Brug af internetbaseret sprogkorpus på sprogstudiet

Det er de færreste studerende der har hørt om korpusundersøgelser og ved hvad et sprogkorpus er, når de starter på et sprogstudium. Ikke desto mindre har mange studerende ganske gode forudsætninger og en latent motivation for at arbejde med internetbaserede sprogkorpora som en del af deres sprogstudium. Vi har på spansk-studiet på Københavns Universitet været positivt overraskede over at brugen af fagspecifikke korpora har virket nærmest magisk på mange studerende. Det skyldes formentlig at internetbaserede sprogkorpora (i højere grad end andre it-værktøjer som Excel, PowerPoint m.m.) åbner en helt ny verden for de studerende og samtidig ligger i naturlig forlængelse af mange studerendes generelle it-kompetencer.

De fleste studerende er vant til fra gymnasiet og fra deres sociale liv at søge information på internettet, bruge søgemaskiner som Google og kommunikere med hinanden via Facebook. Det er generelt set en naturlig del af deres hverdag at bruge internettets faciliteter, herunder digitale informationskilder. Og det er noget de ofte er ret ferme til. Derfor er det en god ide at introducere de studerende til korpusundersøgelser af sproglige fænomener allerede på første semester af deres sprogstudium. Dels fordi de fleste således har gode forudsætninger for at kunne lære at arbejde med sprog ved brug af internetbaseret sprogkorpus – og man derved tager udgangspunkt i en stærk kompetence hos mange unge – og dels fordi det er en helt ny og spændende måde at arbejde med sproget på som de slet ikke kender til i forvejen, og som kan bane vejen for en øget interesse for de sproglige sider af et sprogstudium.

JOHAN PEDERSEN
Lektor i spansk, Københavns Universitet
jhp@hum.ku.dk

Et internetbaseret sprogkorpus er en digitaliseret, ofte meget stor tekstsamling af skrevet og talt sprog der er tilgængelig via internettet. Tekstsamlingen er sammensat med henblik på at være repræsentativ og afbalanceret i forhold til den gældende variation i sprogbrugen, og den har til formål at danne grundlag for sproglige analyser af sprogbrugen. Det er ikke blot sprogbrug og sproglige problemstillinger, men også kulturelle fænomener i bred forstand der vil kunne belyses på grundlag af sådanne analyser.

Ofte er holdningen blandt undervisere på et sprogstudium at særligt interesserede sprogstuderende vil kunne arbejde med korpuslingvistik på sidste del af studiet, ud fra den betragtning at det ikke er så let at gå til og kræver et vist flair for computerprogrammer; med andre ord, at det er meget specialiseret og lidt nørdet. Det har fx betydet at der på spanskstudiet på Københavns Universitet tidligst i forbindelse med bachelorprojekt eller emneopgaver på kandidatniveau har været studerende der har lavet korpusundersøgelser. Problemet har dels været at de først sent og lidt tilfældigt, måske via kontakt til en særligt interesseret underviser, bliver introduceret til dette fagområde, og dels at der har været en diskrepans mellem på den ene side de studerendes metodologiske og redskabsmæssige kompetencer – dvs. basale kompetencer i brug af sprogkorpus – og på den anden side de opgaver de ønsker at lave. Dertil kommer at interessen, hvis den opstår, kommer på et tidspunkt i studieforløbet hvor der til sådanne fagelementer kun tilbydes begrænset individuel vejledning og undervisning på højere fagligt teoretisk niveau (kandidatniveau). Det ligger selvsagt ikke lige for at udbyde basiskurser med omfattende redskabsmæssige elementer på dette niveau.

De senere år har der været gjort en massiv indsats af specialister i korpuslingvistik for at udvikle brugerfladen i de sprogkorpusapplikationer der findes på markedet. Hertil kommer at mange nu er frit og let tilgængelige via internettet. Det betyder at det til forskel fra tidligere nu er relativt let at komme i gang med at lave korpusundersøgelser. For tilrettelæggelsen af et sprogstudium betyder det at man nu kan introducere de studerende til denne måde at arbejde med sprog på allerede i starten af sprogstudiet. Fordelen ved at udvikle disse kompetencer fra starten af er selvfølgelig at de kan bruges i mange af de andre fag de studerende har og skal have senere i studiet, og at kompetencen løbende kan videreudvikles. Hvis de studerende allerede fra første semester udrustes med viden om og basale færdigheder i brugen af forskellige korpusværktøjer – det vil primært sige at kunne søge i et sprogkorpus og analysere og præsentere korpusdata – vil de kunne bruge det undervejs i studiet, og

de vil være rustet til at kunne arbejde med mere komplekse og teoretisk mere krævende opgaver og problemstillinger i forbindelse med bachelorprojektet samt emneopgaver og speciale på kandidatstudiet.

