

Titel: Korpusbaseret lemmaselektion og opdatering

Abstract:

Den Danske Ordbog (DDO) er den første og hidtil eneste korpusbaserede ordbog for dansk. At den er korpusbaseret, afspejler sig på flere planer. På det mikrostrukturelle plan er der etableret en række særlige oplysningstyper der direkte afspejler korpusobservationer, fx statistisk beregnede kollokationsoplysninger. På det makrostrukturelle plan er det selve lemmaselektionen der skal afspejle den sproglige virkelighed sådan som den kommer til udtryk i det anvendte korpus.

I oplægget vil jeg beskrive hvordan lemmaselektionen blev udført for DDO, og kort vise hvilke mangler en strengt frekvensbaseret fremgangsmåde kan have. I den forbindelse kommer jeg ind på hvordan vi har prøvet at undgå disse mangler ved lemmaselektionen for DDO.

Oplæggets hovedvægt vil dog ligge på korpusbaseret opdatering af en ordbog med nye domænespecifikke ord og betydninger. Metoden som jeg vil demonstrere, går ud på 'automatisk' at udtrække domænespecifikke vokabularer fra DDO's korpus. Disse domænespecifikke vokabularer bruges efterfølgende til at domæne-klassificere nyt, 'ukendt' tekstmateriale. Jeg vil her vise hvordan nye domæne-specifikke lemmakandidater (og betydninger) i det nye tekstmateriale kan bestemmes kvantitativt.