse af projektets formål:

Vi har for nylig indledt et pilotprojekt i den danske **EUROTRA**-gruppe, der har til formål at vurdere hvilke muligheder vi har for automatisk/halvautomatisk hhv. manuel tilpasning af ordbogsdata, der er kodet til andre formål end maskinoversættelse. Pilotprojektet koncentrerer sig først og fremmest om at finde frem til de oplysningstyper, hvis bearbejdelse ikke kræver kompliceret programmering.

Metode:

Vi sammenligner de oplysningstyper, som vort oversættelsessystem har brug for i sine leksikalske regler (morfologiske, syntaktiske og på senere tidspunkt semantiske informationer) med oplysningerne i de valgte ordbøger og finder derved frem til, hvilken delmængde af de oplysninger, vi har behov for, der principielt kan findes i disse ordbøger.

Forventet resultat:

1. overblik over, hvilke krav der kan stilles til systematisering af ordbogsoplysninger, hvis genbrug for forskellige formål skal være mulig;
2. rent konkret: lemmaordbogsindgange til brug i oversættelsessystemet.

2 Pilotprojektets baggrund

Da **EUROTRA** arbejder med et niveaudelt oversættelsessystem, bruges der tilsvarende niveauspecifikke etsprogsordbøger til analyse og generering for hvert sprog, samt et sæt transferordbøger for hvert sprogpar.

I stedet for at lave de etsprogede niveauspecifikke ordbøger hver for sig, ønsker vi at fremstille én niveauuafhængig ordbog ved at samle samtlige oplysninger, der er relevante på hvert enkelt niveau, i én ordbog. Herfra skal systemet så for hvert opslagsord selektere de oplysninger, der er behov for på det pågældende niveau i oversættelsesprocessen. En ordbog, der er opbygget på denne måde kalder vi for **lemmaordbog**.

Den viste ordbogsindgang består af to dele, jf. Figur 1:

1. Selve den kørbare leksikalske regel, der indeholder 'de indre oplysninger' der anvendes af oversættelsessystemet;
2. De ikke kørbare 'ydre oplysninger' til brug for leksikografen, indledt med udkommenteringssekvensen '%%', hvilke omfatter
   (a) administrative oplysninger (koder, dato, kilde, kommentarer) og
   (b) 'pragmatiske' oplysninger (definition, eksempler).

Ordbogsindgangene er kodet efter det samme princip som grammatikkens regler, men i stedet for at indeholde generelle oplysninger om fx sætnings- eller

```
1. 'antal_n3' = [cat=n, scat=specifier, part=no, level=zero,
    dalu='antal', darno=n3,
    ers_frame=comp0, dapform1=no, dapform2=no, dapform3=no, dapform4=no,
    daisframe=arg0, dapargl=no, daparg2=no, daparg3=no, dapparg4=no,
    flex_type=fx3, dagd=neut, dcons=1, oc=no, infl=sub_root, term=xx0].
2. %% Coder: anna   22-May-89
   %% Source: Ph3 corpus
   %% DEF: en m[ngde enheder af ngt. t[lleligt
   %% Comments: NDO: uden plur.
   %% Examples: TV-sSt <nm> er en tysk satellit med 3 TV-kanaler og et ukendt
   %%             antal radiofonikanaler. 362 text4
```

Figur 1: Eksempel på en ordbogsindgang fra den danske lemmaordbog.

frasestruktur, indeholder den specifikke leksikalske oplysninger om de enkelte opslagsord.

3 Ordbogskodning

I oversættelsesprojektets tredje og sidste fase, som vi befinder os i, skal det forventede antal ordbogsindgange nå op på ca. 20.000 enheder ialt (nu: 4600), deraf 6.000 tilhørende det generelle ordforråd (nu: 3.900), 14.000 **EUROTRA**-termer (disse kodes ikke helt efter de samme principper som det generelle ordforråd, men dette vil jeg ikke komme nærmere ind på her).

Der er altså en hel del kodningsarbejde tilbage, især hvis vi også tager med i betragtning, at de allerede kodede indgange skal opdateres i takt med grammatikkernes udbygning. I den afsluttende (indeværende) projektfase opprioriteres ordbogsarbejdet.

