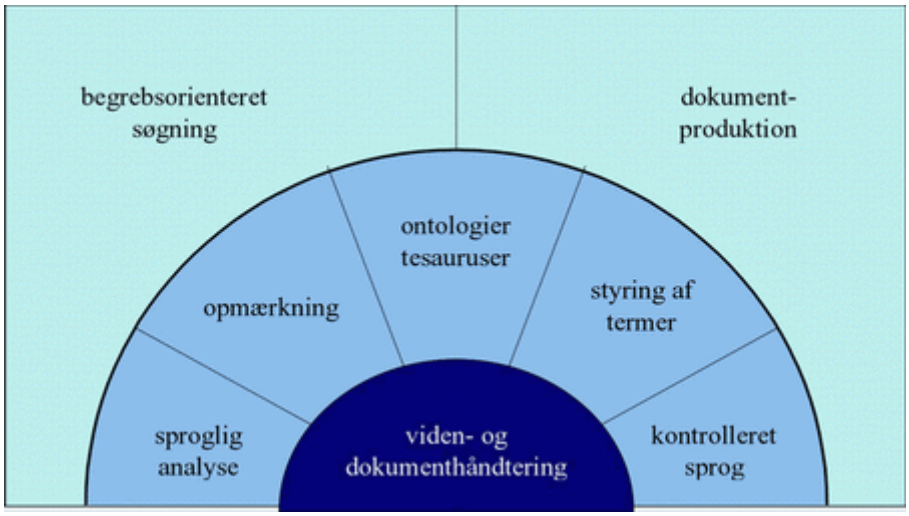

Kontrolleret sprog

Indledende analyse af virksomhedernes regelsæt og sammenligning med eksisterende regelsæt

Lina Henriksen, Bart Jongejan, Bente Maegaard

VID-rapport nr. 1

Center for Sprogteknologi
September 2003

Rapporten kan fås ved henvendelse til CST, cst@cst.dk, eller hentes fra CST's hjemmeside www.cst.dk.

VID-projektet er støttet af Center for IT-forskning (nu overgået til Forskningsstyrelsen).

VID: Viden- og Dokumenthåndtering med sprogteknologi

Der er et udtalt behov hos danske virksomheder for at kunne supplere deres eksisterende sproglige kompetence og viden med sprogteknologiske IT-værktøjer og metoder som dels kan støtte medarbejderne, dels forankre viden og processer i virksomhedens IT-systemer, dels danne grundlag for den udvikling der kræves hvis virksomhederne skal overleve og vokse i den stadigt mere globaliserede økonomi.

VID-projektet er et forsknings- og udviklingsprojekt der har til formål at udforske de forskellige muligheder som sprogteknologi frembyder inden for informationssøgning og dokumentproduktion, og at understøtte de deltagende virksomheder i at udvikle værktøjer til bedre udnyttelse af egen viden, samt til bedre og mere effektiv produktion af dokumentation, herunder flersproget dokumentation. Foruden CST omfatter projektet på den ene side virksomhederne Bang & Olufsen A/S, Zacco A/S og Nordea A/S, som i dette projekt udgør teknologiens brugere, på den anden Navigo Systems A/S og Ankiro, som er teknologiproducenter. Projektet omfatter følgende forskningsopgaver:

- analyse af de tekstuelle data virksomhederne skal kunne håndtere for at kunne fastlægge tesauruser/ontologier for de relevante semantiske domæner, undersøgelse af den bedst egnede formalisme/teknologi til at udtrykke disse;
- afdækning og videreudvikling af sprogteknologiske komponenter til brug for automatisk tekstklassifikation og begrebsorienteret informationssøgning, indbefattende tilpasning af sprogteknologiske 'basismoduler' til opmærkning af tekst;
- udforskning af flertydighed i tekstuelle data som kan vanskeliggøre informationssøgning; ligeledes den omvendte problematik: at samme indhold kan udformes forskelligt rent sprogligt og derfor kan være svær at fremfinde i store datamængder;
- forskning inden for kontrolleret sprog - også set i et flersproget perspektiv - til brug for dokumentproduktion; herunder analyse af den sprogstil og tone som virksomhederne ønsker at anvende, samt opstilling af modeller for dette sprog;
- undersøgelse af hvilke sprogteknologiske metoder der kan anvendes til denne kvalitetssikring af dokumentproduktionen i form af f.eks. termstyring og grammatikkontrol.

Projektet er støttet af Center for IT-forskning og løber i perioden 2003-2004.

Indhold

1 Indledning

Denne rapport er den første i VID-projektets afdækning af kontrolleret sprog for engelsk. Det er på flere måder en delrapport. For det første på den måde at den kun beskæftiger sig med sproglige regler: den vil blive fulgt op af en rapport der omhandler 'tone-of-voice'. For det andet kan man sige at den er delvis på den måde at den ikke har alle de detaljerede regler med som man ville ønske at få med i et implementeret system. Det er dog vores opfattelse at alle væsentlige regler og alle væsentlige regeltyper er beskrevet.

Kapitel 4 i rapporten bygger på materiale der er udarbejdet af medarbejdere i Bang & Olufsen Audio Visual (Anita Thulsted Vestergaard, Anette Johansen, Sussi Tønning, Tina Just Madsen, Christina Lykbak) og Nordea (Sys Bundgaard, Lene Krogshede, Vibeke Rønne, Gitte McKay, Majbritt Duus Asmussen).

Rapporten giver en kort opsummering af de aspekter af fagområdet kontrolleret sprog for engelsk som er relevante i VID, og analyserer herefter det input som Bang & Olufsen Audio Visual og Nordea har bidraget med. Analysen munder ud i nogle første betragtninger over funktionalitet for et støtteværktøj til kontrolleret sprog for de typer tekster som de to firmaer arbejder med.

De karakteristika der kendetegner godt sprog er i hvert fald i nogen grad forskellige fra virksomhed til virksomhed. Det kan afhænge af hvilken kulturel sfære virksomheden tilhører, hvilke relationer virksomheden evt. ønsker at opbygge til kunden samt hvilke værdier virksomheden ønsker at teksterne skal afspejle. Derfor er et første skridt på vejen til en definition af godt sprog, udvikling af en overordnet virksomhedsspecifik sprogpolitik. Dette arbejde er de to virksomheder allerede dybt involverede i og på mange punkter kan man se paralleller. Fx har begge virksomheder et ønske om at muge ud i de konservative og højtidelige udtryksformer. Tekst skal være klar og direkte og på alle måder præget af tilgængelighed. På den anden side er Nordea og Bang & Olufsen to vidt forskellige virksomheder som tilbyder vidt forskellige produkter, og de værdier som virksomhederne ønsker at afspejle, er formentlig også på mange måder forskellige.

Der er således grund til at tro at visse regler for godt sprog vil være fælles for de to virksomheder vi beskæftiger os med, mens andre vil være individuelle. Ydermere er der forskellige typer tekst inden for en virksomhed, fx gælder der typisk forskellige regler for marketingmateriale og tekniske brugsanvisninger.

Derfor skal de metoder og regelsæt der udvikles, være så generelle at de kan anvendes for flere typer af tekst, men samtidig skal de kunne gøres specifikke så de passer til et enkelt firmas behov og til en enkelt teksttype. Det betyder at der sandsynligvis vil blive tale om at vælge regler til/fra, at sætte parametre mv.

2 Baggrund

2.1 Hvad er kontrolleret sprog?

Kontrolleret sprog anvendes i dag af mange virksomheder verden over, især til produktion af dokumentation og brugervejledninger. Formålet med kontrolleret sprog er at producere tekster af en højere kvalitet hvor terminologien er konsekvent og korrekt, hvor skrivestilen er ensartet og dermed tekster som er klare og entydige.

