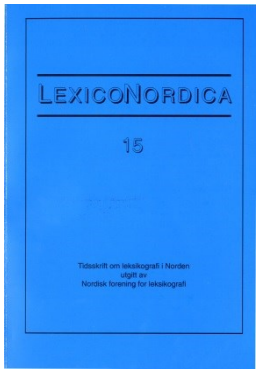
LexicoNordica

Titel: Søgemønstre i logfiler

Forfatter: Richard Almind

Kilde: LexicoNordica 15, 2008, s. 33-55

URL: http://ojs.statsbiblioteket.dk/index.php/lexn/issue/archive



Søgbarhed
Artiklerne i de ældre LexicoNordica (1-16) er skannet og OCR-behandlet. OCR står for ’optical character recognition’ og kan ved tegngenkendelse konvertere et billede til tekst. Dermed kan man søge i teksten. Imidlertid kan der opstå fejl i tegngenkendelsen, og når man søger på fx navne, skal man være forberedt på at søgningen ikke er 100 % pålidelig.

Richard Almind

Søgemønstre i logfiler

Unlike traditional methods for user surveys of printed dictionaries log files have the potential to lead to better user surveys of online reference works. The lexicographer has the possibility to follow the user search by search. Each detail in the search pattern allows for a classification of user types and perhaps even usage situations, the knowledge of which opens up for realistic assessment of a dictionary's design. However, it is necessary to combine the analysis of a log file with user studies to reap the full benefits.

0. Indledning

Et overblik over problemstillingerne ved de gængse former for leksikografiske brugerundersøgelser, deres fordele og især ulemper, gives i Tarp (2008). Generelt konkluderes, at leksikografiske brugerundersøgelser er af ringe kvalitet, oftest grundet mangelfulde metoder eller lige så mangelfuldt empirisk grundlag.

Selv med den mest perfekte undersøgelse kan leksikografen finde det umanerligt svært at studere ordbogsbrug, brugs- og brugersituationer, når ordbogen er i trykt form. Helst vil man observere brugeren uden dennes vidende, men hvordan kan da objektet observeres uden at krænke dets ret til privatliv? Har man imidlertid objektets accept, kan alene dets viden om at blive observeret påvirke brugssituationen. Betænkeligheder og besværligheder af denne art gør sig gældende for alle metoder beskrevet i Tarp (2008).

Anderledes forholder det sig med referenceværker, der stilles til rådighed online på internettet, herefter under et kaldet onlineordbøger. De tekniske betingelser for onlineordbøgernes tilgængelighed gør det muligt at observere ordbogsbrugen gennem undersøgelse af *logfiler*. Ordbogsbrugeren kan her blive gjort til genstand for indirekte observation, både under og efter handlingen, uden dennes vidende og til en vis grad også uden juridiske implikationer, da brugeren langt hen ad vejen forbliver anonym. De indhentede data er metodisk stringente, og den empiriske mængde, der kan studeres, er lig med det samlede antal brugere af en given ordbog.

Det vil i det følgende blive beskrevet, hvad en logfil er, hvilke begrænsninger den har, og hvad den bør optegne for at have leksikografisk værdi.

1. Hvad er en logfil?

Den nærmeste parallel til en logfil i eksisterende brugerundersøgelser er en protokol. I lighed med protokollen optegnes brugerens handlinger linje for linje, men i modsætning til protokollen er der ikke behov for hverken brugerens aktive deltagelse eller hans[1] vidende om optegnelsen. Den sker automatisk ud fra brugerens ordbogshandlinger, som her forstås som aktivering af funktioner i onlineordbogen.

En teknisk præcis beskrivelse af en logfil er, at den er en optegnelse over tid af hændelser og transaktioner enten (a) internt i en enhed, kaldet en systemlog, eller (b) mellem mindst to enheder i et netværk, kaldet en transaktionslog. Det er sidstnævnte, der er genstand for denne artikel.

Type (a), systemloggen, har kun begrænset leksikografisk interesse. Derimod er den såkaldte keylogger, en variant af systemloggen, af åbenlys interesse for leksikografen. Det er et program designet til at optegne alle handlinger foretaget med mus og tastatur. På samme måde som en gammeldags båndoptager optager programmet alle tastetryk, museklik og musebevægelser, fra computeren tændes, til den slukkes. Disse optegnelser kan senere læses og evt. afspilles. Man vil få oplysninger om, hvilke filer der åbnes, hvad brugeren skriver i dem, hvilke søgninger der foretages, om der rettes i søgninger, hvor lange pauser der tages, hvilke programmer der benyttes, hvilke menuer der aktiveres etc. Hvad den ikke kan logge, er fx svaret fra en ordbogssøgning. Den skal genskabes ved at afspille keyloggen. Så smart som keyloggeren kan synes, betragtes den også som en alvorlig sikkerhedsrisiko. At kunne installere den er enhver hackers og industrispions drøm, fordi den optegner alt, dvs. også brugernavne og -koder. Data erhvervet vha. en keylogger uden brugerens vidende falder under kategorien datatyveri.

Type (b), transaktionsloggen, kan betegnes som en optegnelse af den offentligt tilgængelige delmængde af keyloggerens registreringer. For at kunne optegnes skal en hændelse tage form af en transaktion. Fx kan det registreres, hvilke links en bruger aktiverer og hvornår. Det kan også optegnes, hvorfra hændelserne er igangsat, hvilken browser der benyttes, og mange andre systemoplysninger. Men det er altid kun indholdet af transaktionen, der kan registreres, dvs. at man kan optegne den forespørgsel, brugeren har sendt til ordbogen/serveren samt svaret fra serveren til brugeren, men ikke om brugeren har rettet teksten i søgefeltet inden da.

Der er juridiske grænser for, hvad man må registrere, hvor længe

---

[1] Af hensyn til den politiske korrekthed skal det nævnes, at brugerens køn ikke kan registreres i logfilen. At brugeren kaldes et *han*, er en personlig præference.

disse data må gemmes, og hvad de må bruges til, og derfor bliver juridiske overvejelser en nødvendighed, inden man logger brugerdata. Det anbefales, at man sætter sig nøje ind i disse problemstillinger, evt. med bistand fra en specialist på området.

2. Hvad kan logfiler bruges til?

De fleste leksikografer bruger logfiler til overordnede statistiske oplysninger, fx:

- hyppigt søgte ord
- hyppige brugere
- udvikling af søgninger over tid (brugertilgang)
- hyppigst brugte søgeværktøjer
- resultat af søgningen (identifikation af lemmahuller)

Et eksempel på denne type statistik vises i nedenstående eksempel, et lille uddrag af statistikken for *Ordbogen over Faste Vendinger* (2008). Bemærk søgemaskiners andel i det samlede antal søgninger.

