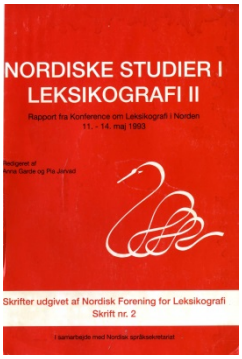
NORDISKE STUDIER I LEKSIKOGRAFI

Titel: Når maskinen tager en på ordet - ordbogsarbejde for maskinoversættelse

Forfatter: Anna Braasch

Kilde: Nordiske Studier i Leksikografi 2, 1993, s. 29-37
Rapport fra Konference om leksikografi i Norden, 11.-14. maj 1993

URL: http://ojs.statsbiblioteket.dk/index.php/nsil/issue/archive



Søgbarhed
Artiklerne i de ældre Nordiske studier i leksikografi (1-5) er skannet og OCR-behandlet. OCR står for ’optical character recognition’ og kan ved tegngenkendelse konvertere et billede til tekst. Dermed kan man søge i teksten. Imidlertid kan der opstå fejl i tegngenkendelsen, og når man søger på fx navne, skal man være forberedt på at søgningen ikke er 100 % pålidelig.

Når maskinen tager én på ordet
- ordbogsarbejde for maskinoversættelse

Anna Braasch

Dette indlæg beskæftiger sig med nogle aspekter af ordbogsarbejdet som er særegne for maskinoversættelse af fagsproglige tekster. (For korthedens skyld bruges i det følgende de engelske forkortelser MT/HT for hhv. maskinel og traditionel (human) oversættelse).

Center for Sprogteknologi deltager i et større forskningsprojekt, Oversættelse af Fagsproglige Tekster (OFT), der støttes af Statens Humanistiske Forskningsråd. CST bidrager med erfaringer fra maskinel oversættelse. Inden for projektets rammer sammenligner vi den viden om sproget, omverdenen og fagområdet, som en oversætter gør brug af i sit arbejde, med de oplysninger, som et MT-program skal have adgang til. Det gøres ved at udføre modelforsøg: et testkorpus, der er velafgrænset med hensyn til både emneområde (domæne) og teksttype, oversættes med et MT-program.

I modelforsøget lægges der særlig vægt på arbejdet med afgrænsing af ordforråd og nødvendige oplysningstyper. Oversættelsesresultaterne synliggør fejl og mangler i ordbogen, som derefter bliver systematisk opdateret og testet. På denne måde bliver effekten af ændringer og tilføjelser målelig. Processen gentages indtil det ønskede resultat er opnået. Pilotarbejdet afsluttes med at sammenfatte de hyppigste problemtyper, der kunne relateres til ordbogen og de informationer, der var nødvendige at få systemet til at oversætte tilfredsstillende. Vore konklusioner vil også kunne inddrages i arbejdet med hjæpemidler til traditionel oversættelse.

Korpus

Teksterne, der tilsammen udgør korpuset, stammer fra domænet automobilmekanik, der hører under fagområdet mekanik, teksttypen er instruktionsbog. Testkorpuset til oversættelse er sammensat af udvalgte afsnit fra vejledninger til forskellige biltyper. Kommunikationsvejen går fra fagfolk (bilproducent) til lægfolk (bruger), hvilket bestemmer teksternes (mellemsvære) fagsproglighedsgrad. Sprogparret er engelsk-dansk. Korpus og domæne er beskrevet i detaljer i Braasch (1992a).

Oversættelsessystemet

Projektet er baseret på EUROTRA prototypen, der et modulært opbygget, transferbaseret MT-system (jf Copeland et al. 1991). Det har to slags sproglige komponenter: grammatikmoduler og ordbogsmoduler, af hvilke vi her vil fokusere på ordbogsmodulet for engelsk-dansk oversættelse. Dette modul består af to ordbogstyper:

- de monolingvale ordbøger: engelsk analyseordbog (bruges i tekstreceptionen) og dansk genereringsordbog (bruges ved tekstproduktionen)
- den bilingvale ordbog (egentlig en ækvivalensliste), der danner bindeleddet mellem den engelske analyse- og den danske genereringsordbog.

