2015/3 · september

tidsskrift *sb.*, -et, -er, *i sms.* tidsskrift- el. tidsskrifts-, *fx* tidsskrift(s)artikel.
tidsspild *sb.*, -et *el.* tids-spilde *sb.*, -t.
tidsspørgsmål *sb.*, -et, tids-spørgsmål, *bf.* -et. *en-e*
tidsstemple *vb.*, -ede.
tidsstempling *sb.*, -en, -er.

Nyt fra Sprognævnet

Jagten på den gode og sikre sprogbruger

Af *Philip Diderichsen* og *Jørgen Schack*

Jagten på den gode og sikre sprogbruger er titlen på et foredrag som blev holdt på Sprognævnets jubilæumsseminar om norm og empiri 23.4.2015. Bag titlen gemmer sig en præsentation af et nyt sprognævnsprojekt med arbejdstitlen *Ortografisk rangering af korpustekster*. Projektet udspringer direkte af en årelang og stadig uafsluttet diskussion om hvad der karakteriserer "den gode og sikre sprogbruger".

"Gode og sikre sprogbrugeres skriftlige sprogbrug"

Hovedprincipperne for dansk retskrivning er ifølge sprognævnsbekendtgørelsens § 1 traditionsprincippet og sprogbrugsprincippet. Ifølge traditionsprincippet skal ordene meget kort fortalt staves som de plejer, "bortset fra justeringer som følge af sprogbrugsprincippet." Og det er hér de gode og sikre sprogbrugere kommer ind i billedet, for ifølge sprogbrugsprincippet skal ord og ordformer i dansk skrives "i overensstemmelse med den praksis som følges i gode og sikre sprogbrugeres skriftlige sprogbrug".

Lad os forestille os at redaktørerne af Retskrivningsordbogen overvejer at indføre ordformen *autencitet* i den næste udgave af ordbogen, enten som dobbeltform til den nuværende form, *autenticitet*, eller som en ny eneform. En indførelse af *autencitet* forudsætter ifølge bekendtgørelsen at ordformen er i overensstemmelse med god og sikker skriftsprogspraksis. Vi kan dokumentere at *autencitet* er forholdsvis udbredt i avissproget: Søger man efter substantiveringer af adjektivet *autentisk* i den store mediedatabase Infomedia, viser det sig at den autoriserede form, *autenticitet*, bruges i ca. 75 % af tilfældene, mens den kortere form, *autencitet*, tegner sig for ca. 25 %. Der melder sig straks to spørgsmål: (1) Er 25 % "nok"? (2) Er

>

de tekster som den omdiskuterede ordform optræder i, skrevet af gode og sikre sprogbrugere? Det sidste spørgsmål er vigtigere end det første, og for at kunne besvare det må vi nødvendigvis have en nogenlunde klar forestilling om *hvad* det vil sige at være en god og sikker sprogbruger. Bekendtgørelsen hjælper os ikke på dette punkt, og det skal den heller ikke: Aben ligger hos Sprognævnet!

Det må være sådan at de gode og sikre sprogbrugere skal identificeres på grundlag af de tekster de producerer, og ikke omvendt: En sprogbruger kan få prædikatet "god og sikker" på grundlag af sin sprogbrug, ikke på grundlag af sit renommé.

Mange har i årenes løb spurgt Sprognævnet hvem de gode og sikre sprogbrugere egentlig er, og svaret har normalt været lidt tøvende. Vi mener selv at vi med rimelig sikkerhed kan udpege hvad der er god og sikker skriftlig sprogbrug, og hvad der ikke er, men vores bedømmelseskriterier bliver sjældent ekspliciteret.

Onde tunger vil måske hævde at Sprognævnets bedømmelseskriterier er kendetegnet ved vilkårlighed og smagsdommeri. Det er de på ingen måde. De hviler på et solidt grundlag af erfaring og empiri. Det sidstnævnte, empirien (dvs. store tekstsamlinger og andre sproglige data), spiller en stadig større rolle i nævnets normerings- og rådgivningsarbejde. Før i tiden måtte vi nøjes med den empiri vi nu engang havde: Det var hovedsagelig excerpterne i nævnets store seddelsamling. I dag har vi adgang til store korpusser (tekstsamlinger). Det er et kæmpe fremskridt, men det rejser visse metodiske spørgsmål. Fx indgår spørgsmålet om frekvens meget ofte i vores beskrivelser og vurderinger af sproglige størrelser af enhver art, og vi må derfor stille strenge krav til de tekstsamlinger som vi søger i og udtaler os på baggrund af: Vi skal vide hvordan de er sammensat, og vi skal have grund til at antage at de er repræsentative for den type sprogbrug som vi vil udtale os om og evt. vejlede eller normere på baggrund af.

Ortografisk rangering: at ordne tekster efter antallet af stavefejl

På et forskningsseminar i 2009 skitserede Sprognævnets daværende formand, Dorthe Duncker, hvordan man kan opbygge et tekstkorpus der gør det muligt at følge udviklingen i den danske skriftsprogsnorm. I et sådant korpus skal de enkelte tekster klassificeres i forhold til en ortografisk "guldstandard", dvs. i forhold til hvordan teksterne i ortografisk henseende forholder sig til normen i Retskrivningsordbogen. Man kan betragte korpusset som en slags ortografisk kvalitetsbarometer: I toppen af barometret finder vi de tekster hvis praksis er nærmest normen i Retskrivningsordbogen, og i bunden finder vi dem der er fjernest fra normen. Hvis det faktisk er muligt at rangere store tekstmængder mht. ortografisk kvalitet, så vil vi kunne afgøre om en form, fx *autencitet*, faktisk er udbredt i gode og sikre staveres skriftlige praksis. Det vil være et godt empirisk grundlag for normeringsarbejdet.