Antal opslag pr. 4. jan. 2008	
Søgemaskiner inkluderet	**407.841**
Antal unikke IP-adresser	32.179
Gnms antal opslag pr. IP-adresse	12,67
Søgemaskiner undladt	**157.551**
Antal unikke IP-adresser	32.145
Gnms antal opslag pr. IP-adresse	4,90
Oppetid (dage)	317
Antal opslag pr. dag	
Gennemsnitlig	1.287
Maksimum	5.187
Minimum	192
Opslag pr. opslagstype	
full	187.618
basic	89.630
forside	61.899

Ill. 1: Uddrag af logfilstatistikken for *Ordbogen over Faste Vendinger*

Oplysninger af denne art kan primært bruges til at oplyse omverdenen om, hvor fortræffelig ens ordbog er i forhold til andres. Det er i disse

popularitetsmålinger især brugermængden, antallet af opslag og den procentvise forekomst af fundne lemmata, der angives. Påstår man imidlertid, at tallene oplyser, i hvor høj grad en ordbog opfylder en specifik funktion, er man på gyngende grund. En overordnet statistik af denne art siger intet om hverken specifikke brugere eller en specifik brug af ordbogen. Den kan dog bruges til at være retningsvisende for det videre studie i at besvare spørgsmål om brug og bruger. En given generalisering behøver derfor ikke være forkert – den skal blot valideres gennem en detaljeret undersøgelse, hvilket Bergenholtz/Johnsen (2007) argumenterer for.

Hvis leksikografen vil analysere en ordbogs brug, kan han med fordel analysere individuelle brugeres søgemønstre, inden han, efter en passende gruppering, begynder at generalisere ordbogsbrugen som helhed. For at kunne gøre det skal han finde brugeren, isolere hans transaktioner og finde formålet med dem. Det vil desuden med en smule fantasi være muligt at gætte sig til brugssituationen i søgningsøjeblikket og, med passende forbehold, at inddrage den gisning i konklusionerne. På denne måde kan leksikografen troværdigt argumentere for at have taget udgangspunkt i brugeren, brugs- og brugersituationen.

En undersøgelse af logfiler er ikke en automatiseret proces, hvor en række statistiske standardfunktioner fører til en forgyldt konklusion, men en langsommelig, koncentrationskrævende og systematisk arbejdsgang, der minder om de årelange forsøgsrækker, man finder i de eksakte videnskaber. At lægge en logfil til grund for en undersøgelse kræver derfor, at leksikografen husker følgende:

- det er ikke muligt direkte at udlede brugersituationer ved hjælp af statistiske standardfunktioner
- man skal også se på hver bruger individuelt
- der kræves sved, knofedt og tålmodighed
- lidt fantasi hjælper også

Belønningen er, at leksikografen gennem tålmodig forskning opnår passende kvantificeret og kvalificeret viden om brugeres adfærd og behov ved at analysere søgemønstre, søgemetodik, søgestier etc. Han vil ideelt set kunne bruge disse konklusioner til at validere eksisterende teorier om brugerbehov og fremover med større sikkerhed koncipere nye og bedre ordbøger eller forbedre eksisterende, så de dækker helt nye, hidtil oversete behov.

3. Hvad skal en logfil logge?

Leksikografen skal være bevidst om, hvilket formål de indsamlede data skal bruges til. Skal logfilen bruges til at finde lemmahuller, er der tale om ganske få datatyper, fx søgestrengen og serverens svar på, om den blev fundet eller ikke. Skal den bruges til at finde søgemønstre, øges mængden af datatyper, som det fremgår nedenfor. Man kan designe en logfil til mere end ét formål, men i så fald må leksikografen acceptere, at programmøren må tage højde for, at selv en relativt enkel protokol kan påvirke systemet i en sådan grad, at det mærkbart belaster serveren. Det er også muligt, at antallet af indsamlede elementer for hver transaktion kan blive så omfattende, at registreringen kan føre til en reduktion af svarhastigheden og dermed ordbogens brugbarhed. I værste fald kan uhæmmet logning føre til, at harddisken fyldes op med et permanent servernedbrud til følge. Det sidste sker dog kun sjældent og skyldes oftere, at serveren på forhånd er underdimensioneret, eller at administratoren ikke tænker sig om.

Det er ikke hensigtsmæssigt at analysere serverens systemspecifikke transaktionslog, da den indeholder oplysninger, der er irrelevante for andre end systemadministratoren. På den anden side er det ikke nok at nøjes med et søgeord og en dato. Det, der er brug for, er at oprette en databasetabel, der er skræddersyet til at optegne brugerens leksikografisk interessante transaktioner. Det forudsætter, at logning programmeres ind i selve hjemmesiden.

For at kunne isolere ordbogens bruger og dennes ordbogsbrug bør en protokol som minimum indsamle oplysninger, der besvarer spørgsmålene hvem, hvad, hvordan og hvornår.

1) **Hvem** – brugeridentifikationen: IP-adresser, cookies eller login.
2) **Hvad** – søgestrengen: De tal og bogstaver, der indtastes i søgefeltet, eller et link, der klikkes på i en allerede funden artikel.
3) **Hvordan** – søgemetodikken: Er en handling sket vha. søgefeltet eller et link? Er der brugt funktioner, der ikke er direkte forbundet med søgefeltet, fx links til omtekster eller værktøjer til præcisering af søgningen? Brugen af disse funktioner, brugerens søgemetodik, er en del af søgestien.
4) **Hvornår** – søgestien: Dato- og klokkeslætsregistrering af brugertransaktioner for at rekonstruere søgninger i rækkefølge.

Kombinationen af hvad, hvordan og hvornår danner et mønster for hver bruger, der kan bruges ligesom et fingeraftryk. Det er vigtigt at holde sig for øje, at man kun får dette mønster ved at analysere over tid. Tidsfak-

toren kan synes banal, men den har afgørende indflydelse på mønsterdannelsen. Rytmen, hvordan ordbogen bruges, hvornår og hvor længe ad gangen, tidspunktet for et eventuelt skift i søgemetodikken eller manglen på samme, bidrager til at identificere søge-, brugs- og brugergrupperinger.

Enkeltbrugere kan identificeres vha. IP-adressen og deres søgemetodik og bestemmes til at være fx professionelle tekstforfattere i et reklamebureau. Leksikografen forventer et bestemt brugsmønster fra denne gruppe og kan nu konstatere, at den opfører sig som forventet, dvs. at den bruger de værktøjer og funktioner, som leksikografen stiller til rådighed. Han kan naturligvis også risikere at skulle stille spørgsmålet om, hvorfor gruppen ikke gør som forventet. Svaret afhænger til dels af leksikografens evne til at sætte sig i brugerens sted og situation, men en del af svarene kan findes i måden, værktøjerne stilles til rådighed på for at dække de forventede behov. Det kan være, at en ombytning af to funktioner kan ændre gruppens metodik fuldstændigt. Logfilen åbner med andre ord op for konkrete og kontrollerede eksperimenter med præcist definerede brugergrupper.

3.1. Brugeridentifikationen

Der er primært tre muligheder for at identificere en bruger. De kan kombineres eller benyttes enkeltvis. Hver af dem har fordele og ulemper, og det anbefales at kombinere mindst to af dem for at sikre en nogenlunde pålidelig identifikation af brugeren.

