TERMINOLOGI OG MASKINOVERSÆTTELSE[1]

Gert Engel og Bertha Toft

Abstract

In this paper, which was originally presented at the Elsnet Bullet Course in Terminology Management in Bergen in 1998, we discuss the possibilities of 1) reusing term bank data in MT and TM systems in order to avoid duplicate efforts and 2) integrating term bank, TM and MT systems; both issues are essential in order to optimize the document production process. The conclusion under 1) is that data from a system of the Danterm for Windows type can be reutilized in both monolingual and bilingual MT dictionaries, albeit with considerable modifications, whereas contextual data from a term bank system can only be reutilized in a TM system in cases where so-called parallel contexts are available. As for 2), it is concluded that integration will have to start from the stage when SL-texts are planned, marked up and written since this is the only way to ensure terminological consistency and quality.

1. Indledning

Sprogteknologiske værktøjer som termbank- og oversættelsessystemer udvikles og anvendes for at opnå rationaliseringsgevinster. Forudsætningen for at nå dette mål er imidlertid, at de sprogdata, som værktøjerne i forbindelse med den aktuelle dokumentproduktion skal arbejde med, faktisk er tilvejebragt. Denne forudsætning skal være opfyldt, før værktøjerne tages i brug. Det er en banal, men desværre ikke sjælden vildfarelse, at tomme systemer kan gøre gavn.

Udarbejdelsen af terminologiske glossarer, mono- og bilinguale ordbøger til et MT (*Machine Translation*) -system og opbygningen af flersprogede tekstarkiver til et

[1] Artiklen er en dansk version af et indlæg på det første Elsnet Bullet Course i Terminology Management, afholdt af HIT-Centeret ved Bergens Universitet, september 1998

translation memory-system er meget ressourcekrævende opgaver. Men indeholder termbanker, TM(*Translation Memory*)- og MT-systemer ikke i stor udstrækning de samme data? Og hvis man allerede har én af databaserne, f.eks. en terminologisk database, må det vel være muligt at lade disse data indgå i de MT-ordbøger, der skal udarbejdes?

I den første del af vores indlæg vil vi derfor diskutere følgende problemstilling:

> *1. Kan terminologiske data genbruges til MT- og TM-systemer, og i givet fald hvordan?*

Lige så vigtigt er det, at alle værktøjer, som er til rådighed, 'spiller sammen'. Der skal derfor etableres et system til effektiv 'work flow' gennem hele produktionskæden, begyndende med dokumentstrukturering, lay-out, skrivning og terminologicheck på kildesproget.

Den efterfølgende oversættelse til et eller flere fremmedsprog kræver under alle omstændigheder supplerende terminologi- og ordbogsarbejde. Selve oversættelsen kan herefter udføres ved hjælp af et MT-system eller med støtte fra et TM-system. Derefter skal der etableres en procedure for den fortløbende vedligeholdelse af alle de sproglige databaser.

Vi vil således i anden del af vort indlæg gøre rede for følgende problemstilling

> *2. Hvordan kan værktøjer som termbanker og TM/MT-systemer spille sammen i den moderne dokumentproduktionsproces?*

Groft forenklet og lidt provokerende formuleret vil det dreje sig om, hvordan vi kan undgå, at der endnu engang opbygges en stor termbank, der først tages i brug 20 år senere. Det vil også dreje sig om genbrug af dyrekøbte arbejdsresultater.

Allerede her vil vi slå fast, at genbrug af arbejdsresultater forudsætter, at man har fuld tiltro til sine egne produkters kvalitet. Ved at genbruge cementerer man jo den standard, som man engang har ment, var god kvalitet.