I projektet *Ortografisk rangering af korpustekster* eksperimenterer vi med at rangere tekster, foreløbig avistekster, i forhold til antallet af afvigelser fra retskrivningsnormen. En sådan rangering kan kun gennemføres hvis vi kan finde normafvigelserne automatisk, for det er meget store tekstmængder

vi arbejder med. Det korpus vi har arbejdet med i den indledende fase, indeholder knap 200 mio. løbende ord fordelt på ca. 370.000 tekster fra 7 landsdækkende aviser. Teksterne dækker perioden 2004-2014 (dog primært 2010-2014). Korpusset er den ene hovedkomponent i projektet. Den anden hovedkomponent er en liste med en særlig type fejlformer. I den indledende fase har vi stort set udelukkende ledt efter *ikke-ord*. Ved et ikke-ord forstår vi en form som ikke svarer til en autoriseret stave- eller ordform, og som derfor principielt er en fejl i en hvilken som helst nutidig dansk kontekst. Det kan fx være "lærene" for *lærerne* og "hierakisk" for *hierarkisk*. Der findes også fejl der kan være korrekte i andre kontekster (tænk fx på "at lærer"), men når vi har koncentreret os om ikke-ord, er det fordi det foreløbig er den eneste type fejl vi kan identificere med relativt stor sikkerhed.

Vores fejlliste indeholder p.t. ca. 20.000 forskellige fejlformer (ikke-ord) fordelt på knap 90 fejltyper, fx typen manglende fuge-*s* før *s* ("elskovsyg" for *elskovssyg* osv.), *-tion* for *-sion* ("reflektion" for *refleksion* osv.). De mange fejlformer er vi nået frem til ved at tage udgangspunkt i ortografiske og morfologiske strukturer som erfaringsmæssigt giver anledning til fejl, bl.a. de to netop nævnte.

Fejllisten er delvist inspireret af automatiske stavekontroller. En af komponenterne i en stavekontrol er en liste med korrekte former, der fungerer som supplement til en komponent som beregner sandsynligheden for om en bestemt form af et ord er den rigtige i en given sammenhæng. De former i en tekst der ikke svarer til en form på listen, og som heller ikke kan dannes af stavekontrollens morfologiske komponent, markeres som fejl (se artiklen *Stavekontrol* i Nyt fra Sprognævnet 2015/1). Det fungerer ofte udmærket, men som de fleste har erfaret, får man alligevel af og til et stort antal falske positiver, dvs. korrekte former der markeres som fejl (især sammensatte ord som stavekontrollen ikke kender og ikke kan danne). For at minimere antallet af falske positiver, tager vi i stedet udgangspunkt i en liste med ukorrekte former. Ulempen ved at arbejde med en fejlliste er selvfølgelig at vi kun finder de fejlformer som optræder på vores liste. Vi har imidlertid valgt at vægte præcision højest i denne fase af projektet: De former vi finder ved hjælp af listen, er med meget stor sandsynlighed ægte fejl.

Vi har i det ovennævnte korpus fundet knap 10.000 fejl ud af knap 10 mio. fejlkandidater. I teksterne er der m.a.o. fejl i 1 promille af de ord- og staveformer som vores fejlliste dækker – som i sig selv kun udgør ca. 5 % af de løbende ord i korpusset.

Bestemmelserne i sprognævnsbekendtgørelsen vedrører først og fremmest ortografi. Bøjnings- og orddannelsesforhold er vel også inkluderet, for der tales her ikke kun om ord, men også om ordformer. Men med til det at være en god og sikker sprogbruger hører jo meget andet end blot at kunne stave og bøje ordene korrekt. Vi er derfor interesserede i om der er en sammenhæng mellem god og sikker ortografi og god og sikker sprogbrug på andre sproglige niveauer. Det første vi har gjort for at undersøge det, er at lave en pilotundersøgelse med vores kolleger, hvor vi bad dem bedømme nogle af de fejltekster vi havde fundet.

Er fejltekster mindre sammenhængende og stilsikre?

Pilotforsøget gik i sin enkelhed ud på at se om nævnets medarbejdere ville være i stand til at skelne tekster med ikke-ord fra tekster uden. Vel at mærke hvis alle ortografiske fejl først blev fjernet fra teksterne. Vi ville se om mere tekstlige og stilmæssige aspekter af teksterne alene (dvs. fx tekstsammenhæng, brug af faste udtryk, billedsprog m.v.) ville få vores kolleger til at foretrække tekster uden vores ikke-ord.

For at øge sandsynligheden for dette resultat brugte vi de mest fejlfyldte tekster vi endnu havde fundet – tre forskellige avisartikler med hhv. 5 og 6 forskellige fejltyper fra vores fejlliste. Disse fejltekster matchede vi med tekster uden ikke-ord fra listen. Til hver fejltekst fandt vi en matchtekst fra samme avis, af nogenlunde samme længde, med samme emne, fra samme år og med en forfatter der lignede fejltekstforfatteren (dvs. med samme køn og skønnet samme alder (ud fra forfatterens navn) samt omtrent samme antal tekster i vores korpus). På den måde lavede vi tre tekstpar med to tekster der lignede hinanden så meget som muligt – lige bortset fra at den ene indeholdt en del fejllistefejl.

Herefter rensede vi teksterne for ortografiske fejl, og så bad vi vores kolleger om at finde den bedste tekst i hvert par. (Se figur 1).

For ikke at sætte barren alt for højt lod vi dog visse typer fejl stå, nemlig: orddannelsesfejl (fx "læsværdige" for *læseværdige*); syntaktiske fejl, typisk kongruensfejl (fx "vor nationale særpræg" for *vort nationale særpræg*); kommafejl og desuden rene slåfejl, fx ekstra bogstaver ("hvillke"), udeladte bogstaver ("kosmopoltiske") eller ombyttede bogstaver ("opmærskomme").