Ordbogsmodulet

Systemarkitekturen er indrettet på en sådan måde, at man ved oversættelse af fagsproglige tekster kan tilkoble én eller flere ordbogsdatabaser. Dette betyder i praksis, at man kan sammensætte ordbogsmodulet af to slags komponenter:

- en stor almensproglig komponent (for hvert af de involverede sprog), der indeholder det generelle ordforråd og
- mindre komponenter (satellitordbøger), der er skræddersyede til bestemte emneområder, fx petrokemi, automobilmekanik, miljøteknik etc.

Der er væsentlige fordele ved denne metode:

- man kan vælge en hensigtsmæssig kombination af almensproglig + fagsproglig ordbog (eller evt. flere fagordbøger); disse udgør tilsammen det aktive ordbogssystem
- man kan prioritere rækkefølgen af ordbogskomponenterne og derved bestemme at et ord først skal slås op fx i den biltekniske ordbog; kun hvis ordet ikke er fundet der søges der i de andre ordbøger i overensstemmelse med den fastlagte prioriteringsrækkefølge.

Ordbogsmodulet bliver altså sammensat og afstemt efter domæne og teksttype, hvilket forbedrer oversættelseskvaliteten betydeligt.

Oversættelse af fagsproglige tekster

Der stilles tre elementære krav til en oversætter:

- at *forstå* kildetekstens ord og mening
- at *kende ækvivalenterne* til tekstens ord på målsproget
- at kunne *danne korrekte og adækvate sætninger* af disse ord på målsproget.

Et maskinoversættelsessystem skal også opfylde disse krav. MT- systemet udfører oversættelsen vha. tre tilsvarende moduler:

- *analyse*, med opslag i den engelske ordbog (reception)
- *transfer*, med opslag i den engelsk-danske ækvivalensliste
- *generering*, med opslag i den danske ordbog (produktion).

Et centralt problem ved oversættelse er tekstens flertydighed. Flertydigheden kan forekomme på flere planer:

- leksikalsk betinget flertydighed, når et ord har flere oversættelser
- strukturel eller syntaktisk flertydighed fx ellipse, anafor
- kommunikativ flertydighed, når kommunikationssituationen er afgørende for valget af oversættelsen.

Disse typer kan også optræde i kombination med hinanden, hvilket komplicerer oversættelsen yderligere.

En humanoversætter kan løse disse problemer ved at læse teksten i dens helhed, se på de eventuelle illustrationer, der visualiserer det faglige indhold, slå op i forskellige ordbøger og sidst, men ikke mindst ved at indhente manglende domæneviden hos fagfolk eller fra fagleksika. Vores MT-system er ikke i stand til selv at uddrage domæneviden fra den tekst, den er i gang med at oversætte (modsat mennesket, der husker det læste!). Illustrationer kan programmet heller ikke gøre brug af. Hvis maskinen altså skal kunne løse sin opgave som oversætter, skal dens ordbog eller database indeholde en stor mængde information, som ikke er repræsenteret i traditionelle ordbøger.

Når maskinen slår op i sit ordbogsmodul, skal den have adgang til de sproglige oplysningstyper, til almenviden om omverdenen og til domæneviden inden for fagområdet. Programmet kan kun arbejde med de informationer der er anført for det pågældende opslagsord i ordbogen - *maskinen tager én på ordet*.

Krav til MT-ordbøger

Kravene til MT-systemets ordbøger kan sammenfattes i nedenstående stikord.

- Generelle krav:
  den leksikalske beskrivelse af ord og andre opslagsenheder skal være eksplicit, entydig, udtømmende (i overensstemmelse med grammatikkens krav til oplysningstyperne) og formaliseret (svarende til systemets brugersprog).
- Specielle krav i EUROTRA MT-systemet:
- entydiggørelse af opslagsenheden skal så vidt mulig ske i analyseordbogen (hvilket her vil sige i den engelske komponent)
- den bilingvale ordbog baseres altid på 1:1 relation mellem kildesprogs- og målsprogsenhed.

