Terminologi på tværs af Danmark – og EU

Om **eTranslation TermBank** - indsamling af terminologi til maskinoversættelse

Sussi Olsen
Lina Henriksen, Bolette S. Pedersen, Claus Povlsen

Center for Sprogteknologi,
Nordiske Studier og Sprogvidenskab

UNIVERSITY OF COPENHAGEN

Indhold

- Om projektet
- Motivation - terminologi og maskinoversættelse
- Forvaltning af terminologi i Danmark
- Resultater af indsamling
- Fremtidigt arbejde

eTranslation TermBank

- EU-projekt, løber sept. 2017- feb. 2019
- Formål: Anspore til indsamling af terminologiske ressourcer inden for specifikke fagdomæner for at forbedre kvaliteten og dækningsgraden af EU's maskinoversættelsessystem CEF eTranslation.
- Ressourcer for alle officielle EU-sprog + norsk (bokmål og nynorsk) og islandsk
- Fokus på domæner knyttet til følgende online-tjenester:
  - Sundhed – eHealth
  - Retsinformation – e-Justice
  - Forbrugerbeskyttelse – ODR Open Dispute Resolution

Co-financed by the European Union
Connecting Europe Facility

Projektet udsprunget af ELRC: European Language Resource Coordination

- Netværk der indsamler sprogressourcer for alle EU-sprog til CEF eTranslation.
- Fokus på parallelle ressourcer, oversatte tekster.
- Man kan selv uploade sprogressourcer til websitet: http://lr-coordination.eu/resources
- Indsamler også terminologi, tung proces -> eTTB-projektet

European Language Resource Coordination

Partnere i projektet

- Tilde, Letland (koordinator)
- Københavns Universitet, Danmark
- Árni Magnússon Institute for Icelandic Studies, Island
- Institute of the Estonian Language, Estland
- Jožef Stefan Institute, Slovenien
- International Network for Terminology, Østrig
- Institute for Language and Folklore, Sverige
- Swedish Centre for Terminology, Sverige
- Institute of the Lithuanian language, Litauen

Hver partner indsamler terminologi for flere sprog.
Danmark er også ansvarlig for norsk og ungarsk.

Hvem er eTTB i DK?

Medarbejdere ved Center for Sprogteknologi

- Lina Henriksen linah@hum.ku.dk
- Sussi Olsen saolsen@hum.ku.dk
- Bolette Sandford Pedersen bspedersen@hum.ku.dk
- Claus Povlsen cpovlsen@hum.ku.dk




Motivation – terminologi og maskinoversættelse

Fejloversættelse af termer – ex: sundhedsdomænet

Eksemplerne er fra professor emerita Bodil Nistrup Madsen, CBS

Dansk (fra begrebsbasen http://sundhedsdata.iterm.dk)	Engelsk (Google)	Definition
engangsdosering	prefilled *Alternativ:* once a day regimen	"enkeltstående dosering af én lægemiddeldosis"
pausering	break ring	"midlertidig afbrydelse af lægemiddelordination efter anvisning af en dertil autoriseret person"

Maskinoversættelse og terminologi

Regelbaseret maskinoversættelse – start i 1940'erne, populære i 80'erne

- Ordbøger og håndskrevne grammatiske regler, terminologi var let integrerbar

Statistisk maskinoversættelse – også kaldet phrase based - overtog i løbet af 90'erne

- Arbejder ud fra store mængder af parallelle tekster på de to sprog, danner sætninger ud fra lign. sætninger i træningsmaterialet. Mere flydende oversættelser. Terminologi – skal håndteres særskilt

Maskinoversættelse og terminologi

Neural maskinoversættelse – siden 2014

- Kontekstbaseret, betjener sig af 'deep learning', arbejder med vektorer, kræver mindre computerhukommelse
- Danner mere flydende sætninger (posteditering nedsat med 25%) og dermed bedre output
- Terminologi skal 'indlæres'

Anvendelse af terminologi i maskinoversættelse

- Monolingual: Præprocessering – domænebestemmelse
- Bilingual: en term skal 'overrule' det ord som ellers typisk optræder i en tilsvarende kontekst

Behov for terminologi inden for alle domæner

Hvad er CEF eTranslation?

CEF står for Connecting Europe Facility – et EU-program

Investerer i og støtter opbygningen af digitale infrastrukturer som kan forbedre dagliglivet for europæiske borgere, samt byggeblokke hertil.

Eksempler på digitale infrastrukturer (DSIs):

- ODR - online dispute resolution portal
- e-Justice
- eHealth

Eksempler på byggeblokke:

- eDelivery
- eInvoicing
- eSignature
- **eTranslation**

eTranslation er en udbygning af EU's tidligere maskinoversættelsesystem.

Terminologi i Danmark

Forvaltning af terminologi i Danmark

- Ingen central organisation for terminologiforvalting, modsat fx Sverige og Norge
- 1998-2015: DanTermCentret arbejdede med at informere om og oplære i systematisk behandling af terminologi (få midler, lukket 2015).
- CBS – havde forskning i terminologi, afd. lukket
- CST, KU deltog i et EU-terminologiprojekt i 1992 hvor termbankerne IATE og EuroTermBank blev udviklet.
- Sprogteknologisk Udvalg – nedsat af regeringen 2018 - skal afklare behovet og perspektiverne for en national termbank.