I det første tekstpar (to litteraturanmeldelser) indeholdt fejlteksten 7 fejl af 6 forskellige typer fra vores fejlliste. Derudover var der 4 ortografiske fejl der ikke var på fejllisten, 1 morfologisk fejl, 1 syntaktisk fejl, 5 slåfejl og 11 kommafejl. Heroverfor var der i matchteksten kun en enkelt kommafejl, se tabel 1.

Fejltype	Fejltekst 1	Matchtekst 1
Ortografiske fejl fra fejlliste	7	-
Yderligere, rent ortografiske fejl	4	0
Morfologiske fejl	1	0
Syntaktiske fejl	1	0
Slåfejl	1	0
Kommafejl	11	1

Tabel 1. Fejlantal i tekstpar 1. Det mørkere område dækker de fejl som vi fjernede fra teksterne.

Resultatet af den første sammenligning kan næppe overraske nogen: 8 ud af 8 bedømte matchteksten som bedre end fejlteksten. Forsøgsdeltagerne blev bedt om at supplere deres bedømmelser med evt. kommentarer. De fleste bemærkede i deres kommentarer de mange kommafejl i fejlteksten, men der blev også peget på andre ting såsom "kongruensfejl", "stilblanding", "usikker brug af billedlige udtryk", "manglende elegance", "svulstigt sprog" og "sjuskefejl". Matchteksten derimod fik ord med på vejen såsom "letløbende", "godt flow", "flyder bedst" og "hænger bedre sammen", alt sammen udtryk for bedre tekstsammenhæng. Desuden blev den bedømt til at være "mindre subjektiv", og de færre formelle fejl blev bemærket.

Figur 1. Skitse af det første tekstpars tilrettelæggelse og bedømmelse. (i) Fejlteksten indeholdt 7 fejllistefejl af 6 forskellige typer, matchteksten ingen. (ii) Fejlteksten indeholdt en mængde øvrige fejl, matchteksten i dette tilfælde kun en kommafejl. (Se også tabel 1). (iii) Alle ortografiske fejl med visse undtagelser blev fjernet før teksterne skulle bedømmes. (iv) Samtlige bedømmere foretrak matchteksten.

Spørgsmålet er om der generelt er en sammenhæng mellem ortografiske fejl, morfologiske fejl osv. og mindre sikker sprogbrug på tekstlig-stilistisk niveau. Som vi skal se nedenfor, understøtter dette pilotforsøg at det kan være værd at se nærmere på.

I det andet tekstpar (to artikler om boligpolitik) fjernede vi 10 fejl fra fejlteksten og 2 fra matchteksten. Til forskel fra det første tekstpar matchede de to tekster herefter hinanden ret godt i antallet af resterende fejl, se tabel 2.

Fejltype	Fejl-tekst 2	Match-tekst 2
Ortografiske fejl fra fejlliste	5	-
Yderligere, rent ortografiske fejl	5	2
Morfologiske fejl	0	0
Syntaktiske fejl	0	0
Slåfejl	0	1
Kommafejl	3	1

Tabel 2. Fejlantal i tekstpar 2. Det mørkere område dækker de fejl som vi fjernede fra teksterne.

Ikke desto mindre bedømte 7 ud af 8 matchteksten som den bedste – selv om de altså her havde langt færre ortografiske problempunkter at holde sig til. Det tyder på at det rent tekstlige her har været det afgørende for bedømmelserne. Det understøttes igen af de kommentarer bedømmerne havde: Fejlteksten blev bedømt som "krukket i en grad så det var forstyrrende for tekstens flow", og der blev påpeget "mislykkede 'friske og sjove' ordforbindelser", "manglende ord" og "dårlig kohærens". Kommafejlene blev også kommenteret. Dog mente en enkelt at fejlteksten havde "mere flydende sprog" og "enklere syntaks". Matchteksten blev karakteriseret som "velstruktureret", "informativ" og "mere sikker og flydende, dog af og til med et lidt pudsigt ordvalg". At disse forskelle trods alt har været ret subtile, indikeres af at en noterede at det var "vanskeligst at finde en vinder" i dette tekstpar.

I tekstpar 3 (to formel 1-reportager) stod de to tekster igen nogenlunde lige i de rensede versioner bedømmerne fik. Men i dette tilfælde var det faktisk matchteksten der havde flest fejl i den originale version (i alt 12 mod fejltekstens i alt 9, se tabel 3).

Fejltype	Fejl-tekst 3	Match-tekst 3
Ortografiske fejl fra fejlliste	5	-
Yderligere, rent ortografiske fejl	1	9
Morfologiske fejl	0	0
Syntaktiske fejl	0	1
Slåfejl	1	0
Kommafejl	2	2

Tabel 3. Fejlantal i tekstpar 3. Det mørkere område dækker de fejl som vi fjernede fra teksterne.

Det kunne lade sig gøre fordi der var fejl i matchteksten der ikke fandtes på vores fejlliste – hovedsagelig særskrivningsfejl.

Det interessante er at det i tekstpar 3 generelt var fejlteksten der blev bedømt som den bedste, nemlig af 7 ud af 8 bedømmere. Det tyder på at antallet af ortografiske fejl m.m. præcis som i de to andre tekstpar hænger sammen med øvrige indikatorer for 'usikker sprogbrug' – også i dette tilfælde tilsyneladende mere tekstlige fænomener. Her var det således fejlteksten der blev karakteriseret som følger: "Mere upåfaldende sprogstil", "bedre flow og struktur", "flydende og næsten fejlfri", men den blev dog også bebrejdet "et par manglende kommaer" og "et par sjuskefejl". Matchteksten var derimod: "Upræcis og fejlfyldt", "hakket op", "usammenhængende", "indforstået", "ikke lige så dramatisk" og "sjusket" og blev desuden klandret for "uklart sprog", "mystisk jargon" og "fejl/mangler i teksten".

På sporet af den (u)sikre sprogbrug(er)

I vores lille materiale finder vi altså den sammenhæng som er en forudsætning for hele projektet med at identificere gode og sikre sprogbrugere: Tekster med mange ortografiske, morfologiske og syntaktiske fejl er også relativt usikre mht. tekstsammenhæng og stil. Der er selvfølgelig tale om en pilotundersøgelse, men tilgangen virker lovende.