Ordbogsindgange i monolingvale MT-ordbøger indeholder i lighed med trykte standardordbøger følgende oplysningstyper:

- opslagsenhed/ord og aktuelt homograf- hhv. tydnummer, den såkaldte reading
- ordklasse og ordklassespecifikke egenskaber (fx ved danske substantiver: køn)
- morfologiske egenskaber (fx bøjningsmønster, fordobling af slutkonsonant)
- syntaktiske egenskaber (fx restriktioner vedrørende ordets syntaktiske funktion i forhold til den pågældende ordklasses generelle egenskaber; eksempelvis hvis et adjektiv

ikke kan optræde i prædikativ stilling)
- kombinatoriske egenskaber (fx valens eller brug af støtteverbum ved verbalsubstantiver, fx : 'foretage et valg')
- semantiske træk (primært ved substantiver, fx 'concrete') og selektionsrestriktioner (dvs. krav, som opslagsordet stiller til de semantiske træk ved sine valensbundne led)

En væsentlig teknisk forskel fra standardordbøgers makrostruktur er, at i denne type MT-ordbøger kodes én ordbogsindgang (dvs. en databasepost) for hver 'reading' (jf. ovenfor) af et givet ord. Medens således i COBUILD verbet *ENGAGE* anføres med 7 betydninger + 1 henvisning i én samlet ordbogsartikel, kodes der 7 (evt.8) indgange i vores ordbase.

Ordbogsarbejdet tager udgangspunkt i de ovenfor opregnede krav. I det efterfølgende gives nogle eksempler på oversættelsesproblemer, der opstår når de kodede informationer i systemets ordbog ikke er præcise eller detaljerede nok.

Problemtyper med relation til maskinens ordbog

Nogle af de hyppigste oversættelsesmæssige problemer er forbundet med maskinens krav om entydige og udtømmende leksikalsk beskrivelse af fx

- sammensætninger
- kollokationer mm.
- flertydige ord

Sammensætninger

I fagsprog er substantiverne hyppigt sammensatte af to eller flere led. I engelsk frembyder dette et specielt problem, idet sammensatte ord ofte består af en række substantiviske elementer adskilt af blanktegn. Problemet har to vigtige aspekter i MT:

- Det er svært automatisk at afgrænse det samlede kompositum, da der ikke optræder fugemærke eller andre specielle markører og da stavemåden er ret svingende: elementer står efter hinanden med blanktegn (ordmellemrum) imellem eller de er forbundet med hinanden med bindestreg, men to eller flere elementer kan også være samskrevne som ét ord. (Problemet kan forværres af at en sammensætning skrives forskelligt i samme tekst !)
- Nogle sammensatte ord kan oversættes kompositionelt, dvs. element for element til dansk (med evt. et indsat fugetegn eller strukturelle ændringer), andre kan kun delvis eller slet ikke håndteres på denne måde. Ingen trykt almen- eller fagsproglig ordbog kan opregne sammensætninger udtømmende, først og fremmest af omfangsmæssige grunde, men også fordi sammensætning er den mest produktive måde at danne nye ord på.

For at illustrere omfanget af dette problem ved udarbejdelse af ordbøger til maskinoversættelse vises her det engelske substantiv *'wheel'* ('hjul') og et udvalg af de sammensætninger med 'wheel' der forekommer i vore tekster.

(1) *'wheel nut'*: Andet led har to ækvivalenter, nemlig den almensproglige 'nød' (frugt) og 'møtrik', hvoraf det andet klart hører til i fagområdet mekanik. I dette tilfælde kan sammensætningen oversættes kompositionelt til *'hjulmøtrik'*.

(2) *'wheel brake'*: Andet led har kun én oversættelse, 'bremse' til dansk, der også hører til i domænet. Det engelske kompositums ækvivalent er *'hjulbremse'*, der kan produceres ved kompositionel oversættelse.

(3) *'wheel brace'*: Andet led har inden for mekanik en hel række oversættelser, fx 'støtte', 'klampe', 'borsving'. Her kan kompositionel oversættelse ikke bruges, da den danske ækvivalent hedder *'svingnøgle'*.

