Nettet som korpus ved flersproglig term- og vidensbearbejdning

Henrik Selsøe Sørensen

Handelshøjskolen i København

Summary

The web as a corpus for multilingual term and knowledge extraction has an important potential in cases where a translator or knowledge engineer needs to identify unknown equivalents. With the expansion of knowledge and parallel creation of new terms in practically all fields, the need to search for unknown equivalents occurs more and more frequently. It is a major barrier multilingual information exchange that relevant multi- or bilingual term repositories and even monolingual resources are usually updated long after new phenomena have been born and baptised. This paper discusses methods and tools for translation-oriented knowledge extraction from the uncontrolled mass of texts on the web. The main focus will be on ways of identifying candidate target language terms in cases where no clues are at hand. Subsequently, smart searches may be used to validate the degree of equivalence between source terms and candidate equivalents. The proposed method is referred to as "multilingual term and knowledge extraction through clustered searches".

1. Indledning

Uanset hvor store ordbøger eller termbaser er, vil der altid være terminologi og viden som ikke er behandlet i dem. Det kan skyldes at denne viden er for ny eller for speciel eller begge dele. Ved viden forstår jeg i denne sammenhæng både begreber og termer samt begrebsrelationer i videste forstand, bl.a. ontologier eller brudstykker heraf. Fagsproglige oversættere og formidlere står meget ofte i den situation, at de forgæves har søgt efter fragmenteret viden, termer eller vidensstrukturer og prøver søgning på nettet for at få besvaret de spørgsmål, som de traditionelle kilder ikke kunne bruges til at besvare. I denne artikel forsøger jeg at kikke systematisk på metoder til vidensindsamling på nettet med særligt henblik på, hvordan man ikke alene finder og ekstraherer énsproglig viden, men også bryder gennem en sprogbarriere og skaffer sig materiale til to- eller flersproglig vidensbearbejdning og efterfølgende kvalitetsvurdering af resultaterne.

2. Ensproget term- og vidensindsamling

Første led i en term- og vidensindsamling er typisk en søgning efter en eller flere termer på ens eget modersmål eller, hvis man oversætter, på kildesproget. Formålet er at finde definitioner og begrebsrelationer samt fraseologi og evt. belæg for brug af stilistiske varianter. Situationen kan være, at man er usikker på ens egen viden, på om den er opdateret, eller opdager, at man er stillet over for noget nyt. Hvis opslagsværker på papir eller i elektronisk form[1] og søgning i nationale korpora[2] ikke har givet tilfredsstillende resultater, er det naturligt at prøve at søge på nettet ved hjælp af en søgemaskine, typisk Google. Andre søgemaskiner kan også være relevante, men i det følgende nævner jeg Google som en alsidig repræsentant for alle søgemaskiner[3].

Årsagen til, at mange begynder på papir eller en cd-rom er, ud over tradition og vane, at der jo her er tale om kilder, hvor viden på forhånd er ekstraheret og bearbejdet af eksperter, hvis troværdighed er veletableret. Ethvert kvalitetsarbejde inden for term- og vidensbearbejdning er meget tidskrævende, og derfor vil det naturligvis være en fordel, hvis man kan få eller købe sig til noget eksisterende. Færdigbearbejdede resultater i form af definitioner og begrebssystemer eller ontologier under en eller anden form er imidlertid en mangelvare inden for mange hurtigt ekspanderende vidensområder. Se Evidence Network (2002) for eksempler hentet fra det samfundsvidenskabelige område.

Selv i de tilfælde, hvor den fagsproglige medarbejder øjensynlig har fundet det søgte på papir, elektronisk eller på anden måde i bearbejdet form, kan det være relevant at tjekke den dagsaktuelle situation i det store ukontrollerede korpus ved hjælp af Google. Det, man kan ønske at finde belæg for i denne situation, hvor der stadig blot er tale om vidensindsamling på ét sprog, er typisk:

a) brug af en given term inden for det relevante vidensområde
b) fraseologi
c) synonymer / varianter: frekvensanalyse
d) synonymer / varianter: stil- og kontekstafhængighed

---

[1] Fx EU's store termdatabase Eurodicautom: http://europa.eu.int/eurodicautom/login.jsp <21.09.2002>.

