Disambiguering i human oversættelse og i maskinoversættelse*

Frede Boje
Eurotra-DK

0. Indledning

Denne artikel udspringer af mit arbejde i Eurotra, hvor jeg især har beskæftiget mig med transfer mellem tysk og dansk, men de problemer, jeg omtaler med udgangspunkt i konkret materiale fra transfer-arbejdet, er temmelig generelle. Emnet ligger i grænseområdet mellem datalingvistik og leksikografi, men selv om jeg ikke kommer meget ind på datalingvistiske formalismer, har jeg forsøgt at anskue problemet fra datalingvistens synsvinkel snarere end fra leksikografens. Det betyder bl.a., at jeg ikke forudsætter fortrolighed med leksikografiske begreber.

Når jeg i det følgende bruger betegnelsen "ordbøger", betyder det almindelige tosprogede ordbøger til brug for mennesker, med mindre jeg udtrykkelig nævner andre former for ordbøger.

1. Aktiv / passiv ordbog

Det er i de senere år inden for tosprogs-leksikografien blevet almindeligt at skelne mellem aktive og passive ordbøger. Betegnelserne, der er indført af Smolik (1969)[1], er ikke særlig velvalgte, men de er vist efterhånden så fastgroede, at det er håbløst at forsøge at udskifte dem. Efter min mening ville det være bedre at tale om produktions- kontra receptionsordbøger eller - med Hausmann (1977) - Hinübersetzungs- kontra Herübersetzungswörterbücher.

Kort fortalt ligger forskellen i oversættelsesretningen i forhold til brugerens modersmål. Ved oversættelse til et fremmedsprog, altså Hinübersetzung, har man brug for en aktiv ordbog, ved oversættelse til sit modersmål, Herübersetzung, for en passiv ordbog. Der er således - principielt - behov for 4 ordbøger for et givet sprogpar, f. eks. mellem dansk og tysk:

1. dansk => tysk for danskere
2. dansk => tysk for tyskere
3. tysk => dansk for danskere
4. tysk => dansk for tyskere.

Hovedtanken i denne opsplitning er, at leksikografen kan - og bør - udnytte brugerens modersmåls-kompetence, når han udvælger de informationer, der skal medtages i ordbogen. Hvilke konkrete konsekvenser dette synspunkt får, vil jeg ikke uddybe her; det er beskrevet indgående i Kromann/Riiber/Rosbach (1984) og flere andre artikler af de samme forfattere. Lidt forenklet kan det siges således: Det er fremmedsprogets ordforråd, der kræver kommentarer (dvs. forklaringer, definitioner, eksempler osv.).

---

* Foredrag ved de Nordiske Datalingvistdage i København 3.-4. november 1987.

Disambiguering i human oversættelse og i maskinoversættelse
Frede Boje
Proceedings of NODALIDA 1987, pages 170-186

Det forudsættes altså, at brugeren ved Hinübersetzung umiddelbart forstår kildesproget, nemlig sit modersmål, mens han kan have behov for kommentarer af en eller anden slags for at kunne vælge den rigtige ækvivalent på målsproget. Omvendt kan han ved Herübersetzung have behov for kommentarer for at forstå den fremmedsprogede kildetekst, mens han forventes at beherske sit modersmål så godt, at han kan vælge den rette målsprogs-ækvivalent, når blot kildeteksten er forstået.

Denne skelnen gælder for ordbøger til mennesker. Ved maskin-oversættelse har hverken maskinen eller de programmer, man putter i den, nogen form for modersmåls-kompetence. Derfor har man brug for en ordbog, som så at sige er dobbelt aktiv:

	Kildesprog	Målsprog
Aktiv ordbog	- komm.	+ komm.
Passiv ordbog	+ komm.	- komm.
Maskin-ordbog	+ komm.	+ komm.

2. Hvor mange - og hvilke ækvivalenter?

2.1 Antallet af ækvivalenter

Den forskel, jeg hidtil har omtalt, er især principiel. I praksis har de fleste eksisterende ordbøger ikke klart definerede målgrupper og derfor ofte en blanding af kommentarer til lemmaerne, altså kildesprogsordene, og målsprogsækvivalenterne.

Derimod er der - både i teori og i praksis - en meget stor forskel på ordbøger til human- og maskinoversættelse m.h.t. hvad der er brug for af det, jeg i første omgang blot har kaldt "kommentarer". Denne forskel vil blive uddybet senere i artiklen, men inden da vil jeg redegøre for en række kvantitative aspekter, der danner baggrund for hovedproblemet.

For at begrænse undersøgelsernes omfang har jeg valgt at koncentrere mig om ord med begyndelsesbogstavet 'f', der i de forskellige værker optager fra knap 5% til godt 8% af pladsen. Eksempler, stikprøveundersøgelser, forklaringer osv. i det følgende er således hovedsagelig taget fra bøgernes afsnit med ord, der begynder med 'f'.

Først har jeg set på, hvor mange oversættelses-ækvivalenter der er pr. lemma. En undersøgelse af de første 500 F-ord i Gyldendals røde ordbog Tysk-Dansk (7) giver det resultat, der er vist i skemaets første dobbeltspalte[2].

