Korpussøgning – Hands-on
http://corp.hum.sdu.dk

Øvelserne er beregnet til det danske cqp-interface (rundt dansk flag), Korpus 2000, men kan også udføres på andre korpora eller for andre sprog. Ved selvstændig brug af interfacet, uden ledsagende kursus, anbefales førstegangsbrugere at se flash-filmen under "Guided Tour". Yderligere information, herunder søgeeksempler og en introduktion til regular expressions, findes i "info"- og "help"-filerne.

Stavevariation

Det danske sprog udvikler sig konstant, og i perioder er det muligt, at der findes flere korrekte staveformer for det samme ord, eller at ældre mennesker kommer til at stave "forkert" når de holder fast ved det, de har lært. Nye ord i dansk skal sommetider igennem en sproglig-"demokratisk" proces, før de får en fast flertalsform eller et fast køn.

Undersøg hyppigheden af følgende ordformer, i et korpus eller på internet. Har retskrivningsordbogen fat i fodfolket?

- tjek – check (betydningsforskel?), linie – linje (RO)
- viruset – virusset

Grammatisk variation

- en (web)site – et (web)site (bøjningsvariation: køn)
- hooliganer (RO) – hooligans (bøjningsvariation: flertal)

Dansk har to konkurrerende måder at danne komparativ på, (a) -ere, (b) mere ...
Forsøg at finde regelmæssigheder for hvilken form foretrækkes. Brug "Refine search" til at søge på adjektiver, sortér på "right edge" og "relative frequency"!

- mere vigtig – vigtigere (!)
- mere vanskelig (!) – vanskeligere

Låneord

Dansk låner konstant ord fra andre sprog. Prøv at uddrage lister af sådanne ord fra et korpus ved hjælp af bestemte bogstavmønstre. Findes der sammensatte ord med blandede "internationale" elementer?

- Latin: -isere, -tion: '.*iserer?', '.*isering'
- Græsk: meta-, geo-, syn-, -log/-logi
- Engelsk: 'th.*', 'sh.*', -ing (hvordan undgår man falsk positive?)

- Tysk: 'fore-' (er det mest verber eller substantiver?)
- Fransk: [áà] (et sæt at bogstaverne 'á' pg 'à' (Sorter efter "right context"!)
- Nordisk: hårde enkeltkonsonanter før endelse, hvor dansk har "bløde" konsonanter: '.*[aeiouy][pk]en' (jf. 'b' og 'g' i dansk)

Undertiden konkurrerer danske sammensætninger med tilsvarende latinske. Sammenlign:

- -icere vs. -gøre, -isering vs. -gørelse
- in- vs. ind- (brug "refine search" til at udelukke andet end verber!)

Engelske ord konkurrerer med danske nydannelser eller låneoversættelser, og ofte tager det en årrække før den éne form har vundet. Du kan søge på 2 eller flere former samtidigt ved at sætte dem i en fælles parentes med '|' imellem: (bærbar|laptop) – så kan køre en sammenlignende statistik med det samme.

- homepage - hjemmeside
- website – websted - webside
- e-mail – e-post
- laptop - bærbar

Nye ord: Forkortelser

Ikke alle nye ord er fremmedord eller sammensætninger. En speciel måde at præge nye ord på er forkortelser, der så selv kan blive genstand for ortografisk variation, sammensætninger eller afledninger.

- Undersøg om forkortelsen foretrækkes med stort eller småt: SMS/sms, DVD/dvd/Dvd, CD-ROM/cd-rom/CD-rom/CD-Rom (CD/cd alene: cave partiet Centrumdemokraterne!)
- Find sammensætninger med forkortelser, eller bøjninger fx. At SMS'e, SMSer: '[A-Z]+er?'

Bandeord og nedsættende udtryk er en sociolingvisk guldgrube. Undersøg områdets sproglige variation og regler!

Bandeordenes grammatik

- Er der grammatisk forskel mellem sammensætninger med hhv. 'skide-', 'møg-' og 'pisse-' på den ene side, og 'lorte-' på den anden side? Sammenlign med 'super-' og 'kæmpe-'! Find andre forstærkerforstavelser!
- Find ud af de syntaktiske regler for brugen af 'sgu'! Hvilke ordklasser kan stå henholdsvis før og efter?
- Også på bandeords-området står engelsk stærkt. Undersøg forskellen i brugen

af 'fuck...'-ord i hhv. Skriftsprogskorpora på den ene side, og internettet på den anden. Er forskellen mellem skrift og tale specifik for dansk, eller gælder den også for engelsk selv?