Et kontrolleret sprog er en delmængde af sproget (sub language) hvor mulighederne for anvendelse af terminologi, syntaks og evt. semantik er begrænsede. Man kan sige at et kontrolleret sprog på mange måder er analogt med skrivevejledninger. Det betyder ikke nødvendigvis at forfatterne tvinges til at anvende et enklere sprog i teksterne, men snarere at virksomhederne ønsker at bruge sproget bevidst til at skabe værdi for læseren og dermed i sidste ende også for virksomheden selv.

2.2 CL-værktøjer til forskellige formål

CL-værktøjer (**C**ontrolled **L**anguage) deles traditionelt op i to kategorier. Værktøjer der er udviklet primært med henblik på at øge læsevenligheden kaldes HOCL (Human-Oriented Controlled Language), og værktøjer der er udviklet til at forbedre resultatet af maskinoversættelse kaldes MOCL (Machine-Oriented Controlled Language). De fleste eksisterende programmer er af typen MOCL.

De programmer der findes af typen MOCL, har generelt regelsæt som i hvert fald i nogen grad adskiller sig fra regelsæt i programmer af typen HOCL. I MOCL-programmer lægger man især vægt på de syntaktiske regler, mens der i HOCL-programmer ofte lægges mere vægt på regler som vedrører informationsstruktur og informationsmængde. Et andet interessant fænomen man kan notere sig, er at forskellige CL-systemer sjældent har identiske eller bare lignende regelsæt. Sammenligner man 5-10 CL-værktøjer, vil kun ganske få regler være identiske. Der er flere årsager hertil, men hovedårsagen er i hvert fald at godt sprog er et fænomen som konstant er i forandring, og som forskellige virksomheder vil have forskellige bud på.

2.3 Metoder til udvikling af CL-værktøjer

Udviklere af eksisterende CL-værktøjer har generelt anvendt to forskellige indfaldsvinkler til udvikling af regelsæt; den negative eller den positive fremgangsmåde. Ved den negative fremgangsmåde udvikles et sæt regler som omhandler de sproglige fænomener man gerne vil undgå. Ved den positive fremgangsmåde udvikles en komplet grammatik som dækker alle de grammatiske konstruktioner som man ønsker skal være mulige.

Der er fordele og ulemper ved begge fremgangsmåder. Ved den positive fremgangsmåde kan man efterhånden udvikle et temmelig sikkert system som vil kunne reagere på rigtig mange typer fejl (selvom intet system er 100% sikkert). Men det er selvfølgelig overordentlig krævende at udvikle en fuldstændig grammatik over alle

tilladte sproglige konstruktioner, især hvis virksomheden i virkeligheden ikke ønsker at sproget skal være særlig restriktivt. Endvidere vil et værktøj af denne type ikke umiddelbart kunne diagnosticere en bestemt fejl og vil altså heller ikke kunne hjælpe brugeren med at finde ud af præcis hvor i sætningen fejlen befinder sig. Det vil kræve at der også tilføjes et negativt regelsæt.

Omvendt er fordelen ved et negativt regelsæt at en bestemt fejl altid kan diagnosticeres, og systemet vil ikke blive så restriktivt. På den anden side overstiger brugeres opfindsomhed altid udviklernes, og det er umuligt at lave regler for alle typer fejl. Et system der er opbygget omkring et negativt regelsæt, vil aldrig kunne blive så sikkert som et system udviklet med et positivt regelsæt.

Et CL-værktøj skal ud over de grammatiske konstruktioner også kontrollere de leksikalske fænomener. Ordforrådet er selvfølgelig et meget vigtigt element i godt sprog. For de fleste eksisterende CL-systemer er der tilsyneladende valgt den positive fremgangsmåde. Dvs. værktøjet indeholder et leksikon over alle virksomhedens tilladte ord, og ethvert ord som ikke findes på denne liste, er derved et ord som virksomheden ikke ønsker at anvende.

En mindre restriktiv metode er selvfølgelig at anvende den negative fremgangsmåde og oprette lister over alle uønskede ord. Men man kan selvfølgelig ikke opregne ALLE uønskede ord, og der kan kun blive tale om lister over uønskede ord som virksomheden af erfaring ved at brugere anvender.

2.4 AECMA SE

AECMA Simplified English (SE) er en guide med et regelsæt af typen HOCL som har været på markedet siden midten af 80'erne (dog med adskillige opdateringer frem til i dag). Guiden blev udviklet til European Association of Aerospace Industries (AECMA, Association Européenne des Constructeurs de Matériel Aérospatiale) for at øge kvaliteten i flyindustriens dokumentation. Denne guide har dannet grundlag for mange af de værktøjer der findes på markedet, heriblandt The Boeing Simplified English Checker. Men virksomheder inden for andre industrier har også tilpasset AECMA regelsættet til deres egne behov. Det skal dog understreges at intet eksisterende CL-værktøj har implementeret hele AECMAs regelsæt, bl.a. fordi det teknologisk er umuligt eller i hvert fald meget svært.

AECMA indeholder et begrænset leksikon og et sæt af regler som skal øge læsevenligheden i tekniske tekster. Leksikonet består af en liste over tilladte almensproglige ord (ca. 1000 ord) som udvides med virksomhedsspecifikke termer, og en liste over uønskede ord med forslag til hvilke ord der i stedet kan anvendes. Ifølge AECMA må hvert ord kun have én betydning og må kun tilhøre én ordklasse. For alle ord der i det almindelige sprog kan tilhøre flere ordklasser og/eller have flere betydninger, defineres en bestemt betydning og ordklasse.

I kapitel 4 hvor virksomhedernes regler beskrives og diskuteres, nævner vi om den pågældende regel også findes i AECMA regelsættet og evt. også i hvilken udstrækning. Vi har valgt AECMA regelsættet som sammenligningsgrundlag af flere forskellige årsager. Regelsættet er anerkendt for at være af typen HOCL, det er formentlig et af de

mest omfattende der er udviklet, og det har som nævnt dannet grundlag for mange af de eksisterende systemer. Derudover er eksisterende systemer i reglen ikke færdige hyldevarer som man nemt kan sammenligne, men systemer som skal tilpasses kundens behov.

3 Metode

3.1 Indledning

For at kunne skabe sig et indtryk af hvilke sproglige fænomener i virksomhedernes tekster der er relevante for teksternes læsekvalitet, er det nødvendigt at have et repræsentativt tekstkorpus til rådighed for hver af virksomhederne. Ved hjælp af et korpus og et passende søgeprogram kan man hurtigt og nøjagtigt fastslå hvordan en given term bruges i virksomhedens eksisterende dokumenter, fx hvor ofte termen bruges, og om den bruges i mere end én betydning. Desuden kan man udpege steder i disse tekstkorpusser hvor læsekvaliteten muligvis er blevet forringet, men det kræver et enormt manuelt arbejde eller et specielt analyseværktøj. Det sidste har vi ikke endnu, men CST har i et tidligere projekt (TQPro) udviklet et tekstanalyseværktøj som kan give oplysninger om hvilke sproglige fænomener det er stødt på som kan give problemer ved automatisk oversættelse fra engelsk til et andet sprog. Det var oplagt at bruge dette analyseværktøj. Selvom en læsevenlig tekst ikke nødvendigvis nemt kan oversættes med et maskinoversættelsessystem (MT), og en tekst som egner sig godt til maskinel oversættelse ikke nødvendigvis er læsbar, er der mange sproglige fænomener som har den samme virkning på læsekvalitet og kvaliteten af en maskinel oversættelse.