3.1.1. IP-adresser

En IP-adresse (IP = internet protocol) identificerer en enhed, der er tilsluttet internettet. Det behøver ikke være en computer. Derfor peger en IP-adresse sjældent på et enkelt individ, men oftere på routere, servere eller netværk, der gemmer sig bag en firewall, som fx biblioteker, skoler, universiteter og større virksomheder.

Det er ressourcekrævende og forbundet med væsentlige tekniske ulemper for internetudbyderen at give hver kunde en fast IP-adresse. Af samme grund tyr udbydere samt større private og offentlige netværk til DHCP (Dynamic Host Control Protocol), som er en metode til dynamisk at tildele computere inden for et netværk de IP-adresser, der er ledige på et givet tidspunkt. Det betyder altså, at samme IP-adresse kan

benyttes af forskellige enheder på et netværk, dog ikke samtidigt. I denne situation bliver arbejdet med logfilen tidskrævende, men det er stadig, med forbehold, muligt at identificere en enkelt bruger ved at se på søgestien. Den er unik for en given person i en given situation og meget lig et fingeraftryk.

Det diskuteres for tiden i EU, om IP-adresser skal klassificeres som private data, fordi de i sjældne tilfælde kan bruges til at identificere en person med navn. Hvis det bliver vedtaget, gælder skærpede krav ved logning og opbevaring af logdata indeholdende IP-adresser.

3.1.2. Cookies

Med udgangspunkt i ovennævnte problemstilling udvikledes cookien. Det er en fil, der kan gemme et lille antal tegn, og som browseren accepterer og gemmer på foranledning af den server, hvorpå en hjemmeside hostes. En cookie er meget tæt på brugeren, men er stadig ikke selve brugeren.

At bruge en cookie er en registrering af den browser, hvorfra en søgning har fundet sted, ikke af hverken brugeren eller computeren. På steder, hvor mange deles om en computer, fx biblioteker, skoler og universiteter, kan logfilens data derfor føre til fejlslutninger. Ved at kontrollere, om IP-adressen peger på en af førnævnte lokaliteter, kan man udlede, om cookien ligger på en offentligt tilgængelig computer, og tage sine forbehold. Det er også muligt, at en bruger benytter to forskellige browsere på den samme maskine. Det vil se ud, som om det er to forskellige brugere. Problematisk er også, at en cookie kan slettes og blive erstattet af en ny, fx under almindeligt vedligehold eller en opgradering eller geninstallation af styresystemet. Det vil ud fra logfilen se ud, som om en ny bruger er kommet til.

Et særligt problem er, hvis en brugers browser er sat til ikke at tillade cookies. Skal en sådan bruger alligevel have lov til at bruge ordbogen? Hvis ja, kan han så gøres til genstand for undersøgelsen, og hvordan?

3.1.3. Brugerregistrering

Den mest præcise brugeridentifikation er at kræve, at brugeren lader sig registrere med en brugerkonto. Metoden er forholdsvis sikker, dog må man ved en undersøgelse af logfilsdata forudsætte, at brugeren ikke de-

ler sin konto med andre. Det er fx normal praksis i skolenetværk at have én konto pr. computer i et computerrum i stedet for at give hver elev sin egen konto. Denne form for flerbrugerkonti findes også hyppigt i private husstande, som kun sjældent har behov for, at mere end én person i husstanden skal bruge ordbogen samtidigt.

Den største ulempe ved metoden er imidlertid, at det ofte afskrækker en potentiel bruger at lade sig registrere. Det skyldes enten, at brugeren ikke bryder sig om registrering som sådan, eller at han ikke vil spilde tid på registreringsprocessen. Man skal derfor regne med, at brugergrundlaget for undersøgelsen er mindre ved denne metode end ved de ovennævnte. Til gengæld vil grundlaget bestå af dedikerede brugere, der selv har defineret et behov for ordbogen. Omvendt vil det også betyde, at man mister den lejlighedsvise, impulsive bruger af en ordbog.

3.2. Søgestrengen

De fleste computerbrugere kender søgestrengen fra søg-og-erstat-funktionen i et tekstbehandlingsprogram, hvor teksten er repræsenteret som en løbende streng af tal, bogstaver og særtegn. De kender den bare ikke med det navn, men plejer at referere til den som fx søgeordet. Den tekniske definition af en (søge-)streng er, at det er et begrænset antal alfanumeriske tegn, dvs. en række tegn, der kan bestå af tal, bogstaver, mellemrum, kommaer etc. Disse tegn er repræsenteret som talkoder for computeren og er ikke betydningsbærende. En søgning forsøger så at finde strengen i en sammenhængende tegnmængde, den løbende tekst.

I en database søges der, modsat i et tekstbehandlingssystem, ikke i en sammenhængende tegnmængde, men i separate dele kaldet felter, der igen er placeret i såkaldte tabeller. En database består af en eller flere indbyrdes forbundne tabeller. En søgning kan begrænses eller udvides vha. søgeparametre, der angiver, i hvilke tabeller og felter der skal søges, hvordan strengen skal sammenlignes med feltets indhold m.m. Denne søgning danner den løbende tekst efter behov.

Brugeren ser ikke sin søgning på denne meget maskinelle måde. For ham er søgestrengen betydningsbærende og nøglen til løsningen af en konkret problemstilling. Man kan sige, at brugeren med søgestrengen forsøger at formulere både spørgsmålet og en vag ide om svaret på samme tid. Mediet, som løser denne knude, er det moderne orakel: computeren. Søgestrengen repræsenterer derfor noget væsentligt mere kompliceret og andet end kun lemmaet. Søgestrengen er det, som brugeren associerer med sit aktuelle problem. Computeren har imidlertid ikke evnen

til at udlede så meget ud fra et antal nuller og ettaller, og leksikografen må derfor forsøge at stille en funktion til rådighed, der tilnærmer sig den ønskede proces. Det kan han ved at splitte ordbogens indhold i sine bestanddele og strukturere dem i passende grupper, som fx lemma og betydning, der tilsammen danner førnævnte løbende tekst ud fra separate dele i en database.

I en onlineordbog indtastes søgestrengen i et søgefelt på hjemmesiden. Feltets indhold sendes sammen med søgeparametrene som en forespørgsel til serveren ved aktivering af en knap eller anden funktion. Parametrene angiver ud over de ovennævnte muligheder også fx, hvordan resultatet skal formateres, hvor mange resultater man ønsker vist pr. side etc. Disse parametre giver svar på brugerens søgemetodik.

Søgninger i databaser kan være meget komplicerede, og det er derfor normalt og hensynsfuldt at prædefinere søgeparametre for brugeren. Det svære er at finde de parametre, der er nemme at forstå, men som samtidig kan dække et muligvis kompliceret behov. Det hænder, at prædefinerede parametre ikke bruges efter leksikografens forestillinger.