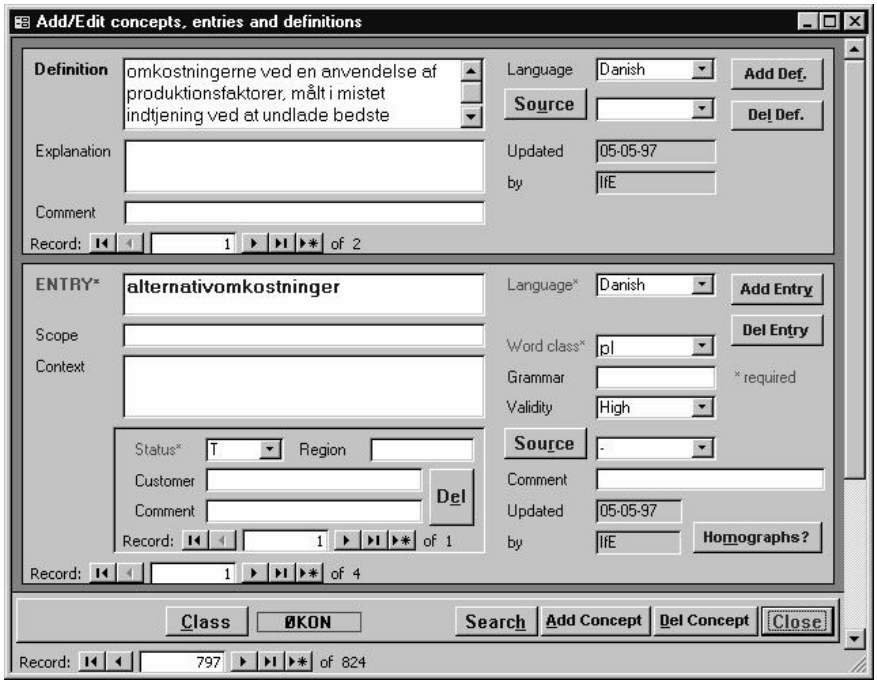
2. Terminologiske data i termbanker, MT-systemer og TM-systemer

2.1 De vigtigste informationskategorier i termbanker

Termbanker indeholder følgende hovedkategorier af data:

- begrebsidentifikation
- emneområde
- definition af begrebsindholdet, evt. med støtte i begrebssystemer
- benævnelser for begrebet med tilhørende 'grammatiske' oplysninger
- kontekster, hvori benævnelserne indgår

Som eksempel vises følgende skærmbillede fra Danterm for Windows:



Figur 1. 'Add/Edit page' i Danterm for Windows-systemet

2.2 Terminologiske data i et MT-system

Et MT-system som METAL/T1 arbejder med monolinguale og bilinguale ordbøger, hvoraf sidstnævnte er sprogretningsbestemte (en for hvert udgangssprog/målsprog-par). De vigtigste informationskategorier i monolinguale ordbøger er følgende:

Eksempel 1:

("arbejde"	NST
ALO	"arbejde"
C-ALO	"arbejds"
CL	(P-R S-T)
GD	(N)
KN	(MS-CNT)
SX	(N)
TYN	(ABS)

Forkortelsernes betydning:

NST	=	Substantivstamme
ALO	=	Allomorf
C-ALO	=	Sammensat allomorf
CL	=	Ordklasse (Bøjning):
		P = pluralis
		S = singularis
GD	=	Køn (grammatisk)
KN	=	substantivkategori:
		MS = ikke-tælleligt
		CNT = tælleligt
SX	=	Køn (naturligt)
TYN	=	Semantisk type
		ABD = abstrakt

En sammenligning af dataene i de to typer af systemer giver følgende resultat:

Termbanker indeholder ligesom monolinguale MT-ordbøger benævnelser (=ord), men de grammatiske oplysninger, der er knyttet til disse benævnelser, er helt utilstrækkelige til MT-formål. Især de manglende valensoplysninger og semantiske træk er en alvorlig ulempe set fra et MT-synspunkt.

Forklaringen på disse mangler er ganske ligetil: termbanker er primært udviklet til brug for sprogkyndige, men ofte ikke fagkyndige, oversættere.