Samme forhold gælder, når *'wheel'* er andet led i et kompositum,

(4) *'front wheel'*; første led oversættes til for(side) og ækvivalenten er *'forhjul'*.

(5) *'replacement wheel'* er lidt mindre gennemskueligt, da den første komponent oversættes med 'erstatning', 'udskiftning', når den står alene, men sammensætningen hedder *'reservehjul'* på dansk.

(6) *'steering wheel'* kan til gengæld slet ikke håndteres kompositionelt, idet ækvivalenten er et simpelt enstavelsessubstantiv: *'rat'*.

Der er desuden et stort antal flerleddede sammensætninger i vores korpus, fx *'spare wheel mounting bracket'*. Langt de fleste af disse leksikalske enheder hører til inden for domænesproget.

Kollokationer

En anden type ordkombination er fx verbale udtryk, der består af et verbum med et substantiv eller en nominalfrase som objekt. Disse danner ofte en leksikalsk-semantisk enhed, en kollokation. Et typisk eksempel fra korpuset på en domænespecifik kollokation er:

(7) *'change wheel'* der altid oversættes til *'skifte hjul'*. I dette udtryk kan 'change' ikke oversættes med 'bytte' eller 'veksle' som i mange andre sammenhæng er synonyme med 'skifte'.

De ovenfor nævnte to grundlæggende problemer rejser spørgsmålet om, hvilke sammensætninger og kollokationer (og andre flerordsenheder) der principielt skal medtages i en MT-ordbog. Da en datamatisk ordbogs omfang ikke behøver at begrænses på samme måde som en trykt ordbogs, er det hensigtsmæssigt at medtage alle sammensætninger og flerordsenheder, der ikke kan håndteres kompositionelt.

Flertydige ord - flertydige sætninger

De fleste ord har flere inhærente betydninger der tilsammen udgør ordets betydningsomfang. I en given kontekst realiseres normalt kun én af disse betydninger; hvilken én, det bestemmes af den givne begrebsmæssige relation mellem nærkontekstens elementer. I kontrastiv henseende betyder dette at ordet har flere oversættelsesmuligheder. En sammensat betydningsstruktur gengives i ordbogsartiklen i almensproglige ordbøger ved at anføre nummererede betydninger, der evt. belyses med parafraser, eksempler osv.

I det følgende vil jeg koncentrere mig om sådanne problemer i oversættelsen af vore tekster, hvor leksikalsk flertydighed bliver årsag til overgenerering (dvs. for mange oversættelsesforslag) eller fejloversættelse, hvis den nødvendige domæneviden ikke er integreret i MT-systemets ordbog.

I en del tilfælde har dette problem to aspekter:
(A) Et engelsk ord (eller leksikalsk enhed) har flere almensproglige oversættelser
(B) Samme ord har også en række domænespecifikke oversættelser

(A) Et godt eksempel på dette fænomen er verbet '*remove*', der hører til det centrale ordforråd i vores korpus:
(8) Remove the wheel nuts!
(9a) Remove the jack from the engine compartment!
(9b) Remove the hubcap from the wheel!
(10) Remove any corrosion on the mounting surfaces with a brush!
(11) Remove the wheel chocks and ensure that all items ...
(12) Remove the brackets!

Verbet oversættes på forskellig vis i disse sætninger:

(8) '*fjerne*'; (9a) '*tage frem + fra*'; (9b) '*tage af + fra*';
(10) '*fjerne + fra* (med); (11) '*tage væk*'; (12) '*flytte*';

afhængigt af kontekstens struktur (valensmønstret) og leksikalske indhold (semantiske træk), dvs. antallet og arten af verbets argumenter. I disse tilfælde kan MT-systemet håndtere valget af ækvivalent ved at sammenholde resultatet fra den sætningsstrukturelle analyse med ordbogens oplysninger om verbets valensstruktur, der i dette tilfælde altid inkluderer akkusativobjekt og i nogle tilfælde også et præpositionsled.