Virksomhederne – B&O, Nordea og Zacco – forsynede CST med et stort antal engelsksprogede tekster som repræsentanter for alle de typer af tekster som virksomhederne ønsker at underlægge regler for kontrolleret sprog. Disse tekster blev konverteret til 'flade' tekster (dokumenter kun indeholdende bogstaver, mellemrum og læsetegn, blottet for afsnits- og tekstformatering og billeder). De flade tekster blev behandlet af et program som bestemmer ordklassen for hvert ord. Tekst og ordklassebestemmelserne blev derefter lagt sammen i tekstkorpusser som er tilgængelige for diverse søgeprogrammer. Ydermere brugte vi førnævnte analyseværktøj til at lave rå analyser af teksterne. Analyseresultaterne blev lagt ud på nettet (med passwordbeskyttelse).

3.2 Indsamlede tekster

Virksomhed	Beskrivelse	Tekst-format	Antal tekster	Antal ord
B&O	Marketingmateriale, træningsvejledninger, brugervejledninger, PR-materiale og internetsider for fem produkter: Beocom 4 (telefon) BeoLink PC2 (pc-software + hardware) BeoSound 3200 (musiksystem)	PDF	25	90.218

	BeoSound 9000 (musikunderholdningscenter) BeoVision 5 (fjernsyn)			
Nordea	Fire grupper af filer: 1) customers 2) internal 3) investor relations 4) public at large	PDF, DOC, HTML, PPT	92	239.576
Zacco[1]	Patentrelaterede dokumentskabeloner.	DOC	202	47.341
(Idem)	Varemærkerelaterede dokumentskabeloner	DOC	74	10.911

3.3 Konvertering til flad tekst

For at kunne bestemme et ords ordklasse, men også for at tjekke at en tekst overholder standarder for kontrolleret sprog, er det nødvendigt at arbejde på både ord- og sætningsniveau. For et computerprogram er det ikke altid ligetil at bestemme hvor et ord starter og slutter. Problemet er endnu større ved bestemmelsen af start og slut på en sætning. Normalt kan et program nøjes med at søge efter blanktegn og punktummer, men det er ikke altid nok. Fx afslutter et punktum normalt en sætning, men det kan også være en del af en forkortelse. Og nogle 'sætninger' slutter ikke med et punktum, fx overskrifter. Ved konvertering fra et dokument med lay-out oplysninger til et dokument uden disse er det derfor vigtigt at udnytte lay-out oplysninger til bestemmelse af ord- og sætningsgrænserne, før lay-out oplysningerne bliver smidt ud. Fx kan overskrifter normalt kendes ved at de er sat i en anden skrifttype eller skriftstørrelse end brødteksten. Særdeles svære er 'bullets'. Disse er symboler hvorom det kan siges at de står forrest på en linje og gentages mindst én gang, men udformningen kan variere fra specielle tegn til kombinationer af tegn som også bruges andre steder, fx en bindestreg efterfulgt af et blanktegn. Og det står hen i det uvisse om 'bullets' skal bibeholdes i den flade tekst, og hvis ja, om de skal repræsenteres på en ensartet måde.

Det viste sig at de standardapplikationer som er i stand til at eksportere PDF, DOC, HTML og PPT format til flad tekst (TXT), ikke selv var i stand til at opdele teksten i sætninger. Ej heller var det muligt, på grund af tabt lay-out information, at opdele teksterne i sætninger ved en efterbehandling. Derfor udviklede vi et program som kan konvertere RTF-tekster til flad tekst - med én sætning per linje. RTF er et format som de anvendte dokumentformater kan konverteres til med bibeholdelse af al relevant lay-out.

[1] Zacco har leveret tekster i to omgange. Den sidste pulje, bestående af 75 Word skabeloner, er endnu ikke behandlet. Tallene i tabellen refererer til teksterne i den første pulje.

Samtidig er RTF en åben standard, om end med en komplicerende udviklingshistorie på bagen. Programmet løser problemet med punktummer som afslutter en forkortelse ved at slå kandidat-forkortelser op i en liste af kendte 'forkortelser-med-punktum'. Programmet er også i stand til at spotte flerordsenheder ved hjælp af opslag i en ekstern liste.

3.4 Ordklassebestemmelse

Vi brugte et offentligt tilgængeligt værktøj til at bestemme ordenes ordklasse, Eric Brills POS-tagger (POS = ordklasse) (http://www.cs.jhu.edu/~brill/RBT1_14.tar.Z). Vi har ændret programmet for bedre at håndtere overskrifter og ord som starter med stort bogstav. POS-taggeren tager flad tekst som input, med én sætning per linje og alle ord og tegn adskilt med blanktegn.

Taggerens engelske lingvistiske resurser stammer ligeledes fra Eric Brill, men er tilpasset i TQPro-projektet. Til tagging af danske tekster (Zaccos) bruger vi også Brill Taggeren, men med CST's egne danske lingvistiske resurser.

3.5 Opbygning af korpusser

De flade, POS-taggede tekster blev kopieret efter hinanden i store filer, én fil per virksomhed (Zacco har to). Disse store filer blev lagt ind i et søgesystem, XKWIC. Programmet gør det muligt at søge på ord, men også avancerede søgestrenge, indeholdende (dele af) ord, jokertegn og ordklasser, er mulige. Resultatet af søgningen vises som konkordans, hvilket vil sige at man får en liste af alle tekstdele hvori søgemønstret forekommer, med søgemønstret i midten af resultatvinduet, omgivet af konteksten af ordet. Men kan vælge at få vist hvilken ordklasse ordene tilhører og fra hvilke dokumenter linjerne stammer.

Via internettet – med brug af passwords – kan man søge i virksomhedernes korpusser med en forenklet udgave af XKWIC (http://cst.dk/cgi-bin/defisto/).

3.6 Rå analyser

Vi har brugt et værktøj fra TQPro-projektet (Translation Quality for Professionals), til at producere nogle analyser af virksomhedernes tekster. Programmet tager engelsk-sproget POS-tagget tekst som input og laver to typer analyser:

- analyser som tager udgangspunkt i tekststruktur/pragmatiske regler (konkret: ordtællinger)
- analyser som tager udgangspunkt i syntaktiske regler (med part-of-speech tags som værktøj)
- analyser som tager udgangspunkt i leksikalske regler (opslag i ordlister).

Programmet er udviklet med et andet formål end implementering af kontrolleret sprog, nemlig at bedømme hvorvidt et oversættelsesprogram ville kunne lette oversætterens arbejde ved oversættelsen af et givet dokument. Sådanne oversættelsesprogrammer har oversættelsesvanskeligheder i sætninger med visse sproglige fænomener. Ikke helt tilfældigt er mange af disse sproglige fænomener også skyld i at mennesker har svært

ved at forstå sådanne sætninger. Som eksempel kan nævnes at lange sætninger generelt er sværere at oversætte end korte, og at sekvenser af flere end tre substantiver kan give oversættelsesproblemer. Disse fænomener opfatter man generelt også som hæmmende for forståelsen af en tekst.

Her er en liste af de fænomener som analyseprogrammet kan finde i en engelsk tekst:

Fænomener som besværliggør maskinoversættelse, fundet ved analyse af tekststruktur:

- Sætningen er meget kort (færre end seks ord).
- Sætningen er meget lang (flere end 25 ord).

Fænomener som besværliggør automatisk oversættelse, fundet ved syntaktisk analyse:

- Sætningen mangler et verbum.
- Sætningen mangler et finit verbum.
- Sætningen indeholder en eller flere følger af tre eller flere substantiver.
- Sætningen indeholder to eller flere konjunktioner.
- Sætningen indeholder præpositioner eller underordnende konjunktioner.
- Sætningen indeholder flere end tre interpunktionstegn (,;:-).

Fænomener som besværliggør maskinoversættelse, fundet ved leksikalsk analyse:

- Sætningen indeholder et eller flere ord som ifølge ordbogen kan være såvel adjektiv, som substantiv og som verbum.
- Sætningen indeholder et eller flere ord som ifølge ordbogen kan være både adjektiv og verbum (fx 'long').
- Sætningen indeholder et eller flere ord som ifølge ordbogen kan være både substantiv og verbum. (fx 'tap').