[2] Danmark, **Korpus 2000**: http://korpus.dsl.dk/korpus2000/indgang.php <21.09.2002>
Norge, **Tekstlaboratoriet**: http://www.tekstlab.uio.no/norsk/bokmaal/index.html <21.09.2002>
Sverige, **Språkbanken**: http://spraakdata.gu.se/ <21.09.2002>
England, **British National Corpus**: http://www.hcu.ox.ac.uk/BNC/ <21.09.2002>.

[3] En god portal med en vis "varedeklaration" til de forskellige søgemaskiner findes på http://allsearchengines.com/ <21.09.2002>.

e) forekomst i tekster (fx som genvej til at finde tekster, der angriber samme emne fra en anden synsvinkel).

Metoder til at foretage denne vidensindsamling eller til at verificere og evt. opdatere kendt viden beskrives i forbindelse med præsentation af eksempler i afsnit 4 nedenfor.

Har man ved hjælp af Google fundet forekomster af relevante termer i en eller flere tekster, som jo godt kan være lange tekster, skal der afsættes tid til at læse eller analysere teksterne. Den fremsynede fagsprogsmedarbejder vil formentlig, hvis teksten er relevant og kvalitetsmæssigt egnet til det, ønske at ekstrahere termer og viden med henblik på lagring i en term- og vidensbase, hvor også ensproglig vidensbearbejdning tillægges værdi, bl.a. fordi vidensdeling herved muliggøres.

Der findes gode sprogteknologiske redskaber til at effektivisere denne monolingvale proces, og et par af disse redskaber skal nævnes kort.

WebCorp[4] giver mulighed for gratis og on-line at få lavet en enkel frekvensanalyse af en web-adresse og ikke mindst at kombinere en Google-søgning med en konkordansfunktion.

WordSmith Tools[5] tilbyder det samme plus flere avancerede faciliteter specielt skræddersyet til engelsk.

Hvis der ønskes en egentlig term- og vidensekstraktion fra en større tekstmængde, anbefales System Quirk Language Engineering Workbench[6] til engelske tekster. Systemet udvikles dog også til andre sprog end engelsk, men arbejder stadig monolingvalt. Software kan downloades gratis fra det angivne site. Denne såkaldte Language Engineering Workbench indeholder forskellige redskaber, først og fremmest Tracker, som er et vidensindsamlingsredskab, der kan finjusteres alt efter ens personlige behov, og som derefter kan køre i baggrunden og scanne nettet for tekster indeholdende ønsket terminologi. Når ens personligt indsamlede korpus er hentet ned fra nettet, kan en stribe Quirk-redskaber bruges til at identificere fagtermer, hvorefter en ny scanning med

4 WebCorp: http://www.webcorp.org.uk/wcadvanced.html <21.09.2002>.

5 WordSmith Tools: http://www.liv.ac.uk/~ms2928/ <21.09.2002>.

6 System Quirk Language Engineering Workbench: http://www.mcs.surrey.ac.uk/SystemQ/ <21.09.2002>.

Tracker kan iværksættes med udgangspunkt i de identificerede termer. Metoden og redskaberne er bl.a. dokumenteret i forbindelse med termvalidering (Ahmad, Gillam 1997, Ahmad, Rogers 2001).

Til analyse af franske tekster findes bl.a. Intex[7], som også er et meget omfattende og kraftfuldt redskab udarbejdet af Laboratoire d'Automatique Documentaire et Linguistique. Skal Intex anvendes, forudsætter det en aftale med ophavsmændene om brug til forskningsformål.

Resultater af ensproget term- og vidensindsamling kan sammenfattes som en identifikation af fragmenteret viden og eventuelt sammenhængende vidensstrukturer i løbende tekst og en efterfølgende bearbejdning inklusive etablering af begrebsrelationer. Normalt vil etableringen af begrebsrelationer ske ved at indføje ny fragmenteret viden i eksisterende vidensstrukturer i en term- og vidensbase. De fragmenter, som identificeres i løbende tekst, leverer ofte betydningsfulde data om, hvorledes fragmenterne spiller sammen med andre fragmenter i den dynamiske kommunikationsproces. Når de enkelte fragmenter og deres kontekst derimod er indfanget, bearbejdet og lagret i en database, antager de i databasen en statisk karakter. Samspillet mellem de blandede vidensfragmenter i det dynamiske tekstunivers og de samme fragmenters pladser i statiske strukturer er udgangspunkt for den klyngemetode, som foreslås anvendt.