(B) Nedenstående eksempler illustrerer vigtigheden af, at den systemet har adgang til domæneviden. Den er kodet i ordbogen i form af semantiske træk og seletionsrestriktioner og styrer valget af den fagsproglige oversættelse i nedenstående sætning, hvori begge indholdsord er flertydige.
(13) '*Remove the brackets!*'
Sætningen fortolkes meget forskelligt af forskellige fagfolk, fx

'Ophæv parenteserne!' (matematiker); 'Flyt lampetterne!' (byggningselektriker); 'Tag knæpladerne af!' (sømand); 'Fjern takstklasserne!' (jurist/økonom). Med opslag i ordbogen for domænet automekanik vil oversættelsen være: '*Flyt/Fjern konsollerne*!'

En anden mulighed er, at maskinen søger i den almensproglige ordbase (svarende ca. til Kjærulff-Nielsens Engelsk-danske ordbog) ved oversættelsen af følgende korpussætning

(14) '*Remove the jack from the engine compartment!*'

Resultatet vil være (med tilfældigt valgte ækvivalenter) fx 'Ryd fyren fra førerrummet af vejen!' eller 'Fjern pengene fra maskinrummet!' sammen med hel række andre, ligeså forkerte sætninger.

Problemet er at både '*remove*' og '*jack*' har flere almensproglige oversættelser, som maskinen ikke kan skelne imellem på grundlag af de oplysninger der står i ordbogen. (Et yderligere problem er at sammensætningen '*engine compartment*' ikke findes i den almensproglige ordbog, dvs. systemet forsøger at oversætte den kompositionelt.)

Men, som det allerede er blevet omtalt i afsnittet om systemarkitektur, foretages der prioriteret ordbogsopslag (først i domæneordbogen og derefter i den almensproglige komponent) og sætningen kan dermed oversættes korrekt til '*Tag donkraften frem fra motorrummet!*'

Ordbogsarbejdet

Ordbogsarbejdet - også inden for MT - handler i høj grad om at genbruge eller udnytte tilgængelige oplysninger i eksisterende opslagsværker. Foruden at gennemgå og bearbejde det genbrugelige materiale udfører vi også korpusbaseret ordbogsarbejde. Korpusundersøgelser fokuserer på et ords nærmeste omgivelser; i vores arbejde betyder dette altid analyse inden for den aktuelle sætnings grænser.

Arbejdsmetode

Den maskinelle korpusanalyse kan gøre brug af mere eller mindre sofistikerede metoder, afhængig af, hvor store tekstmængder der skal analyseres, til hvilket formål, og hvilket programmel man har til rådighed. Vores korpus på ca. 50 A4 sider fylder ikke ret meget i maskinlæsbar form.

Vi anvendte først et større konkordansprogram (WordCruncher), der har gode søge- og sorteringsfaciliteter, men er noget omstændeligt i brug. Programmet er blevet benyttet i den indledende fase til at fremstille en samlet liste over alle ordformer samt deres forekomsttal i korpuset. Desuden blev der fremstillet en samlet nøgleordskonkordans (KWIC). Dette materiale dannede grundlaget for det mere målorienterede arbejde, der har involveret en række manuelle og automatiske processer der beskrives nedenfor.

Hertil valgte vi at bruge programmet Excerper, der er mere egnet til hurtige detaljeundersøgelser på grund af sin smidighed og enkle, men effektive virkemåde.

Vi har foretaget:

- selektion af leksikalsk materiale til videre behandling med hensyn til de tidligere opregnede oversættelsesmæssige aspekter (fx led i sammensætning hhv. kollokation; flertydighed mm.)
- syntaktisk analyse af nøgleordets forekomster med henblik på valensmønstre, kollokationer
- sortering af forekomsterne i valens- og kollokationstyper
- semantisk analyse af nøgleordets kontekst (primært af de valensbundne led, jf afsnittet om 'Krav til MT-ordbøger')
- leksikalsk analyse der omfatter sammensætninger, brug af synonymer, antonymer mm.
- morfologisk analyse (bøjningsformer, kongruens, stavevarianter)

Foruden at dele de observerede sproglige træk op i typer og dermed forberede dem til beskrivelse i systemets formelle sprog er der også behov for at beskrive visse grundlæggende, generelle begrebsrelationer inden for domænet (jf eksemplerne 8 til 12, 'remove').