Denne fordeling har jeg derefter sammenlignet med den samme ordbogs oversættelse af F-ordene i Heinz Oehlers "Grundwortschatz Deutsch"(15), der rummer de 2500 hyppigste ord på tysk.

	1 (500 første) F-ord uden hensyn til hyppighed		2 F-ord hørende til hyppigste ord		3 F-ord hørende til næsthyppigste ord		4 Ikke hørende til hyppigste eller næsthyppigste ord	
Antal ækvivalenter	Antal lemmaer	Procent	Antal lemmaer	Procent	Antal lemmaer	Procent	Antal lemmaer	Procent
1	340	68	10	25	26	29,9	335	69,5
2	99	19,8	5	12,5	19	21,8	97	20,1
3	27	5,4	4	10	10	11,5	27	5,6
4	15	3	5	12,5	7	8,1	14	2,9
5	8	1,6	1	2,5	9	10,3	6	1,2
6	3	0,6	3	7,5	4	4,6	3	0,6
7	3	0,6	3	7,5	2	2,3	-	
8	-		1	2,5	2	2,3	-	
9	3	0,6	2	5	2	2,3	-	
10	1	0,2	-		3	3,5	-	
11-15	-		2	5	3	3,5	-	
15	1	0,2	4	10	-		-	
I alt	500	100,0	40	100,0	87	100,1	482	99,9

Fordelingen af de 40 F-ord, der efter Oehlers angivelser hører til de 600 almindeligste tyske ord, er vist i skemaets 2. dobbeltspalte.

Resten af F-ordene i grundordforrådet (i alt 87 ord) havde den fordeling, der er vist i skemaets 3. dobbeltspalte.

Endelig har jeg sammenlignet med de ord fra første dobbeltspalte, der ikke hører til grundordforrådet. Deres fordeling er vist i skemaets sidste dobbeltspalte.

Det skal også nævnes, at det af forordet til ordbogen tydeligt fremgår, at man bevidst har "udeladt det meste af det ordforråd, der er umiddelbart forståeligt". Der er her i høj grad tale om lavfrekvente ord, der kun har én ækvivalent, f. eks. et stort antal uproblematiske fremmedord. Det vil sige, at procenten af ord med én ækvivalent snarere er højere, end optællingen i ordbogen viser.

Hvis vi generaliserer ud fra denne stikprøve - og det mener jeg man roligt kan tillade sig - kan vi se, at langt over halvdelen af det samlede ordforråd (i vores tilfælde 68%) ikke giver disambigueringsproblemer, fordi der kun er én oversættelsesækvivalent pr. lemma. Jo sjældnere et ord er, desto større er sandsynligheden for, at det kun har én ækvivalent, og omvendt: jo almindeligere et ord er, desto flere ækvivalenter har det som regel, og desto sværere er det derfor at disambiguere sikkert.

Lingvistisk og leksikografisk er det altså sværere at få et maskinoversættelsessystem til at oversætte de 5-600 almindeligste ord rigtigt i en rimelig procentdel af tilfældene end at udvide et velfungerende systems ordforråd fra f. eks. 2.000 til 20.000 ord eller for den sags skyld til 100.000 ord. Den sidste store udvidelse er ikke et lingvistisk problem, men et datamatisk, primært et kapacitetsproblem.

2.2 Hvilke ækvivalenter skal vælges?

2.2.1 Sandsynlighed

Foreløbig har vi kun set på antallet af ækvivalenter, som en almindelig ordbog "tilbyder". Når man skal tage stilling til, hvilke ækvivalenter der skal bruges i et maskinoversættelsessystem, kan man med fordel inddrage en anden form for statistik, nemlig sandsynlighedberegning. Når et ord i ordbogen f. eks. har et stort antal ækvivalenter, er det rimeligt at undersøge, om antallet kan reduceres til noget mere overkommeligt. Man kan bl.a. overveje, hvor sandsynlig en oversættelsesmulighed er.
Lad os f. eks. se, hvad der står i ordbogen (7) om "fahrbar":

Fähnlein *n* -. trop
Fähnrich *m* -e. officersaspirant
Fahr-abteilung *f* (mil.) motoriseret afdeling; **-ausweis** *m* rejsehjemmel; *SZ* kørekort; **-bahn** *f* kørebane; **-bahnmarkierung** *f* kørebaneafstribning; **-bahnrand** *m* rabat;
→ **-bar** *adj* transportabel, som kan køres; (gld) farbar, fremkommelig; *-er Tisch* rullebord; **-bereich** *m* aktionsområde; **-bereit** klar til afgang/start, køreklar

De 4 ækvivalenter hører betydningsmæssigt sammen to og to, hvilket er angivet ved semikolon. Den betydning, som udtrykkes med ækvivalent 3 og 4 er markeret som (gld.), altså gammeldags, en vurdering, som deles af ordbøger som Wahrig (11), DudGW (13) og DUW (14). Vi kan altså med god samvittighed se helt bort fra disse to ækvivalenter; tilbage bliver 1 og 2, som jeg ikke vil kommentere yderligere i denne omgang.