3.3. Søgemetodikken: brug af ordbogens funktioner

Logfilen kan også indeholde oplysninger om, hvilke andre links end lige netop de søgningsrelaterede der er blevet brugt. Alle de links, knapper m.m., der er på en ordbogs hjemmeside, kan i princippet logges, forudsat at man tager højde for det i programmeringen.

Funktionen af links i en artikel kan fx være at henvise til antonymer, synonymer eller tekster på andre hjemmesider end ordbogens. At registrere aktiveringen af disse har potentiel værdi for leksikografens samlede vurdering af ordbogens brug. Knapper, der angiver, om der alene skal søges i en given delmængde af databasen, fx kollokationer eller eksempler, kan give et fingerpeg om, hvad en bruger har behov for, men også om brugssituationen.

Brugeren vil på et eller andet tidspunkt aktivere alle de funktioner, en ordbog er udstyret med. Hver af dem enkeltvis eller kombineret dækker et behov. Enkelte af disse funktioner vil brugeren aktivere i bestemte kombinationer alt efter præferencer eller brugssituation.

En given loglinje kan afsløre, at en bruger har klikket på linket til fx brugervejledningen. Det er i sig selv interessant, at en bruger konsulterer brugervejledningen, men det er dog mindst lige så interessant, hvornår han gør det. Sker det efter en søgning, viser det, at interessen ikke kun er rent akademisk, men at den konsulteres, enten fordi der er opstået et

fortolkningsproblem i forbindelse med artiklens opbygning, eller fordi søgningen ikke gav det forventede resultat. Et andet valg af funktioner kan afsløre, om brugeren er interesseret i at maksimere resultatet, dvs. at få vist alle elementer i en artikel, eller om han nøjes med den minimerede visning. Det kan også ske, at han får den ene visning, men derefter vælger en anden. Måske er der sammenhæng i, hvornår det sker. Gør han det, hver gang artiklen har et vist omfang, eller gør han det, når et bestemt element mangler? Organiseret over tid udgør protokollen over aktiverede funktioner brugerens søgemetodik.

3.4. Søgestien: tidspunkt for transaktionen

Hver linje i en logfil er en brødkrumme i brugerens søgesti gennem databasen. Krummerne gør det muligt at følge brugerens forsøg på at få dækket sine muligvis subjektive, ekstra-leksikografiske behov opslag for opslag. Det sker, fra han først kommer ind i ordbogen, til han forlader den igen. Hver handling sker på et bestemt tidspunkt, og afstanden mellem tiderne giver et overblik over, hvad brugeren foretager sig. Er der lange pauser i aktiveringen af funktioner, eller sker det med korte mellemrum? Hvor længe er brugeren aktiv på siden? Er han det fra morgen til aften eller midt på dagen?

Kombineret med brugeridentifikation og søgemetodik er søgestien en afgørende faktor i belysningen af brugerens situation og behov. Søgestien gør det bl.a. muligt at se, om en bruger konsulterer en omtekst, som fx brugervejledningen, før eller efter søgningen har givet resultat.

Imidlertid kan søgestien udviskes og føre til fejlslutninger. I de tilfælde, hvor en computer bruges af flere i løbet af dagen, fx når den står bag en firewall på et bibliotek eller i en virksomhed, da bliver det vanskeligere at finde et entydigt søgemønster. Det er derfor nødvendigt at vide, hvorfra søgningen foregår, dvs. at kende IP-adressens ejer på tidspunktet for søgningen.

4. Hvordan læses en logfil?

De fleste ville her ønske at få en matematisk formel, hvis resultat er en statistisk oversigt over brugere, brugergrupper, brugssituationer og ordbogsbrug. Det er derfor på sin plads at minde om de i afsnit 2 nævnte huskeregler og straks forkaste utopiske tanker om universelle formler som en særlig grov form for dovenskab.

En logfil kan på en enkelt dag optegne tusinder, titusinder eller måske endda millioner af søgninger. At underkaste den en automatisk undersøgelse eller en undersøgelse baseret på tilfældigheder vil ikke føre til et brugbart resultat. Det er nødvendigt at foretage udvælgelse.

I det følgende vil man kunne få det indtryk, at nogle brugere, brugssituationer eller typer ordbogsbrug har større værdi end andre. Intet kunne være mere forkert. Alle opslag i en ordbog dækker over et behov. Behovet er bare ikke altid i overensstemmelse med ordbogens overordnede funktioner.

Undersøgelsen må derfor have et standpunkt, et fast udgangspunkt, et spørgsmål, der skal besvares. Hvis man som leksikograf vil gennemføre en brugerundersøgelse, vil man helt naturligt stille sig selv spørgsmålet: Hvad ønsker jeg at belyse med denne undersøgelse? Ønsker han at undersøge, om ordbogen opfylder en bestemt funktion efter forventningerne, skal han bruge én metode. Vil han undersøge, hvor mange forskellige brugertyper der bruger en given ordbog, skal han bruge en anden metode. En tredje problemstilling kan medføre en tredje metode og så fremdeles. Fælles for metoderne er imidlertid, at brugere og deres søgemønstre skal isoleres og kvantificeres.

Formålet med den her beskrevne metode er ikke at skabe en komplet katalogisering af samtlige brugertyper, men at belyse om brugeradfærden svarer til forventningerne, dvs. om brugerne benytter sig af ordbogens funktioner, sådan som det var forventet, at de skulle bruge dem. Som sidegevinst finder man også brugere, der benytter ordbogens funktioner på uforudsete måder. Denne adfærd kan pege på ukendte problemstillinger, der kan undersøges nærmere på anden vis. Uanset formålet vil det altid være nødvendigt at isolere søgemønstrene.

Ikke alle søgemønstre er lige interessante. Enkelte mønstre, som fx dem, der er genereret af søgemaskiner, er direkte uheldige, da de ikke beskriver en menneskelig bruger, og andre mønstre beskriver en ikkegenuin brug af ordbogens overordnede funktion. Disse sidste er typisk lejlighedsvise søgninger foretaget af kedsomhed. Dermed ikke sagt, at denne type brug ikke bør undersøges, men blot at man skal være opmærksom på, hvilke brugergrupper man ønsker at målrette undersøgelsen mod. For at kunne udarbejde en komplet brugerundersøgelse på basis af en logfil er det nødvendigt at

1. finde interessante søgninger
2. analysere transaktionerne
3. opsøge kilden
4. validere transaktionen

Punkterne (3) og (4) ligger uden for denne artikels ramme. I det følgende vil derfor kun punkterne (1) og (2) blive belyst.