4 Regler for kontrolleret sprog

I dette kapitel opstiller vi de forskellige typer af regler der foreløbig er blevet uddraget af materialet. Vi diskuterer reglerne og muligheden for at automatisere dem. I gennemgangen af reglerne forholder vi os som nævnt også til om reglen findes i AECMA regelsættet.

Gennemgangen af virksomhedernes regler er i det nedenstående ikke helt udtømmende fordi materialet er meget stort (vi har bl.a. ikke medtaget Nordeas 'How to write letters in English'). Det betyder ikke at reglerne er glemt, eller at vi har udelukket at kunne håndtere dem.

4.1 Regler og anbefalinger

Der er dels tale om obligatoriske regler, dels om regler der i udgangspunktet mere har karakter af anbefaling. Eksempler på obligatoriske regler er 'Anvend britisk engelsk' og 'Højst to præpositionsforbindelser lige efter hinanden'. Sådanne regler vil enten være opfyldt eller ikke opfyldt, og det er muligt at pege præcist på det sted hvor de ikke er opfyldt. Hvis man altså skal udvikle et system der skal teste om reglerne er opfyldt, kan dette system pege direkte på det sted hvor der skal rettes noget, svarende til den måde et stavekontrolprogram virker på.

Regler kan endvidere være obligatoriske eller være fakultative. Obligatoriske regler kan aldrig fraviges, mens fakultative regler kan fraviges hvis der er grunde hertil.

Eksempler på anbefalinger er 'Brugen af passiv begrænses' og 'Brugen af superlativer begrænses'. Her er der to problemer ved en automatisering. Dels kan man undersøge om det er muligt at definere hvad 'begrænses' betyder i hver enkelt sammenhæng. Man kan fx antage at det er muligt at definere hvad der er et acceptabelt niveau for passiv. Det kan gøres fx ved at definere at højst 15% af verballeddene må være i passiv. Anbefalingen kan herefter omformuleres til en regel: 'Højst 15% af verballeddene må være passiv'. Men det andet problem er at det, selv med en præcis definition af grænseværdien, ikke er muligt at pege præcist på det sted hvor reglen ikke overholdes. Brugeren må selv vælge hvilke passivformer der bedst kan omskrives til aktiv. Det betyder at et programs rapportering om denne type regler må være fx af denne type: 'Der er 55 forekomster af passiv i denne tekst, det svarer til 20%. Skriv mindst 14 af passiverne om til aktiv'.

Det andet eksempel på anbefalinger ovenfor, nemlig 'Brugen af superlativer begrænses', er måske af en type der ikke kan specificeres til en regel. Det giver formentlig ikke mening at sætte en grænseværdi for antallet af superlativer. For dette fænomen kan man altså udelukkende få en rapport der angiver antallet af superlativer, og eventuelt også superlativerne som % af alle ord, eller som % af adjektiver.

Vi har altså set at der dels må skelnes mellem regler og anbefalinger. For en regel gælder det at det er muligt at afgøre om den er opfyldt eller ej. Anbefalinger har ingen faste grænser, og der er derfor ikke et ja/nej svar på om de er opfyldt. Nedenfor angives

i hvert enkelt tilfælde om der er tale om en regel eller en anbefaling, og om vi ser muligheder for at omforme en anbefaling således at den bliver en regel. Vi mener at det er hensigtsmæssigt at forsøge at udforme så meget som muligt som regler, da det giver bedre muligheder for at vejlede brugeren.

Vi har også set at det for funktionaliteten gælder at manglende opfyldelse af visse typer af regler kan markeres på det sted i teksten hvor reglen brydes, mens der for andre typer af regler er tale om mere statistiske forhold, som savner opfyldelse for teksten som helhed, og at man derfor kun kan rapportere om disse forhold i en separat rapport.

<table>
<tr><td rowspan="2">Regler (obligatoriske eller fakultative)</td><td>Markering i teksten</td></tr>
<tr><td rowspan="2">Rapportering separat fra teksten</td></tr>
<tr><td>Anbefalinger</td></tr>
</table>

En anden måde at klassificere reglerne på er at inddele dem efter deres indhold, dvs. efter lingvistiske og andre tekstnære kriterier. Typisk skelnes mellem leksikalske regler, syntaktiske regler og tekststruktur-/pragmatiske regler. Et første forslag hertil, baseret på det indkomne materiale, ses nedenfor.

Problemområderne er således grupperet i fire grupper, nemlig Leksikalske regler, Enkle syntaksregler, Syntaktiske regler og Opsætning og tekstniveau. For hver gruppe af regler er der en kort indledning, derefter en beskrivelse af reglerne/anbefalingerne, sammen med en diskussion af om man kan forestille sig at implementere reglerne i et automatisk system og i givet fald hvordan.

Som man vil se, er der en del overvejelser om hvordan brugen af positivlister og negativlister evt. vil kunne løse det enkelte problem. Det skal dog bemærkes at det værktøj som måtte blive valgt, vil have indflydelse på hvad der er mest praktisk i en implementering.

Koder:
R betyder regel
A betyder anbefaling
AECMA skrives efter kommentaren, og det specificeres hvorvidt den pågældende regel findes i AECMA regelsættet.

4.2 Ordforråd

Et af de fænomener som alle skrivevejledninger indeholder, er overvejelser om ordforrådet, dvs. hvilke ord man må bruge, særlige stavekonventioner som skal overholdes mv.

Som en første hypotese er vi gået ud fra at det ligger uden for rammerne af dette projekt at opstille en fuldstændig liste over alle de ord som må anvendes i den enkelte virksomhed.

Dette valg er foretaget fordi der allerede findes store generelle ordbøger til kontrol af stavning og ordforråd, og fordi vi mener at der er andre aspekter af kontrolleret sprog som det er mere vigtigt at udforske i dette projekt.

Til håndtering af ordforrådet i dette projekt kan nævnes flere løsninger som evt. kan anvendes i kombination. I nogle tekstbehandlingssystemer er der mulighed for at anvende en generel stavekontrol, og de ord der accepteres af denne, er principielt tilladte, med mindre andet er defineret. Mængden af tilladte ord kan i disse systemer i reglen udvides ved oprettelse af en brugerordbog. Mængden af tilladte ord vil vi i dette projekt indskrænke ved at definere en negativliste over forbudte ord.

4.3 Leksikalske regler

Ved leksikalske regler forstår vi regler der regulerer det ordmateriale der kan bruges og dets stavning.

4.3.1 Britisk engelsk

R B&O: *Anvend britisk engelsk. Amerikansk engelsk bruges til en enkelt type opgave.*
R Nordea: *Anvend britisk engelsk.*

Kommentar: Denne regel testes ved stavekontrollen, intet særligt mht. kontrolleret sprog.

AECMA: Denne regel kan håndteres idet kun ord fra leksikonet må anvendes.

4.3.2 Bindestreger og sammensatte ord

A Nordea: *i Nordea Style Guide findes nogle få regler for hvordan man bruger bindestreger, eksempler på tilladt brug af bindestreger, eksempler på ikke-tilladt brug af bindestreger samt lister med eksempler på termer der skal i et eller to ord.*

Kommentar: Hvad angår bindestreger kan anvendes en kombination af positivliste og negativliste. Ved en automatisering må listerne gøres så udtømmende som muligt, da systemet ikke kan generalisere ud fra eksempler, og det skal specificeres hvad der skal ske med ord som ikke står på nogen af listerne.

Man kan også vælge kun at anvende en positivliste. Denne løsning, hvor man afviser alle de bindestreger der ikke er godkendte, er den sikreste og bør vælges hvis det er vigtigt at undgå forkert placerede bindestreger.