3. Flersproget vidensbearbejdning

Når man beskræftiger sig med fagsproglig oversættelse og formidling på tværs af sproggrænser, er vidensindsamling og -bearbejdning på ét sprog som regel den mindste af udfordringerne. Interessante problemer, som kan være svære at løse, opstår, når viden og termer indsamlet på ét sprog skal holdes sammen med ækvivalenter på målsproget. I mange tilfælde eksisterer der ækvivalenter, blot har ingen endnu identificeret og bearbejdet dem i en to- eller flersproget database. I andre tilfælde findes ækvivalenterne slet ikke, og i disse tilfælde er det måske specielt vigtigt ved hælp af en flersproglig ækvivalensundersøgelse at skaffe sig belæg for erkendelsen af den manglende ækvivalens. Først herefter kan en løsning på det deraf følgende formidlingsproblem findes.

For den, der skal løse et oversættelsesproblem her og nu, som har søgt forgæves i ordbøger etc., er det oplagt at benytte nettet. Hvis der ikke kan findes et

[7] Intex: http://ladl.univ-mlv.fr/index.html <21.09.2002>.

troværdigt færdigbearbejdet svar i en on-line termliste her, må man dels selv at prøve at identificere nogle sproglige data og lave den tværsproglige bearbejdning på stedet. Samtidig hermed foretages bevidst eller ubevidst den nødvendige kildevurdering. Studiet af redskaber og metoder til at udføre disse processer uden at råde over kontrollerede korpora er et nyt og spændende forskningsområde, som forener informations- og sprogvidenskab, og som er blevet aktuelt i takt med, at flere og flere tekster er tilgængelige som et kæmpestort ukontrolleret korpus på nettet.

Kontrollerede korpora er tekstsamlinger sammensat ud fra et eller flere kriterier, som ofte går ud på at gøre samlingerne repræsentative. Flersprogede kontrollerede korpora kan være parallelle eller sammenlignelige.

Ved parallelle korpora forstås samlinger af oversatte tekster, der i princippet er helt ækvivalente. Med hensyn til sammenlignelige tekster, benyttes Fungs definition: tekster med en emnemæssig og funktionel lighed mere end en oversættelsesmæssig ækvivalens (Fung 1998). Fung benytter i sin forskning manuelt fundne sammenlignelige - også kaldet ikke-parallelle korpora - for gennem statistiske metoder at søge at identificere oversættelsesækvivalenter med henblik på at udarbejde ordbøger til maskinoversættelse. Der er en udbredt interesse for ikke kun at studere termer og disses kollokationer i møjsommeligt sammenstillede bilingvale parallelle korpora (Picchi and Peters 1998) men også for at studere metoder til først at finde mere eller mindre sammenlignelige korpora fra nettet og derefter benytte disse, som altså må betegnes som kontrollerede korpora, til at udtrække flersprogede ordbogs- og termdata (Grefenstette and Kilgarriff (Eds.) Forthcoming). Inden for området "tværsproglig informationssøgning" arbejder Adriani og van Rijsbergen (1999) med ad statistisk vej at måle lighed mellem termer i et forsøg på at disambiguere bilingvale ordbøgers forskellige betydninger. De ønsker at effektivisere det, de kalder 'dictionary-based cross-language information retrieval (CLIR) method'. Selv om endemålet i alle de nævnte tilfælde ligesom i mit er fremstilling af bilingvale eller multilingvale ordbøger / termbaser og afklaring af ækvivalensforhold, så adskiller den nedenfor beskrevne klyngestrategi sig fra de nævnte tilgange ved at angribe ikke-sammenlignelige korpora uden først at gøre forsøg på at parallellisere dem.

Klyngestrategiens formål er udfylde et specifikt videnshul på den anden side af en sproggrænse gennem søgning i nettets ukontrollerede tekstmængder af svingende kvalitet efter udvalgte klynger af kendte ækvivalenser. Hensigten er så præcist som muligt at skyde sig ind på relevant tekstmateriale, som efter

bearbejdning kan forsyne fagsprogsoversætteren eller fagsprogsformidleren med data til at lukke videnshullet, samtidig med at der finder en kvalitetssikring sted.