Der er to indvendinger som vi kort vil berøre her: (1) Eftersom avisartikler jo redigeres og i mange tilfælde også korrekturlæses, hvordan kan vi så tale om "den gode og sikre sprogbruger" som om ansvaret for avisernes sprogbrug ligger hos enkeltpersoner? (2) Der kan (bl.a. pga. korrekturen) næppe være ret mange fejl i aviserne. Så er korpusset ikke bare én stor omgang god og sikker sprogbrug? Dette kunne være et par gode grunde til ikke at bruge avistekster. Men vi mener nu alligevel at vores data er brugbare (omend meget gerne i samspil med andre typer tekster med tiden).

Den første indvending har fat i noget, men rammer i sidste ende forbi. Vores aviskorpus vil næppe kunne bruges til at sige noget om den enkelte skribents stavefærdighed – men det er i virkeligheden heller ikke det vi er ude efter, titlen uagtet. Vi er meget mere interesserede i den gode og sikre sprog*brug*. Og hvis en bestemt forfattersignatur er en pålidelig indikator for enten sikker eller usikker sprogbrug, så vil vi selvfølgelig bruge den i vores søgen efter relevante tekster. Så gør det ikke så meget at diverse redigerende, som det hedder i fagjargonen, plus evt. korrekturlæsere har været inde over teksterne. Hvis tekster "af" en bestemt skribent systematisk er relativt fejlfyldte, ja, så skal vi selvfølgelig lede efter flere fejl i tekster fra denne skribent snarere end fra en signatur hvis tekster vi har kunnet konstatere er fejlfri. Flere menneskers indflydelse på tekster har i øvrigt altid været et vilkår for Sprognævnets normeringsarbej-

de. Før som nu blev store forfattere benyttet som sproglige forbilleder – men det giver næsten sig selv at heller ikke deres værker bare er "deres", men selvfølgelig også er et produkt af en udgivelsesproces der typisk involverer adskillige andre personer (forlagsredaktører, korrekturlæsere osv.).

At der ikke er ret mange fejl i aviserne, er helt rigtigt. Det vil vi faktisk gerne fremhæve, for i vores arbejde indtil nu er det vi har set, at fejlprocenterne generelt er meget lave, og endda faldende over tid. Når vi alligevel mener at vi vil kunne få noget ud af vores korpus, er det for det første fordi det er så stort. Trods lave fejlprocenter kan vi pga. korpussets størrelse finde så mange fejl at vi kan se meningsfulde statistiske mønstre i dem. For det andet er vores tekster ikke bare tekster. De indeholder også ret omfattende metaoplysninger såsom emnekategorier og geografiske kategorier foruden kilde-, dato- og forfatteroplysninger. I vores arbejde indtil nu har vi fx allerede set tegn på at der er en sammenhæng mellem artiklernes emne og antallet af fejl. Metaoplysninger af denne type vil vi selvfølgelig gerne udnytte.

Spørgsmålet er nu om sammenhængen mellem antal ortografiske fejl og overordnet tekstkvalitet kan påvises mere generelt. Her venter der stadig et stort stykke arbejde. Kigger man nærmere på resultaterne ovenfor, vil man se at vores procedure til at finde fejltekster har en væsentlig brist. Der er ganske vist en tendens til at antallet af ortografiske fejl (altså dem der blev fjernet fra teksterne) hænger sammen med bedømmelserne (flest ortografiske fejl <=> dårlig bedømmelse). Men antallet af fejllistefejl hænger ikke altid sammen med antallet af øvrige fejl. I det tredje tekstpar er der således samlet set flest øvrige fejl i matchteksten. Det betyder at vi nok ikke skal regne med at finde en pålidelig tekstrangeringsmetode hvis vi udelukkende benytter os af fejllisten som den ser ud på nuværende tidspunkt.

Philip Diderichsen *(f. 1977) er videnskabelig medarbejder i Dansk Sprognævn.*

Jørgen Schack *(f. 1961) er seniorforsker i Dansk Sprognævn.*

Det er der heldigvis råd for. I næste fase af projektet vil vi dels udbygge fejllisten så meget vi kan, dels udvide fejlsøgningen fra rene ikke-ord til også at inkludere betingede fejl. Betingede fejl er ord- og ordformer som er fejl i den konkrete kontekst, men som er korrekte former i andre kontekster, fx *hvis* i "et fænomen som har en hvis udbredelse" og *anføre* i "han er deres nye anføre". Betingede fejl er langt sværere at finde automatisk fordi det forudsætter at korpusset er forsynet med ordklasseangivelser og andre grammatiske kategorier. Fx vil former som *anføre* næsten kun kunne identificeres som fejl på en generel måde hvis det automatisk kan registreres at de – trods det manglende *r* – står på et substantivs plads i sætningen. Annoteringen med grammatiske kategorier foregår i >

samarbejde med Eckhard Bick fra Syddansk Universitet i Odense.

Der begynder allerede nu at tegne sig et billede af hvor man kan kigge efter god og sikker hhv. mindre sikker sprogbrug i forbindelse med normeringsarbejdet. Hvis vi vil identificere mindre sikker sprogbrug, skal vi formentlig kigge hos de skribenter der har de højeste fejlantal. Det vil projektets næste faser vise. Hvis vi vil se god og sikker sprogbrug, skal vi sammensætte det størst mulige korpus af stort set fejlfri tekster – et korpus som samtidig er balanceret så diverse dagblade og emner er fordelt repræsentativt.

Hvis denne metode viser sig at være effektiv, er det planen at gå videre med andre tekster der kan tænkes at være væsentligt mere fejlfyldte end dem vi arbejder med nu: tekster fra lokale aviser, fra det offentlige Danmarks websider og måske endda fra de sociale medier.