Alternativt kan man anvende en negativliste, dvs. en liste med de mest typiske fejl. Denne metode til håndtering af forkert placerede bindestreger vil selvfølgelig betyde at kun kendte fejl vil blive fundet.

AECMA: Denne regel håndteres i nogen udstrækning idet ord med bindestreg kun accepteres, hvis de findes på listen over tilladte ord. Termer der er skrevet i to ord, men som burde have haft bindestreg, vil ikke nødvendigvis blive identificeret. I øvrigt findes i AECMA regler for hvordan man gør termer mere læsbare ved at bruge bindestreger.

Fx angives det i Nordea Style Guide at man ikke må anvende bindestreg ved en term som 'interest rate risk'. Her ville AECMA regler foreslå 'interest-rate risk'.

4.3.3 Forkortelser

R B&O: *Forkortelser benyttes ikke.*

Kommentar: Hvis denne regel skulle vise sig at være for restriktiv, kan der evt. etableres en positivliste. Forbudte forkortelser, dvs. forkortelser som ikke findes på evt. positivliste, identificeres enten vha. punktum eller ordklasse-tagger. Alternativt kan oprettes negativliste med alle forkortelser der ikke må anvendes.

R Nordea: *Forkortelser forklares første gang de anvendes, hvis de ikke er almindeligt kendte.*

Kommentar: Hos Nordea anvendes forkortelser særdeles ofte, både i interne og eksterne tekster. Nordea ønsker ikke at anvendelsen af forkortelser skal reduceres, forkortelserne skal blot forklares hvis de ikke er kendte. Derfor er opgaven i denne sammenhæng at udlede hvornår en forkortelse er 'almindeligt kendt'. Det vil formentlig i nogen grad være afhængigt af teksttype og af om kommunikationen er intern eller ekstern. Her må i så fald oprettes forskellige positivlister, alternativt negativlister, til forskellige teksttyper og målgrupper alt efter hvad der må forventes at være almindeligt kendt.

R Nordea: *Der anvendes ikke punktum ved forkortelser (eg, ie).*
A Nordea: *Forkortelserne eg og ie bør ikke anvendes så ofte.*
R Nordea: *Forkortelser som m (million) og bn (billion) anvendes, men uden mellemrum, fx 34bn.*
A Nordea: *Forkortelsen for thousand (k) anvendes sjældent.*
R Nordea: *Der skal være mellemrum mellem valuta og beløb, fx EUR 34bn.*
R Nordea: *Punktum anvendes ikke ved initialer i navne, fx Peter J Johnson.*
R Nordea: *I øvrigt findes i Style Guide flere regler om forkortelser, bl.a. liste over tilladte forkortelser af måleenheder.*

Kommentar til regler om forkortelse: Det er tydeligt at der er forskel på de to firmaers behov for og brug af forkortelser. Regler om forkortelser kan altså ikke være fælles.

I forbindelse med reglen om at der ikke anvendes punktum ved initialer i navne må det undersøges om reglen kan generaliseres til at der ikke anvendes punktum efter et enkelt stort bogstav.

AECMA: Regler om forkortelser håndteres i begrænset omfang idet kun forkortelser der findes i leksikonet kan anvendes.

4.3.4 Akronymer

R B&O: *Undgå akronymer. 'Bang & Olufsen' og produktnavne skrives fuldt ud.*
R Nordea: *Akronymer skrives med versaler (hvis akronymet kun er dannet over forbogstaver). Akronymer der kan udtales, skal ikke have bestemt artikel foran. Andre organisationer (ikke firmaer) skal i reglen have 'the' foran (the EU).*

Kommentar: Det er formentlig generelt fornuftigt at etablere en positivliste med akronymer og deres stavemåde. Forbudte akronymer kan nok kun identificeres automatisk hvis de indeholder flere versaler.

AECMA Reglen håndteres idet der kun kan anvendes akronymer som findes i leksikonet.

4.3.5 Tal

R Nordea: *komma og punktum skal angives korrekt.*

Kommentar: Korrekt anvendelse af komma og punktum i forbindelse med tal må nærmere specificeres i forbindelse med automatisering.

R Nordea: *En sætning må ikke starte med et tal. I så fald skal tallet skrives med bogstaver*

Kommentar: Dette gælder muligvis kun for 'rigtige' sætninger. Det skal nærmere specificeres om tal må indlede fx en dato eller andet som ikke er en rigtig sætning. Ved automatisering må der i så fald skelnes mellem sætninger og ikke-sætninger.

R Nordea: *Fra 1-10 skrives tal med bogstaver. Undtagelser: Tal anvendes altid ved decimaler, procent eller anden måleenhed, fx 2%, 5 kg, i liste hvis et af tallene er højere end 10, fx 14, 9 og 6.*

Kommentar: Reglen kan automatiseres. Undtagelsen kan automatiseres i et vist omfang.

Der findes flere regler om tal i Nordea Style Guide.

AECMA: Har ikke disse detaljerede regler om tal.

4.3.6 Pronomen

R B&O: *Anvend ikke possessivt pronomen sidst i en sætning, fx 'yours'.*
R B&O: *Undgå pronomenet 'one'.*

Kommentar: Den første regel kan automatiseres ved at etablere en liste med ord der ikke må forekomme sidst i en sætning.

Den anden regel kan også automatiseres med ret stor sikkerhed, idet det burde være muligt at skelne pronominet 'one' fra talordet 'one' i de fleste kontekster (talordet vil ofte være fulgt af et substantiv).

AECMA: Har lignende regel vedr. pronomenet 'one'.

R Nordea: *Anvend I/du/De i stedet for 'the customer' og vi i stedet for 'the bank'.*

Kommentar: Denne regel handler om at skrive direkte til modtageren, og om at mindske afstanden mellem afsender og modtager. Derfor anvendes hellere et pronomen eller evt. 'Nordea' i stedet for 'the bank' og 'you' eller evt. firmanavnet i stedet for 'the customer'.

En mulig måde at håndtere denne regel kunne være at oprette en liste over ord som 'the bank' og 'the customer'. Det vil nok være urealistisk at forbyde brugen af disse ord, men systemet kunne i stedet markere de steder hvor ordene er brugt med en påmindelse til forfatteren om at anvendelsen skal tjekkes.

AECMA: Denne regel findes ikke.

4.3.7 Please

A B&O: *Anvendes sjældent*.

Kommentar: Der vil være behov for at præcisere hvad der menes med sjældent. Alternativt kan man markere alle anvendelser.

AECMA: 'please' er ikke med i leksikonet og kan derfor ikke anvendes.

4.3.8 Modalverber

R Nordea: *'shall' må ikke bruges, i stedet anvendes 'must' eller 'is/are to'.*
'can' anvendes sjældent og refererer til en færdighed.
'must' refererer til et krav og 'may' til en mulighed.

Kommentar: Man kan kontrollere for 'shall'. Det må defineres hvad 'sjældent' betyder i forbindelse med 'can' som i øvrigt er meget almindelig hos B&O, så dette er måske endnu et eksempel på forskelle i teksttyperne.

Det er ikke umiddelbart indlysende hvordan man kan kontrollere den rette anvendelse af 'must' og 'may'.

AECMA: 'shall' findes i AECMA's negativliste med angivelse af at 'must' anvendes i stedet, 'can' må kun anvendes i betydningen 'to be able to', 'must' må kun anvendes i betydningen 'obligation' og 'may' må ikke anvendes.

4.3.9 Sætningsindledere

A B&O: *Sætningsindledere som 'However' og 'Nevertheless' skal helst undgås. Man bør også undgå at indlede sætninger med præpositionsled og adverbielle led*

Kommentar: Den første anbefaling kan kontrolleres automatisk. Anden anbefaling med en vis usikkerhed.

AECMA: 'However' og 'Nevertheless' findes begge i AECMA's negativliste med angivelse af at 'but' anvendes i stedet.

