Abstract til oplæg i LEDA den 26.04.2017

Korpusbaseret udvælgelse og prioritering af lemmakandidater

I de kommende år skal der tilføjes flere tusinde nye ord til Den Danske Ordbog på ordnet.dk. Hovedkilden til ordbogen er korpusset *Tidsmaskinen*, som indeholder ca. 1 milliard ords løbende tekst fra perioden 1983 frem til i dag. I dette korpus gemmer der sig givetvis mange ord, som burde optages i ordbogen. Men hvordan finder man dem? Hvilke relevanskriterier kan man opstille? Og hvordan prioriterer man lemmakandidaterne? Ud over de ord, som man direkte kan trække ud af korpus, er der andre, som stammer fra en række korpuseksterne kilder, for eksempel ord, som der er blevet søgt forgæves på på ordbogens website, brugerindberetninger og redaktionsinterne forslag. Hvordan kan man bestemme relevansen af ord fra korpuseksterne kilder og eventuelt prioritere dem? Jeg vil prøve at give nogle forsøgsvise svar på spørgsmålene ved at vise, hvilke fremgangsmåder og værktøjer der bruges til udvælgelse og prioritering af lemmakandidater på Den Danske Ordbogs redaktion.

Jørg Asmussen, DSL, marts 2017