Ord og IT

Esben Alfort

With so many new digital language tools being developed these years, it is easy to fear for the immediate future of traditional lexicography. However, no digital tool can be perfected without lexicographic knowledge being supplied by someone who really understands the way of words. In fact, the need for lexicographic skills has never been greater.

1. Et nyt syn på tingene

Da jeg sidste år var blevet inviteret til at sidde i et panel som skulle diskutere ordbogens fremtid, gjorde jeg det meget klart at jeg som repræsentant for den sprogteknologiske virksomhed Ankiro føler mig overbevist om at de traditionelle ordbøger også har en rolle at spille fremover. Ganske vist dukker der hele tiden nye smarte digitale værktøjer op som giver unikke muligheder for at udforske sprog og finde frem til den helt rigtige betydningsnuance, men det ændrer ikke ved det faktum at det traditionelle ordbogsformat, hvad enten det er på papir eller som elektronisk dokument, har sine styrker og klare fordele, og derfor vil der utvivlsomt også fremover være efterspørgsel efter dem. Alligevel er der ingen tvivl om at brugen af især papirordbøger er faldet mærkbart med den øgede adgang til mobile elektroniske alternativer. Derfor sker det også let at man som leksikograf bukker under for en ulmende frygt for at de digitale værktøjer helt vil overtage markedet og tage livet af ordbogsbranchen. Hvis man udelukkende ser de elektroniske sprogværktøjer som en trussel, går man imidlertid glip af et stort potentiale for gensidigt givende samarbejder, for der er stor brug for folk med leksikografisk og anden sproglig viden og kunnen inden for den sprogteknologiske branche. Computere er notorisk uintelligente, så uden denne viden og indsigt bliver værktøjerne aldrig tilfredsstillende.

Selv er jeg uddannet teoretisk lingvist fra Københavns Universitet og havde egentlig forestillet mig en helt anden karriere da jeg ved et tilfælde havnede på Ankiro for 7-8 år siden, hvor der var brug for en svenskkyndig til at lave det der skulle blive vores svenske digitale ordbog. Siden lavede jeg

en finsk ordbog, og med tiden fik jeg ansvar for en række forskellige projekter, bl.a. udtræk af forskellige slags information ved hjælp af natursprogsparsing. I dag er jeg medejer og leder af forskningsafdelingen, som opretholder nære kontakter til universitetsverdenen. Jeg blev meget hurtigt fænget af atmosfæren i Ankiro, både fordi firmaet har et stærkt fokus på fordybelse og forskning, og fordi der opstår en utroligt kreativ atmosfære når man sætter netop sprogfolk og programmører sammen. Det er to meget forskellige mennesketyper med et vidt forskelligt syn på næsten alt, men ikke desto mindre med en fælles grundlæggende fascination af *systemer* og en udviklet strukturforståelse. Forskellene betyder i praksis at man er tvunget til at nytænke og redefinere alting før man bliver i stand til at formidle sine tanker på tværs af afdelingerne. Det lyder måske besværligt, men resultatet er at begge parter forstår fænomenerne langt bedre, hvad enten det drejer sig om sprog eller IT.

Ved at fortælle lidt om hvad vi laver i Ankiro og hvad det er for nogle sproglige udfordringer vi dagligt brydes med, håber jeg at det vil lykkes mig at formidle det store behov der er for sprogfolk som os – inkl. leksikografer – i den digitale verden.

Ankiro lever af at lave sprogteknologi – et begreb der dækker over en lang række forskellige løsninger og produkter, alt efter hvad vores kunder har brug for. Fælles for alle vores løsninger og produkter er at de involverer kombinationen af IT og sprog, hvilket nærmest pr. definition kan være temmelig udfordrende, og fælles for os der arbejder i Ankiro, er derfor også at vi elsker sproglige udfordringer. Opgaverne skal faktisk helst være på grænsen til det umulige. Vi opererer primært på dansk, svensk, bokmål, nynorsk og finsk, samt i skrivende stund engelsk, tysk, hollandsk og fransk for de kunder der har brug for det, men Norden har altid været vores kerneområde.