Desværre er det de færreste tilfælde, hvor man har så enkle metoder til at afgøre, hvad der kan skæres fra. Normalt må skønnet baseres på ens erfaring og common sense. Et tilstrækkelig stort relevant korpus ville være en stor hjælp. Det korpus, vi arbejder med i Eurotra i øjeblikket, er, dels p.g.a. sit ringe omfang, dels p.g.a. sin tilblivelse, ikke egnet til at give tilstrækkelig sikre kriterier.

2.2.2 Generalisering

Det er ikke kun hyppigheden, der spiller en rolle. En væsentlig faktor er også den enkelte ækvivalents betydningsomfang. Ofte har man i ækvivalentrækken en generel glose, der kan bruges i alle eller næsten alle tilfælde, plus nogle mere eller mindre synonyme udtryk, der hver for sig har en mere begrænset anvendelse. Her vil man normalt vælge det mest generelle udtryk, selv om man derved går glip af nogle nuancer.

Et eksempel herpå er oversættelsen af det danske "bestanddel", som ifølge Vinterberg & Bodelsen (1) har fire engelske oversættelses-ækvivalenter:

1. component
2. constituent
3. element
4. ingredient

Ifølge vores engelske samarbejdspartnere (der jo har ansvaret for oversættelserne til engelsk) vil "component" altid kunne bruges, mens de tre andre forslag har hver deres begrænsning. Man kan altså nøjes med at bruge én oversættelse: "bestanddel" => "component".

2.2.3 Tilføjelser

Selv om vi i Eurotra-projektet i øjeblikket af praktiske grunde prøver at begrænse antallet af ækvivalenter mest muligt, kan vi også blive nødt til at tilføje oversættelsesmuligheder, som ikke står i de gængse ordbøger. Det hænger sammen med, at vi i den nuværende fase arbejder korpus-baseret. De ord, der står i de tilsvarende korpora på de andre sprog, skal mindst kunne oversættes i den eller de betydninger, der optræder i disse korpora, og det giver undertiden oversættelser, der ikke findes i ordbøgerne. Et ekstremt eksempel her - som ikke er med F - er verbet "ansprechen", som kun optræder én gang i det tyske korpus. Det har i konteksten en betydning, som bedst gengives med "vedrøre". Den tysk-danske ordbog har i alt 25 oversættelsesforslag, men ikke "vedrøre" eller dermed beslægtede udtryk.

3. Monolingval/bilingval/multilingval disambiguering - Readings

Når man i lingvistisk litteratur har beskæftiget sig med ambiguitet og disambiguering, har det - såvidt jeg har kunnet se - normalt været i et monolingvalt perspektiv, og det er mit indtryk, at mange datalingvister regner med, at en monolingval disambiguering er tilstrækkelig også i oversættelsessammenhæng.

Tager vi en tekststreng - med F - som:

<fortaler>,

vil mange måske mene, at opgaven er løst, når man har fundet ud af, om det drejer sig om en form af:

1. "fortaler", cat=n, f. eks.: "Han er fortaler for en ny politik".
2. "fortale", cat=n, f. eks.: "Han har skrevet fortaler til begge bøger" eller
3. "fortale sig", cat=v, f. eks.: "Han fortaler sig let, når han er nervøs".

I oversættelsessammenhæng er disambigueringen imidlertid kun færdig, hvis det viser sig, at hvert af disse tre udtryk har lige præcis én oversættelsesækvivalent. Ved oversættelse til tysk eller engelsk er det tæt på at være rigtigt, men lad os tage en anden sammensætning med "-tale", nemlig:

<aftale>

Her vil man ud fra et rent dansk synspunkt mene, at det må være tilstrækkeligt at skelne mellem verbet og substantivet, for "en aftale er en aftale". I hvert fald er der hverken i NDO (9), ODS (10) eller Dansk Sprogbrug (11) så meget som en antydning af en skelnen mellem flere betydninger af substantivet.

Ikke desto mindre får man ved oversættelse til engelsk og tysk problemet med at vælge mellem ækvivalenterne[3]:

engelsk:	tysk:
1. agreement	1. Abkommen
2. appointment	2. Absprache
3. arrangement	3. Rücksprache
4. collusion	4. Termin
5. date	5. Verabredung
	6. Vereinbarung
	7. Vertrag

Denne liste kan muligvis reduceres med et par ækvivalenter efter de principper, jeg tidligere har nævnt, men tilbage bliver den ubehagelige kendsgerning, at f. eks. i forbindelse med oversættelse til tysk er substantivet "aftale" først disambigueret, når der kan skelnes mellem 5-7 readings af ordet.

Reading

Begrebet "reading", som jeg nu brugte, er desværre også problematisk, ikke kun fordi vi savner et godt dansk ord for det. I de fleste tilfælde svarer det meget godt til ordet "betydning", men langtfra altid. F. eks. virker det i vores eksempel lidt mærkeligt at sige, at substantivet "aftale" alt efter synsvinkel har én, fem eller syv betydninger.