Et let tilgængeligt undervisningsemne inden for sociolingviskken er "sproglig kønsdiskrimination". Er der forskel på "kvindelige" og "mandlige" ord eller deres kontekst?

Sprog og køn

- Find ud af, hvilke substantiver der hyppigst står til høje for hhv. 'hans' og 'hendes' (Brug statistikknapperne 'freq' og 'rel' med "right context"). Søg først i Korpus 90, og så sammenlignende i Korpus90 og 2000 på én gang. Frekvenssortering vil så vise 2 parallele lister.
- Find sammensætninger med 'mand.+' og 'kvind', evt. 'dreng.+' og 'pige.+', og sammenlign resultaterne statistisk! Brug [a-zæøå]{4,20} i stedet for .+ for at udelukke simple bøjningsformer.

Det danske ordforråd er på en måde – en morfologisk måde, vel at mærke - mere dynamisk end det franske eller engelske, fordi ordene kan sammensættes.

Afledninger

- Brug korpus-interfacet til at finde så mange afledninger som muligt for rødderne *binde, trække, skrive, sætte.* Brug "left-edge" sortering med frekvens-knappen ("freq").
- Find det længste danske ord! Man kan bruge et antal prikker '...' for et tilsvarende andtal bogstaver, eller – smartere – et udtryk som [a-zæøå]{20}. Brug ikke (!) .+ efter mange prikker – dette overbelaster maskinen. Kan du slå *'Menneskerettighedskonventionen'* (fra den korrigerede del af Korpus90/2000) og *'Tændstiketiketsamlersammenslutningen'* (fra Korpus 2000)?
- Hvad er tilsvarende længste ord for andre sprog?

Ledstilling

Når flere ord af samme ordklasse følger på hinanden, er det som regel ikke ligegyldigt, i hvilken rækkefølge de står. Brug de statistiske redskaber i Corpuseye: "left edge" vs. "right edge", og almindelig frekvens-sortering ("freq").

- Find flest mulig adjektiver ved siden af hinanden: Har de faste pladser alt efter deres betydning?
- Find flest mulig adverbier i en række. Er der nogen, der kun kan stå først, eller kun sidst, eller kun det éne sted i sætningen?

- Find flest mulig verber oven i hinanden: Hvad kendetegner det første og det sidste verbum i kæden, henholdsvis?

Syntaktisk sprogudvikling: Dansk som festsprog

Normalt placerer dansk objekter efter hovedverbet – bortset fra festsange, hvor den gamle, "tyske" ordstilling er et yndet virkemiddel. En anden undtagelse er fokuseret materiale i sætningen, der rykkes frem i forfeltet. Find ud af under hvilke betingelser man også i moderne dansk kan finde objekter *før* hovedverbet.

- Skriv [func="ACC>"] i søgefeltet og vælg "cqp-speak", for at finde direkte (akkusativ-) objekter med verbet til højre. Det er bedst at bruge den korrigerede del af Korpus90/2000, fordi konstruktionen er sjælden og behæftet med en høj fejlrisiko i automatisk analyse.
- For at udelukke relative pronominer, skriv [pos="N" & func="ACC>"] . *Pos* er "part of speech" (ordklasse), *func* er (syntaktisk) funktion.
- Søg på "har" + "ingen" + substantiv + Predicator/main verb (brug "refine search" - For at tillade "intet" på samme tid, prøv med "ingen" som "base form" i stedet for som "word form".

Flere korpus-øvelser findes på VISL's undervisnings-server, i URKAS-sektionen under menupunktet "Almen Sprogforståelse" på http://beta.visl.sdu.dk, eller direkte: http://beta.visl.sdu.dk/urkas_pearls.html. Korpus-øvelserne er ordnet efter grammatiske emner, og findes i den blå perlerække, (E) "Test & data". For undersøgelser på egne tekster, brug "Text Painter" (http://beta.visl.sdu.dk/visl2/texttyping.htm).

Der henvises i øvrigt til følgende publikationer, der kan hentes i pdf-format på: http://beta.visl.sdu.dk , under link'et "Publications":

Bick, Eckhard (forthcoming), CorpusEye:Et brugervenligt web-interface for grammatisk opmærkede korpora, In: Peter Widell & Mette Kunøe (eds.), 10. Møde om Udforskningen af Dansk Sprog 7.-8.okt.2004, Proceedings. pp.xx, Århus University

Bick, Eckhard (forthcoming), Live use of Corpus data and Corpus annotation tools in CALL: Some new developments in VISL, In: Henrik Holmboe (red.), Nordic Language Technology, Årbog for Nordisk Sprogteknologisk Forskningsprogram 2000-2004 (Yearbook 2004). pp.xx. Copenhaguen: Museum Tusculanum