De klassiske strategier, som oversættere og andre sprogformidlere benytter til at krydse sprogbarrieren, vil kan kort beskrives som trin 1 og 2 i nedenstående oversigt.

Strategier

1) **Klassisk informations- og dokumentationssøgning (I&D)**

a) **Opslag i to- eller flersprogede opslagsværker**

Kan ske på papir eller i elektronisk form. Omfatter søgning efter mulige over- og underbegreber samt synonymer. Lykkes det at få et svar, skal dette ofte verificeres, og til det formål er strategierne 1b) og 3) velegnede. Såfremt klassisk I&D ikke giver et tilfredsstillende og pålideligt svar, prøves naturligvis en af de følgende stategier.

b) **Systematisk søgning efter ukendt term på målsproget**.

Når to- eller flersprogede opslagsværker ikke giver svar, søges fragmenteret eller struktureret viden identificeret på målsproget fx gennem systematisk gennemgang af en ontologi eller tesaurus eller opslag i kilder på papir som fx artikler og bøger med gode indekser etc. Af andre muligheder for systematisk at fremsøge terminologiske data i ensproget materiale kan nævnes den indlysende strategi, der går ud på fx at søge firmaspecifik terminologi på et fremmedsprog ved at besøge det pågældende firmas hjemmeside på det relevante lands sprog - eller læse i firmaets eget dokumentationsmateriale på det pågældende sprog. Mulighederne er utallige og velkendte og alle er selvfølgelig ikke nævnt her.

2) **Trial and error**

Har strategi 1a) og 1b) ikke givet resultater, kan der laves en hypotese om, hvad den ækvivalente term på målsproget kunne være, og hypotesen skal naturligvis derefter verificeres nøje gennem klassisk I&D eller søgning i kontrollerede eller ukontrollerede korpora på målsproget, herunder også søgninger på nettet.

3) **Klyngestrategi**
   Der tages udgangspunkt i en tekst på kildesproget, der indeholder den eller de termer, til hvilke det ikke er lykkedes at finde ækvivalenter på målsproget. I denne tekst identificeres et mindre antal andre termer, hvis ækvivalenter allerede kendes på målsproget, og som samtidig er centrale i forhold til det hovedemne, teksten behandler. Ved hjælp af en klynge af disse målsprogsækvivalenter, foretages søgning efter en sammenlignelig tekst på målsproget, og den eller de fundne tekster analyseres for at søge at identificere den ækvivalent, der ikke kunne opspores ad anden vej.

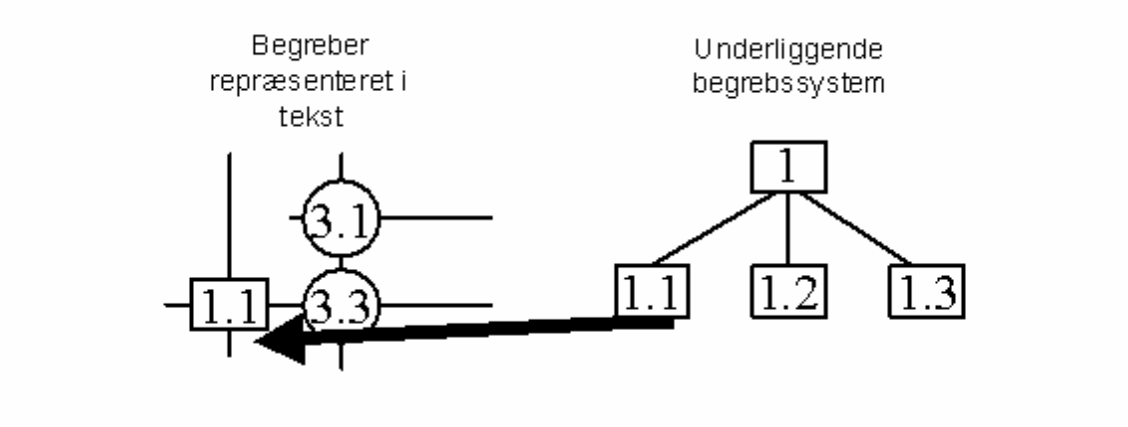
Trin 3 er den såkaldte klyngestrategi, som jeg arbejder med at forfine, således at søgning efter ukendte ækvivalenter i nettets milliarder af tekster kan effektiviseres. Klyngestrategien udspringer af en pragmatisk funderet hypotese om, at en god del af de (faglige) emner, som er genstand for tekstuel kommunikation i ét sprogområde også er det i de fleste øvrige sprogområder, og at det derfor efter al sandsynlighed er muligt at finde en sammenlignelig tekst. Den offentligt tilgængelige tekstmængde på nettet inden for langt de fleste vidensområder, også de fagsproglige, antages inden for de større sprog i hvert fald at være rimeligt repræsentativ, så længe man ikke bevæger sig ned på et meget specialiseret niveau (Grefenstette 2001), samtidig med at andre sprog end engelsk bliver bedre og bedre repræsenteret.