Resten af denne artikel giver et praktisk eksempel på anvendelsen af en overordnet metode til undersøgelse af logfiler og tager udgangspunkt i logfilen fra *Ordbogen over Faste Vendinger* (2008). Logfilen blev dannet over en periode på ca. tolv måneder. Ordbogens hjemmeside er programmeret i HTML og anvender PHP til opslag i en MySQL-database. Andre onlineværker benytter andre programmeringsmetoder, men de samme principper kan gøres gældende. Selve logfilen er en tabel i databasen, og data deri kan bearbejdes direkte vha. SQL-kommandoer og/eller eksporteres til et regneark.

log_key	searchstring	searchoption	searchtype	searchresult	user_ip	searchdate	searchtime	is_klik	sidevalg
26336				0	[illegible]	2007-04-01	17:35:11	0	
26339	hjælpsom	contains	ext	3	[illegible]	2007-04-01	17:36:05	0	
26341	hjælpsom	contains	full	6	[illegible]	2007-04-01	17:36:35	0	
26342	gammel	contains	ext	60	[illegible]	2007-04-01	17:37:16	0	
26345	blander sig	contains	ext	17	[illegible]	2007-04-01	17:38:30	0	
26347				0	[illegible]	2007-04-01	17:38:44	0	brugervej
26351				0	[illegible]	2007-04-01	17:40:16	0	
26353	parkere aben	contains	ext	1	[illegible]	2007-04-01	17:41:04	0	
26354	nisse	contains	ext	4	[illegible]	2007-04-01	17:41:53	0	
26360	stresset	contains	ext	2	[illegible]	2007-04-01	17:45:29	0	

Ill. 2: Udtræk af logfil fra *Ordbogen over Faste Vendinger*

Felterne angiver i rækkefølge: søgenummer (log_key), søgestreng (searchstring), søgningsvalg (searchoption), søgetype (searchtype), antallet af fundne forekomster (searchresult), brugeridentifikation i form af en IP-adresse (user_ip), søgedato (searchdate), søgeklokkeslæt (searchtime), om søgestrengen er indtastet (is_klik = 0), eller om det er en søgning via et link (is_klik = 1), og endelig hvilken omtekst brugeren evt. har valgt at se på (sidevalg). Loglinjen for søgenummer 26339 viser fx, at brugeren sendte strengen *hjælpsom* til ordbogen fra søgefeltet med ønsket om at få en udvidet visning (ext), og at han har fået tre artikler som resultat. Den næste linje viser, at en ny søgning blev skrevet i søgefeltet (is_klik = 0), og at resultatet skulle vises med så mange felter som muligt (full). At antallet af fundne forekomster er større ved full (searchresult = 6) end ved ext (searchresult = 3) skyldes, at valget af søgetype (searchtype) ikke kun påvirker visningen, men også hvilke felter/tabeller der søges i.

Alle disse data kan virke uoverskuelige for den uindviede, men hvert felt har et defineret formål, som kan bruges alene eller i kombination med andre. Overordnet består metoden til isolering af søgemønstre af følgende skridt:

1. Statistik over søgefrekvenser
2. Udvælgelse af en interessant delmængde
3. Identificering af brugeren
4. Isolering af brugerens søgninger
5. Tidsbaseret analyse

I de efterfølgende analyser anvendes der SQL-kommandoer, der for de fleste er uforståelige og også irrelevante, da de kun kan finde anvendelse på den konkrete tabel. Det vil derfor kun blive forklaret, hvad eksemplet viser, men ikke hvordan resultatet er opnået.

4.1. Statistik over søgefrekvenser

Et typisk spørgsmål for tilhængere af funktionsteorien er, om der er brugere, der benytter ordbogen efter dens genuine formål? Eftersom datamængden i logfilen kan være uoverskuelig, skal den reduceres. Det er derfor nærliggende at vende spørgsmålet om og spørge, hvilke brugere der ikke bruger ordbogen efter dens formål?

Svaret er desværre ikke helt så nærliggende. De brugere, der kan udelades, er typisk dem, som ingen brugbare data genererer. Metoden forudsætter, at der er en søgestreng, en søgemetodik og en søgesti, og at disse kan tildeles en identificerbar bruger. Hvis en af disse fire informationer mangler, falder metoden til jorden.

Leder man efter søgemønstre, er søgestien nemmest at isolere, idet det kræves, at der som minimum er en begyndelse og en afslutning på en søgesti, dvs. to tidsmæssigt adskilte linjer i logfilen. Derfor vil alle linjer med kun én IP-adresse kunne udelukkes, da de ifølge denne definition ikke er en egentlig søgesti. Ekstreme søgestier med flere tusinde opslag over relativt kort tid kan også undlades, da det er usandsynligt, at en enkeltperson har foretaget dem. Det er mere sandsynligt, at det er en automatiseret proces, enten en søgemaskinerobot som fx Google, der leder efter oplysninger til sit indeks, eller en såkaldt spamcrawler, der leder efter mailadresser til spammails.

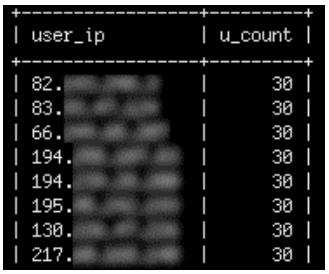
Der gælder derfor den statistiske regel om at fjerne yderpunkterne i mængden for at finde sammenlignelige delmængder. De præcise grænser må leksikografen selv fastsætte ud fra stikprøver og sund fornuft. Er den samlede mængde loglinjer stor nok, hvilket den ofte vil være ved onlineordbøger, er det ingen katastrofe, hvis en enkelt bruger overses.

For at finde frem til interessante søgninger og de tilhørende IP-adresser skal man først isolere de enkelte brugere og det antal søgninger,

de hver især har foretaget. Dernæst træffes beslutning om, hvor man ønsker at sætte grænsen for søgefrekvensen. I det her beskrevne eksempel blev en gruppe med 30 søgninger vilkårligt udtaget, men i en komplet undersøgelse må et langt større felt underkastes vurdering.

4.2. Udvælgelse af en interessant delmængde

Der findes i logfilen en gruppe på 30 IP-adresser, der hver for sig har søgt 30 gange i ordbogen. En søgning, der viser disse IP-adresser, kan se ud som følger:



user_ip	u_count
82.	30
83.	30
66.	30
194.	30
194.	30
195.	30
130.	30
217.	30

Ill. 3: Unikke IP-adresser med 30 søgninger hver

Et detaljeret gennemsyn af hver af disse 30 brugere viser, at ingen af dem har det samme søgemønster. Enkelte af dem har søgemønstre, der lader én tvivle på, om der er tale om en menneskeskabt søgning. Andre igen viser, at brugeren muligvis havde misforstået brugen af ordbogens funktioner eller måske endog selve ordbogens formål. De er alle vigtige, men ikke alle lige vigtige.

4.3. Identificering af brugere

Den efterfølgende udvælgelse kan ske vilkårligt, hvis formålet med undersøgelsen er at lave stikprøver, som det er sket i dette tilfælde. I praksis skal hver IP-adresse i en gruppe identificeres for at fastslå, om enheden bag IP-adressen er en bruger, eller om den er en computer, der bruges af mange. Til det formål findes et antal værktøjer og programmeringskneb, men hjemmesiderne www.who.is og www.ripe.net er værd at konsultere, hvis man ikke har adgang til de forkromede løsninger.