Terminologinetværk i Danmark

FORVIR – Forum for Videnmodellering i Offentligt Regi

- Et netværk for ansatte i den offentlige administration. Formålet er at styrke terminologisk begrebsarbejde kvantitativt og kvalitativt

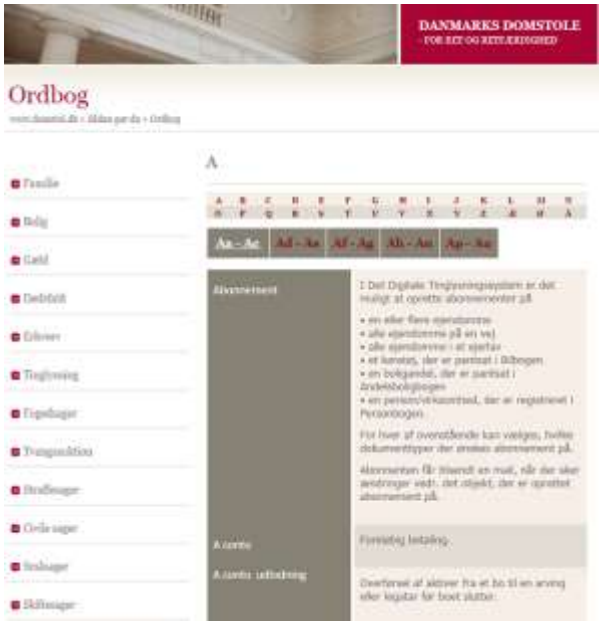
Terminologigruppen

- Et netværk for både den private og offentlige sektor. Formålet er at udveksle erfaringer og deltage i nordisk og internationalt terminologiarbejde

Terminologiarbejde i danske offentlige institutioner

Visse offentlige institutioner gør meget ud af terminologi- og begrebsarbejde, fx

- Sundhedsdatastyrelsen
- Socialstyrelsen
- Domstolsstyrelsen



Terminologiarbejde i danske offentlige institutioner

MEN

- Ca. 75 % af de offentlige institutioner som vi har henvendt os til, udfører ikke terminologiarbejde og har ingen termsamlinger.
- Kun 25% udfører terminologiarbejde, de fleste kun på dansk
- Resten 'outsourcer' deres oversættelser til oversættelsesbureauer eller
  klarer sig med google!

Indsamling

Hvem kan levere terminologi til eTTB?

I princippet alle.

Vi har primært henvendt os til offentlige institutioner:

- De er forpligtede til at levere deres data if. PSI-direktivet (Lov om videreanvendelse af den offentlige sektors informationer)
- Adgang til CEF eTranslation er lige nu forbeholdt offentlige institutioner (det forlyder at systemet bliver åbent for alm. borgere og private virksomheder fra 2020)

Vi har også henvendt os til halvoffentlige og private institutioner som vi har hørt skulle råde over terminologi.

Manglende interesse for og vilje til at levere terminologi

Det kræver meget tid og gentagne henvendelser at få leveret terminologi.

- Største problem: manglende kendskab til og interesse for CEF eTranslation

Mange spørgsmål og megen usikkerhed

- Hvem vil bruge terminologien og hvordan?
- Hvor lagres den?
- Kan data blive misbrugt?
- Hvad nu hvis vi opdaterer vores data og denne opdatering ikke sker I det vi afleverer?
- Kan vores data ende med at gøre mere skade end gavn?

Foreløbigt resultat af indsamlingen

Vi har henvendt os til godt 30 institutioner

Vi har modtaget terminologi fra nedenstående:

- Socialstyrelsen
- Beskæftigelsesministeriet
- Københavns Universitet, CIP 
- Forsknings- og Uddannelsesministeriet
- Sundhedsdatastyrelsen
- LO
- Privat terminolog

Konklusion - og de sidste måneder i projektet

Erfaringer fra indsamlingen – udfordringer

- Uvidenhed om hvor værdifulde tekster og termlister er for forskning og udvikling i dansk sprogteknologi
- Mange offentlige institutioner anvender hverken terminologiske resurser el.lign. som kan sikre ensartet tekstkvalitet
- sprogarbejde er ikke prioriteret - udliciteres
- Derfor ingen stor interesse for CEF eTranslation
- Usikkerhed om IPR, opdatering af versioner etc.
- Vi har ikke en klar fælles sprogpolitik og fælles fokus på sproglige problemstillinger i det offentlige (som andre nordiske lande)

Fremtidige opgaver i eTTB

- Flere ressourcer meget velkomne, så hvis I har kendskab til terminologi – så kontakt os endelig
- Det samme gælder hvis I kender til større parallelsprogede publikationer
- Alle domæner er relevante

Ellers vil vi nu koncentrere os om norske og ungarske ressourcer – for

- indsamling af dansk terminologi har vist sig at være temmelig tungt, men det er en leg sammenlignet med indsamling af ungarsk terminologi.

Tak for jeres opmærksomhed! Spørgsmål?