Når der er fundet en eller flere sammenlignelige målsprogstekster ved hjælp af klyngestrategien, som beskrives nærmere nedenfor, er der en vis sandsynlighed for i denne eller disse at kunne

a) identificere den ukendte ækvivalent
b) finde grundlag for at antage, at der ikke findes nogen ækvivalent
c) finde links i direkte eller overført forstand til nye kilder, som kan analyseres og evt. føre til den ukendte ækvivalent.

Det er klart, at der langtfra i alle tilfælde eksisterer ækvivalenser, men det kan klyngestrategien også kaste lys over; der er som bekendt tre muligheder:

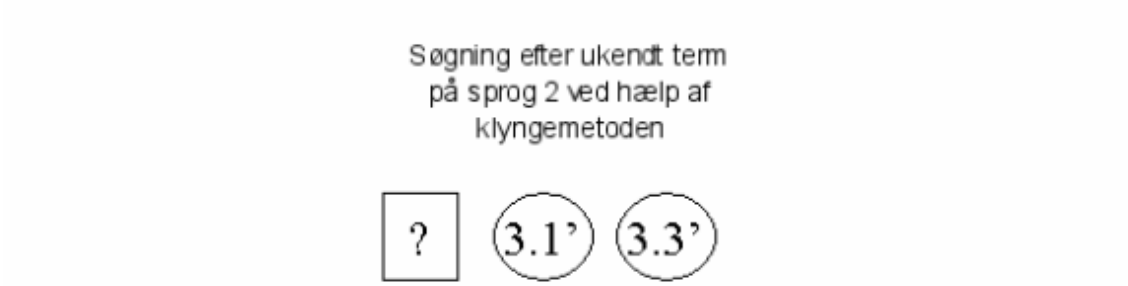
- termen / begrebet har en fuld ækvivalent
- termen / begrebet har en delvis ækvivalent
- termen / begrebet har ingen ækvivalent.



Figur 1 Samspil mellem termer/begreber i dynamisk kontekst (løbende tekst) og statisk lagret i en term- og vidensbase

Grundlaget for klyngestrategiens operationer, som forbinder det dynamiske tekstunivers og den statiske viden, kan fremstilles skematisk som i figur 1-3. Udgangspunktet er den statiske viden, som kan gøres til genstand for vidensmodellering og lagres i en term- og vidensbase eller i form af en ontologi, jf. i figur 1 '*underliggende begrebssystem*'. Når en tekst konstitueres, hentes enkeltfragmenter af statisk viden frem og blandes dynamisk i teksten, hvor de danner et som regel unikt mønster, der bestemmes alene af forfatterens kommunikationshensigt. Dette mønster skitseres grafisk i figur 1 som '*begreber repræsenteret i tekst*'.

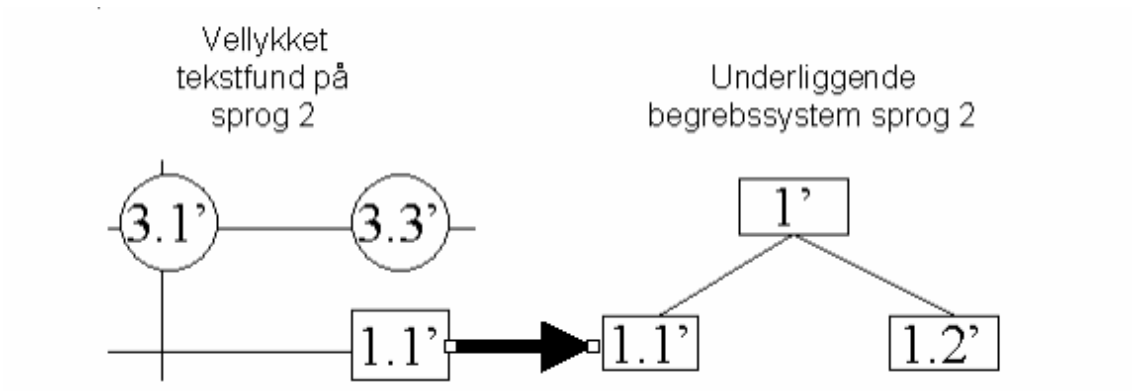
Såfremt ækvivalenten til term 1.1 i figur 1 ikke kendes på målsproget, starter oversætteren eller terminologen med at identificere en klynge centrale termer,