I faget *Sprog og it* har vi på spanskstudiet på Københavns Universitet i samarbejde med Center for Sprogteknologi med succes lavet forsøg med allerede på første semester at introducere de studerende til arbejdet med sprogkorpora. Ordningen er nu gjort permanent. De bliver bl.a. undervist i følgende elementer: Hvad er et sprogkorpus? Søgninger i forskellige korpustyper, fx monolingualt korpus og parallelkorpora (der fx består af originaltekster og oversatte versioner af teksterne); bearbejdning af data: simple korpusanalyser og frekvensberegninger; brug af fagspecifikke korpora, fx Det Spanske Akademis leksikalske sprogkorpora (CREA og CORDE) med avancerede, men let anvendelige søgefaciliteter der bl.a. giver mulighed for at undersøge dialektal variation, teksttypevariation og sprogforandringer; og korpora som Corpus del Español der giver mulighed for at anvende en mere kompleks søgestreng (ordene er fx annoteret for grundbetydning, ordklasse m.m.). Som eksamensprojekt skal de studerende i en opgave eller rapport arbejde med at dokumentere et bestemt sprogligt fænomen der eventuelt afspejler et kulturfænomen. Det kan fx være en bestemt sprogbrug: ordsprog, slang, dialektal sprogbrug, eller det kan være et grammatisk fænomen. De skal desuden redegøre for hvilke korpusfaciliteter og andre it-faciliteter de har til rådighed på sprogstudiet og for hvorledes de har kunnet bruge dem, og hvilke overvejelser de i øvrigt har gjort sig i forbindelse med arbejdet. Endelig skal de vise at de kan formidle deres resultater i en kort PowerPoint-præsentation der vedhæftes opgaven samt redegøre for deres overvejelser i forbindelse med at lave en sådan præsentation.

Vi ønsker med dette fagelement at sætte fokus på de metodologiske og indholdsmæssige aspekter i praksis, dvs. hvordan man kan bruge et korpus som datakilde til at belyse et sprogligt/kulturelt fænomen på et sprogstudium. Det er altså ikke meningen at de studerende skal bruge den første måned til at finde ud af hvad der interesserer dem, hvad de skal skrive om, problemformulering osv. Fokus ligger i højere grad på at de studerende skal opøve færdigheder i at bruge sprogkorpus og reflektere over og redegøre for de muligheder, udfordringer og problemer der ligger i at belyse et sprogligt/kulturelt fænomen i forbindelse med brug af sprogkorpora. Derfor hjælper vi på kurset de studerende med en masse gode ideer til hvad der kunne være interessant og muligt at undersøge.

Mange emner er velegnede til at blive belyst ved hjælp af Det Spanske Akademis korpora der hovedsageligt lægger op til søgninger på bestemte ord eller udtryk, men samtidig er meget avancerede med hensyn til at facilitere fremskaffelsen af omfattende og strukturerede data om den sproglige variation. De studerende har eksempelvis lavet opgaver hvor de dokumenterer brugen af bestemte spanske navne i forskellige områder af den spanskprogede verden med inddragelse af disse sprogområders særegenhed og kultur. Hvilke navne er mest hyppige? Er der dialektale forskelle? Hvordan har navnebrugen ændret sig? Nogle grupper har belyst forskelle i brugen af »kælenavne« (kortformer) i de forskellige sprogområder, fx *Paco* for *Francisco*, *Pepe* for *José*, *Tere* for *Teresa*, eller *Maribel* for *María Isabel*. Det har også været populært at dokumentere ændringer i forekomsten af bestemte ord og faste udtryk, herunder slang, typisk med henblik på regionale forskelle.

Andre grupper af studerende har undersøgt hvordan *det at være noget* udtrykkes i det spanske sprog, og hvordan brugen har ændret sig. På spansk bruger man to forskellige kopulative verber for 'at være': *ser* og *estar*. Brugen af *ser* er klassificerende idet subjektet ses i forhold til generelle normer og kategorier: *María es católica* 'María er katolik'. Med *estar* anskuer man subjektet i en given situation set i forhold til andre potentielle situationer: *María está contenta* 'María er glad'. I visse tilfælde kan man bruge begge verber med samme betydning: *María es/está casada* 'María er gift'. Grupper af studerende har på basis af korpussøgninger lavet empiriske undersøgelser af hvorledes denne sprogbrug kan variere regionalt, historisk og i forskellige teksttyper. De har fx kunnet dokumentere at man i et historisk perspektiv kan observere en stigende tendens til at bruge *estar* i stedet for *ser*. *Estar* har som grundlæggende betydning 'at være et sted', eller 'være i en bestemt situation/fase'. I forbindelse med udviklingen fra *María es casada* til *María está casada* er en hypotese der ligger lige for, selvfølgelig at en sådan ændring i sprogbrugen kan afspejle en kulturel udvikling i de spansktalende (katolske) lande i retning af en opfattelse af ægteskabet som en fase i livet af længere eller kortere varighed, snarere end et ubrydeligt forhold der etableres for resten af livet. De studerende har også undersøgt brugen af *ser feliz* versus *estar feliz*, der betyder 'at være glad, lykkelig'. Her har de dokumenteret at *estar feliz* i højere grad forekommer i talesprog end i skriftsprog, og at udtrykket i højere grad end *ser feliz* bruges i betydningen 'at være glad/tilfreds'. Et besnærende og vel også kvalificeret gæt er at udviklingen i denne sprogbrug i en eller anden forstand kan afspejle en udvikling i den gængse opfattelse af det at være lykkelig, i retning af at være en følelse af situationsbundet tilfreds-

◀ INDHOLD

hed, en lykketilstand af kortere eller længere varighed, snarere end at være en mere permanent lykketilstand. Det er selvfølgelig ikke meningen at undersøgelserne skal kunne holde vand i videnskabelig forstand – dette er vigtigt at tage op og diskutere med de studerende, også for at sikre at der ikke opstår en misforstået, forsimplet opfattelse af hvad videnskabelighed er; de studerende skal dokumentere mulige tendenser i sprogbrugen og gerne opstille og diskutere mulige, interessante fortolkninger af data, inklusive relevante forbehold.

