Lidt fakta om *Kukkuniiaat*

Kukkuniiaat er en del af Microsofts *Office* helt som de andre delprogrammer, der indgår i *Office*. Den skal derfor lige som alle andre Windows programmer først installeres på den enkelte maskine. Dette er helt uproblematisk, for programmet leveres som en installationspakke, man bare skal dobbeltklikke på for at klargøre programmerne.

Kukkuniiaat består først og fremmest af den *automat*, som kan læse og analysere enorme mængder af grønlandske ord og fremmedord ('enorme mængder' skal tælles i milliarder), og som kan generere nye ord. Et ord er defineret som de sammenhængende bogstaver og tal, der befinder sig mellem to skille- eller blanktegn. Derudover kræver *Kukkuniiaat* selvsagt et system af programmer, der håndterer trafikken mellem fx *Word* og automaten og dermed mellem brugeren og automaten. Men det er automaten, som er systemets "intelligens".

Kukkuniiaat fungerer på den måde, at automaten læser det ord, brugeren skriver på skærmen. Hvis den kan analysere ordet går den videre til næste ord. Hvis den derimod ikke kan analysere ordet beder den *Word* om at sætte en rød streg under det uanalyserede ord og forsøger samtidig at bygge en liste med ord, der ligner det uanalyserede ord bortset fra en enkelt karakter, den såkaldte 'forslagsliste'. En fejlskrivning som *agivoq* vil således få automaten til at danne en forslagsliste med ord som *angivoq*, *anivoq*, *akivoq* etc. etc.

Kukkuniiaats version 1.0 har en dækningsgrad og funktionalitet nogenlunde som de første versioner af den danske stavekontrol. Dækningsgraden er på publiceringstidspunktet lidt over 80% af alle løbende ord, hvilket må kaldes fuldt tilfredsstillende. Vi vil dog hurtigt nå op på en bedre dækning, fordi vi har opnået en meget fordelagtig aftale med Lingsoft, som gør det muligt for os løbende at opdatere *Kukkuniiaats* ord- og tilhængslister uden om Lingsoft og uden at ændre den eksisterende installationspakke. Dette er teknisk muliggjort af at installationsprogrammet ikke leveres med leksikonfilerne, men derimod med indbyggede links til Oqaasileriffiks server, hvorfra den henter leksikonfilerne på installationstidspunktet. Vi opdaterer ordlisterne løbende og påregner inden for et halvt års tid at nå en dækningsgrad omkr. 90% inden for rammerne af version 1.0

Kukkuniiaat indeholder nogle fejl og mangler, det ikke har været muligt at udbedre i version 1.0. Som den væsentligste kan det nævnes, at *Kukkuniiaat* ikke indeholder den såkaldte 'korrekturliste', som er en liste over ord og dobbeltformer, der kræver særlig opmærksomhed (fx nye staveformer der adskiller med mere end én karakter som i gl. *biili* vs. ny *bil* eller nyere udviklinger i sproget som i (ældre) *ukallisut* vs. (nyere) *ukalertut*). Vi afventer fortsat en sådan liste fra Grønlands sprognævn, men håber at kunne medtage den i version 2. Desuden har vi fundet et par programmeringslus, som først bliver udbedret med version 2. Fx har automaten problemer med vokalerne i navneord med såkaldt metatese så at de accepterer både korrekte former som *qeqqa* og ukorrekte som *qiqqa.*

Automaten er udviklet af Oqaasileriffik medens det er Lingsoft a/s i Helsingfors, der har fået den til at fungere inde i Office (eller med et teknisk udtryk har forestået kompileringen). Lingsoft er det firma, der har kompileret stavekontrollerne til de øvrige nordiske sprog og til mange andre sprog i Europa og Asien.