4.3.10 Valuta

R Nordea: *ISO standard anvendes for valuta. Valutaen skrives før selve tallet. I øvrigt også enkelte andre regler om valuta.*

Kommentar: Man kan etablere en positivliste med valutaer og en liste med hyppige fejlskrivninger og den rette version; men en 100% kontrol kan være svær at gennemføre.

AECMA: Reglen håndteres i begrænset omfang idet kun ord fra leksikonet kan benyttes.

4.3.11 Ordspil, slang

R B&O: *Ordspil, slang undgås helt i User's Guide. Er tilladt i marketingmateriale.*

Kommentar: Slangord kan medtages på negativlisten, jf. nedenfor. Ordspil kan næppe fanges automatisk.

AECMA: Denne regel findes.

4.3.12 Afdelingsnavne

R Nordea: *Visse afdelingsnavne skrives og evt. forkortes på en ganske bestemt måde.*

Kommentar: Her må anvendes en negativliste over hyppige fejlskrivninger.

AECMA: Reglen håndteres idet kun ord fra leksikonet kan anvendes.

4.3.13 Jargon

A Nordea: *Anvendelsen af specialistjargon begrænses. Brug hellere et almindeligt ord end et fremmedord.*

Kommentar: Afhængigt af teksttype. Kontrollen kan evt. gøres ved en negativliste, jf. næste punkt.

AECMA: Reglen håndteres i begrænset omfang idet kun ord fra leksikonet kan benyttes.

4.3.14 Negativliste over ord

R Begge virksomheder har generelt brug for en liste over ord der ikke bør anvendes. Her tænkes på ord som ikke falder ind under nogen af de andre grupper som er nævnt i denne rapport.

AECMA: Har negativliste over ord med angivelse af mulige alternativer.

4.4 Enkle syntaksregler

Enkle syntaksregler er regler der omhandler et enkelt ord eller en lille gruppe af ord.

4.4.1 Sammentrækninger

R B&O: *Sammentrækninger som fx' isn't', 'don't' er forbudt i User's Guide. Må gerne anvendes i marketingmateriale*

Kommentar: Negativliste over sammentrækninger kan etableres. For hver teksttype må det defineres om reglen gælder eller ej.

AECMA: Sammentrækninger er ikke tilladt.

4.4.2 NP'er

R B&O: *Max 3 substantiver i User's Guide, max 4 i marketingmateriale*

R Nordea: *Ikke for mange substantiver efter hinanden.*

Kommentar: Som man kan se, er der forskel på kravene i de forskellige teksttyper. Der er behov for en præcisering af udtrykket 'ikke for mange'. Reglen kan automatiseres.

AECMA: Max 3 substantiver.

4.4.3 Verbum i sætningen

R B&O: *En sætning skal indeholde et bøjet verbum.*

AECMA: Denne regel findes ikke direkte, men der findes et antal andre regler om verber som vil betyde at en sætning næsten altid til indeholde et bøjet verbum.

4.4.4 Genitiv

R B&O: *Anvend altid of-konstruktion i forbindelse med produktnavne.*

Kommentar: Denne regel skal formentlig formuleres som et 'forbud' i stedet for et 'påbud', for at lette automatiseringen, dvs. '*genitiv udtrykt ved apostrof-s er ikke tilladt i forbindelse med produktnavne*'. Reglen kræver at man etablerer en positivliste med produktnavne.

AECMA: Denne regel findes ikke.

4.4.5 Præpositionsforbindelser

R B&O: *Højst 2 lige efter hinanden.*

Kommentar: Kan automatiseres.

AECMA: Denne regel findes ikke.

4.4.6 Datoer og tid

R Nordea: *Der findes regler for datoer og tid i Style Guide*

Kommentar: En stor del af reglerne om datoer og tid vil kunne håndteres.

AECMA: Der findes ikke regler for dette.

4.4.7 Tempus

A B&O: *Præsens participium og perfektum participium undgås helst.*
Brug helst de usammensatte former: Nutid, datid, infinitiv og imperativ.

Kommentar: Det er endnu ikke helt klart om dette kan udtrykkes som en regel eller om det vil være nødvendigt med en negativliste.

AECMA: Samme regel findes.

4.4.8 Imperativ

A B&O: *Anvendes gerne.*
A Nordea: *Anvendes ved instrukser.*

Kommentar: B&O giver i sit materiale nogle eksempler på hvornår imperativ skal anvendes. Måske kan det også for B&O koges ned til at imperativer anvendes ved instrukser. Derfor vil der være behov for at systemet kan finde instrukser i teksten. Denne regel vil muligvis skulle gælde for specielle teksttyper med specielle layout-mæssige fænomener (fx punktopstilling).

AECMA: I forbindelse med instrukser skal verber skrives i imperativ. Det nævnes ikke i hvilke forbindelser imperativ ellers skal/kan anvendes.

4.4.9 Verbum i singularis/pluralis

R Nordea: *Ved kollektive substantiver tages stilling til om hovedvægten er på de enkelte individer eller helheden [I style guide under 'Singular or plural'].*
R B&O: Samme regel (især vedr. 'Bang & Olufsen').

Kommentar: denne regel vil være meget vanskelig at automatisere.

AECMA: Lignende regel findes.

4.4.10 -ingformer for verber

A B&O: *Undgå brug af –ingformer for verber.*
A Nordea: *Begræns brugen af –ingformer*

Kommentar: -ingformer er acceptable hvis de tilhører andre ordklasser end verber, fx 'settings', 'interesting', også efter præpositioner 'without asking for permission'. Det må undersøges om der skal etableres en positivliste med kendte substantiver og adjektiver der ender på -ing, eller om det kan klares automatisk.

AECMA: Samme regel findes.

4.4.11 Passiv

A B&O: *Brugen af passiv undgås helst i User's Guide.*
A Nordea: *Brugen af passiv begrænses. 80-90% af finitte verber skal være aktive.*

Kommentar: For at man kan automatisere undersøgelsen af passivbrugen, må der sættes en grænse, fx 85%. Eventuelt kan grænsen afhænge af teksttypen.

AECMA: Samme regel i AECMA. Passiv anvendes kun yderst sjældent og kun til beskrivelse.

4.4.12 Nominaliseringer

A B&O: *Undgås helst.*
A Nordea: *Undgås helst.*

Kommentar: Nominaliseringer svarer meget til passiv, og det er naturligt at de to anbefalinger følges ad. Eksempel: 'While using this function, distribution of other sources is not possible'. Her svarer 'distribution' til 'you cannot distribute'.

Det kan imidlertid være svært at bestemme automatisk hvilke substantiver der er nominaliseringer som kunne være undgået. 'Distribution' i andre sammenhænge kan jo være i orden.

AECMA: Denne regel findes.

4.4.13 Superlativer

A B&O: *Begræns brugen af superlativer.*

Kommentar: Dette kan kontrolleres automatisk. Alle superlativer kan markeres i teksten.

AECMA: Denne regel findes ikke direkte i AECMA, men håndteres fordi kun ordformer der er angivet i leksikonet, kan anvendes

4.4.14 Sætningsindledere

A B&O: *Undgå præpositionsforbindelse og infinitiv som sætningsindleder.*

Kommentar: Det er muligt at kontrollere at det første ord i en sætning ikke er en præposition eller en infinitiv.

AECMA: Denne regel findes ikke.

4.5 Syntaktiske regler

Syntaktiske regler er regler der beskriver hvordan man danner sætninger.

4.5.1 Ordstilling

A Nordea: *Subjekt skal stå tidligt i sætningen. Subjekt og verbum skal stå så tæt på hinanden som muligt.*
A B&O: *Ligefrem sætningskonstruktion skal anvendes hvor det er hensigtsmæssigt.*

Kommentar: Uden en fuldstændig analyse kan det være svært at se automatisk om denne regel er opfyldt. Man kan evt. bestemme det finitte verbum og undersøge om der er et led med passende ordklasse som står i en passende position i forhold til verbet.