Figur 2 Termerne 3.1' og 3.3' er kendte ækvivalenter på sprog 2, som eftersøges i en klynge med henblik på at identificere den ukendte 1.1'.

fx 3.1 og 3.3, for hvilke ækvivalenterne 3,1' og 3.3' er kendt på målsproget, jf. figur 2. Formålet er ved hjælp af en søgning efter disse ækvivalenter at opspore en potentiel sammenlignelig tekst på målsproget, som ud over den samme klynge af dynamisk sammenbragte termer også kunne indeholde ækvivalenten 1.1'. Findes der slet ikke nogen brugbar tekst på målsproget, kan det enten være fordi 3,1' eller 3.3' er en forkert ækvivalent, eller fordi klyngen er blevet for stor og kravene til målsprogsteksten dermed for skrappe. I så fald forsøges med henholdsvis en ændret klynge og en reduceret klynge. Findes der for mange svar, tilføjes der yderligere termer til klyngen. Fundne tekster skal naturligvis vurderes nøje med hensyn til kildens kvalitet Hvis det efter gentagne forsøg ikke giver resultat, findes 1.1' måske ikke på målsproget. En sådan konklusion er det naturligvis risikabelt at drage uden yderligere undersøgelser, som kunne gå ud på at forfølge links og spor i forbindelse med de fundne målsprogstekster. Der kan være tale om links til andre sites eller om spor i form af navne, lovparagraffer etc. Sådanne spor kunne også føre til andre tekster, der kan indeholde den ukendte ækvivalent.

Lykkes det imidlertid at identificere ækvivalenten 1.1' (se figur 3), kvalitets-vurderes den gennem søgning efter præcis denne term på nettet, og fundene vurderes med hensyn til kilde, frekvens etc. Når den fundne term er bekræftet, kan denne i bedste fald lede til en statisk kilde, der behandler den som del af et begrebssystem. Strategien illustreres nedenfor med et eksempel på konkret anvendelse.



Figur 3 Term 1.1' er fundet sammen med klyngen 3.1' og 3.3', og en ækvivalensbearbejdning kan indledes. 1.1' kan evt. føre til en statisk kilde

4. Eksempel

Udgangspunktet er en fransk økonomisk tekst, hvori termen 'servitude pour dettes' indgår:

> Dimension raciale de la traite des personnes, en particulier des femmes et des enfants [8]
>
> […] Les trafiquants cherchent généralement à exercer un contrôle sur l'identité légale de la victime en lui confisquant son passeport ou ses papiers. Son entrée ou son séjour dans le pays de destination est généralement illégal, ce qui la met dans une situation de dépendance accrue à l'égard des c. Le système de la **servitude pour dettes** est largement utilisé. Il permet de contrôler les victimes et de tirer indéfiniment profit de leur travail. Le recours à la force physique, aux brutalités et à l'intimidation est fréquemment signalé. […]

Ingen af strategiens første to trin har givet en ækvivalent til termen på engelsk og dansk. Termen kan ikke slås op, og trial and error-oversættelse har ikke givet resultater. En Google-søgning efter 'servitude pour dettes' har givet godt 800 forekomster, og gennemlæsning af nogle af konteksterne har givet en ide om, hvad termen dækker, omend der ikke er fundet nogen egentlig definition. Opgaven går nu ud på at finde en engelsk og en dansk ækvivalent ved hjælp af klyngestrategien. Engelsk først, idet det antages at være lettere at komme fra engelsk til dansk end fra fransk til dansk.