Det hele startede i 1999 med intelligent søgning. Idealet har fra starten været at vores søgemaskiner skal arbejde på brugernes præmisser og finde det de leder efter og ikke bare det de skriver, og det er kun muligt ved at tilknytte en række sprogressourcer i form af synonymordbøger, ontologier og forskellige slags parsere og regelværktøjer. Der er jo ingen garanti for at brugerne formulerer sig på samme måde som forfatteren af de dokumenter de leder efter, eller for den sags skyld at de overhovedet tilstræber at efterligne sproget i det søgte idealdokument.

Når man arbejder med sprog, må man desuden altid være forberedt på at ordene har en ekstra overraskelse i baghånden. Hvad er f.eks. forskellen på *lakridsstang* og *stanglakrids*? I hvilken udstrækning er man interesseret i det ene hvis man søger efter det andet? Den slags materiale-og-form-ord (i Ankiro kaldet "stanglakridsord" – andre eksempler er *sandstrand* og *kursus-uge*) er bare en af mange sproglige finurligheder der må håndteres med en særlig regel specialdesignet til netop det formål.

Ret hurtigt fandt man i Ankiro på at bruge de ressourcer man havde udviklet til søgning til også at lave nogle store jobportaler. Hver nat gennemsøger vi det danske internet for alt hvad der ligner et jobopslag, og derefter gør vi det muligt for jobsøgende at søge i databasen med semantiske hjælpemidler. Man kan f.eks. søge på *leksikograf Amager* og få foreslået et job som ordbogsredaktør i København, fordi systemet ved at en leksikograf næsten er det samme som en ordbogsredaktør og at København ligger tæt på Amager. Projektet blev en stor succes og udgør i dag Danmarks største jobportal.

En anden ting vi bruger vores sprogressourcer til, er informationsudtræk af forskellig art. Vi analyserer store tekstkorpora automatisk og udtrækker bestemte oplysninger ved hjælp af møjsommeligt udformede abstrakte sproglige regler. Igen er det kun muligt fordi vi kan trække på de ordbøger og ontologier m.m. som vi har udviklet gennem årene og fortsat løbende forbedrer. Man kan nemlig ikke gøre reglerne tilstrækkelig abstrakte uden at tilføje viden om synonymi, taksonomi og syntaktiske variationsmuligheder.

Vi ligger inde med de sidste mange års jobopslag fra hele landet, og det er selvfølgelig en guldgrube af viden om hvad man forventes at kunne hvis man søger bestemte stillinger. Jeg er derfor begyndt at anvende vores informationsudtræksmetoder på disse jobdata og løbende udtrække kompetencekrav fra jobopslagene for fagbevægelsen så de jobsøgende kan blive vejledt endnu bedre. Arbejdet bliver dog aldrig færdigt, for det er et evigt kapløb med forfatternes sproglige kreativitet og variationslyst. Man når lige at lære systemet at fortolke sætninger som *desuden skal du være interesseret i og helst de sidste mindst fem år have arbejdet intenst med rørtrådssvejsning*, før man mødes af formuleringen *du har en ivrig rørtrådssvejser i maven*…

2. Udfordrende ord

Også søgemaskiner er som sagt afhængige af sproglig viden hvis de skal leve op til idealet om at fungere på brugernes præmisser. Søgning er jo kommunikation mellem en bruger og et system, og derfor kommer der også alle mulige sproglige forhold ind som komplicerer processen og som man er nødt til at finde løsninger på. For det første indeholder 10-20 % af alle søgninger mindst én stavefejl, afhængig af domæne og målgruppe, og de kan sommetider være meget alvorlige. Hvad menes der f.eks. med ordet *tredbrøn*? Man kan måske få en idé når man ser den fulde søgestreng *trådløs tredbrøn*, især når man ved at søgningen blev udført på hjemmesiden for et firma der leverer bredbåndsforbindelser, men det er ikke nemt for en computer at afgøre.

For det andet udtrykker folk sig som nævnt ikke nødvendigvis på samme måde som den virksomhed eller institution der ejer det site de søger på. Meget ofte vælger de et andet ord end det der bruges i det dokument de leder efter, og tit bruger de desuden andre morfologiske former og syntaktiske konstruktioner. Her følger to eksempler.