AECMA: Lignende regel findes.

4.5.2 Bisætninger

R B&O: *Max 3 efter hinanden*.

Kommentar: Kan automatiseres.

AECMA: Reglen findes ikke direkte.

4.5.3 Indskudte sætninger

R B&O: *Max 2 efter hinanden*.

Kommentar: Kan være svært at bestemme automatisk om der er tale om indskud eller bisætninger. Måske kan man sætte antallet af bisætninger til 2.

AECMA: Reglen findes ikke direkte

4.5.4 Parenteser

R B&O: *Undgå forklaringer/sætninger i parenteser. Begreber i parentes er OK.*

Kommentar: Denne regel kan muligvis håndteres automatisk, i hvert fald delvist. Vi har brug for en bedre forståelse for hvilke størrelser 'begreber' kan være. Det kan her nævnes at Nordea ikke kan bruge denne regel om parenteser eftersom de anbefaler anvendelse af parenteser ved forklaring af forkortelser.

AECMA: Lignende regel om begrænset anvendelse af parenteser findes.

4.6 Opsætning og tekstniveau

4.6.1 Sætningslængde

R B&O: *4 – 25 ord*
R Nordea: S*kriv korte sætninger*.

Kommentar: Sætninger er i denne sammenhæng brødtekst, ikke fx overskrifter. Det må defineres hvad en kort sætning er, hvilke elementer der kan afgrænse en sætning (fx kolon, punktum og tankestreg) og hvilke elementer der tæller med ved ordoptælling (fx også indhold i parentes). Med denne præcisering kan reglen kontrolleres automatisk.

AECMA: Samme regel findes.

4.6.2 Afsnitslængde

R B&O: *Max 13 sætninger i et afsnit.*
R Nordea: *Korte afsnit.*

Kommentar: Det må defineres hvad et kort afsnit er. Måske skal den tilladte afsnitslængde også være afhængig af teksttypen. Med disse præciseringer kan reglen kontrolleres automatisk.

AECMA: Lignende regel findes, men det kan nævnes at man i dette regelsæt højst accepterer 6 sætninger i et afsnit.

4.6.3 Informationsmængde i en sætning

A Nordea: *En sætning skal være klar med et hovedpunkt, måske med et underpunkt, i hver sætning.*
A B&O: *Begræns antallet af ideer i en sætning.*

Kommentar: Denne regel er meget svær at håndtere automatisk.

AECMA: Lignende regel findes.

4.6.4 Overskrifter

A Nordea: *Korte overskrifter*.

Kommentar: Overskrifter kan genkendes ud fra layout/meta-data. Det må defineres hvad en kort overskrift er, fx op til 5 ord eller 10 ord.

AECMA: Denne regel findes ikke.

A Nordea: *Opfordrer til at der i en tekst skal indsættes overskrifter med jævne mellemrum.*

AECMA: Kræver at der er overskrifter eller nøgleord foran alle afsnit.

4.6.5 Humor

R B&O: *Humor undgås helt i User's Guide.*
Kommentar: Vanskeligt at kontrollere automatisk.

AECMA: Reglen findes ikke direkte.

4.6.6 Orddeling

R B&O: *Ingen orddeling.*
R Nordea: *Ingen orddeling.*

Kommentar: Automatisk indsatte orddelinger vil ændre sig når layoutet ændres. Derfor skal automatisk orddeling slås fra hvis denne funktionalitet findes i det anvendte tekstbehandlingssystem. Manuelt indsatte orddelinger vil kunne forhindres automatisk.

AECMA: Denne regel findes ikke.

4.6.7 Kursiv og versaler

A B&O: *Brug af kursiv og versaler frarådes generelt, men anvendes til at fremhæve tekst der skal oversættes i User's Guide, fx menutekster.*

Kommentar: Måske kunne B&O med fordel markere denne tekst på anden vis.

A Nordea: *Brugen af stort begyndelsesbogstav begrænses. Kun det første ord i en overskrift har stort begyndelsesbogstav. I Nordeas style guide findes regler samt liste over hvordan versaler anvendes ved titler etc.*

Kommentar: Nordea vurderer at det vil være en meget stor opgave at oprette en positivliste over ord der skal have stort begyndelsesbogstav, og det er muligvis lige så vanskeligt at udarbejde en negativliste, i hvert fald hvis det skal gøres manuelt.

Da Nordea også vurderer at for hyppig brug af stort begyndelsesbogstav er en af de største fejlkilder, skal der overvejes forskellige løsningsmodeller. Man kunne evt. oprette en positivliste over ord der skal have stort begyndelsesbogstav automatisk på baggrund af allerede godkendte tekster, dvs. lade et program finde alle ord med stort begyndelsesbogstav i eksisterende tekster. Man kunne også overveje om det vil være muligt at videreudvikle en eksisterende navnegenkender, dvs. en komponent der kan genkende navne (både på personer, organisationer mv.), således at den kunne anvendes på Nordeas tekster.

AECMA: Denne regel findes ikke direkte, men håndteres alligevel i begrænset omfang idet fx produktnavne og andre navne vil findes med stort begyndelsesbogstav i leksikonet.

A Nordea: *Kursiv anvendes ved henvisning til publikationer samt Nordeas 'rules, terms and conditions'. Kursiv anvendes også ved latinske/franske ord som ikke er almindeligt anvendt på engelsk.*

Kommentar: Dette kræver at programmet kan genkende kursiv. Det er ikke umuligt; men det er ikke alle systemer der kan genkende denne type layout-information.

AECMA: Denne regel findes ikke.

4.6.8 Tegnsætning

R Nordea: *Akronymer i pluralis har ikke apostrof, fx SMEs.*
Årstal i pluralis har ikke apostrof: 1990s.

Kommentar: Det kan blive svært automatisk at skelne mellem SMEs i pluralis og SME's i genitiv. Man kan lave en generel advarsel hver gang man møder en apostrof i denne kontekst og spørge om det er pluralis.

R Nordea: *Kolon efterfølges ikke af stort begyndelsesbogstav (gælder dog ikke overskrifter).*

Kommentar: Kan kontrolleres automatisk.

A Nordea: *Normalt ikke komma foran 'and'. Der er flere regler vedr. komma i Style Guide.*

Kommentar: Komma foran 'and' kan kontrolleres. Det må overvejes om der skal flere kommaregler med.

AECMA: Disse regler findes ikke. AECMA har kun få regler om tegnsætning.

4.6.9 Lister

A Nordea: *har beskrevet forskellige måder at lave lister i Style Guide*

Kommentar: Der er foreløbigt ikke taget stilling til dette.

AECMA: har følgende regler: Hvert punkt starter med stort begyndelsesbogstav, kun det sidste punkt slutter med punktum, hvis sekvensen er vigtig nummereres punkterne. Punkterne skal endvidere hænge grammatisk sammen – også med listens indledning.

4.7 Regler i CL, som ikke er nævnt ovenfor

Ovenfor har vi markeret hvilke af de opregnede regler der kendes fra AECMA. I dette afsnit vil vi kort gennemgå væsentlige typer af regler der *ikke* er nævnt ovenfor, men som findes i nogle CL-værktøjer (de fleste nedenstående regler findes bl.a. i AECMA regelsættet).

Et emne man kunne nævne, er synonymi-polysemi. Ofte frarådes synonymi, altså at man anvender forskellige ord for samme begreb, idet man kan hævde at det forvirrer at en ting skifter betegnelse. B&O mener at synonymer kan være nødvendige, men samtidig at de skal kontrolleres, og at kun de godkendte navne på begreber må anvendes (Words & Technologies).