Det, der især gør "reading" til et problematisk begreb, er, at mange bruger ordet, som om det var en nøje defineret størrelse, skønt der lægges forskellige betydninger i ordet. Det samme gælder "lexical unit", forkortet LU. F. eks. er afgrænsningen mellem, hvad der er to LU'er og hvad der er to "readings" af én LU, ikke særlig klar.

Jeg vil gerne sætte de to begreber i relation til begreberne "homonymi" og "polysemi", men først lige afklare, hvilke betegnelser jeg bruger for nogle fænomener, der selv er simple, men hvor terminologien tilsyneladende ikke er fast, i hvert fald ikke på dansk. For de tre engelske betegnelser:

feature: attribute = value (a feature is an a/v pair)

bruges:

feature: træk = værdi.

eks: number = plural

Som bekendt er det vanskeligt at give sikre kriterier for en skelnen mellem homonymi (her som homografi) og polysemi, og der har været argumenteret for, at denne skelnen inden for datalingvistikken skulle være irrelevant, at der f. eks. ikke skulle være grund til at lægge mere vægt på ordklasse end på alle andre træk i en feature-beskrivelse.

I én forstand er det rigtigt: I en unificeringsgrammatik er det ligegyldigt, om de værdier, der i en given regel skal (eller ikke må) matche, er værdier for ordklasse eller f. eks. for tællelighed, komparation eller tempus.

Det er muligt, at man i en analyse, der kun skal bruges monolingvalt, ikke har behov for en skelnen, men til oversættelsesformål mener jeg, at det er i det mindste praktisk, og måske nødvendigt at opstille nogle klare operationelle kriterier for en skelnen, der ikke nødvendigvis falder sammen med en grænsedragning efter f. eks. etymologiske principper. Det kan også være relevant at operere med forskellige kriterier for forskellige analyseniveauer. Mit forslag til skelnen lyder således:

Ord, der (i deres opslagsform) skrives ens, men tilhører forskellige ordklasser, er homografer og dermed forskellige LU'er. Er det substantiver, anses de for homografer - og altså forskellige LU'er, hvis de har forskelligt køn. Bortset herfra gælder, at hvis de tilhører samme ordklasse, men har forskellige værdier for mindst ét træk, er de polysemer og dermed forskellige readings af samme LU.

Denne opdeling er foretaget til praktiske formål og kan selvfølgelig forfines, hvis man føler behov for det. Det vil således nok være naturligt at inddrage i hvert fald bøjning som yderligere kriterium. Intuitivt virker det f. eks. mærkeligt at behandle 'die Bank' (= bank) og 'die Bank' (= bænk) som readings af samme LU[4].

4. Hvordan disambiguerer man?

Der skal ikke her siges så meget om, hvor disambigueringen finder sted; det uddybes i Annelise Bech og Poul Andersens artikel i dette nummer af Lambda[5]. I stedet vil jeg koncentrere mig om, hvilke midler man har.

Man kan se disambigueringsprocessen som en form for regelmatchning: Hvis en tekstenhed opfylder nogle bestemte betingelser, udløses et bestemt valg mellem ækvivalenter. Disse betingelser kan være angivet i en ordbog, en grammatik eller en anden form for opslagsværk, eller de findes inde i oversætterens hoved, som viden eller intuition. Det sidste tilfælde unddrager sig beskrivelse, så vi holder os til elektroniske og trykte medier.

For at der kan finde en matchning sted, skal det altså for det første være muligt at opstille nogle klare betingelser, for det andet skal det være muligt at konstatere, hvornår betingelserne er opfyldt. Det lyder banalt og selvindlysende, men det er alt andet end simpelt at anvende principperne i praksis. Ved maskinoversættelse er det meget sværere at opfylde krav nr. 2 end ved human-oversættelse, og det får også indvirkning på, hvilke betingelser der kan opstilles.

Human-oversætteren kan inddrage sin viden om verden og sin forståelse at teksthelhed, situationskontekst m.m. og kan drage

alle mulige analogislutninger, når han skal afgøre, om betingelserne for et givet valg er opfyldt.

Ved maskinoversættelse kan der stort set kun bruges betingelser, som går på forekomsten af bestemte fænomener i teksten. Mere konkret drejer det sig om tilstedeværelsen af størrelser med bestemte features eller bestemte værdier på trækkene.

Et afgørende spørgsmål er nu: Finder man i ordbøgerne oplysninger af denne type?

4.1 Disambigueringskriterier i ordbøgerne

Et fællestræk ved de fleste ordbøger, både mono- og bilingvale, er, at de vigtigste og undertiden næsten eneste midler til disambiguering er forklaringer og eksempler. "Forklaringer" er her brugt som overbegreb for definitioner, parafraser, synonymer osv. Uanset hvor relevante og velvalgte disse informationer måtte være set fra et almindeligt leksikografisk synspunkt, er de totalt ubrugelige til maskinoversættelse, vel at mærke direkte. Maskinen kan ikke stille noget som helst op med en oplysning i ordbogen om, at et givet ord f. eks. er synonymt med et andet. Derimod kan gode forklaringer og eksempler selvfølgelig være en hjælp for de mennesker, der laver ordbøgerne til maskinen.