Vi har behov for effektivisering af arbejdet, hvilket kan gøres på forskellige måder, dels ved organisationsmæssige ændringer, dels vha. automatisering af kodningsarbejdet på de områder, hvor det praktisk er mulig.

Kodningen kan til en vis grad automatiseres ved brug af **makroer** i inddateringsproceduren. Dette letter leksikografens arbejde ved tastatur og skærm.

Den manuelle ordbogskodning rummer muligheder for både skrivefejl og indholdsmæssige fejl. For at begrænse disse fejl har vi udarbejdet faste ordbogsmakroer i UNIX-editoren **emacs**. Disse små programmer omfatter forskellige faciliteter, såsom konsistens- og validitetscheck, prompt for oplysningtype til inddatering, liste af lovlige værdier for hver informationstype (attribut) osv.

Desuden anvendes ETS (**EUROTRA** Translation System) regelfortolkeren til at kontrollere, at ordbogsindgangen er kodet i overensstemmelse med regelsættet for formel og indholdsmæssig beskrivelse af leksikalske regler.

Der gives fejlmeddelelse om syntaksfejl (manglende separatorer, parenteser etc.), samt hvis et attributs kodede værdi ikke er element i det pågældende attributs værdiliste (dvs. ikke lovlig værdi). Der gives ingen fejlmeddelelse, hvis attributtets værdi er en fri streng (fx opslagsord eller præposition).

En anden metode til at effektivisere ordbogsarbejdet er at udnytte oplysninger om ord fra maskinlæsbare ordbøger. Dette er emnet for pilotprojektet. Formålet er at undersøge, hvilke muligheder vi har for at udnytte maskinlæsbare ordbogsdata ved udbygningen af lemmaordbogen. Vi arbejder jo i forvejen på den måde, at vi definerer opslagsordet og derefter 'slår op' manuelt i forskellige trykte, monolingvale ordbøger, primært i Nudansk Ordbog, Retskrivningsordbogen og Dansk Sprogbrug. Dette betyder at vi henter de relevante informationer fra ordbøger, der er lavet til andre formål end maskinoversættelse.

Arbejdet med disse ordbøger viser, at oplysningerne ikke behøver at være formaliserede, og at de ikke altid er eksplicit til stede. Andre gange følger ordbogen ikke helt den prædefinerede formatbeskrivelse. Der kan fx være afvigelser fra den erklærede forkortelsesliste, eller det kan forekomme, at den søgte oplysning ikke er repræsenteret i den ordbog, man har slået op i.

Dette er oftest ikke noget større problem for ordbogsbrugeren, men et maskinoversættelsessystem skal have entydige, eksplicitte og udtømmende oplysninger til rådighed i sine ordbøger, udtrykt i den valgte formalisme, som i vort system hedder E-(**EUROTRA**)formalismen.

Ordbogskoderens arbejde ved anvendelse af de traditionelle, trykte ordbøgers oplysninger til udbygning af lemmaordbogen består altså af flere trin:

Søgning og udvælgelse, systematisering, validitetscheck, konvertering til E-formalisme og selve indtastningen af ordbogsartiklen.

EUROTRAs ordbøger har naturligvis en række særlige træk, der er afhængige af maskinoversættelsessystemets krav, men de grundlæggende leksikalske og kontekstuelle oplysninger svarer til dem, der også findes i de nævnte trykte ordbøger.

4 Pilotprojekt — indledende overvejelser

Da vi rent faktisk 'genbruger' en del oplysninger fra trykte ordbøger, er spørgsmålet nu, hvordan vi kan inddrage maskinlæsbart ordbogsmateriale i arbejdsgangen for ordbogskodning: Hvilke trin i processen kan automatiseres, i hvor høj grad kan vi inddrage maskinkraft til at udføre opgaver automatisk eller interaktivt, og hvilke oplysninger skal indføres rent manuelt?