I eksemplet blev der ikke foretaget nærmere undersøgelse af IP-adresserne. I stedet blev hver brugers søgninger listet for at finde et interessant mønster eller i hvert fald et mønster, der med en vis sandsynlighed ikke stammer fra en søgemaskine. Da der ikke var flere end 30

adresser, var det muligt inden for nogle få minutter at finde frem til en enkeltbruger, der var tilstrækkelig interessant til en stikprøvekontrol. I det konkrete tilfælde dækker IP-adressen over en antenneforening. Det er derfor ikke sikkert, at der er tale om en enkeltperson, men som det vil blive tydeligt, er det en stærk sandsynlighed.

4.4. Isolering af brugerens søgninger

Det er nu muligt at isolere en enkelt af disse IP-adresser og vise, hvad vedkommende har søgt på i detaljer:

searchstring	searchoption	searchtype	searchresult	user_ip	searchdate	searchtime	is_klik	sid
			0	195.	2007-03-18	04:36:07	0	
katten sækken	words	basic	2	195.	2007-03-18	04:36:56	0	
			0	195.	2007-03-19	03:16:31	0	
bageriet	words	basic	2	195.	2007-03-19	03:17:11	0	
			0	195.	2007-03-19	16:41:37	0	
			0	195.	2007-03-19	16:59:51	0	
			0	195.	2007-04-04	11:13:24	0	
alen	contains	basic	48	195.	2007-04-04	11:13:38	0	
to alen af et stykke	isequal	basic	2	195.	2007-04-04	11:14:03	1	
			0	195.	2007-04-05	15:06:54	0	
			0	195.	2007-07-17	11:35:05	0	
apostlenes heste	contains	basic	1	195.	2007-07-17	11:36:08	0	
			0	195.	2007-07-19	15:42:32	0	
			0	195.	2007-07-19	15:43:09	0	
			0	195.	2007-08-17	14:56:58	0	
skæppen	contains	basic	1	195.	2007-08-17	14:57:13	0	
opsang	isequal	full	8	195.	2007-08-17	15:12:04	1	
skæppen	contains	basic	1	195.	2007-08-17	15:13:11	0	
skæppen fuld	contains	ext	1	195.	2007-08-17	15:14:13	0	
			0	195.	2007-09-05	09:51:32	0	
skæppe	contains	basic	6	195.	2007-09-05	09:51:46	0	
			0	195.	2007-11-04	04:07:48	0	
loppe	contains	basic	7	195.	2007-11-04	04:08:01	0	
			0	195.	2007-11-05	09:05:46	0	
søbe	contains	basic	1	195.	2007-11-05	09:05:56	0	
medicin	contains	basic	6	195.	2007-11-05	09:06:17	0	
			0	195.	2007-11-06	13:26:39	0	
medici	contains	basic	6	195.	2007-11-06	13:26:48	0	
			0	195.	2007-12-08	18:05:15	0	
ballelars	contains	basic	2	195.	2007-12-08	18:05:28	0	

Ill. 4: Detaljeret visning af en brugers søgesti

Datoangivelserne skrives som år (4 cifre), bindestreg, måned (2 cifre med foranstillet nul), bindestreg, dag (2 cifre med foranstillet nul). Sorteringen er dato før tid med nyeste dato sidst og nyeste tid sidst inden for hver datogruppe. Alle tidsangivelser er angivet i serverens lokaltid. Eksemplet her har 11 søgestier fordelt over 30 opslag/loglinjer. Den første sti er fra den 18. marts 2007 og den sidste fra den 8. december 2007. Linjer med en blank søgestreng (searchstring) er de gange, hvor brugeren har åbnet hovedsiden inden en søgning.

4.5. Tidsbaseret analyse

Ved at rekonstruere en brugers søgninger vises bl.a., hvilke ordbogsfunktioner han finder behov for. I det ovennævnte eksempel søger brugeren næsten udelukkende vha. søgetypen basic. Funktionen har til formål at finde faste vendinger i brugssituationen forstå en tekst. Den foretager en meget snæver søgning i faste vendinger samt en minimalvisning bl.a. indeholdende den fundne faste vendings betydning og dens varianter. Kun en enkelt gang forsøger brugeren sig med andre muligheder. Den 17. august klikker han på linket "opsang" i artiklen "skæppen" med muligheden full, som er en udvidet søgning i tre felter med en komplet resultatvisning. Efterfølgende søger han gennem søgefeltet på "skæppen fuld" med søgetypen ext, som er en søgning i to felter med en udvidet resultatvisning. Det er derfor rimeligt at antage, at det kun er af nysgerrighed, eftersom det er eneste og sidste gang, disse søgetyper bliver taget i brug. Bemærk, at brugeren kan være uvidende om, i hvilke felter der søges, eftersom det står i omteksterne, og disse har ikke været konsulteret. Det burde dog ikke forhindre ham i at se, at resultatvisningen er en udvidet samling data, eller at han undertiden får et større resultat. Alligevel vælger han ikke fremover funktioner, der kan udvide svarmængden. Leksikografen burde ud fra søgemetoden konkludere, at brugeren har et behov for at forstå en tekst, men er det tilfældet? Måske er brugeren i virkeligheden blot tilfreds med de mest rudimentære funktioner, ordbogen har at byde på, fordi funktionen basic er valgt som standard. Denne påstand understøttes af, at brugeren altid søger med indeholder (contains), dvs. at søgestrengen som helhed skal indgå i resultatet, hvilket er standardmetoden siden den 1. april 2007.

Noget andet, denne bruger viser, er, at han kun søger én gang for hver gang, han konsulterer ordbogen. Det blanke felt, hvor søgestrengen (seachstring) burde stå, viser, at han har aktiveret ordbogen, den efterfølgende linje viser, hvad han har søgt i den. Med den 17. august som undtagelse ses, at hver ordbogskonsultation kun består af én søgning. Oven i købet er søgningerne før den 4. november kun søgninger på eksempler, der på den tid nævntes på forsiden af ordbogen. Følgende spørgsmål kunne derfor trænge sig på: Hvis denne adfærd er typisk for brugerne, vil det så betyde, at to tredjedele af alle søgninger i virkeligheden blot foretages af nysgerrighed? En overordnet statistik ville kunne føre til denne konklusion.

Før man præmaturt konkluderer, at kun ca. 30% af alle opslag er problembaserede søgninger og de resterende 70% ren nysgerrighed, må man undersøge flere brugere. Imidlertid viser tankerækken, at det er

dybt problematisk at basere konklusioner på overordnede statistikker alene.

En enkelt detalje, man ikke må glemme, er at tage klokkeslættet med i sine betragtninger. Den aktuelle bruger har ikke de mest spændende karakteristika, men andre brugere er meget mere aktive, og for disse gælder, at man med fordel kan se på, hvornår de aktiverer søgninger. Af og til aktiverer en bruger ordbogen tidligt om morgenen og lader flere timer gå, før en ordbogskonsultation finder sted. En sådan adfærd antyder professionel brug, som det vil fremgå i det næste afsnit.