1. enddentitetstyveri
2. id tyv
3. id tyveeri
4. id tyveri
5. idenditets tyveri
6. idenditetstyver
7. idenditetstyveri
8. idenditettyveri
9. ideniitets tyveri
10. idensitets tyveri
11. idensitetstyveri
12. identitesttyveri
13. Identitestyveri
14. identitet tyveri
15. identitetestyveri
16. identitets tyveri
17. identitetsstyveri
18. identitetstyveri
19. identitettgyveri
20. identitettyveri
21. identitstyveri
22. identitstyveri’
23. ididentitetstyveri
24. IDtyveri
25. idtyveri
26. indentitet tyveri
27. indentitetstyveri
28. indetitetstyver

1. rotte
2. rotter
3. rotter på loftet
4. skadedyr rotter
5. rotteplage
6. anmeldelse af rotter
7. rottefanger
8. rottefænger
9. rottemand
10. rottejagt
11. rotteudryddelse
12. rotte bekempelse
13. rotte bekæmpelse
14. rottebekaempelse
15. rottebekmpelse
16. rottebekæmbelse
17. rottebekæmpelse
18. bekæmpelse af rotter

Det første materiale stammer fra en søgelog knyttet til en stor dansk virksomheds intranet, hvor jeg fandt 28 forskellige måder at referere til identitetstyveri på. Ikke alle er stavefejl; for at fortolke disse udtryk skal systemet både bruge information om synonymer, forkortelser og afledninger, foruden en kraftig intelligent stavekontrol.

Det andet eksempelmateriale er hentet fra en kommunal søgelog og består af 18 forskellige måder at søge om hjælp til rottebekæmpelse på.

Der er en hel del sproglige udfordringer bare i dette lille materiale – stavefejl, morfologisk variation (ental vs. flertal), syntaktisk variation (*rottebekæmpelse* vs. *bekæmpelse af rotter*), synonymvalg (*bekæmpelse* vs. *udryddelse*), og ren og skær forskel i tilgang (*rotteplage* vs. *rottefænger* vs. *rottebekæmpelse*). De personer der har foretaget alle disse søgninger har formentlig alle sammen haft omtrent samme intention – de vil finde information om rottebekæmpelse, og sandsynligvis også anvende denne information til at tilkalde hjælp, men de går til opgaven på helt forskellig vis.

Stavekontrollen er bygget op omkring en kombination af tastetekniske og fonetiske regler. De tastetekniske regler tager højde for at man kan ramme forkert, så bogstaver der ligger tæt ved hinanden, eller f.eks. i samme finger i modsat hånd, kan blive forbyttet, mens de fonetiske regler afspejler det enkelte sprogs lydsystem og specificerer de mest oplagte muligheder for ortografisk forveksling af ens eller næsten ens lyde. Disse regler er det første vi laver, når vi tager hul på et nyt sprog.

Selv med sådan et kraftfuldt stavekontrolapparat er man dog aldrig helt sikker, for konteksten kan også spille ind i fortolkningen, som vi så med

bredbåndseksemplet. Hvis man som menneske ser søgestrengen *kartofler carutter*, så er man ikke i tvivl om at der skulle have stået *karotter*, men hvis søgestrengen havde været *cigarer carutter*, så ville man have tolket det samme ord som *cerutter*.

Hvis man vil være i stand til at tage højde for konteksten i den slags situationer, er man nødt til at modellere det relevante emneområde. Til det formål opbygger man en ontologi, dvs. en formel struktur af begreber bundet sammen af forskellige relationer, f.eks. *kartofler* og *karotter* som underbegreber til én knude og *cigarer* og *cerutter* under en anden. Ofte vil kun den ene betydning være relevant for en given kunde, fordi den anden falder uden for domænet. Så kan konteksten være underforstået på det sproglige plan fordi den er knyttet til sitet som helhed.

Den mest klassiske måde at opbygge ontologier på er som en taksonomi af over- og underbegreber; når en zebra er et pattedyr og et pattedyr er et dyr, så er det måske oplagt at en der søger efter information om dyr også potentielt vil være interesseret i information om zebraer. Helt så simpelt er det desværre ikke, for information kan være generel eller specifik, og i nogle tilfælde er det decideret problematisk at inddrage underbegreber i søgemaskineriet.