Vi har valgt p.t. at inddrage to forskellige datasæt i pilotprojektet, som lægger hovedvægten på forskellige oplysningstyper, svarende til deres leksikografiske koncept og anvendelsesområde:

Vi har erhvervet den maskinlæsbare version af **Retskrivningsordbogen** (herefter forkortet: **RO2**). Dette materiale er med få undtagelser indholdsmæssigt identisk med den trykte udgaves alfabetiske afsnit 'Ordbog a-å', som omfatter opslagsord tilhørende det almindelige danske ordforråd. Materialet indeholder hovedsagelig oplysninger om selve opslagsordet uden (større) kontekst, fx ordklasse, bøjningsformer, sammensætningsformer, sideformer. Der er dog artikler, der også indeholder kommentarer og eksempler, fx ved forholdsordene 'af' og 'ad' på grund af disses komplekse betydningsstruktur.

Den maskinlæsbare version (**RO2**) adskiller sig på to væsentlige punkter fra den trykte udgave:

1. Hver oplysning indledes af en feltkode, der angiver det pågældende felts oplysningstype. (Feltkoden overtager dermed skriftgradsskiftets rolle som adskiller af oplysningstyper.)

2. Der er yderligere nogle få oplysninger i den maskinlæsbare version (fx ordklassebetegnelse for navneord og udsagnsord), der kun implicit er til stede i den trykte version.

Vi har af ordbogsgruppen på Handelshøjskolen i København (HHK) fået stillet et udvalg af ordbogsposter fra **Dansk-fransk ordbase** (herefter forkortet: **DFOB**) til rådighed. Ordbasen indeholder ordbogsindgange fra Blinkenberg-Høybye's Dansk-fransk Ordbog, der er en oversættelsesordbog med kildesproget dansk og målsproget fransk.

Materialet er på HHK blevet optisk indlæst og derefter lagt ind i databaseposter. Det af os udvalgte materiale omfatter kun et mindre antal substantiver begyndende med 'afv-' og transitive verber begyndende med 'af-'.

Ved dette materiale er det tale om kontrastiv behandling af de danske opslagsord; foruden basisoplysninger vedrørende ordklasse og bøjning er der til de fleste ord yderligere materiale, omfattende betydningsopdeling, definitioner, semantiske oplysninger (emneområde), brugsrestriktioner, frekvens, eksempler og naturligvis de tilsvarende franske oversættelser.

Den franske del af materialet er p.t. ikke inddraget i pilotprojektet, men bliver det på længere sigt.

Begge datasæt foreligger som strukturerede ASCII-tekstfiler og vi har fået detaljeret formatbeskrivelse af dem begge.

Det første trin i pilotprojektet var at skaffe overblik over, hvilke af de oplysninger, **EUROTRA**-oversættelsessystemet har brug for, der er repræsenteret

```
'antal_n3' = [cat=n, scat=specifier, part=no, level=zero,
 dalu='antal', darno=n3,
 ers_frame=comp0, dapform1=no, dapform2=no, dapform3=no, dapform4=no,
 daisframe=arg0, dapargl=no, daparg2=no, daparg3=no, dapargl=no,
 flex_type=fx4, dagd=neut, dcons=1, oc=no, infl=sub root, term=xx0].
%% Coder: anna   22-May-09
%% Source: Ph3 corpus
%% DEF: en m[ngde enheder af ngt. t[lleligt
%% Comments: NDO: uden plur.
%% Examples: TV-sSt <nm> er en tysk satellit med 3 TV-kanaler og et ukendt
%%           antal radiofonikanaler. 362 text4
```

HORD:	an‖tal	--->	HORD:	tal
HOKL:	no		HOKL:	no
HTGN:	,		HENT:	-let,
HSMS:	antals-,		HNFL:	tal -lene;
HSMX:	antalsbegr[nsning		HEKS:	1200-tallet

Figur 2: Lemmaordbogsindgang og tilsvarende poster fra **RO2**.

i de to datasæt, hvordan de er kodet, og på hvilken måde vi kan udnytte eller overføre disse til lemmaordbogen.

I fremstillingen i Figur 2 vil jeg primært koncentrere mig om **RO2**, og for overskuelighedens skyld viser jeg kun substantivernes kodning som eksempler.