5. Typologi af en bruger – et eksempel i praksis

Ovennævnte bruger er svær at typologisere. Der er for få opslag og for få reelle oplysninger om vedkommende. Søgestrengene giver ingen fingerpeg om, hvem brugeren er, og hvorfor han slår op i ordbogen. Det eneste, vi ved, er, at IP-adressen peger på en boligforening, at han har tid til at bruge ordbogen, når det passer ham, og at han kun lejlighedsvis finder det nødvendigt at konsultere ordbogen i dybden. Søgningsmønstrene som sådan er ikke særligt opmuntrende for leksikografen, men der er dog det lyspunkt, at standardsøgningen viser sig at være nyttig. Anderledes spændende er denne bruger:

searchstring	searchoption	searchtype	searchresult
midt	contains	basic	4
centrum	contains	basic	1
central	contains	basic	1
midt	contains	basic	4
by	contains	basic	88
storby	contains	basic	1
beliggende	contains	basic	1
			0
bløde	contains	basic	7
det bløde	contains	basic	1
			0
			0
			0
			0
brød	contains	basic	29
Man skal ikke slå større brød op, end man kan bage.	isequal	basic	1
slå brød op	isequal	basic	1
brød	contains	ext	34
mad	contains	basic	43
være den rene barnemad	isequal	basic	1
være barnemad	isequal	basic	1
kok	contains	basic	7
køkken	contains	basic	8
familie	contains	basic	10
middag	contains	basic	11
			0
grund	contains	basic	39
fast grund under fødderne	isequal	basic	1
grund	contains	basic	39

Ill. 5: Professionel tekstforfatter (et reklamebureau)

Illustrationen er kun et udsnit af de samlede data og viser ikke alle søgemønstre, men dem, der er, er ikke atypiske for brugeren. Der vises de første 56 linjer ud af ca. 200 søgninger fordelt over 21 dage i perioden 23. februar til den 4. december 2007. Illustrationen er SQL-resultatet eksporteret til et regneark og formateret med tomme linjer efter hvert søgemønster.

IP-adressen (her fjernet fra statistikken) peger på et dansk reklamebureau. Ud fra firmaets offentligt tilgængelige oplysninger og dets hjemmeside forudsættes, at der er mange ansatte. Én af disse ansatte bruger *Ordbogen over Faste Vendinger*, og det er usandsynligt, at der er flere end denne bruger. Han er sandsynligvis professionel tekstforfatter.

Den efterfølgende analyse tager udgangspunkt i en overordnet betragtning af brugerens adfærd efterfulgt af en gennemgang af hvert søgemønster. Det kan ikke blive en komplet analyse i denne artikel. Dertil er data for omfattende, men det kan give en antydning af de guldkorn,

man kan finde i en logfil, hvis man anvender flid og fantasi i en god kombination.

Det første indtryk, man får, er, at brugeren er vant til at benytte søgemaskiner. Fra den allerførste søgning er det tydeligt, at han ikke er i tvivl om, hvad han forventer at finde i denne ordbog. Der er ingen fumlen og forsøgen. Her er et redskab, og det bliver brugt. Det understøttes bl.a. af, at han aldrig aktiverer en omtekst, som fx brugervejledningen.

Brugeren arbejder målrettet og koncentreret. Søgemønstrene er klart adskilte i hver deres tidsramme. Ingen søgestier varer længere end 30 minutter, de fleste mellem 10 og 15 minutter. Enkelte gange konsulteres ordbogen to gange i løbet af dagen adskilt af en længere pause midt på dagen. Den tidligste søgning påbegyndes klokken 9:49, den seneste afsluttes klokken 15:52 med en enkelt undtagelse klokken 16:30. Der er anomalier i søgemønstrene, der kan tyde på, at en anden bruger tilgår ordbogen fra denne IP-adresse, men det er ikke entydigt. Den sene søgning 16:30 er dobbelt uden for normen, da søgestrengen er uden for sammenhæng eller rettere, i en sammenhæng, der er hændt fire måneder tidligere. Desuden sker det to gange, at siden aktiveres uden søgning med få sekunders mellemrum. Det kan skyldes en teknisk fejl.

Hvert søgemønster har en tematisk/semantisk sammenhæng, som de sidste 31 linjer i illustration 5 viser. Tilsvarende temablokke kan ses i samtlige mønstre. De nævnte 31 linjer har tematisk sammenfald med ord som fx *grund, hus, bygge* og *landskab*. Det er nærliggende at forestille sig en tekst om fast ejendom. Det foregående mønster fra den 17. april 2007 er variationer over temaet *brød, mad* og *middag*. En køkkenreklame? Et bagerfirma? Fælles for det tematiske aspekt er, at det har et tydeligt fingeraftryk set i forhold til de meget snævre tidsrammer, hvori søgningerne foregår. Hustemaet ender ca. midt i maj og efterfølges af *natur* indtil den 5. september, hvor det bliver til *hus* igen efterfulgt den 7. september med temaet *revy*, etc. Det kunne være interessant at se, om man kan finde reklameteksterne i medierne i de efterfølgende perioder for at få det endelige resultat, eller blot at adspørge forfatteren. Uanset hvad viser tematiseringen, at der er tale om en enkelt bruger med specifikke tekster og tekstproblemer.

Søgetypen er stort set altid sat til standardvalget basic, som er et værktøj til at forstå en tekst. En tekstforfatter burde forventes at vælge funktionen ext, som er beregnet til dem, der har problemer med at skrive en tekst. Men mod forventning bruges ext kun syv gange og funktionen full aldrig. Det kan skyldes flere faktorer, og de kan kun bestemmes ved at adspørge vedkommende direkte. En gisning kunne være, at de udvi-

dede søgninger giver for mange resultater, og at søgningerne tager længere tid, ikke alene pga. den større resultatmængde, men også fordi der søges i flere felter. Det kan en professionel i reklamebranchen ikke vente på, da der er en deadline at overholde, og denne deadline er pr. definition dagen forinden. De gange, hvor resultatmængden bliver stor, går der kun kort tid, inden en ny søgning er sat i værk. Det ses tydeligt i den første blok i illustrationen: *by* efterfølges hurtigt af *storby*. De sidste tre linjer tilsvarende: *enge* følges af *fyn* følges af *hjem* på 59 sekunder.

Som det ses, er der meget, man kan udlede af denne bruger på basis af relativt få oplysninger. Endnu mere kan udledes, hvis man begynder at kombinere og krydsreferere med andre brugere. Der er imidlertid også meget, der kan overfortolkes, og man løber den risiko at gisne ud fra gætterier og ufunderede antagelser. Alligevel kan denne overfladiske og ikke nødvendigvis tidskrævende gennemgang af objektets søgestier give os en anvendelig brugerprofil:

- han er professionel tekstforfatter
- søgningerne er hurtige, nærmest hektiske, og tematisk sammenhængende
- han benytter sig af og til af de indbyggede links, men oftest lader han være
- han bruger ikke funktionerne efter leksikografens forventning

Det store spørgsmål er, hvorfor en tekstforfatter ikke bruger mulighederne for tekstproduktion. Svaret kan ligge i resultaterne. De er for lange og for længe om at dukke op. Tekstforfatteren bruger ordbogen for at finde inspiration, måske en personlig form for brainstorming. Det er naturligvis en gisning. Han er ikke blevet adspurgt, men det burde han blive. Så kunne logfilen blive til en brugerundersøgelse.