4.1.1 Bilingvale ordbøger

De bilingvale ordbøger, oversættelsesordbøgerne, indeholder normalt mindst én oplysning, der direkte kan angives som et feature, nemlig ordklassen; for substantivers vedkommende angives tit også køn, for verbernes vedkommende transitivitet.

Desuden er der én type forekomst-relation, der ofte angives, nemlig den relation, der på engelsk kaldes "co-occurrence", men som vist ikke er navngivet på dansk. Det drejer sig typisk om forholdet mellem adjektiv og substantiv, altså den kendsgerning, at oversættelsen af et adjektiv ofte afhænger af, hvilket substantiv det modificerer. F. eks. kan man se i den dansk-tyske ordbog, at "høj" hedder "hoch" om et hus og "groß" om et menneske. Det store problem ved den slags oplysninger er, at det som regel ikke klart fremgår, hvilken form for co-occurrence der er tale om, altså om en given oversættelse af adjektivet er knyttet til ét bestemt substantiv, eller om substantivet står som repræsentant for en klasse - og i givet fald hvorledes denne klasse er afgrænset.

Selv hvis der er gjort forsøg på en afgrænsning, er den ofte vag og skal tages med forbehold. I eksemplet med "høj" står der således:

> 3. (om mennesker) groß (1,80 g.),

Ikke desto mindre finder man under "hoch" eksemplet

> "ein hoher Beamter",

og embedsmænd er dog normalt også en slags mennesker. Længere nede i artiklen står der:

5. (stofpåvirket) high [hai].

Det er normalt også kun relevant i forbindelse med mennesker.

Den slags inkonsekvenser spiller tit ikke nogen særlig rolle for den hærdede ordbogsbruger, fordi han er vant til at skulle hente supplerende disambigueringskriterier i eksemplerne og i øvrigt bruge sin sunde fornuft, men det gør det vanskeligt at omsætte betingelserne til features, som kan bruges i en maskine.

Ud over de ovennævnte oplysninger er der i de fleste oversættelsesordbøger næsten ingen informationer, der kan bruges som kriterier ved ækvivalentvalget, når de skal kunne udtrykkes ved feature-beskrivelser. Med andre ord: Oversættelsesordbøger kan bruges som en hjælp til at finde oversættelsesækvivalenter, men kun i ringe omfang til ud fra formelle kriterier at afgøre valget mellem dem, altså til at disambiguere maskinelt.

4.1.2 Monolingvale ordbøger

Ser vi nu på de monolingvale ordbøger, kan vi konstatere, at der er ret stor forskel på, hvilke informationer de giver, og hvilken form informationerne har. Som i de bilingvale ordbøger indtager eksemplerne en fremtrædende plads, men i de monolingvale ordbøger er forklaringsdelen som regel mere omfattende og rummer ofte egentlige definitioner.

Forskellen mellem ordbøgerne ligger især i, hvilke formaliserede eller direkte formaliserbare informationer de giver.

Næsten alle har oplysning om ordklasse, om substantivers køn (selvfølgelig ikke på engelsk) og mere eller mindre udtømmende oplysninger om bøjning[6]. Men derudover har nogle ordbøger systematiske oplysninger om opslagsordets omgivelser. Det gælder først og fremmest to ordbøger, Longmans Dictionary of Contemporary English (16) og dtv-Wörterbuch der deutschen Sprache (12). Longman har et meget udbygget kodesystem med bogstaver og tal, der angiver kombinationer af bøjningstype og distribution, mens dtv-Wörterbuch udelukkende bruger talkoder til at udtrykke lignende informationer.

Når man undersøger, hvorledes man kan udnytte disse oplysninger ved maskinoversættelse, konstaterer man, at de - med undtagelse af en vigtig gruppe informationer - som regel ikke er til særlig stor nytte. De er så sproginterne, at de er absolut relevante, når det drejer sig om tekstproduktion på det pågældende sprog, her altså henholdsvis engelsk og tysk, men de udgør sjældent kriterier ved disambiguering.

Undtagelsen er verberne. Her er der efterhånden udviklet metoder til en ret detaljeret beskrivelse af verbernes valens, og det ser ud til, at disse valensbeskrivelser kan gøre stor nytte ved udformningen af oversættelseskriterier.

4.2 Monolingvale readings og oversættelsesækvivalenter

I monolingvale ordbøger foretages der en inddeling i readings, der normalt begrundes i forskelle i ordenes betydning, og i oversættelsesordbøger er den væsentligste grund til at angive flere ækvivalenter, at de udtrykker forskellige betydninger. Det ville derfor være nærliggende at antage, at den skelnen mellem betydninger, der foretages monolingvalt, kunne bruges til at vælge mellem forskellige oversættelsesmuligheder.

Det kan desværre ikke uden videre lade sig gøre. Der er en række problemer forbundet med det.