Hvis man søger efter dokumenter der handler om *kæledyr*, er det ikke utænkeligt at man i mange tilfælde også kunne være interesseret i at finde information om katte. Man kunne derfor lægge *katte* ind i ontologien som underbegreb til *kæledyr*. Man kunne også lægge en række andre dyrebetegnelser som *hund*, *marsvin* og *rotte* ind, men man skal passe meget på, for det kan hurtigt komme til at skade mere end det gavner. Hvis man søger efter information om kæledyr på Københavns Kommunes hjemmeside, er man f.eks. næppe interesseret i at høre om rotter, for med kommunens briller er en rotte altid et skadedyr, og hvis man søger efter information om kæledyr, er man mindre interesseret i at få at vide hvordan man bedst slår dem ihjel…

Det er således kontekst- og domæneafhængigt hvad der skal inkluderes som underbegreber. Hvad der er farligt at have med, skifter fra domæne til domæne, så man er nødt til at opbygge forskellige ontologier til forskellige domæner, og i det hele taget være meget forsigtig og altid tænke på hvad informationen skal bruges til i et konkret tilfælde.

I Ankiro har vi derfor en række *kundeordbøger* designet til at passe den enkelte kundes behov. De er baseret på nogle mere generelle *domæne-*

ordbøger, som igen er dannet på baggrund af *almensproglige ordbøger*, én for hvert sprog, og tilsvarende med ontologier. De almene ordbøger er 100 % manuelt gennemarbejdede i løbet af de sidste 15 år (for den danske ordbogs vedkommende, kortere for de andre sprog). Vi kunne også have brugt de 15 år på at udvikle et automatisk statistikbaseret system, men så ville vi have stået med et langt mindre solidt og pålideligt materiale i dag. Det har derfor altid været vores filosofi at alt skal være tjekket af sprogfolk, i hvert fald i de almensproglige ordbøger som bruges på tværs af de enkelte projekter, og hvis indhold man derfor skal kunne stole på.

Eksemplerne ovenfor viser i øvrigt at man ikke bare kan importere ontologier og tesauri fra eksterne ressourcer, for så får man alt for mange relationer. Man kan lade sig inspirere af dem, men ikke importere dem ukritisk. Da den engelske ordbog i sin tid blev etableret, købte man sig til en synonymordbog, og så måtte man ellers bruge enorme ressourcer på at slette semantiske relationer, fordi det simpelthen ikke er holdbart at man får resultater om alt muligt som på en eller anden måde er relateret til det man har søgt på, eller synonymt med det i en eller anden begrænset kontekst. Der skal meget strengere kriterier til, baseret på den konkrete anvendelse materialet er tiltænkt. Det kan jo ikke nytte noget at man får tekster om kartofler når man søger efter information om kæledyr, bare fordi *fritter* både kan være kæledyr og franske kartofler. Så man må være meget tilbageholdende og bruge sin sproglige fornemmelse.

Når stavekontrol og semantiske relationer er på plads, kommer næste trin i forståelsesprocessen, nemlig det vi kalder *konstruktionsgenkendelse*. Det er semantiske relationer på konstruktionsniveau, eller "syntaktisk synonymi". Visse konstruktioner er synonyme eller beslægtede, f.eks. *bekæmpelse af rotter* og *rottebekæmpelse*. Selvom dette er en meget oplagt synonymirelation, kan en computer ikke uden videre se det; man er simpelthen nødt til at lære systemet at visse typer af verbalsubstantiv fulgt af visse præpositionsforbindelser er synonyme med det omvendte kompositum. Den slags følger i vid udstrækning faste regler, men sprog er jo altid fulde af undtagelser, så man er alligevel nødt til at lægge en hel del sproglig information ind for ikke at få for meget støj. Hvordan skal systemet f.eks. vide at det hedder *forbud mod* og ikke *forbud af*, og hvad indebærer det for et ord som *havbeskyttelse* at samme korpus kan indeholde ordet både i betydningen *beskyttelse AF havet* og *beskyttelse MOD havet*? Den slags spørgsmål vil det altid kræve

ægte sprogfolk at besvare. I sidste ende gør man således klogt i at lade de automatiserede metoder være vejledende snarere end bestemmende i sig selv.

