HUMANIORA
KØBENHAVNS UNIVERSITET

# **Data og sprogressourcer**
# ved Center for Sprogteknologi, Nordisk Forskningsinstitut, Københavns Universitet

Bolette Sandford Pedersen
Institutleder v. Nordisk Forskningsinstitut, KU

# Data og sprogressourcer

Center for Sprogteknologi ved Nordisk Forskningsinstitut, KU, koordinerede registrering og analyse af data og sprogressourcer til hvidbogen om dansk sprogteknologi

”Det danske sprog i den digitale tidsalder” 2012

Planlægger opdatering af rapporten

# Data og sprogressourcer

Indsamling, udvikling og berigelse af sprogressourcer til sprogteknologi har i mange år været en del af Center for Sprogteknologis mission

- Både leksikografiske og tekstresurser
- Både ensproglige og flersproglige/parallelle korpora
- Alignering af korpora
- Både ”rene” data og opmærkede data
- Manuel og automatisk opmærkning

# Data og sprogressourcer

- Vi anvender primært to platforme til opbevaring og deling af data: CLARIN og METASHARE

- CLARIN (Common Language Resources and Technology Infrastructure): primært for forskere, inklusiv værktøjer
- META-SHARE: bredere målgruppe, sprogteknologiske interessenter

# Parallelle CLARIN-ressourcer

- Juridiske tekster fra JRC Acquis: Parallelle tekster for dansk, engelsk og tysk med 52 mio. tokens for dansk, 56 mio. tokens for engelsk og tilsvarende tekster for tysk.
- Generelle tekster fra **Rapid** (pressemeddelelser): Parallelle tekster for dansk, engelsk og tysk.  3,0 mio. tokens for dansk, 3,2 mio. tokens for engelsk, og delvist også disse tekster for tysk.
- **Årsberetninger** for dansk og engelsk med 2,3 mio. tokens for dansk og 2,1 mio. tokens for engelsk.
- Søren Kirkegaard, to ungdomsjournaler. Parallelle tekster for dansk, engelsk og tysk til forskningsformål. Der er ca. 90.000 tokens pr sprog.

# Language Technology Observatory


Center for Sprogteknologi deltager i EU-projekt der skal skabe en oversigt over tilgængelige sprogressourcer der kan bruges af firmaer der udvikler maskinoversættelse og af oversættelsesfirmaer.

Udvikling af metoder og best practice omkring indsamling og brug af tekster til brug for udvikling af maskinoversættelse

# Udfordringer ved tekstindsamling

Observerede problemer ved indsamling af tekster til maskinoversættelse:

- Der bruges ikke altid direkte oversættelser; man tilrettelægger i høj grad teksterne efter målgruppen.
- Eksempel fra parallelle hjemmesider: KU har en masse personaletekster på både engelsk og dansk, men teksterne er tilrettet to forskellige målgrupper.
- Ofte byttes der om på rækkefølgen af afsnit etc.

HUMANIORA
KØBENHAVNS UNIVERSITET

# Sprogteknologisk netværk

Fagrådet for fagsprog og sprogteknologi ønsker at samle et bredt felt af sprogteknologiske interessenter i et sprogteknologisk netværk

Tanken er at mødes et par gange om året for at videndele

- Udviklere/udbydere
- Offentlige og private institutioner der arbejder med sproglige data
- Sprogforskere sprogteknologiforskere
- Andre sprogteknologiske interessenter

Interesseret?

Kontakt Bolette S. Pedersen bspedersen@hum.ku.dk eller
Sabine Kirchmeier-Andersen, Dansk Sprognævn sabine@dsn.ku