Lemmaordbogsindgangen i Figur 1 er forsynet med markering af de værdier, der indsættes af systemet (ubrudt understregning), samt oplysninger der er automatisk indsat fra **RO2** (markeret med stiplet linje). Figuren viser desuden sammenhængen mellem en ordbogsindgang i E-formalisme og den artikel i datasættet, som oplysningen hentes fra. (I dette tilfælde kræves 2 opslag i **RO2**, da et sammensat opslagsords bøjning som regel findes i ordbogen under det sidste led.)

5 Pilotprojektets indeværende fase

Efter at have lavet en liste over de oplysningstyper for de enkelte ordklasser, som vi skal have adgang til i lemmaordbogen, og efter at have sammenlignet denne liste med de oplysninger, der er til stede i de maskinlæsbare materialer, har vi vurderet at mulighederne for 'genbrug' er følgende:

1. En del oplysninger kan konverteres direkte fra det maskinlæsbare materiale til oplysninger i E-formalisme vha. editor-makroer (emacs), fx ordklassebetegnelse i feltet *HOKL* (='ordklasse'): no → cat (='ordklasse') = n;

2. Oplysninger kan konverteres vha. programmeret opslag i **RO2** og sammenligning af den maskinlæsbare oplysning med listen over de værdier, der kan repræsentere informationstypen (attributtet) i en lemmaordbogsindgang; fx kan de kodede bøjningsendelser for substantiver i feltet *HNFL* (='navneord flertal') sammenholdes med listen over værdierne for attributtet *flex_type*(='bøjningstype').

Første kolonne i tabellen i Figur 3 angiver koden i E-formalisme (udsnit: *flex_type*=fx1 ... fx5 for regelmæssigt bøjede substantiver), anden kolonne: eks. på opslagsord i ental, ikke genitiv. Kolonnerne 3 og 4 viser, hvordan **RO2** anfører de pågældende bøjningsendelser. Kolonne 5 indeholder den automatisk ekspliciterede flertalsendelse i bestemt form (ved regelmæssig dannelse er denne ikke til stede i **RO2**). Kolonne 6 viser attributtet *d_cons* (='fordobling af slutkonsonanten'); denne oplysning bliver ekspliciteret i konverteringsforløbet, udskilt fra felt *HENT* eller *HNFL*.

3. Udledning af en oplysning der kun er implicit til stede i materialet, fx kønnet for et substantiv, udledes af *HENT*-feltets (='navneord ental') sidste tegn og konverteres til eksplicit oplysning i E-formalisme i attributtet *dagd* (='Danish gender').

Foruden automatisk konvertering hhv. udledning af oplysninger har vi andre muligheder:

```
*****************************************************************************
      s.ind-nge      s.def          pl.ind-nge       pl.def-nge
FX1:  stol           -en            -e               -ene      (d_cons=no)
      hat            -ten           -te              -tene     (d_cons=t)
      bord           -et            -e               -ene      (d_cons=no)
      blik (1)       -ket           -ke              -kene     (d_cons=k)
*****************************************************************************

*****************************************************************************
      s.ind-nge      s.def          pl.ind-nge       pl.def-nge
FX2:  rede           -n             -r                -rne
      vindue         -t             -r                -rne
      NB! ingen fordobling af konsonant
*****************************************************************************

*****************************************************************************
      s.ind-nge      s.def          pl.ind-nge       pl.def-nge
FX3:  virksomhed     -en             -er              -erne     (d_cons=no)
      anorak         -ken            -ker             -kerne    (d_cons=k)
      abstrakt       -et             -er              -erne     (d_cons=no)
      stakit         -tet            -ter             -terne    (d_cons=t)
*****************************************************************************

*****************************************************************************
      s.ind-nge      s.def          pl.ind-nge       pl.def-nge
FX4:  film            -en                -              -ene      (d_cons=no)
      bit             -ten               -              -tene     (d_cons=t)
      forslag         -et                -              -ene      (d_cons=no)
       ar (2)         -ret               -              -rene     (d_cons=r)
*****************************************************************************

*****************************************************************************
FX5:   s.ind-nge      s.def           pl.ind-nge       pl.def-nge
       bager          -en               -e              -ne
       NB! Ingen fordobling af konsonant
*****************************************************************************
```

Figur 3: Udsnit af tabel til sammenligning af koder i E-formalisme og kodningen af bøjningsendelser i **RO2**.