Sådanne relationer er den såkaldte "is_a" relation (forklaring ved nærmeste overordnede begreb, fx 'wheel_nut' is_a 'fastener'; "is_part_of" relation (del relateres til helhed fx 'hubcap' is_part_of 'wheel') og "functions_as" (der relaterer genstanden til en funktion inden for domænet, fx 'jack' functions_as 'lifting_tool'). En del af disse oplysninger kan uddrages fra selve korpusteksten, resten kan findes i de tilhørende illustrationer eller indhentes fra eksterne kilder.

Disse oplysninger sættes i system: begrebernes indbyrdes relationer registreres i et hierarki (som over/under- hhv. sideordning) og til slut opstilles en taksonomi over domænet. De relationer, som taksonomien er baseret på kan udtrykkes i maskinens formelle sprog. På denne måde tilføjes en slags domæneviden til ordbogen på linje med de sproglige oplysninger. Herved kan entydiggørelse af kildeteksten optimeres og valget af oversættelsesækvivalent styres. På samme måde fungerer definitioner og forklaringer i en tosprogsordbog.

Konklusion

Den maskinelle oversættelse stiller krav om fuldstændig og systematisk beskrivelse af tekstens ordforråd. Opfyldes disse krav ikke, bliver oversættelsen mangelfuld eller forkert. De her præsenterede oversættelsesproblemer har væsentlige lighedspunkter med de vanskeligheder, en oversætter kan komme ud for, når der mangler opslagsord eller informationer i den anvendte fagordbog. Et andet problem kan være, at den anførte information er flertydig eller ikke er detaljeret nok; selv ved hjælp af kreativ tænkning kan man ikke altid finde frem til den korrekte oversættelse.

Ved at overføre relevante erfaringer fra arbejdet med maskinel oversættelse til planlægning af fagordbøger burde man kunne nå frem til en mere målrettet valg af ordforråd

og oplysningstyper. Samtidig kunne præsentationen blive mere konsistent end det er tilfældet med de ordbøger, vi har undersøgt. Man kunne med fordel integrere begrebsmæssige relationer i ordbogsartiklernes definitionsdel, i lighed med dem der bruges i MT.

Hensigten med at beskrive nogle centrale problemer i MT med relation til ordbogen har været at vise, hvorledes traditionel fagleksikografi kan drage nytte af det lingvistiske arbejde, der udføres i forbindelse med forskningsprojektet OFT.

Litteratur

Braasch, A. (1992a.) *Valg af tekstsort - Korpus - Undersøgelsesaspekter.* In: Ark 65, Oversættelse af fagsproglige tekster. HHK. København. 1992.

Braasch, A. (1992b). *Text based dictionary work for a domain-specific language.* In: Papers in Computational Lexicography. Budapest. 1992.

Copeland, C., Durand, J., Krauwer, S. & Maegaard, B. (eds.) *Studies in Machine Translation and Natural Language Processing* Volume 1: The Eurotra Linguistic Specifications. Luxembourg: CEC. 1991.

Sager, J. C. *A Practical Course in Terminology Processing.* Amsterdam/Philadelphia: John Benjamins Publishing Company. 1990.

Ordbøger

Clausens tekniske ordbøger. Engelsk-dansk.

Collins COBUILD English Language Dictionary. London & Glasgow, 1987.

Kjærulff-Nielsen, B.: Engelsk-Dansk Ordbog. Gyldendal, København, 1985.

Skibsted, S.: Teknisk engelsk-dansk ordbog. København, 1971.

Skjerk, Ebbe: Bilteknisk ordbog. Engelsk-Dansk. Teknisk Forlag, 1991.

Teknisk ordbog. Engelsk-dansk. L&H Ordbøger. København, 1991.

950 amerikanske-engelske automobil-fagudtryk oversat til dansk. General Motors International A/S

Programmer

Excerper, Version 2.0. Produceret af Norling•C Dataleksikografi, København

WordCruncher, Version 4.23. Produceret af Brigham Young University. ElectronicText Corporation, Provo, Utah.