Postboks 980
DK-3900 Nuuk
+299 345833
oqaasileriffik@gh.gl

Härkätie 371
FIN-21490 Marttila
+358 (0)24846062
pela@gh.gl

Automatens inddata kommer fra håndkodede ordlister med godt 12.000 navne, ca. 40.000 navneord, ca. 35.000 udsagnsord samt lidt under 1.000 småord. Desuden indeholder automaten en talgenerator og en akronymgenerator for også at kunne håndtere de uendeligt mange 'ord' som *SIK-mi, 2006-imi* og *KNAPK*. Ordlisternes basisdata er kodninger på basis af *Erik Fleischers* 350.000 ords store samling af grønlandske enkeltord, som Erik donerede til Oqaasileriffik for små 4 år siden. Fremtidige opdateringer vil i stigende grad ske ved automatisk processer, idet vi nu har nået et punkt, hvor vi kan lade automaten læse den nye *Sermitsiaq* eller *Atuagagdliutit* og lave en maskinlæsbar liste med de ord, den ikke forstår.

Automaten fungerer ved at stammerne fra ordlisterne først går gennem et leksikon, der om nødvendigt kan sætte et eller flere tilhæng på stammerne. De herved nydannede ord sendes dernæst videre til en slags digital grammatik, der tilføjer de nødvendige endelser. Endelig sendes ordene gennem et sidste leksikon, der om nødvendigt tilføjer et eller flere tilhængspartikler. Til sidst køres det således dannede ord gennem et system af lyd- og skriveregler, der håndterer de forandringer, der sker når de forskellige sproglige elementer sættes sammen. Tag et ord som *qimmeqaravinngooq* ('Det er fordi du har hunde siger man') som eksempel:

Ordet består af qimmeq+qar+gavit+gooq, der efter turen gennem regelautomaten bliver til det ønskede *qimmeqaravinngooq* og således bliver et muligt grønlandsk ord, som *Kukkuniiaat* kan acceptere og altså ikke sætte streg under.

Automaten er udviklet af Oqaasileriffik over en periode på lidt under to år. Dette er usædvanligt kort tid, hvilket ikke mindst skyldes, at arbejdet har fået stor hjælp fra udenlandske forskerkolleger. Frem for nogen er universitetslektor Trond Trosterud fra Universitetet i Tromsø en helt central figur i arbejdet. Trond er datalingvist på meget, meget højt niveau og har investeret flere hundrede ulønnede timer i arbejdet og på adskillige andre måder været os en god mand. Først og fremmest ved at give os adgang til sit arbejde med den samiske automat og dermed muligheden for at kunne tilpasse en færdig model i stedet for at gå den møjsommelige trial-and-error vej selv, og det er da også en server på Universitet i Tromsø, der gemmer vore sikkerhedskopier og styrer det komplicerede versionskontrolsystem,. Men ikke nok med det. Trond har også helt konkret deltaget i programmeringen, fx ved at skrive den teknisk meget krævende talgenerator.

Projektet har også været i den heldige situation at have haft *Tero Avellan* i hånden i hele perioden. Tero er en finsk stjerneprogrammør (stadig polyteknisk student, men allerede fastansat i Nokias inderkreds). Han har været guru på den komplicerede Linux-platform, der er en forudsætning for arbejdet og det er ham, der har taget over, når projektets Perl- og PHP programmer har oversteget Oqaasileriffiks egne medarbejderes programmørkompetence.

Det er seniorrådgiver Per Langgård, der har udformet programmet og skrevet den særlige grammatik, som systemet forudsætter, medens kodningen af Erik Fleischers ordlister især er udført af studentermedhjælper Marianne Hansen og forskningsassistent Elisa Isaksen med Erik som altid tilgængelig "orakel" på telefonen. Det tekniske arbejde med opdateringen vil fremover i vid udstrækning blive forestået af forskningsassistent Aviaq Tobiassen med AC fuldmægtigene Nuka Møller og Lisathe Møller Kruse som lingvistiske tutorer.

Udviklingen af *Kukkuniiaat* har modtaget støtte fra Nunafonden til den indledende fase (i 2002) og fra Nordens Sprogråd til selve udviklingsfasen. Bortset fra det er arbejdet finansieret af Oqaasileriffiks bevillinger i 2005 og 2006, dog således at projektet fik en øremærket TB på 380.000 i 2005.