4. Automatisk opslag og selektering af visse oplysninger og automatisk el. interaktiv overførsel af disse;

5. Mulighed for interaktiv kodning: valg af eksempler samt tilpasning af eksempler (tilføjelse, sletning etc).

De sidste to muligheder bliver især aktuelle ved import af **DFOB**'s danske oplysninger: **DFOB** indeholder naturligvis langt flere oplysninger end **RO2**, bl.a. kommentarer og eksempler. Det har dog vist sig, at selv om dette materiale er fyldigt, har vi brug for begge datasæt, da vi har fundet visse tilfælde, hvor selve

```
HORD:   afviser#sten                OPSL      afviser/sten
HOKL:   no                          UNDERK    c,
HENT:   -en,                        HENV      v. afviser  (2).
HNFL:   afvisersten -ene            ADM       #OPSL (afvisersten)
                                              #DATO 1975/10/01
                                              #BEAR (o)
```

Figur 4: Post fra **RO2** *og* **DFOB**.

a)

```
'afvigelse_n1' = [cat=n, scat=no, scat=no, level=zero,
 dalu='afvigelse', darno=n1,
 ers_frame=f30, dapform1=no, dapform2='fra', dapform3=no, dapform4=no,
 dalsframe=arg12, dapargl=no, daparg2='fra', daparg3=no, dapargq=no,
 flex_type=fx2, dagd=comm, d_cons=no, oc=yes, infl=root, term=xx0].
%% Coder: anna   21-Apr-89
%% Source: makrotest
%% DEF:
%% Comments:
%% Examples: afvigelser fra regelen (NDO)
```

b)

afvigelse -n, -r.

HORD:	afvigelse
HOKL:	no
HENT:	-n,
HNFL:	-r

c)

afvigelse *c* (*-r*) (= *vigen til side, fjernen sig fra*) (1) déviation* (2) écart (*fra* de, d'avec) (3) (*om magnetnål, opt.*) déclinaison*; (*astron.*) (4) (*i bane*) perturbation* (5) (*stjerners*) aberration*; (6) (*drt.*) dérogation* (*fra* à). II) (= *forskellighed*) (7) divergence* (8) différence* (9) (*i meninger, opt.*) dissidence*.
~ fra betingelser (*fx. i romburs*) ¶ discordance*; *~ fra dagsordenen* (6), incident; *danne en ~ fra n.* faire dérogation à qc.; *~ fra emnet* digression*; *en ~ mellem ... og ...* un écart entre ... et ..., un écart qui sépare ... de ...; *~ fra princippet* (*undert.*) tempérament du principe; *projektils ~* (1) (2); *~ fra regelen* (1) (2), anomalie*, exception*; (*undert.*) tempérament de la règle; *udvise en ~ fra* présenter un écart avec.

c)

```
OPSL     afvigelse
UNDERK   c
FLEKS    -r
GLOS     = vigen til side, fjernen sig fra
NUM      1
OVERS    déviation *
NUM      2
OVERS    écart
PART     fra  de, d'avec
NUM      3
GLOS     om magnetnåle, opt.
OVERS    déclinaison * ;
GLOS     astron.
NUM      4
GLOS     i bane
OVERS    perturbation *
NUM      5
GLOS     stjerners
OVERS    aberration * ;
NUM      6
ENHE     drt.
OVERS    dérogation *
PART     fra  à
BET      ¶
GLOS     = forskellighed
NUM      7
OVERS    divergence *
NUM      8
OVERS    différence *
NUM      9
GLOS     i meninger, opt.
OVERS    dissidence * .
DAORDF   slåk fra  betingelser  (fx. i romburs)
ENHEEDO  com.
FRORDF   discordance * ;
DAORDF   slåk fra  dagsordenen
FRORDF   (6), incident;
DAORDF    danne  en slåk fra n.
FRORDF   faire dérogation à qc.;
DAORDF   slåk fra  emnet
FRORDF   digression * ;
DAORDF   en slåk  mellem  ... og...
FRORDF   un écart entre... et..., un écart qui sépare
         ... de ...;
DAORDF   slåk fra  princip et
GLOSEDO  undert.
FRORDF   tempérament du principe;
DAORDF    p roje ktils  slåk
FRORDF   (1) (2);
DAORDF   slåk fra  regelen
FRORDF   (1) (2), anomalie * , exception * ;
GLOSFFO  undert.
FRORDF   tempérament de la règle;
DAORDF    udvise  en slåk fra
FRORDF   présenter un écart avec.
ADM      $OPEL (afvigelse)
         $DATO 1975/10/01
         $BEAR (o)
```