I første forsøg vælges fra teksten en klynge bestående af fire termer, hvis oversættelse til engelsk ikke volder problemer: 'trafiquants', 'dettes', 'pays', og 'destination' og derefter laves en Google-søgning på de engelske ækvivalenter 'smugglers debts countr* destination'. Formålet er at finde en sammenlignelig tekst på engelsk, som kunne kaste lys over, hvad ækvivalenten på engelsk til 'servitude pour dettes' er. Hvis de først valgte termer ikke giver det ønskede resultat, udskiftes de én efter én indtil et brugbart resultat er opnået - fx kunne 'passeport' anvendes, men termer fra andre afsnit i teksten er selvfølgelig også kandidater.

Resultatet blev faktisk godt allerede i første forsøg, idet der kun blev fundet to websider, og den første af disse viste sig at indeholde et bud på den eftersøgte ækvivalent. En søgning på 'debt' i denne tekst førte hen til tekstafsnittet:

[8] Citeret fra: http://www.un.org/french/WCAR/e-kit/issues.htm <21.09.2002>.

> ”Human trafficking on the rise” [9]
>
> In many countries in South Asia and Central and South America, millions are living in conditions of **debt bondage** -- tied to their employers by unpayable debts.”

Via engelsk var det derefter muligt at finde noget dansk tekstmateriale via 4 ækvivalenter på dansk, men ingen af disse tekster indeholdt, hvad der skulle vise sig at hedde ‘gældsslaveri’. En reduktion af klyngen til kun at omfatte ‘menneskesmugling’ og ‘gæld’ førte til nogle tekster, der også indeholdt ‘gældsslaveri’. Yderligere analyser bragte potentielle synonymer frem på alle tre sprog, se figur 4.

En første gennemgang af et udvalg af de i figur 4 viste fund indikerer, at både mennesker og lande kan være gældsslaver, men at det oftest er mennesker, særligt kvinder og børn. På dansk gives definitionen “*Gældsslaveri, dvs. at man stiller sin egen eller en af én afhængig persons arbejdsydelse som sikkerhed for gæld.*”[10]

Til det videre arbejde med ækvivalensverificering og bearbejdning af begrebssystemer benyttes primært klassisk I&D men evt. igen klyngestrategien. I en tekst som den følgende findes allerede fragmenter til den videre bearbejdning (fremhævet med fed):

[9] Citeret fra: http://www.freedomsite.org/pipermail/fs_discussion/2001-May/001855.html <21.09.2002>.
[10] http://www.menneskeret.dk/menneskeretieuropa/konventionen/artikel4/slaveriart4/ <04.11.2002>.

Frekvensoversigt forekomster fundet ved hjælp af Google 21.09.2002	
servitude pour dette	245
servitude pour dettes	804
esclavage pour dette	26
esclavage pour dettes	75
debt bondage	7860
debt slavery	1990
debt servitude	426
bonded labour	8870
bonded labor	7250
gældsslaveri	796
gældsbinding	4
gældsbundet arbejde	2

Figur 4 Frekvensanalyse med Google

> "Debt Bondage - Possible Solutions
>
> The impression given by much of the information presented to the Working Group about **bonded labour** by Anti-Slavery International over the past 25 years is that we are an organisation which monitors violations of international standards on this form of **slavery**, referred to by international organisations variously as **debt slavery**, **servile status**, or a form of **forced labour**. Consequently we give the impression of being preoccupied primarily with abuses and with cases where States are failing to prevent abuse and to implement their own standards and laws."[11]

I hvilken grad der er større eller mindre betydningsforskelle mellem de fundne kandidater på de enkelte sprog, må undersøges nærmere, men diskuteres ikke her. Det har blot været meningen at illustrere de første resultater tilvejebragt

[11] http://www.antislavery.org/archive/submission/submission2000-WGCFS.htm

ved hjælp af klyngestrategien i en situation, hvor man helt manglede en ækvivalent, men hvor man efter noget arbejde kan ende med at stå med en tresproget bearbejdning af overbegrebet 'slaveri i moderne tid'.

Oversætteren kan ved hjælp af klyngemetoden ret hurtigt sikre sig, at den fundne ækvivalent er anvendelig i en given kontekst og fortsætte med sin oversættelse. Til gengæld kræver en egentlig vidensbearbejdning som den skitserede naturligvis betydelig mere tid. Foretages der en seriøs definitionsjagt og flersproglig bearbejdning, bør resultatet naturligvis gemmes i en term- og vidensbase.