Det er tydeligt at et intelligent søgesystem har brug for en hel del sproglig information af alle mulige slags for at kunne håndtere selv de simpleste søgninger. Faktisk er der behov for endnu flere slags information fra den sproglige verden end man måske umiddelbart skulle tro. Ud over de nævnte fænomener er der nemlig nogle ret uventede som man også skal tage stilling til. Hvis man f.eks. går ind på sitet *visitlondon.com* og søger efter information om *Arsenal*, men kommer til at skrive *Arsenol*, svarer den med spørgsmålet *Did you mean arsehol?* Hvis man på Google søger på *perlersvin* og glemmer et mellemrum, så bliver man tilsvarende spurgt om man i virkeligheden mente *perkersvin*. Den slags forslag er ikke altid lige heldige, især ikke hvis det site man befinder sig på, tilhører en offentlig myndighed, for så kan det opfattes som en blåstempling af det foreslåede udtryk. Ankiros søgemaskiner ved godt at der er visse udtryk man skal holde sig fra at smide i hovedet på brugerne. De er ganske enkelt markeret som potentielt anstødelige i vores ordbøger.

Der skal altså også lægges stiloplysninger i ordbøgerne, og helst også information om genre og domæne, hvis man vil vide hvad der er passende i en given sammenhæng. Det forudsætter selvfølgelig at der er nogen der lægger informationen ind, og her kommer den leksikografiske viden endnu engang på banen. Der skal altid lidt ekstra til for at få ord og IT til at fungere sammen.

3. Intention

Man kan spørge sig selv hvad det er der får folk til at udtrykke sig forskelligt. Hvorfor udtrykker de sig som de gør, og hvad er egentlig forskellen på de forskellige udtryk? Ligger der en forskel i intention bag en forskel i udtryk? Det er vigtigt at vi stiller disse teoretiske spørgsmål, for hvis vi kunne forstå hvad der får folk til at udtrykke sig som de gør, så kunne vi måske også vende processen om og forstå hvad de har behov for ud fra den måde de udtrykker sig på. Men det er aldrig helt let.

Hvad er det for eksempel der får visse brugere til at skrive *har glemt min pinkode*, mens andre skriver *ny pinkode* og atter andre *lav ny pinkode*? Det er tre forskellige former for syntaks der anvendes, men ligger der så også en tilsvarende forskel i intention bag dette valg? Når det første er en beskrivelse af problemet, mens de andre to er forsøg på at finde en løsning, betyder det så at brugerne konceptualiserer problemet forskelligt, og måske også har lidt forskellige behov, eller interagerer de bare forskelligt med systemet? Og hvordan er det lige med deiksis – ligger der noget bag perspektivvalget i udtrykkene *betale MIN regning*, *betal DIN regning HER* og *kan man betale SIN regning*?

Det var sådan nogle spørgsmål der fik mig til at kaste mig ud i det forskningsprojekt der mundede ud i en ph.d.-afhandling, som var finansieret af Ankiro og CBS i fællesskab, hvor jeg analyserede nogle af vores kundesites' søgelogs for at se hvordan folk udtrykker sig og i første omgang udvikle et framework og en notationsform som gør det muligt at diskutere variationen på et videnskabeligt plan.

Noget af det der gør det så vanskeligt at fortolke søgestrenge, er at folk er notorisk kortfattede når de formulerer sig i et søgefelt. Typisk består en søgestreng af 1-2 ord, og det giver jo ikke systemet (og analytikeren) ret meget at arbejde med, hverken hvad angår sproglig substans eller kontekst. Alligevel forventer brugerne at få et fornuftigt svar på deres forespørgsel. Det er egentlig ret meget forlangt, for man går jo ikke bare ind i en butik og siger "Hurtig pc" og forventer at få fornuftig betjening på basis af det cue, og man ringer heller ikke til et supportcenter og siger "Langsom pc". Ikke desto mindre er det sådan folk udtrykker sig når de søger, og de forventer at det virker, og det skal være inden for få millisekunder.