6. Konklusion

Nogle brugere udnytter alle ordbogsfunktioner, muligvis endda efter hensigten. Det er typisk de brugere, der har tid til det. Andre brugere, som fx professionelle tekstforfattere, burde gøre det, men gør det ikke. Vi må antage, at det skyldes et tidspres, at de bruger standardvalget og kun sjældent har tid til at uddybe. Det kan også være, at de bare ikke vil have for mange data.

Vigtigst: Hver enkelt bruger har sin helt egen strategi for at finde det, han leder efter. Denne banalitet viser kompleksiteten i at lave en

brugerundersøgelse og nødvendigheden af at undersøge enkeltbrugeres søgeadfærd i elektroniske opslagsværker.

Det vigtigste, som søgemønstre viser, er, hvorfor en tidsbaseret analyse bedre beskriver brugen af en ordbog end en automatiseret, overordnet statistik. Statistikken giver ganske vist et fingerpeg om, hvad den gennemsnitlige brug af ordbogen kunne være, men ikke hvad den reelt er. Det svarer stort set til at tage en række billeder af en sommerferie, lægge dem oven på hinanden og se, at det meste af tiden var himlen blå. Deraf kan man ikke konkludere, at alle feriegæster var glade.

Ved at følge den enkelte brugers søgninger og resultater vil man kunne sætte sig i brugerens sted og begrænse antallet og arten af mulige brugersituationer. Jo flere brugere, der isoleres, jo bedre bliver billedet. Ud fra typologiseringen vil det være muligt at identificere brugersituationer og -typer og holde dem op imod funktionsteorien. Man vil så enten finde brugere, der overordnet passer ind i teorien, eller brugere, der åbenlyst falder udenfor. Det vil derfor uanset resultatet være muligt at drage nytte af en detaljeret betragtning af enkeltbrugeres søgemønstre.

En logfil gør ingen brugerundersøgelse. Logfilen og dens slægtning, keyloggeren, har deres begrænsninger. Ønsker man indblik i brugerens tankegang, brugs- og brugersituationer, kan hverken transaktionsloggen eller keyloggeren give endegyldige svar. Til det formål må den enkelte bruger inddrages direkte.

Man må derfor supplere med direkte kontakt. Evt. kunne man forsøge sig med et spørgeskema, som kan udfyldes af en bruger via hjemmesiden, eller man kan bede brugere om at blive inddraget, igen via hjemmesiden, og evt. gennemføre et interview via e-mail, et chatprogram, eller på gammeldags vis pr. telefon. Alle disse metoder kan gennemføres med bibeholdelse af brugerens anonymitet.

Alternativt kan man foretage et feltstudium, hvor man interviewer og/eller observerer brugeren på dennes arbejdsplads. Man kan da på stedet validere de foreløbige konklusioner af logfilsundersøgelsen. Metoden har en anden fordel: Det er svært at vurdere, hvor stor en indflydelse observation har på brugerens opførsel i en given situation. Ved at sammenligne logfilens data fra før, under og efter observationen burde eventuelle påvirkninger af observationen blive synlige i søgemønsteret.

Uanset hvilken metode man supplerer logfilsundersøgelser med, er de indledende procedurer til kvantificering af logfilsdata de samme. Man skal altid være vagtsom ved automatisk genererede statistiske sammenfatninger baseret direkte på logfilen. De kan ikke stå alene, men kan

primært bruges som retningslinjer for, i hvilken retning man vil søge sine oplysninger.

Desværre kommer man som seriøs leksikograf ikke uden om at bruge megen tid og flid, når man undersøger logfiler. En enkelt relativt lille logfil på 200.000 linjer vil muligvis tage to år at studere. Det kan også være, det kan gøres hurtigere. Det kan ikke vides, da det aldrig er blevet gjort. Den gode nyhed er, at man ikke behøver at gennemgå alle linjer, især ikke, hvis det drejer sig om logfiler på en million linjer eller mere. Det, som er vigtigt, er, at man målrettet forsøger at finde brugertyper. Når man har en repræsentativ mængde af dem, kan undersøgelsen siges at have opnået sit formål.

En mulighed, som kun har været nævnt i forbigående, er at benytte logfiler til udarbejdelsen af eksperimenter. Det er teknisk muligt at stille forskellige hjemmesider til rådighed for forskellige IP-adresser. Man kan forestille sig tre tekstforfattere med lignende søgemønstre i hver deres virksomhed. Den ene kan bruge hjemmesiden uændret, de to andre justeres med en eller to variabler. Det kræver ikke megen fantasi at forestille sig, hvilke døre der står åbne.

Logfilsundersøgelser, med eller uden brugerinddragelse, har en stor værdi for leksikografen. Det må anses som sandsynligt, at tilstrækkeligt mange og detaljerede undersøgelser kan danne grundlag for udarbejdelsen af nye ordbøger på basis af erfaringer fra gamle. Interessant ville være at samle oplysningerne i en vidensbank indeholdende en lang række logfilsundersøgelser, der havde en fælles grundform. En leksikograf ville kunne benytte sig af den viden til at udarbejde en ny ordbog eller forbedre en eksisterende.

Desværre deler mange leksikografer her meningsfællesskab med ordbogsforlagene, idet de færreste er villige til at dele ud af selve logfilerne. Der kan være juridiske hindringer, men disse kan overvindes. Uanset hvad ville et internationalt repositorium for den type undersøgelser, logfiler og resultater fra evt. eksperimenter være et vigtigt og nyttigt redskab til udarbejdelsen af bedre (online-) referenceværker.

Det forbliver dog vigtigt at huske, at fortsætter leksikografen med at foretage brugerundersøgelser på samme vis, som det er sket hidtil, vil ordbøger kun kunne forbedres gennem tilfældigheder.

Litteratur

Bergenholtz, Henning og Mia Johnsen 2007: Log Files can and should be prepared for a functionalistic approach. I: *Lexikos 17*, 1–20.

Bergenholtz, Henning, Vibeke Vrang og Esben Bjærge: *Ordbogen over Faste Vendinger*. Database og design: Richard Almind. Århus: Handelshøjskolen i Århus, Aarhus Universitet 2008. (www.idiomordbogen.dk)

Tarp, Sven 2008: Kan brugerundersøgelser overhovedet afdække brugernes leksikografiske behov? I: *LexicoNordica 15*, 5–32.

Richard Almind
Forskningsmedarbejder med ansvar for IT-udvikling
Center for Leksikografi
Aarhus Universitet, Handelshøjskolen
Fuglesangs Alle 4
DK-8210 Århus V
rab@asb.dk