Jagten på "den gode og sikre sprogbruger" er kun lige begyndt. Og vi er på sporet!

Maskinoversættelse

Af *Philip Diderichsen*

Artikelserie:
Hvordan virker sprogteknologien i din hverdag?

Hvad karakteriserer egentlig en god oversættelse? I bund og grund er det hovedsageligt to faktorer: 1) hvor direkte en oversættelse det er (dvs. hvor loyal mod originalsproget den er), og 2) hvor flydende resultatet er (dvs. hvor loyal mod målsproget den er). For oversættere er det en grundlæggende udfordring at tilgodese begge hensyn, og det kan ofte være tæt på umuligt. Se fx dette eksempel fra "Bøfsiden: bommerter og fusere på tv og tryk"[1], hvor det dog er mislykkedes mere end nødvendigt:

Original sætning (i tv-film)	It's nice to have an early bird dinner here.
Undertekst på tv	Jeg nyder denne fjerkræsmiddag.
Det burde fx have været	Jeg nyder denne tidlige (evt. billige) middag. (En "early bird dinner" er en middag, som visse restauranter tilbyder til nedsat pris sent om eftermiddagen før almindelig spisetid. Selskabet spiste i øvrigt bøffer!)

Kunsten er at finde en oversættelse af *early bird dinner* der ikke enten går ud over målsproget (*tidlig fugl-middag*) eller går ud over originalsproget (*tidlig middag* eller *billig middag*). Det ene er tydeligvis ikke vellykket på dansk, og det andet mister enten nuancen

1 Se http://www.titlevision.dk/boeuf.htm.

'billig' eller nuancen 'tidlig' (hvor *tidlig, billig middag* allerede begynder at blive for kluntet på dansk).

Udfordringen er den samme for maskinoversættelse, som nok stadig er det som flest forbinder med begrebet sprogteknologi. En for direkte oversættelse er grinagtig fordi den lyder som cirkusdansk – en for flydende oversættelse kan være ubrugelig fordi meningen pludselig er blevet en helt anden. Maskinoversættelser kan lide af begge disse dårligdomme i varierende grad. Det kan bl.a. have at gøre med hvilken grundide maskinoversættelsessystemet er opbygget efter. Som det har været berørt i de tidligere artikler i denne serie, er der overordnet set to forskellige tilgange til sprogteknologi, og de går igen i maskinoversættelse: den regelbaserede tilgang og den statistiske tilgang. Her er et rids af hvordan de virker.

Regelbaseret maskinoversættelse

Hvad vil det sige at maskinoversættelse er regelbaseret – hvordan ser reglerne ud? Tænk på ordet *regne*, der fx kan kræve et af disse ord i en engelsk oversættelse (se Bick 2009): *rain*, *calculate, expect, include*. Hvis der er et formelt subjekt i sætningen (fx *det* i *det regner*), så vælges vejrbetydningen, og oversættelsen bliver *rain*. Hvis ikke, så udløser det oversættelsen *calculate*. En trumfende regel lyder at hvis *regne* efterfølges af *med*, så bliver oversættelsen *expect* – dog kun hvis *med* er en præposition, ikke hvis det er en adverbialpartikel; så bliver oversættelsen i stedet *include*. Disse regler giver fx en fornuftig oversættelse af begge disse sætninger: *De regner ofte **med** et underskud* (*med* er præposition, altså: *They often expect a deficit*) og *De regner ofte et underskud **med*** (*med* er adverbialpartikel, altså: *They often include a deficit*).

Reglerne kan stilles lidt mere overskueligt op i et hierarki som det nedenstående. Hierarkiet vil selvfølgelig i praksis være langt større – der er ganske mange flere oversættelser til *regne*, ja, alene til *regne med* er der mange flere oversættelser end vist.

Regne:

- Formelt subjekt? (= *Det*)
  - Ja => ***rain***
  - Nej => Efterfølgende *med*?
    - Ja => *Med* er præposition?
      - Ja => ***expect***
      - Nej => ***include***
    - Nej => ***calculate***

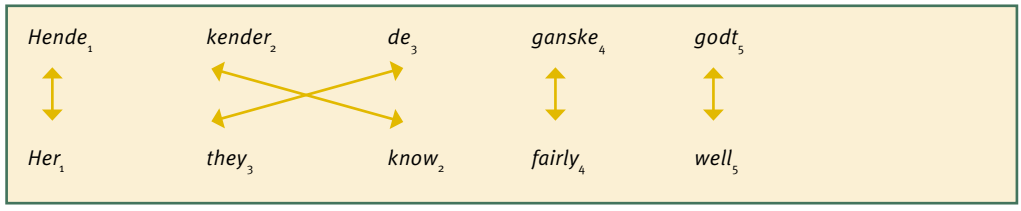
Denne opstilling giver en ide om hvordan et regelbaseret maskinoversættelsessystem har brug for en endog meget lang og kompleks liste med regler for ordoversættelser –

Artikelserie: sprogteknologi

Dette er den tredje artikel i Nyt fra Sprognævnets serie om sprogteknologi. Artikelserien handler om hvordan sprogteknologi egentlig virker. Nævnet vil i de kommende år beskæftige sig mere med sprogteknologi, ikke bare med fokus på dens betydning for retskrivningen, men også med henblik på at sikre dansks status i alle sammenhænge – fx international handel. Her kommer maskinoversættelse ind i billedet.

ikke bare en liste over de danske ord og en liste over de tilsvarende udenlandske ord og så færdig.

Også ordrækkefølgen i en sætning skal ofte laves om. Et simpelt eksempel kunne være sætningen *Hende*$_1$ *kender*$_2$ *de*$_3$ *ganske*$_4$ *godt*$_5$ (=> *Her*$_1$ *they*$_3$ *know*$_2$ *fairly*$_4$ *well*$_5$), se illustrationen nedenfor. Det er der andre regler for.