(1) Det første er det tekniske problem, jeg allerede har nævnt, at forskellen mellem readings ofte kun angives ved forklaringer og eksempler. Det vil f. eks. være svært at formalisere forskellen på de første to readings under substantivet "fault" i Longman:

> **1** a mistake or imperfection: There are several faults in that page of figures.| a small electrical fault in the motor
>
> **2** a bad point, but not of a serious moral kind, in someone's character: Your only fault is that you won't do what you're told.| I love her for her faults as well as for her virtues.

(2) Det næste problem er, at "betydning" og "betydningsforskel" er så subjektivt bestemte begreber, at det er svært at blive enige om, hvor mange betydninger et ord har, og hvorledes forholdet er mellem disse betydninger. Det ses tydeligt, hvis man sammenligner beskrivelsen af det samme ord i forskellige monolingvale ordbøger. Ikke blot er antallet af readings og afgrænsningen mellem dem forskellige, men undertiden ser man også, at et eksempel, der i én bog skal illustrere én betydning, er identisk med eller svarer nøje til et eksempel, der i en anden bog illustrerer en anden betydning.

Hvis man ved oversættelsen vil tage udgangspunkt i en monolingval disambiguering i kildesproget, må man derfor tage stilling til, hvilke kriterier der skal lægges til grund og altså bl.a., om man vil gå ud fra en enkelt ordbog, eller om man vil gå eklektisk til værks. For behandlingen af dansk som kildesprog stiller sagen sig lidt anderledes. Vi har ikke noget at vælge imellem, da der ikke findes noget værk på dansk, der svarer til de andre sprogs Longman, Wahrig, Petit Robert osv., men vi må i stedet finde ud af, hvordan de sparsomme oplysninger i NDO skal suppleres.

(3) Det tredie problem er hovedproblemet. Det hænger nært sammen med det foregående, og det er oven i købet ekstra kompliceret i Eurotra-sammenhæng, fordi projektet ikke er bilingvalt, men multilingvalt.

Problemet er, at det ville være lettere at lave simpel transfer, hvis kildesprogsanalysen leverede lige præcis de readings, der udløser forskellige oversættelser, men samtidig ved vi, at man ved oversættelse til forskellige sprog har brug for forskellige betydningsafgrænsninger. Desværre er der mig bekendt ikke nogen, der kan sige noget særlig konkret om dette sidste fænomen.

For at belyse problemet kan vi se på Longman-eksemplet ovenfor. Vi kan på den ene side konstatere, at det fra en dansk synsvinkel er ligegyldigt, om man kan skelne mellem reading 1 og 2, for de skal begge oversættes ved "fejl". På den anden side kan vi ikke vide, om de måske skal oversættes forskelligt til f. eks. fransk eller græsk, så det alligevel kunne være relevant at kunne redegøre for forskellen.

Lad os se på et dansk eksempel: Verbet "føre" har i NDO 4 readings. Nr. 2 er forklaret som "stå i spidsen for" og rummer følgende eksempler, som jeg har nummereret:

1. føre en hær
2. Niels førte (ɔ: var forrest) under slutspurten
3. føre bog over sine udgifter
4. føre hus
5. føre krig
6. føre en sag
7. de førende (ɔ: toneangivende) kredse
8. føre ordet
9. føre en samtale
10. føre et rædsomt sprog
11. det er et skrækkeligt liv, han fører
12. refleksivt: hun forstår at føre sig ɔ: optræder smukt.

Man kan godt spørge sig selv, om "stå i spidsen for" er en rimelig parafrase for verbet i de anførte eksempler - jeg ved f. eks. ikke, hvordan man kan "stå i spidsen for et rædsomt sprog". Sagt på en anden måde: Det er et spørgsmål, hvor meget betydningsfællesskab verbet har i disse eksempler. Men lad os nu se på, hvad der sker ved oversættelse. Jeg har ikke fået 'native speakers' til at checke det følgende, men hvis jeg har brugt ordbøgerne rigtigt, og hvis man kan stole på dem (?!), skal (eller kan) "føre" i de anførte eksempler oversættes således:

	Engelsk:	Fransk:	Italiensk:	Tysk:
1.	command	conduire	?	führen
2.	lead	mener le peloton	essere in testa	führen
3.	keep	faire (registre)	tenere	führen (Haushalt)
4.	keep	tenir	governare	führen
5.	make	faire	fare	führen
6.	conduct	plaider	difendere	führen
7.	leading	prédominant	predominante	führend
8.	act as spokesman	porter	essere portavoce	führen
9.	carry on	soutenir	?	führen
10.	use	tenir	?	führen
11.	lead	trainer	condurre	führen
12.	carry oneself	se conduire	avere un bel portamento	sich führen

Oversættelsen til tysk er her ret simpel - jeg kan endda tilføje, at også samtlige eksempler under reading 1 og 3 naturligst oversættes med "führen". Hvis man omvendt prøver at oversætte de eksempler med "führen", der står i Wahrig, får man til gengæld brug for en halv snes danske ækvivalenter.