Nogle studerende har været interesseret i den regionale brug af bestemte former af ordene der ligeledes kan afspejle væsentlige kulturelle elementer; fx de såkaldte diminutivformer i spansk – fx *abuela* 'bedstemor' → *abuel**ita*** 'lille bedstemor'. I dette eksempel skal diminutivformen ikke forstås bogstaveligt (at bedstemor er 'lille'), men snarere som en ømhedsmarkør. Man kan påvise varierede mønstre i sprogbrugen; fx har forskellige regioner et forskelligt inventar af diminutivformer der ligeledes optræder med varierende hyppighed og betydning. Andre grupper har kunnet dokumentere et ændret mønster i brugen af særlige femininumsformer i arbejdstitler, hvis køn ellers er maskulinum (fx den kvindelige læge: el médico → la médica). Også disse data vil kunne fortolkes som en sproglig spejling af en kultur og et samfund i forandring – i dette tilfælde med hensyn til den øgede tilstedeværelse af kvinder på arbejdsmarkedet i spansktalende lande.

Også mere grammatisk orienterede problemstillinger har vakt interesse hos de studerende, fx brugen af pronomenet *vos* 'du/I' i et historisk og regionalt perspektiv – Spanien sammenlignet med Argentina, Uruguay og andre latinamerikanske lande. Eller empiriske undersøgelser af *leisme*-problematikken, dvs. brugen af *le*(*s*) (pronomen i dativ) for direkte objekt, i forskellige regionale varianter, og/eller set i en historisk sammenhæng.

Det har været en stor udfordring for de studerende i dokumentationsprojekterne at vurdere hvad man egentlig kan sige ud fra data, og at tage højde for hvad den tekstmæssige sammensætning af et sprogkorpus betyder for fortolkningen af data. Vi forventer ikke at de studerende på første semester skal kunne lave en korrekt designet korpusanalyse der fx regulerer for en eventuel skæv sammensætning af korpus, men vi forventer at de diskuterer problemet i deres opgave. Nogle af de dygtige studerende har ud over at diskutere problematikken konkret forholdt sig til den i den måde de har designet deres undersøgelser på, fx ved at præsentere data som tal der er vægtet i forhold til korpussammensætningen. Generelt har det vist sig at det vanskelige for de studerende i mindre grad har

været af teknisk art, men snarere har været refleksionen over hvad data kan sige om et sprogligt fænomen, hvilket omfang data bør have for at man kan konkludere noget fornuftigt, hvilke karakteristika, herunder søgefaciliteter der gør et bestemt sprogkorpus velegnet som datakilde til ens projekt, tekstsammensætningen i korpus osv. Disse udfordringer repræsenterer på sin vis generelle forskningsmæssige udfordringer som vi netop ønsker de studerende skal møde på et universitetsstudium, og gerne så tidligt som muligt.

De studerende vil på et senere tidspunkt i deres studium kunne kikke tilbage i deres afrapportering/dokumentationsopgave fra første semester for at se hvordan de helt grundlæggende kan bruge et sprogkorpus, hvis de igen ønsker at lave korpusundersøgelser; kurset på første semester har også den funktion at det giver de studerende en masse ideer til emner og typer af realistiske problemstillinger de vil kunne undersøge mere indgående senere i deres studium. Dokumentationsopgaven kan være et praktisk udgangspunkt for at videreudvikle deres færdigheder og forfølge interesser inden for dette område. Dertil kommer at flere studerende giver udtryk for at de i adskillige af fagene i de følgende semestre konstant bruger de sprogkorpora de er blevet introduceret til på kurset i første semester.

I de studerendes evalueringer af kurset 'sprog og it' har der været meget forskellige meninger om hvilke it-kompetencer der skal undervises i, og hvilken vægt de skal have i undervisningen. Der har dog været udbredt enighed om at viden om og træning i brug af korpusværktøjer er meget nyttig og vigtig på et sprogstudium, og at det skal være kernen i kurset. Mange studerende oplever at de dermed får nogle it-kompetencer de kan bruge fremadrettet i både studium og job.

Litteratur

Corpus del Español – Mark Davies: http://www.corpusdelespanol.org/

CREA/CORDE – Real Academia Española: http://www.rae.es/rae/gestores/gespub000019.nsf/voTodosporId/D55F5BFB05D63980C1257164003F02E5?OpenDocument&i=2