Dertil kommer omfattende sproglige analyser af det sprog der skal oversættes (i dette tilfælde dansk), analyser der også foregår vha. regler. Én sådan regel er allerede antydet i det ovenstående: Oversættelsen *calculate* kræver et menneskeligt subjekt, der i givet fald skal findes i den danske originalsætning. En regel for dansk for om et ord refererer til et menneske, kunne være noget i retning af: Hvis ordet er et personligt pronomen (de er hurtigt opregnet), eller hvis udtrykket er et personnavn (langt sværere, det kræver et særligt navnegenkendelsesmodul i systemet), så refererer udtrykket til et menneske. At identificere ordet som sætningens subjekt kan opnås bl.a. med syntaktiske analyseregler, fx "Hvis X står før verbet, så er X sætningens subjekt" (en regel der selvfølgelig har masser af undtagelser). Og at *regner* hedder *rains* når originalsætningen fx er *Det regner hvert 10. år* (=> *It rains every 10 years*), men *rain* hvis originalsætningen er *Frøer regner ned fra himlen hvert 10. år* (=> *Frogs rain from the sky every 10 years*) – ja, det kan som regel klares med morfologiske analyseregler der afgør om det danske verbums subjekt står i singularis eller pluralis (fx "Hvis subjektet ender på *(e)r*, så står det i pluralis" – selvfølgelig igen med mange undtagelser).

Regler, regler, regler. Regelbaseret maskinoversættelse kræver store mængder regler, der møjsommeligt skal formuleres af leksikografer og lingvister, og denne tilgang er derfor meget dyr – selvfølgelig desto mere jo flere sprogpar der skal kunne oversættes indbyrdes. Til gengæld kan der i princippet opnås oversættelser af meget høj kvalitet. Fejl er nemlig ofte forholdsvis gennemskuelige og kan lokaliseres præcist til en bestemt regel, der så kan rettes (hvis der ellers er penge og ekspertise til det).

Heroverfor står statistisk baseret maskinoversættelse, der ikke kræver nær så meget leksikografisk-lingvistisk ekspertarbejde, men til gengæld avanceret matematisk ekspertise.

Statistisk baseret maskinoversættelse

Statistisk maskinoversættelse af en sætning foregår groft sagt ved at der genereres et antal oversættelser af forskellige delmængder af sætningen, hvorefter en sandsynlighedsberegning afgør hvilken kombination af de resulterende byggeklodser der er den bedste. Hvilken oversættelse der er bedst, kan som

nævnt bedømmes på hvor direkte og hvor flydende en oversættelse det er (dvs. hvor loyal den er mod originalsprog og målsprog). Disse to faktorer kan tages helt bogstaveligt og beregnes som sandsynligheder der ganges med hinanden for at give et mål for den maskinelle oversættelseskvalitet. Jo større de begge er, des bedre er oversættelsen.

Hvordan beregner et maskinoversættelsessystem loyaliteten mod original- og målsprogene – og hvorfor beregnes de som sandsynligheder? Lad os vende tilbage til underteksterens bøf ovenfor, således at originalsproget nu er engelsk og målsproget dansk.

Hvad målsproget angår, så kan man benytte sig af en af de allermest udbredte komponenter i statistisk sprogteknologi, nemlig *sprogmodellen*, som også er blevet nævnt i de to foregående artikler i serien om sprogteknologi. En sprogmodel bygger på en liste af ordhyppigheder. *Og*, *i, at* og *jeg* er ekstremt hyppige i løbende tekst; *fjerkræsmiddag* er ikke. Udover hyppigheder for enkeltord udnytter sprogmodellen også hyppigheder for flere ord efter hinanden. *Jeg nyder denne* vil være hyppigere end *nyder denne fjerkræsmiddag*, som igen vil være hyppigere end *i at fjerkræsmiddag* (som højst sandsynligt aldrig er forekommet før nu).

Ordenes hyppigheder svarer til rå sandsynligheder for at støde på dem som det næste ord i en tekst. Ligesom sandsynligheden for at slå to seksere efter hinanden er 1/6 x 1/6, kan man bruge ordsandsynlighederne til at udregne sandsynligheden for en sætning ved at gange kæden af ordsandsynligheder med hinanden. En hel sætnings sandsynlighed kan så bruges som et mål for hvor god en sætning det er: jo højere sandsynlighed, des bedre. I praksis tager sprogmodellen dog flere ord i betragtning ad gangen, og en sætnings sandsynlighed kan således beregnes ved at gange sandsynligheden for hver af rækkerne i følgende skema med hinanden (jf. det tilsvarende skema i artiklen om stavekontrol i Nyt fra Sprognævnet 2015/1).

<start>	Jeg	nyder			
	Jeg	nyder	denne		
		nyder	denne	fjerkræsmiddag	
			denne	fjerkræsmiddag	<slut>

Fjerkræsmiddag er i sig selv et meget lidt hyppigt ord, meget mindre hyppigt end ordfølgen *tidlige middag*. Så den ringe sandsynlighed for at møde *fjerkræsmiddag* i en dansk sætning vil i sig selv gøre oversættelsesbøffen i eksemplet til en lavere rangerende oversættelse i et maskinoversættelsessystem til fordel for *Jeg nyder denne tidlige middag*.

Der er dog mange oversættelser der er "gode" hvis "god" alene betyder at en oversættelse består af hyppige ord i hyppige kombinationer. Så hvis systemet ikke skal ende med at oversætte *It's nice to have an early bird dinner here* til *I eftermiddag kun lidt sol, og de fleste steder byger af og til* eller en anden helt urelateret, men i sig selv meget sandsynlig dansk sætning, så er loyaliteten overfor originalsproget nødt til at indgå i beregnin- >

gen. Hvilket den selvfølgelig også gør.

Originalsproget tilgodeses ved hjælp af flere sandsynlighedsberegninger. Statistisk maskinoversættelse går ud fra at sandsynligheden for at to sætninger svarer til hinanden på original- og målsproget, stiger, jo flere ord og ordforbindelser der er oversat direkte. Denne sandsynlighed beregnes af en oversættelsesmodel.