De to nævnte udtryk, *hurtig pc* og *langsom pc* er i øvrigt ret interessante, for selvom de er helt identisk opbygget, så dækker de over vidt forskellige intentioner, hvilket da også blev afspejlet i mit valg af eksemplificerende situationer: mens *hurtig pc* er en produktsøgning, er *langsom pc* en supportsøgning hvor man søger en løsning på et problem. Hele forskellen ligger i det leksikalske valg af adjektiv. Det er måske ikke så underligt at et adjektiv med den modsatte betydning kan ændre betydningen af et udtryk drastisk, men hvordan skal computeren vide at der pludselig er tale om en helt anden type søgning og ikke bare et andet emne? Svaret er selvfølgelig, at nogen er nødt til at fortælle den det. Nogen med forstand på ord og sprog.

4. Reference, modalitet og aspekt i søgning

Hvis man søger på *leksikograf Amager* i en af Ankiros jobportaler, er systemet som nævnt i stand til at fortolke *leksikograf* som en stilling og *Amager* som et geografisk sted og kan finde relaterede jobs i nærheden. Hvad systemet ikke kan vide, er om brugeren mente at hun *vil være* leksikograf på Amager, eller at hun *er* leksikograf fra Amager og gerne vil have et relevant job i nærheden. Det kan jo være afgørende, især fordi vi går ind og analyserer på hvilke kompetencer der står i jobopslagene at man skal have og kan anbefale kurser hvis der er noget vigtigt der mangler på cv'et.

Man kan i den forbindelse spørge sig selv hvad søgestrenge overhovedet refererer til. Kan man gøre sig a priori antagelser om hvad folk vælger at referere til? Den traditionelle opfattelse er at søgestrenge refererer til *emner*, eller "topics", men så enkelt er det faktisk ikke. I en vidensbase er det ganske rigtigt oplagt at emnesøgninger er hyppige, men i en job- eller boligportal møder man dem kun sjældent, og i en webshop er de så godt som ikke-eksisterende; her det andre typer af søgninger der dominerer. Brugerne har med andre ord et bredt spektrum af referencemuligheder til rådighed.

Grundlæggende viser det sig at man kan referere enten til det jeg kalder Stadie 1 – den problematiske situation – eller Stadie 2 – den ønskede situation hvor problemet er løst. Et eksempel er jobsøgningssituationen fra før, hvor man kan beskrive den nuværende tilstand (hvad man kan og hvor man bor) eller den ønskede tilstand (det job man er interesseret i og det sted man ønsker at arbejde). Et andet eksempel er supportsituationen, hvor man kan beskrive problemet eller den løsning man forestiller sig kunne findes og som man ønsker information om. Søgestrengen *langsom pc* er en Stadie 1-søgning der beskriver problemet; en tilsvarende Stadie 2-søgning kunne f.eks. være *hurtigere pc*.

Med lingvistiske briller kan man sige at stadievalget er en form for modal distinktion mellem det deklarative som gælder før søgningen og det optative som man ønsker skal gælde i fremtiden. Man kan så referere til en af disse to situationer. Desværre kan et søgesystem aldrig vide om brugeren refererer til stadie 1 eller stadie 2, fordi referencen udføres på nøjagtig samme måde. En løsning kunne selvfølgelig være at spørge brugeren direkte, men folk er som regel ikke ret glade for at skulle svare på spørgsmål når de søger. De vil bare have et resultat. Og det skal være nu.

Det viser sig at der er flere interessante egenskaber ved den måde der refereres på. Foruden den *modale* forskel mellem den eksisterende problematiske situation og den ønskede eller potentielle løsning er der også *aspektuelle* forskelle som afhænger af hvordan brugeren anlægger sit perspektiv på de to situationer. Det er noget der er meget fremtrædende i forbindelse med jobsøgning, hvorfor jeg vil tage mine eksempler fra det domæne.