5. Konklusion

Målet på langt sigt er at tilvejebringe en samling redskaber og metoder, som kan fungere som en term- og vidensbase "on demand", hvor den kontrastive behandling foretages på stedet med nettets tekster som råmateriale. Disse redskaber og metoder skulle give fagsprogsmedarbejderen (der stilles i sagens natur store kompetencekrav til brugeren) mulighed for, når behovet opstår, effektivt at:

- ekstrahere flersproget relevant term- og vidensmateriale fra nettets ukontrollerede tekstmængder
- udvinde flersproget viden fra dette materiale ved hjælp af egen kompetence og pragmatisk tilgang
- verificere produktet og lagre det på en måde, så det kan deles med andre.

Jeg ser dette som et attraktivt supplement til de mange forsøg, der gøres på at etablere afbalancerede og repræsentative kontrollerede korpora, hvoraf de flersprogede kan være enten parallelle eller sammenlignelige. Disse korpora er anvendelige i mange sammenhænge, men for fagsprogsoversætteren eller fagsprogsformidleren har de som oftest én stor fejl: de indeholder ikke lige den viden eller de termer, der var brug for i en given situation.

Bibliografi

I) Artikler

Adriani, Mirna & van Rijsbergen, C.J. (1999) Term Similarity-Based Query Expansion for Cross-Language Information Retrieval. In: *Proceedings of Research and Advanced Technology for Digital Libraries, Third European Conference, ECDL'99*. Paris: Springer Verlag. 311-322.

Ahmad, Khurshid & Rogers, Margaret A. (2001) Corpus Linguistics and Terminology Extraction. In: Sue-Ellen Wright & Gerhard Budin (Eds.) *Handbook of Terminology Management (Volume 2)*. Amsterdam & Philadelphia: John Benjamins Publishing Company. 725-760.

Fung, Pascale (1998) A statistical view of bilingual lexicon extraction: From parallel corpora to nonparallel corpora. In David Farwell, Laurie Gerber & Eduard Hovy (Eds.). *Machine Translation and the Information Soup. Third Conference of the Association for Machine Translation in the Americas*. Berlin: Springer. 1-16.

Grefenstette, Gregory (1998) The Problem of Cross-Language Information Retrieval. In: Grefenstette, Gregory (Ed.) *Cross-language information retrieval*. Boston: Kluwer Academic Publishers. 1-9.

Grefenstette, Gregory & Kilgarriff, Adam (Eds.) Forthcoming. *Web as Corpus*. Special Issue of Computational Linguistics. URL http://www.itri.bton.ac.uk/~Adam.Kilgarriff/wac_cfp.html <21.09.2002>.

Picchi, Eugenio & Peters, Carol (1998) Cross Language Information Retrieval: A System for Comparable Corpus Querying. In Grefenstette, Gregory (Ed.) *Cross-language information retrieval*. Boston: Kluwer Academic Publishers. 81-92.

Wagner, Andreas (2000) *Enriching a lexical semantic net with selectional preferences by means of statistical corpus analysis*. In: Proceedings of the ECAI-2000 Workshop on Ontology Learning, Berlin, August 2000. URL http://www.sfb441.uni-tuebingen.de/~wagner/publications/ontology.ps.gz <21.09.2002>

II) Internet

Ahmad, Khurshid & Gillam, Lee (1997) *The Interval Validation ToolKIT*. URL http://www.computing.surrey.ac.uk/ai/new_interval/documents/wp4/t41/t41-0006.doc <21.09.2002>.

Evidence Network (2002) *Searching in the Social Sciences*. URL http://www.evidencenetwork.org/searching.asp <21.09.2002>.

Grefenstette, Gregory (2001) *Multilinguality in the in the Web. Search Engines: Diversity and Controversy*. The Sixth Search Engine Meeting, April 9-10, 2001, Boston, Massachusetts. URL http://www.infonortics.com/searchengines/sh01/slides-01/grefen.pdf <21.09.2002>.

Sørensen, H. Selsøe (2002) 'Find Enhance Store Share Information'. *TNP2 Workshop on New Learning Environments*, AEGEE (Association des Etats Généraux des Etudiants de l´Europe). Abstract og PowerPoint-præsentation publiceret på http://www.taalnet.rug.ac.be/tnp/fs_copenhagen.htm <07.11.2002>.