Konklusionen er, at monolinguale MT-ordbøger godt kan suppleres med fagspecifikke termposter, men disse poster bliver ikke anvendelige, før der er tilføjet en hel del yderligere information.

2.2.1 De væsentligste datakategorier i bilinguale MT-ordbøger

Bilinguale terminologiske glossarer er begrebsorienterede, hvilket vil sige, at kernen i en 'terminologisk enhed' er et begreb; til dette begreb anføres alle de benævnelser, som findes på hvert af de relevante sprog.

Begrebsorientering kan opfattes som helt irrelevant for MT-systemer, som simpelthen 'overfører' (konverterer) ord fra kildesproget til ord på målsproget. Bilinguale MT-systemer indeholder altså ikke begreber, men 'ordligninger', f.eks. (DA) bord = (EN) table.

Eksempel 2:

"mellemrum" NST --> "Leerzeichen" NST
    Pref S.0.0.00 Tag (DP)
"mellemrum" NST --> "Abstand" NST
    Pref S.0.0.00 Tag (GV)

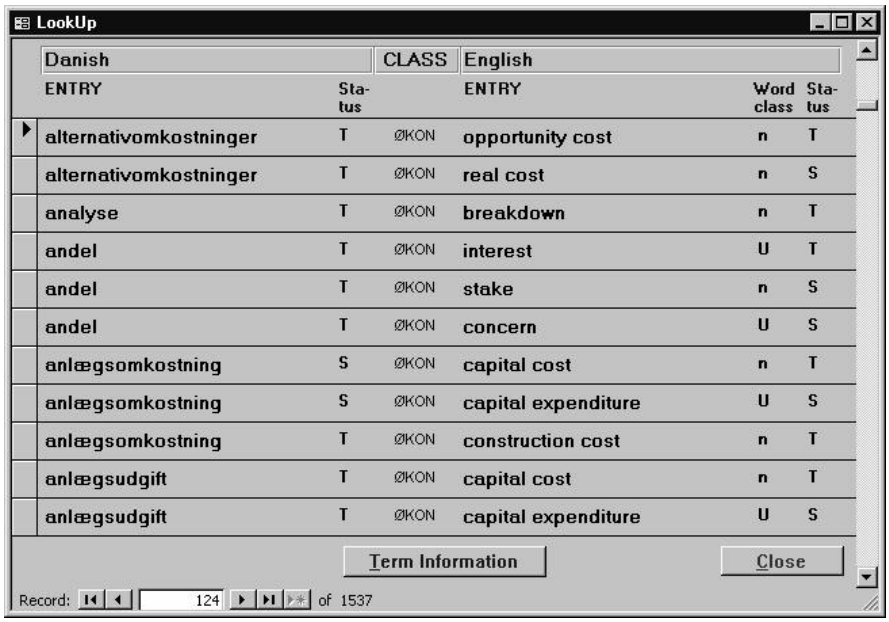
Eksempel 3:

"aflæsning" NST --> "Abladen" NST
    Pref H.1.1.03 Tag (CTV)
    Tests (XT-NST-PP :CAN "af" :TYN-ALL CNC)

"aflæsning" NST --> "Erfassung" NST
    Pref S.0.0.00 Tag (CTV)

2.2.2 Ordligninger i termbanker

En termbank kan imidlertid også være opbygget på en sådan måde, at resultatet af en bilingual søgning er umiddelbart anvendeligt for en oversætter. Danterm for Windows omfatter således en 'look-up-page', der indeholder en række ækvivalente termer:



Figur 2. 'Look up-page' i Danterm for Windows-systemet

Bemærk, at termbanken ikke kun angiver status for hver ækvivalent term, den gentager også termen på kildesproget i de tilfælde, hvor der til samme begreb findes flere benævnelser på målsproget. Dette vil gøre det muligt for et MT-system at vælge den foretrukne term på målsproget, også i de tilfælde, hvor søgeordet på kildesproget har status 'S' (synonym) i stedet for status 'T' (term).