Polysemi refererer til at et ord har flere betydninger, fx det engelske 'tap' der både kan være et substantiv og et verbum, og herunder bl.a. kan betyde 'vandhane' og 'tappe' og 'aflytte'. Det kan virke forvirrende at et ord kan have flere betydninger, og derfor anbefales det ofte enten at undgå sådanne ord, eller at definere den brug der gøres af dem i den specifikke kontekst. I en tekst om sikkerhedstjenester ville 'tap' således kun kunne bruges om aflytning, og i en tekst om fadølsanlæg ville det kun kunne bruges om aftapning. AECMA SE undgår polysemer på den måde at de definerer en bestemt betydning og udelader de andre.

En anden regeltype er at man kan kræve at relativpronominer ikke udelades. I henhold til denne regel hedder det altså 'the system which you have bought', ikke 'the system you have bought'. Dette foreslås fordi det skulle give klarere tekst, - men måske også tungere.

Endvidere vil man ofte undgå ellipser, dvs. konstruktioner hvor man undlader et ord eller et led der kan underforstås. Denne regel er en generalisering af reglen om relativpronominer. Eksempler er 'provide an extensive and enveloping viewing experience'. Her skulle 'an' gentages. Tilsvarende skulle 'min onkel og tante' skrives 'min onkel og min tante'.

Man kan også nævne negation: hvilke ord må anvendes til at negere med. Undgå dobbelt negation. Fx 'Jeg skal ikke undlade at gøre opmærksom på' udtrykkes hellere 'jeg skal gøre opmærksom på'.

Ovenfor under reglerne har vi en regel om at en sætning ikke må slutte med et possessivt pronomen, fx 'yours'. På samme måde kan man anbefale at undgå at præpositioner 'hænger' alene i slutningen af sætningen: 'the glasses you read with' skulle i stedet være 'the glasses with which you read'.

Vi har medtaget disse regeltyper som en kilde til inspiration.

5 Opsummering

Vi har set at de to virksomheders forslag til regler og anbefalinger har mange fællestræk, og at de samtidig er forskellige. Vi har også set at mange af de foreslåede regler genkendes helt eller delvis fra AECMA. Virksomhedernes politik på området må altså siges at være helt på linje med hvad der i øvrigt foregår internationalt.

Mange af de regler og anbefalinger som virksomhederne ønsker at implementere, kan klares ved brug af lister, mens andre kræver mere avanceret datamatisk analyse. Lister er på mange måder den enkleste løsningsform, men det kræver en del vedligehold altid at have de rette positiv- og negativlister.

De næste skridt i projektet er nu dels at arbejde med begrebet tone-of-voice og dels at arbejde videre med en undersøgelse af hvad eksisterende systemer kan tilbyde og hvilke muligheder der er for at udvide eksisterende systemer med de regeltyper som ønskes.

Litteratur

Det skal bemærkes at en stor del af den litteratur der er brugt, er fundet på internettet. Denne litteratur er nævnt i Bilag A.

Mitamura, Teruku, Kathryn Baker, Eric Nyberg, David Svoboda: Diagnostics for Interactive Controlled Language Checking, in: *Controlled Language Translation, Proceedings of the EAMT-CLAW 03 Conference,* Ireland 2003

Møller, Margrethe H.: Grammatical Metaphor, Controlled Language and Machine Translation, in: *Controlled Language Translation, Proceedings of the EAMT-CLAW 03 Conference*, Ireland, 2003

O'Brien, Sharon: Controlling Controlled English. An Analysis of Several Controlled Language Rule Sets, in: *Controlled Language Translation, Proceedings of the EAMT-CLAW 03 Conference*, Ireland, 2003

Reuther, Ursula: Two in One – Can it work? Readability and Translatability by means of Controlled Language, in: *Controlled Language Translation, Proceedings of the EAMT-CLAW 03 Conference*, Ireland, 2003

Bilag A Links mv. til værktøjer og projekter vedr. kontrolleret sprog

I dette afsnit giver vi den liste over værkstøjer og projekter som vi har kendskab til. Disse værktøjer, samt evt. andre, vil blive nærmere undersøgt med henblik på om de kan have interesse i VID-sammenhæng.

AECMA-systemer

Disse fire AECMA[2] SE[3] 'compliant' systemer er så vidt vi kan se de eneste Simplified English produkter som er på markedet i august 2003.

- MAXit

  Smart Communications, http://www.smartny.com/
  Dette er næppe et hyldeprodukt. Se
  http://www.raycomm.com/techwhirl/archives/9805/techwhirl-9805-01208.html
  På hjemmesiden kan man se en del materiale som kan give et indtryk af hvad programmet kan og laver.

- Xplanation

  http://www.xplanation.com/. Hvis man indtaster http://www.lant.com bliver man linket videre til den første adresse, hvilket betyder at Xplanation er en videreudvikling af et andet produkt, LANTMASTER, som selv er et kommercielt spin-off af SECC-projektet.

  I SECC-projektet deltog følgende udviklingspartnere: Siemens Nixdorf, Leuven Universitet, Cap Gemini Innovation. SECC benytter sig af COGRAM (Controlled Grammar).

  B&O har set demonstration.

- HyperSE

  Tedopres hjemmeside: http://www.controlledenglish.com/

[2] Association Européenne des Constructeurs de Matériel Aérospatial
[3] AECMA Simplified English, PSC-85-16598 *A Guide for thePreparation of Aircraft Maintenance Documentation in the International Aerospace Maintenance Language*

- Asset

  Piper Group Plc, http://www.piper-group.com/

  Det eneste af produkterne hvor det har været muligt at få et demoprogram. Programmet giver indblik i regler og ordbog. Man kan skelne mellem approved og unapproved ord i forskellige ordklasser. Det arbejder sammen med Adobe Framemaker.

Ved søgning på internettet dukker flere navne op, men de synes ikke at være aktive:

- Language Manager (LM; X.Systems)
- ClearCheck (Carnegie Group)
- Simplified English Checker (SEC)
- Simplified English Checker/Corrector (SECC; Campbell Systems)
- CLarity (Cap Gemini)
- Boeing Simplified English Checker (Lernout & Hauspie)
- Boeing Technical English Checker og Boeing Plain Language Checker
- SE Exceptions Dictionary (Eclipse Tbs)

Endvidere en kontakt som ikke arbejder med AECMA, men kan være interessant alligevel:

CLAT (IAI, Tyskland). Baseret på KURD – A Formalism for Shallow Post Morphological Processing. CST har tidligere samarbejdet med IAI og har kontakt med CLAT. Multilint er CLAT for tysk.

Andre links

http://www.plainlanguagenetwork.org/

http://www.userlab.com/SE.html

Her kan man hente et dokument med AECMA-regler (SE.pdf).

http://www.aecma.org/Publications.htm

http://www.air-transport.org/public/publications/display1.asp?nid=924

Her kan man købe AECMA Simplified English Guide for the Preparation of Aircraft Maintenance Documentation

http://www.susanne-goepferich.de/itw_sw.html

Dokumentations- und übersetzungsrelevante Software – Ein Überblick

http://www.shlrc.mq.edu.au/masters/students/raltwarg/clindex.htm

Controlled Languages: An Introduction

http://appling.kent.edu/ResourcePages/LTStandards/Chart/standards.chart.htm

Standards for Content Creation and Globalization

http://ogden.basic-english.org/

http://www.raycomm.com/techwhirl/archives/0107/techwhirl-0107-01781.html

Indlæg i diskussionsforum som omtaler adskillige produkter

http://www.raycomm.com/techwhirl/archives/9805/techwhirl-9805-01208.html

Indlæg i diskussionsforum som omtaler måden hvorpå MAXit implementeres i en virksomhed.

http://www.sec.gov/investor/pubs/englishhndbk.htm

U.S. Securities and Exchange Commission: A Plain English Handbook