Bortset fra det tyske er det tydeligt, at eksemplerne skal oversættes så forskelligt, at der ikke er vundet noget som helst ved at lave en betydningsgruppering af det danske verbum.

Nu kunne man tænke sig, at verbet "føre" var et særlig ondskabsfuldt eksempel, og det er rigtigt, at det er værre end gennemsnittet, men det er ikke ekstremt med hensyn til antal ækvivalenter. Til gengæld er det meget almindeligt forekommende, så vi er nødt til at kunne gøre noget ved det. Det tyske "führen" hører til de 600 hyppigste ord, og jeg vil tro, at hyppigheden af "føre" er nogenlunde den samme.

Nogle vil måske mene, at ovenstående eksempel især viser, at der er behov for at få gjort noget ved kollokationer. Det er også rigtigt.

Det generelle spørgsmål, som sættes i relief af eksemplet med "führen", er, hvor stor overensstemmelse der er mellem de readings, man finder frem til ved den monolingvale analyse, og de oversættelses-ækvivalenter, der skal bruges.

Endnu er vores erfaringsmateriale ikke så stort, at vi kan sige ret meget om det, men et eksperiment, som vi har udført sammen

med den tyske Eurotra-gruppe, tyder på, at en opdeling af de enkelte verber efter valens er en stor hjælp i transfer-arbejdet, primært efter syntaktisk valens, men også efter semantiske restriktioner. I de tilfælde, hvor syntaktiske oplysninger ikke var tilstrækkelige, var der dog en tendens til, at oversættelseskriterierne faldt i to grupper: enten var trækket +/- HUM(AN) afgørende, eller også måtte man opfinde ad hoc-regler. Eksperimentet er beskrevet i Boje & al. (1986).

5. Konklusion

I abstract'et til dette foredrag opstillede jeg en tese. Jeg vil slutte med at gentage tesen i udbygget og kommenteret form.

Jeg hævdede, at "en væsentlig del af de oplysninger, der er relevante" for maskinoversættelse, "enten ikke findes i de almindelige ordbøger eller højst er implicitte", og jeg mener at have vist, at det gælder fuldt ud for oversættelses-ordbøgerne. I nogle monolingvale ordbøger findes en række formaliserede oplysninger; det endnu uafklarede spørgsmål er, i hvor høj grad de er relevante for oversættelse.

Hvorledes "maskinoversættelse stiller særlige krav til oplysningernes formaliserbarhed", mener jeg også at have vist, selv om det næppe har været nogen overraskelse for ret mange.

Endelig siger jeg i abstractet, at "formaliseringen ... med fordel vil kunne udnyttes i fremtidige ordbøger til human-oversættelse". Den slags påstande er svære at dokumentere, især så længe der er så få resultater at fremlægge, men lige så vel som den øgede formalisering har gjort monolingvale ordbøger bedre, er der for mig ingen tvivl om, at det samme vil kunne ske med oversættelsesordbøgerne, efterhånden som resultaterne af formaliseringsarbejdet viser sig.

Jeg har ikke givet svaret på to vigtige spørgsmål, som ikke står i abstractet, men måske ligger der - implicit:

(1) Hvordan konkretiserer og formaliserer man implicitte oplysninger?

(2) Hvordan fremtryller og formaliserer man de oplysninger, som er nødvendige for fuldstændig disambiguering, men som ikke findes i nogen opslagsværker?

Der kan endnu kun gives meget ufuldkomne svar på disse to spørgsmål. Det arbejde, der i øjeblikket udføres i Eurotra, omfatter bl.a. forsøg på at finde frem til et fælles system til kodning af semantiske træk. Forhåbentlig vil erfaringerne fra dette arbejde snart kunne udmøntes i en redegørelse, der rummer i hvert fald nogle af svarene på disse to spørgsmål.

Noter

1. Henvisning til ordbøger sker ved et tal i parentes. Dette tal angiver værkets nummer i litteraturlisten. "Anden citeret litteratur" angives på traditionel vis (som her): forfatter + årstal.

2. Optællingen skal selvfølgelig tages med et vist forbehold, ikke kun fordi det er en stikprøve, men især fordi kriterierne for optællingen altid kan diskuteres.
a. Der er kun medregnet ord og komplekse udtryk, der er anført som danske ækvivalenter til lemmaet alene, mens komplekse tyske udtryk, som indeholder lemmaet, men ikke er oversat kompositionelt, er holdt uden for beregningerne. Det gælder altså bl.a. idiomatiske udtryk.
b. Der er ikke gjort forsøg på at vurdere, hvilke ækvivalenter der er helt eller delvis synonyme. Til gengæld er ord (eller mere korrekt: tekststrenge), der er anført flere gange som ækvivalent for samme lemma, kun talt med én gang, uanset at de (i hvert fald efter ordbogsforfatterens mening) må anses for polysemer eller homografer. Antallet af ækvivalenter er derfor ikke automatisk lig med antallet af "betydninger".