Mit eksempel kommer fra WorkInDenmark, som er et offentligt site for udlændinge der søger job i Danmark. En analyse af søgeloggen viser med det samme at der er store forskelle på hvordan man griber en jobsøgning an. Kun i godt halvdelen af tilfældene refererer brugerne til det job de søger, og selv blandt dem er der stor variation. Nogle beskriver stillingen, det sted de vil arbejde eller det de vil arbejde med. Engelsksprogede eksempler er *driver* (stillingsbetegnelse), *cleaning* (arbejdsopgave), *farm* (arbejdsrammer) og *pig* (materiale m.m.). Det er alt sammen Stadie 2-søgninger med et "neutralt" aspekt, men af forskellige undertyper. Der findes imidlertid også en anden slags Stadie-2 søgninger som har hvad man kunne kalde et *prospektivt* aspekt; det er folk der beskriver det søgte job som set fra deres egen synsvinkel, som noget der udfylder et eksisterende behov. Eksempler på dette er *student job, seasonal work* og *copenhagen*. Sådan nogle søgninger fortæller os ofte mere om den søgende selv end om jobbet, for hvis man søger job i København, er der stor sandsynlighed for at det er fordi man bor (eller har tænkt sig at bo) i eller nær København og derfor søger noget i nærheden, og hvis man søger studiejob, er det nok tilsvarende fordi man er studerende.

Omvendt er der også nogle der beskriver sig selv som en potentiel arbejdsgiver ville se dem, som værende personer med forskellige kompetencer. Det er Stadie 1-søgninger med et *retrospektivt* aspekt. Eksempler fra WorkInDenmark er *no experience, biology, java* og *spanish speaking*. Man kan selvfølgelig argumentere for at disse søgninger er eksempler på at folk leder efter jobopslag der indeholder disse strenge, men den mulighed er altid til stede uanset hvordan søgningerne er udformet og er derfor helt uinteressant fra et analysesynspunkt. Hvis der er en anden mulighed, bør man som loganalytiker næsten altid vælge den, fordi det tilfører mere information om den søgende. Alternativet er ganske enkelt at konstatere at loganalyse er umuligt, og det bliver vi jo ikke klogere af.

Den sidste hovedtype af jobsøgningsreferencer repræsenteres af dem der beskriver sig selv uden nogen som helst reference til arbejdsgiveren. Det er

folk der ganske enkelt refererer til sig selv som *spanish, young* eller *unemployed*. Man kunne måske tro at denne type var sjælden, men faktisk er det nogle af de hyppigste søgninger på hele sitet. Tallene er dog lidt misvisende, for en betegnelse som *spanish* vil naturligvis oftere være at betragte som en sproglig kompetence end en nationalitetsbeskrivelse. De fem hyppigste søgninger på sitet er *spanish, engineer, french, russian, farm* og *english,* og det er tydeligt at de nævnte nationaliteter netop er nogle der tit vil optræde som sprogkompetencer, men lidt længere nede på listen finder man mange der ikke umiddelbart kan tolkes på den måde, f.eks. *nepal*, *mexican* og selvfølgelig *young*.

Der er tydeligvis stor forskel på hvordan man går til selv en så klart defineret opgave som at søge job. Faktisk opererer jeg i min afhandling med 22 undertyper af jobsøgning, så man skal først og fremmest være forberedt på stor diversitet i brugen af enhver søgemaskine. På den baggrund siger det sig selv at en mindre klart defineret søgning som f.eks. søgeboksen på et site eller et intranet har et meget komplekst og indirekte forhold til den intention der ligger bag en given søgning.

5. Konklusion

Tilbage er kun at konstatere at ord og IT er en vanskelig kombination. Faktum er at der mere end nogensinde er brug for sproglig ekspertise og digitale sproglige ressourcer ude hos os der laver de digitale værktøjer, og det får man kun med sprogfolk som os. Jeg vil derfor benytte lejligheden til at opfordre til nye frugtbare samarbejder. De digitale værktøjer er ikke en trussel mod leksikografien; de repræsenterer nye muligheder for os alle sammen. Udfordringen er så at identificere de nye behov, og de dækker i vid udstrækning stadig over traditionel leksikografisk viden og kunnen.

Litteratur

Alfort, Esben (2013): *The expression of a need – Understanding search*. (PhD thesis). Frederiksberg: Copenhagen Business School, Department of International Business Communication.

Esben Alfort
leder af Afdelingen for Forskning og Innovation, ph.d.
Ankiro
ealfort@ankiro.dk