I ikke så få tilfælde er ordligningerne i en MT-ordbog kun korrekte under ganske bestemte betydningsmæssige forudsætninger. Systemet må derfor udføre en test af, om disse betingelser er opfyldt, for eksempel om der er angivet et specifikt emne- eller fagområde. I termbanker findes der sådanne oplysninger, men om de umiddelbart kan anvendes i en MT-ordbog afhænger af, om de to systemer anvender samme klassifikationssystem.

En sådan test kan også baseres på en undersøgelse af de kollokationer, som er typiske for ordene på hver side af ordligningen. I termbanker findes sådanne kollokationer i det TEXT-felt, som er knyttet til alle benævnelser (termer).

I de tilfælde, hvor testen baseres på valensforhold eller semantiske træk, kan en termbank derimod ikke være til nogen hjælp. MT-systemet kan heller ikke gøre brug af termbankens bedste bud på et redskab til afklaring af betydningsforskelle, nemlig de definitioner, som anføres til hvert begreb.

Det er endnu ikke undersøgt, om en disambigueringstest vil kunne baseres på oplysningen om, i hvilket begrebssystem, en bestemt benævnelse anvendes, og i bekræftende fald hvor pålideligt, resultatet af en sådan test måtte være. Det er naturligvis en forudsætning, at MT-systemet i så fald har adgang til oplysninger om, hvilke begrebssystemer der refereres til i den tekst, som skal oversættes.

Det kan sammenfattende konstateres, at der er et betydeligt sammenfald mellem informationskategorierne i termbanker og bilinguale MT-ordbøger (transfer-ordbøger).

Det betyder, at der er store muligheder for at genbruge terminologiske data til MT-formål.

Tilbage står så opgaven med at eksportere terminologiske data til MT-systemer. Løsningen kan meget vel være så omkostningskrævende, at rationaliseringsgevinsten forsvinder op i den blå luft.

2.3 Terminologiske data i TM-systemer

Et TM-system indeholder kun én type informationsenheder, nemlig tekstsegmenter – oftest sætninger, sjældnere afsnit, på kildesproget – som er oversat og derefter 'linket' til ækvivalerende tekstsegmenter på målsproget.

Når et nyt dokuments tekstsegmenter skal oversættes, checker systemet, om de enkelte segmenter tidligere er blevet oversat. Er det tilfældet, indsættes det tidligere oversatte tekstsegment i det aktuelle dokument.

En termbank vil indeholde enkelte, særligt udvalgte tekstsegmenter i form af kontekster, der betragtes som karakteristiske for den enkelte benævnelse. Men det vil kun i et begrænset antal tilfælde være muligt at linke disse tekstsegmenter til tilsvarende segmenter på to eller flere målsprog.

Der findes én lykkelig undtagelse, nemlig de såkaldte parallelle kontekster. Her vil der foreligge en autentisk tekst på kildesproget, medens de parallelle tekster på målsprogene er oversættelser. Sådanne parallelle tekster kan umiddelbart genbruges i et TM-system.