Figur 5: Overblik: En lemmaordbogsindgang — og det tilsvarende opslagsord fra de valgte ordbøger.

den søgte oplysning (fx et sammensat substantivs flertalsendelse, eksempelvis ved ordet 'afvisersten') mangler i **DFOB** (se Figur 4).

Om vi vil bruge begge datasæt parallelt, eller bruge **RO2**'s ordklasse- og bøjningsangivelser til at verificere hhv. komplettere oplysningerne vi har hentet fra **DFOB**, har vi endnu ikke taget stilling til.

Figur 5 omfatter foruden en autentisk lemmaordbogsindgang (a), den tilsvarende trykte ordbogsartikel fra Retskrivningsordbogen og ordbogspost fra **RO2** (b), samt ordbogsartiklen fra Dansk-fransk Ordbog og databaseposten — uden typografiske styretegn — fra **DFOB** (c). (Bemærk det specielle tegnsæt i **DFOB**-posten!)

Vi har en del uløste grundlæggende spørgsmål i forbindelse med automatisk udfyldning af lemmaordbogsindgangen, som det fremgår af markeringerne i Figur 5:

- Kan et homografnummer hhv. et betydningsnummer fra **RO2** eller **DFOB** udnyttes til at definere opslagsordets læsninger for oversættelsessystemet, svarende til attributtet *darno* (='Danish reading number');

- Kan valensrammerne (*ers_frame* hhv. *daisframe*) til et ord udledes af de danske ordforbindelser der er anført i artiklen som brugseksempler for ordet og udtrykkes i E-formalisme vha. en konverteringsrutine;

- Kan det antages, at alle obligatoriske valenser er repræsenteret, at de valensbundne præpositioner (*daparg/dapform1–4*) er til stede, at eksemplerne er valgt og listet ud fra et bestemt leksikografisk princip, som kan danne grundlag for den påkrævede (sandsynligvis interaktive) formalisering af oplysninger?

I indeværende fase af pilotprojektet har vi undersøgt materialet fra **DFOB** for at skaffe os overblik over, hvilke muligheder vi har for at løse disse spørgsmål. Materialet indeholder kontekstseksempler for næsten hvert kodet dansk opslagsord. Eksemplerne er ofte sætninger, hvor ordet optræder med alle sine valensbundne led (både med og uden præpositioner). Det tosprogede materiale er kodet udfra et kontrastivt synsvinkel med fokus på danske ord og disses betydningsopdeling (og deres franske ækvivalenter).

Ofte er der knyttet definitioner og kommentarer til ordet, især fx vedr. syntaktisk subkategorisering, semantiske selektionsrestriktioner og pragmatiske forhold. I **DFOB** er disse oplysninger ikke formaliserede. Vi mener dog at kunne hente de manglende oplysninger for kodningen af lemmaordbogsindgange efter at have udfoldet ordbogsartiklen helt, dvs. oprettet en ny artikel for hver nummererede betydning af opslagsordet, hvor ordet automatisk er blevet indsat (eller dets korrekt bøjede form) i stedet for tilde osv.

Vi er i gang med at udfolde artiklerne og vælge de oplysninger fra, som vi ikke vil bearbejde på nuværende tidspunkt. Desuden udarbejdes der en liste over de søgekriterier, samt tabeller mm. der bliver brug for når vi henter og konverterer oplysningerne fra DFOB. I første omgang udføres aktionerne interaktivt.