Endelig vil antallet af ækvivalenter selvfølgelig også afhænge af bl.a. ordbogens omfang. Sammenligner man f. eks. de to engelsk-danske ordbøger fra Gyldendal, er mængden af ækvivalenter i den røde ordbog (3) (i hvert fald normalt) en ægte delmængde af ækvivalenterne i Kjærulff Nielsens store ordbog (2).

Eksempel:
fahrbar (adj): transportabel, som kan køres; (gld) farbar, fremkommelig; -er Tisch: rullebord

er talt som 4 ækvivalenter, "(fahrbar)-er Tisch" er ikke medregnet.

3. Fundet i hhv. (1) og (7).

4. På den anden side er forholdet mellem bøjning, syntaks og betydning ikke altid så enkelt. F. eks. er der i velplejet tysk en syntaktisk og betydningsmæssig forskel på det stærkt og det svagt bøjede verbum 'hängen' ('er hängte' er transitivt, 'er hing' er intransitivt), mens en tilsvarende skelnen på dansk er næsten opgivet. De fleste danskere bruger 'hængte' og 'hang' i flæng, både transitivt og intransitivt.

5. De to foredrag er udarbejdet uafhængigt af hinanden, og ingen af os udtaler sig på Eurotras vegne.

6. Der er her tale om et forsømt område inden for leksikografien, se f. eks. Kromann (1985).

LITTERATURLISTE

1. Bilingvale ordbøger

(1) **Dansk-engelsk Ordbog** v. H. Vinterberg og C.A. Bodelsen, 2.udg. 7. opl. (1985)

(2) **Engelsk-dansk Ordbog** v. B. Kjærulff Nielsen, 2. udg. 3. opl. (1985)

(3) **Engelsk-dansk Ordbog** af Jens Axelsen (Gyldendals røde ordbøger) 10. udg., 7. opl. (1985)

(4) **Dansk-fransk Ordbog** v. A. Blinkenberg og P. Høybye, 3. udg. (1976)

(5) **Dansk-Italiensk Ordbog** v. P. Høybye og J. Mengel, 2. udg., 2. opl. (1979)

(6) **Dansk-tysk Ordbog** v. E. Bork (Gyldendals røde ordbøger) 8. udg. (1980)
(7) **Tysk-dansk Ordbog** v. E. Bork (-"-) 11. udg. (1982)

2. Monolingvale ordbøger

2.1. Danske

(8) **Dansk Sprogbrug** v. E. Bruun (-"-) 1. udg. (1978)

(9) **Nudansk Ordbog** v. E. Oxenvad 11. udg. (1982) (NDO)

(10) **Ordbog over det dansk Sprog** Bd. 1-28 (1919-56) (ODS)

2.2. Udenlandske

(11) **Deutsches Wörterbuch** v. G. Wahrig, Neuausgabe (1980) (Wahrig)

(12) **dtv-Wörterbuch der deutschen Sprache** v. G. Wahrig (1978) (dtv-Wb)

(13) **Duden: Das große Wörterbuch der deutschen Sprache in sechs Bänden** (1976-81) (DudGW)

(14) **Duden: Deutsches Universalwörterbuch** (1983) (DUW)

(15) **Grundwortschatz Deutsch** v. H. Oehler (1966)
(frekvensoplysningerne taget fra den danske bearbejdelse:
Tysk-dansk Grundordbog v. O. Børløs Jensen (1970))

---
(16) **Longman Dictionary of Contemporary English** (1985) (Longman)

(17) **Petit Robert 1:** Dictionnaire de la Langue Francaise (1986)

(Der er ikke lagt vægt på at give fuldstændige bibliografiske oplysninger; kun de vigtigste data er medtaget. Udgaverne er de faktisk anvendte, selv om der i nogle tilfælde findes nyere udgaver)

3. Anden citeret litteratur:

Boje & al. 1986: Frede Boje, Birgit Weck and Hanne Ruus: The Choice of German and Danish Target LUs, based on Governor and Complement Information. (Internt Eurotra-papir, maj 1986)

Hausmann 1977: Franz Josef Hausmann: Einführung in die Benutzung der neufranzösischen Wörterbücher. Tübingen 1977. (Romanistische Arbeitshefte 19).

Kromann 1985: H.-P Kromann: Zur Selektion und Darbietung syntaktischer Informationen in einsprachigen Wörterbüchern des Deutschen aus der Sicht ausländischer Benutzer. In: Lexikographie und Grammatik. Akten des Essener Kolloquiums zur Grammatik im Wörterbuch 28.-30.6.1984. Hrsg. H.Bergenholtz & J. Mugdan. Tübingen 1985.

Kromann/Riiber/Rosbach 1984: H.-P. Kromann/Th. Riiber/P. Rosbach: "Active" and "Passive" Bilingual Dictionaries: The Scerba Concept Reconsidered. In: R.R.K. Hartmann (ed.): LEXeter '83 Proceedings Vol. II. Bilingual Lexicography and the Learner's Dictionary. Tübingen 1984.

Smolik 1969: W. Smolik: "Aktives" Wörterbuch Deutsch-Russich. In: Nachrichten für Sprachmittler H.3. 1969, 11-13.