3. Samspillet mellem termbanker og TM/MT-systemer

3.1 Samspillet bør begynde på et tidligere stadium

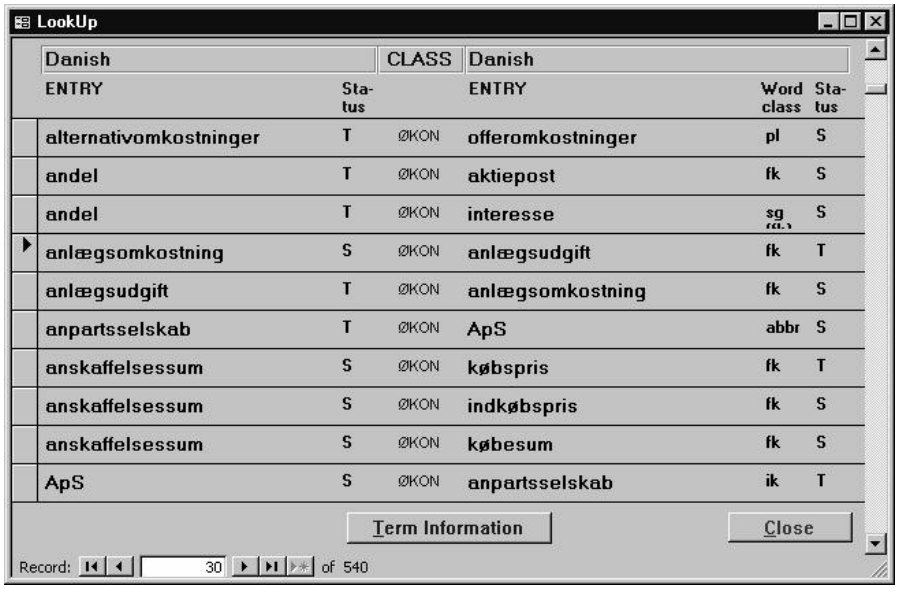
Formuleringen i dette kapitels overskrift er i virkeligheden udtryk for en traditionel tankegang. Den proces (det work-flow), der slutter med præsentation og udskrivning af multilingual dokumentation, begynder naturligvis ikke med terminologiarbejdet, men på det stadium, hvor den enkelte teksts lay-out samt dens formelle og indholdsmæssige struktur fastlægges og forsynes med elektronisk 'mark-up'.

Det er på dette tidlige stadium, der er mulighed for at skabe et solidt grundlag for et TM-system, hvis tekstsegmenter alene er opmarkeret på basis af deres indhold, dvs. uden at

være underlagt de begrænsninger, som udgøres af sætningsgrænser eller arbitrære afsnitsinddelinger.

Sikringen af den sproglige kvalitet bør ikke alene bestå i revision af allerede oversatte tekster, men skal tværtimod sætte ind allerede ved formuleringen af teksten på kildesproget. Det faktum, at kildesproget som regel er forfatterens modersmål, medfører – måske en smule overraskende – at fristelsen til sprogligt sløseri er særlig stor.

Termbanken skal altså konsulteres allerede på dette stadium. Det er de færreste termbanker, som er 'gearet' hertil, og Danterm for Windows er ingen undtagelse i den henseende. Dog giver den brugeren mulighed for at foretage en monolingual søgning mellem dansk og dansk.



Figur 3. Resultat af monolingual søgning i Danterm for Windows med dansk som udgangs- og målsprog

Hvis forfatteren er i tvivl om den virksomhedsinternt fastsatte korrekte terminologi, kan vedkommende foretage kontrolopslag for at konstatere, om den benævnelse, han har tænkt sig at anvende, rent faktisk har status som term.

3.2 Vedligeholdelse er lige så vigtig som anvendelse

Hvis en termbanks nytteværdi skal opretholdes, må vedligeholdelse og opdatering af termbanken indgå som et fast led i dokumentproduktions-processen. En forfatter eller en oversætter, som har søgt i termbanken uden resultat, er tvunget til selv at finde frem til en løsning på sit problem. Løsningen (resultatet) skal ikke blot sættes ind i det aktuelle dokument, men naturligvis registreres i termbanken med henblik på senere brug. Det er vigtigt at gøre sig klart, at resultatet ikke bare er den benævnelse, man har fundet frem til, men også dens kontekst samt en præcis kildeangivelse til den pågældende kontekst.

Hvis man begynder sin arbejdsproces med opstart af termbanken og valg af det relevante sprogpar og derefter åbner for tekstbehandlingssystemet, vil det ved hjælp af de gængse Windows-faciliteter være muligt hurtigt at skifte (switche) og kopiere fra termbank til tekstbehandlingssystem og omvendt.