6 Pilotprojektets næste fase

Det andet trin i pilotprojektet skal koncentrere sig om at undersøge forholdet mellem på den ene side den nuværende manuelle del af kodningsarbejdet i forbindelse med udbygning af lemmaordbogen, og på den anden side den mulige effektivisering der kan påregnes ved overførsel af oplysninger fra maskinlæsbare ordbøger. Ved vurderingen skal der naturligvis også tages hensyn til kvalitetsaspekter.

7 Konklusion

Pilotprojektets emne og formål er bestemt af et konkret behov i **EUROTRA** oversættelsessystemet. Det grundlæggende spørgsmål er, om behovet til dels kan dækkes ved at automatisere arbejdet med eksternt ordbogsmateriale, der kan erhverves i maskinlæsbar form.

Det er ikke uproblematisk at forsøge at sammenføre oplysninger fra maskinlæsbart ordbogsmateriale, der dels er kodet til forskellige formål og ud fra forskellige leksikografiske principper, dels fremtræder teknisk forskelligt, fx hvad tegnsæt (jf. posten fra **DFOB**) eller datastruktur/format angår.

Andre problemer er fx, at de indhentede oplysninger kan være inkompatible indbyrdes eller med **EUROTRA**s system, eller at det maskinlæsbare materiale indeholder sammenlægning af forskellige indholdstyper i ét oplysningsfelt (i **RO2**). Selve selektionen af relevante oplysninger fra et meget omfattende materiale er også forbundet med væsentlige principielle beslutninger.

Disse aspekter kan selvsagt ikke glemmes, kun gemmes, da vi først og fremmest ønsker en generel vurdering af mulighederne for udnyttelse af eksisterende maskinlæsbare ordbogsdata i **EUROTRA-DK**'s ordbøger.

Vi har behov for en ny type ordbog, der indeholder udtømmende, systematiserede og formaliserede oplysninger, der kobler morfologiske, syntaktiske, semantiske og evt. pragmatiske oplysninger sammen til en helhed — den fulde beskrivelse af opslagsordet. Dette vil kunne danne basis for en etsprogsordbog, der kan bruges som kildesprogsordbog ved opbygning af to- eller flersprogsordbøger.

Da den ordbog vi har beskrevet her, omfatter veldefinerede oplysningstyper og formaliserede oplysninger, vil den efter automatisk kontrol og konvertering (der kan indbygges i systemet) også kunne blive velegnet til andre formål.

En tak til:

Bodil Nistrup Madsen, HHK, for at venligst have stillet **DFOB**-materialet til rådighed; Bente Maegaard og Henrik Selsøe Sørensen, EUROTRA-DK, for at deltage i drøftelserne af pilotprojektet, samt Ole Norling-Christensen, Gyldendals Ordbøger, for omhyggelig korrekturlæsning og kommentarer til manuskriptet.

Litteratur

Erik Brun. 1980. *Dansk Sprogbrug.* Gyldendal, København.

Dansk-fransk Ordbog. 1975. [Ved] Andreas Blinkenberg og Poul Høybye. Nyt Nordisk Forlag Arnold Busck, København.

DFOB: Et udvalg af poster fra Dansk-fransk Ordbogsbase i maskinlæsbar form, fra Institut for Datalingvistik ved Handelshøjskolen i København.

Lauterbach, Birgitte. 1988. Dansk-fransk Ordbogsbase, manual. Red: Dansk-fransk Leksikografi. Internt papir ved Handelshøjskolen i København..

Nistrup Madsen, Bodil. 1987. Dansk-fransk Ordbogsbase. Ordbøger i Danmark, En oversigt [Udgivet af DANLEXgruppen]:124–131, København.

Nudansk Ordbog. 1987. [Red.:] Becker-Christensen, Christian m.fl. 13. udgave, 2. oplag. Politikens Forlag, København.

Retskrivningsordbogen. 1988. Udgivet af Dansk Sprognævn, 1. udgave 5. oplag. Gyldendal, København.

Retskrivningsordbogen på edb. 1988. Udgivet af Dansk Sprognævn. København.

RO2. 1988. Den maskinlæsbare version af Retskrivningsordbogens alfabetiske del (1. udgave 5. oplag med rettelser). København, Dansk Sprognævn/Klokker & Bro Aps.

EUROTRA-DK
Københavns Universitet
Njalsgade 80
DK-2300 København S.