En oversættelsesmodel minder lidt om sprogmodellen fra før. I bund og grund er det en statistisk tosprogsordbog – en liste med ord og ordforbindelsers oversættelser og disses sandsynligheder. Hvor ordet *regne* i regelbaseret maskinoversættelse har en betydelig mængde oversættelsesregler, så har hver mulig oversættelse (*rain*, *expect*, *include*, *calculate* etc.) i statistisk maskinoversættelse en sandsynlighed baseret på hvor hyppigt den enkelte oversættelse forekommer i oversat tekst. Oversættelsesmodellen har også oversættelser af flerordsforbindelser (hvis systemet ellers har haft adgang til passende mængder oversat tekst hvor de forekommer). *Det regner* vil således have oversættelsen *it rains* med en høj sandsynlighed, og fx *regne med* vil både have oversættelsen *expect* og *include* med hver sin sandsynlighed (*regne med et underskud* => ***expect*** *a deficit*; *regne underskuddet med* => ***include*** *the deficit*).

Disse sandsynligheder kan dog ikke som i sprogmodellen findes ved simpelthen at tælle oversatte ord og flerordsforbindelser – for hvilke hører sammen? Ord og ordforbindelser i originalteksten skal først kobles sammen med ord og ordforbindelser i målteksten. Dette kræver sin egen sandsynlighedsberegning, der i sidste ende afhænger af hvor ofte de enkelte ord i en samling oversatte tekster forekommer sammen med deres oversættelse indenfor den tilsvarende sætning på det andet sprog.

En sidste, vigtig del af oversættelsesmodellen sørger for at trække ned i sandsynligheden for en bestemt oversættelse af et ord eller en ordforbindelse hvis den ender for langt væk fra sin pendant. Et eksempel kunne være sætningen i følgende skema:

Gør de	noget	**ved**	den	nye	diskussion	**om**	læring?
Are they doing	anything	about	the	new	discussion	on	learning?

Ofte oversættes *om* til *about* – men ikke i denne sætning. Sandsynligheden for at det er *om* der skal oversættes til *about*, mindskes fordi *ved* og *about* står tættere på hinanden.

Guidet af sprogmodellen og oversættelsesmodellen orkestrerer et statistisk maskinoversættelsessystem oversættelsen af en given sætning ved at generere deloversættelser af alle de enkeltord og ordkombinationer der kan tænkes (dvs. har en rimelig sandsynlighed for) at være oversættelser af hinanden. Det følgende konstruerede eksempel er lavet ved at indsætte hvert ord, hver toordsforbindelse osv. i Google Oversæt. Det er antydet hvordan der kan findes adskillige forskellige observerede oversættelser af såvel enkeltordene som flerordsforbindelserne. Det man altså skal se for sig, er at Google Oversæt har genereret en endnu større tabel af mulige deloversættelser ved hele tiden undervejs at 'slå op i' sprogmodellen og oversættelsesmodellen. Til sidst vælges den bedste vej gennem alle kom-

binationsmulighederne ved at vælge de ord og ordforbindelser der maksimerer sandsynligheden for en god (dvs. originalsprogs- OG målsprogsloyal) oversættelse.

Vi	regner	med	et	overskud
We	rains	with	a	profits
Marry	calculate	by	an	excess
	...	...	...	...

Vi	regner	med	et	overskud
We expect				
...				
	plan			
	...			
		with a		
		...		
			an excess	
			...	

Vi	regner	med	et	overskud
We expect				
...				
	counting on a			
	...			
		with an excess		
		...		

Vi	regner	med	et	overskud
We expect a				
...				
	expects a profit			
	...			

Vi	regner	med	et	overskud
We expect a profit				
...				

Sandsynligheder, sandsynligheder og atter sandsynligheder, beregnet i et hierarki af komponenter – det er sådan der ser ud når man kigger ned i maskinrummet på et statistisk maskinoversættelsessystem.

God statistisk baseret oversættelse:

- Høj sætningssandsynlighed via sprogmodel (= vellykket sætning på målsproget)
- Høj oversat ord-/ordforbindelsessandsynlighed via oversættelsesmodel (= direkte oversættelse)
  - o Høje oversættelsessandsynligheder for ord og ordforbindelser
    - Høje sandsynligheder for sammenfaldende enkeltord
    - Kobling af oversatte enkeltord til oversatte ordforbindelser
  - o Høj sandsynlighed for korrekt kobling (tættere sammen = højere sandsynlighed)

Sandsynlighederne kommer fra ordhyppigheder, og i sidste ende er den statistiske tilgang til maskinoversættelse således baseret på optælling af ord. Det sproglig-grammatiske arbejde kommer i anden række. Det kan virke fremmedgørende for faglingvister, og tilgangen fik derfor i starten øgenavnet 'antilingvistik'. Selv om selve maskinoversættelsessystemet i princippet er støvsuget for lingvistisk-leksikografisk viden, skal man dog huske på at der stadig er indlejret en enorm sproglig viden i systemets input i form af de store mængder oversat tekst som systemet afhænger af. Dertil kommer at der i praksis også vil ligge avanceret sproglig indsigt til grund for de finjusteringer der foretages af datalingvisterne bag et sådant system.

Men eftersom statistisk maskinoversættelse hovedsageligt afhænger af statistisk ekspertise, rå regnekraft og store samlinger af oversat tekst, giver det næsten sig selv at det er forholdsvis billigt at overføre teknologien på nye sprogpar – så længde der altså findes oversatte tekstsamlinger. Bl.a. på grund af ophavsretsmæssige problemstillinger er dette ikke en selvfølge, især ikke for mindre sprog som dansk (for slet ikke at tale om virkelig små sprog som fx grønlandsk eller færøsk). Jo færre tilgængelige parallelle tekster, des dårligere statistisk maskinoversættelse.