Vedligeholdelse og opdatering af et TM-system kan foregå samtidig og på samme måde. Hvis systemet udelukkende suppleres med de tekststrenge, der registreres som kontekster i termbanken, vil der naturligvis – mildt sagt – være tale om selektiv vedligeholdelse. Samtidig er det dog værd at overveje, om ikke en brugerkontrolleret udvælgelse af væsentlige tekstsegmenter på lidt længere sigt kan bidrage til at reducere mængden af støj i TM-systemet, et velkendt problem ved de gængse systemer.

For MT-systemernes vedkommende er vedligeholdelse af ordbøgerne integreret i selve oversættelsesproceduren. Her vil en mulighed for at skifte over til termbanken være meget nyttig.

Det kan sammenfattende konkluderes, at samspil mellem systemerne i løbet af vedligeholdelsesprocessen er af størst betydning, når det gælder tekstbehandlings- og termbanksystemet.

3.3 Opbygning af systemer

Hvis brugeren selv skal opbygge systemerne, foreligger der en helt anden situation. Alle systemer har brug for samme fundament, nemlig en fuldtekstdatabase, hvis indhold er repræsentativt for de dokumenter, som fremover skal produceres med støtte fra termbank-, TM- og MT-systemerne.

Den mest velegnede type fuldtekstdatabaser er den, som indeholder dokumenter, der foreligger på flere sprog. Situationen er helt ideel, hvis der allerede er foretaget segmentering af tekstsegmenterne og en 'alignment' (sammenkobling) af parallelle segmenter på de forskellige sprog. I sådanne tilfælde vil man faktisk råde over memory-delen af et TM-system.

Endnu er denne idealsituation dog undtagelsen, der bekræfter reglen. Den første opgave, der skal udføres, vil normalt være at gennemføre en segmentering og derefter 'alignment' af teksterne.

Det næste skridt vil bestå i at ekstrahere en liste over de benævnelser, der kan betragtes som 'termkandidater', samt at få dem godkendt af fageksperter. Derefter registreres alle de godkendte termkandidater i termbanken, sammen med typiske og gerne også forklarende kontekster.

Næste skridt er at beslutte, hvilken status, de registrerede benævnelser skal tildeles, hvorefter man i sit korpus kan erstatte alle synonymer med de benævnelser, som har fået termstatus. På denne måde kan man sikre den terminologiske konsistens i sit korpus, som nu kan fungere som tekstarkiv (dvs. memory-del) i TM-systemet.

Efter opbygningen af termbanken og TM-systemet kan korpus anvendes som grundlag for opbygningen af et MT-system. Det kan sammenfattende konstateres, at integreret (samlet) opbygning af systemerne ikke blot betyder sparede omkostninger, men også kan bidrage væsentligt til kvalitetssikringen.

4. Integrerede systemer

Da vi i det foregående ihærdigt har argumenteret for det størst mulige samspil mellem termbanker og MT/TM-systemer, vil det være naturligt at spørge: hvorfor ikke tage skridtet fuldt ud og integrere dissse systemer?

Et meget beskedent forsøg på integration er gjort i Danterm for Windows, hvor der som 'entries' også kan lagres tekststrenge, som vi har kaldt 'phrase'. I enkelte specialer, udarbejdet af studerende ved danske handelshøjskoler, opereres der desuden med 'parallelle kontekster', som svarer til de testsegmenter, der indgår i MT-systemer.

TRADOS og TERMSTAR har endvidere gjort vellykkede forsøg med integration af termbanker og TM-systemer. Imidlertid stiller håndteringen af disse systemer store krav til brugerne, samtidig med at de befinder sig på et prisniveau, som små og mellemstore virksomheder vanskeligt vil kunne klare.

T1 tilbyder i sin professionelle udgave såvel et MT- som et TM-system. Endelig er der via EU's Othello-program skabt grundlag for integration af terminologiske glosarer og MT-ordbøger.