Introduktion til Sprogteknologi

Forår 2020
Lecture 6

Dagens mål

- Ting fra i går, men i det analyserede output
  - Forkortelser
  - Proprier
- Stavefejl, slang, eller udvikling i sproget?
- Kort kig på andre metoder

Gammel Teknologi

Kalaallisut-analysatoren er bygget på gammel regel-baseret sprogteknologi

- Two-Level Morphology (TwolC) fra 1983
- Finite State Morphology (LexC, XFST) fra 2003
- Constraint Grammar (CG) fra 1990

Der findes nyere statistiske metoder, machine learning, neutral networks, deep learning, etc, som Google, Microsoft, og andre har stor success med. Kan de fungere for Grønlandsk?

Nyere Teknologi

Google Translate

- Første version fra 2005 var trænet på milliarder af ord fra både mono- og bilinguale corpora.
- En fordobling af tekstmængde gav kun 0.5% forbedring af kvalitet.

Google Transformer

- Fra 2017. Nyeste modeller trænet med 3 milliarder engelske ord på 1024 processorer giver 96% korrekte morfologiske analyser.

...på Grønlandsk?

- Vi har corpora med sammenlagt 20 millioner ord
- Af dem er kun 10 millioner bilinguale

Der er langt til en milliard ord.

Derudover er grønlandsk et morfologisk rigt sprog - engelsk er morfologisk næsten trivielt. De matematiske modeller virker dårligt for komplekse sprog.

Så nej, af flere grunde er regel-baseret teknologi stadig det bedste for grønlandsk.