Maskinoversættelse i din hverdag

Efterhånden dukker maskinoversættelse op i sammenhænge hvor vi ikke kan undgå at møde den. Det har i adskillige år været muligt at få en tålelig oversættelse af alt fra afrikaans til zulu på Google Oversæt. Hvis man har udenlandske venner på Facebook, vil man somme tider se linket "Se oversættelse" (dog endnu kun for udvalgte statusopdateringer og udvalgte sprog, for mit eget vedkommende: ikke engelsk og italiensk, men derimod svensk).

Internetgiganter som Google og Facebook er åbenlyst privilegerede idet de hele tiden kan høste sproglige data fra deres milliarder af brugere. Denne overflod af data er som skabt til den statistiske tilgang til sprogteknologi, herunder maskinoversættelse, som derfor nok må siges at have været den dominerende i de senere år. I sprogområder uden store mængder oversat tekst som fx det grønlandske og det samiske arbejder man dog støt og målrettet ud fra den regelbaserede tilgang, ikke bare for at hjælpe med at føre disse sprogområder ind i den digitale tidsalder,

Philip Diderichsen ***(f. 1977) er videnskabelig medarbejder i Dansk Sprognævn.***

men også for at forhindre sprogene i at uddø. Så længe der endnu findes modersmålstalende af disse små sprog, er det muligt at nedskrive deres sproglige intuitioner i form af grammatikker, der så igen kan overføres til regelbaseret sprogteknologi, der fx kan bruges i sprogundervisning.

Helt aktuelt er der kommet fornyet fokus på maskinoversættelse i forbindelse med EU-initiativet det digitale indre marked. Initiativet skal bl.a. gøre det lettere for små og mellemstore virksomheder i EU at drive e-handel. Det var dog i første omgang blevet overset hvor stor en rolle sprogbarrierer spiller i den forbindelse, men det er lykkedes en kreds af europæiske sproginstitutioner at sætte på den europæiske dagsorden hvor stort et potentiale der faktisk er i maskinoversættelse (se http://www.meta-net.eu/projects/cracker). Man har fx peget på at under 5 % af de små og mellemstore virksomheder i EU i øjeblikket sælger varer på andre sprog end deres eget samtidig med at e-handelskunder er seks gange mere tilbøjelige til at købe varer hvis handlen foregår på deres eget sprog. Man ønsker derfor et *flersprogligt* digitalt indre marked hvor virksomhederne hjælpes til at bruge maskinoversættelse m.m. Sprognævnet tager aktivt del i disse bestræbelser. En vigtig opgave er at anspore virk-

somheder og institutioner til at donere deres data i form af oversatte tekster for at forbedre kvaliteten af maskinoversættelse.

Sprognævnets interesse i sprogteknologi begrænser sig altså ikke til retskrivning i forbindelse med stavekontrol og taleteknologi, men omfatter også lidt mere vidtrækkende perspektiver: den gode oversættelse i et internationalt marked.

Litteratur

Bick, Eckhard (2009). *Maskinoversættelse – en sammenligning af to forskellige metoder*. Hentet 17.6.2015 fra http://sprogmuseet.dk/teknologi/maskinovers%C3%A6ttelse-%E2%80%93-en-sammenligning-af-to-forskellige-metoder.

Jurafsky, Daniel og Martin, James H. (2009). *Machine Translation*. Kap. 25 i Speech and Language Processing, 2. udg. London m.fl.: Pearson Education.

Generelle abonnementsvilkår

Levering

Nyt fra Sprognævnet udkommer 4 gange om året, i marts, juni, september og december. Hvis du midt i måneden endnu ikke har modtaget bladet, bedes du henvende dig til Dansk Sprognævn for at få tilsendt et erstatningseksemplar.

Betaling

Abonnementet betales forud, og opkrævning sker i januar, enten ved at Dansk Sprognævn sender dig et indbetalingskort (til betaling på posthuset, i banken eller via netbank) eller via BetalingsService (BS).

Som ny abonnent modtager du de numre der er udkommet i indeværende år, samt et indbetalingskort for et fuldt årsabonnement.

Pris

Prisen er 120 kr. om året. Du betaler for et år ad gangen, så selvom du opsiger dit abonnement midt på året, vil du stadig modtage og betale for resten af årets udgivelser.

Varighed og opsigelse

Abonnementet løber indtil du opsiger det skriftligt, enten per brev eller mail. Seneste frist for opsigelse er 10. december. Det er ikke gyldig opsigelse blot at undlade at betale. Ved opsigelse af abonnementet skal du huske selv at opsige en evt. betalingsaftale med BS.

Flytning

Flytning skal du meddele per telefon, mail eller brev. Ved henvendelser om abonnementet skal du have dit abonnementsnummer klar.

Dansk Sprognævn
Worsaaesvej 19, 4. sal · DK-1972 Frederiksberg C

POST B
PP DANMARK
47372

Nyt fra Sprognævnet

2015/3 · september

Næste nummer udkommer i december 2015

Ansvarshavende redaktør: Sabine Kirchmeier-Andersen

Redaktionssekretær: Jørgen Nørby Jensen

ISSN 0550-7332

Layout: Falcon Grafisk Design · Dtp: Jannerup A/S

Nyt fra Sprognævnet udgives af Dansk Sprognævn. Det udkommer 4 gange om året og koster 120 kr. (inkl. moms og forsendelse) for en årgang. Man kan kun tegne abonnement hos Sprognævnet.

Usignerede artikler og artikler med initialer giver udtryk for Sprognævnets mening. Artikler med navn giver ikke nødvendigvis i enhver henseende udtryk for Sprognævnets mening.

Abonnement mv.:
Telefon: 33 74 74 11 eller
mardal@dsn.dk

Spørgetelefon: 33 74 74 74
(Mandag-torsdag 10-12 og 13-15)

www.dsn.dk
www.sproget.dk