



Fact sheet about *Kukkuníiaat*

Kukkuníiaat is an application in Microsoft's *Office* suite like all other parts of *Office*. Therefore it has to be installed on the local machine before use. This is simple since you download *Kukkuníiaat* as an installation wizard, you just have to doubleclick.

Kukkuníiaat is basically a *finite state automaton* that knows how to read and analyse enormous amounts of Greenlandic words and loanwords ('enormous' count in billions) and with the capacity to create new words automatically. A word is defined as a coherent string of numbers and characters between white spaces or punctuation marks. Furthermore *Kukkuníiaat* needs a system of programmes to handle the traffic between *Word* and the automaton and thus between the automaton and the user. But the system's "intelligence" is solely situated in the automaton.

Kukkuníiaat basically works this way: The automaton reads the user's input. If it is able to analyse the input it passes on to the next word. If on the other hand it does not understand the word in question it asks *Word* to underline the unanalysed word and simultaneously tries to build a list of words that differ by one character. The so called 'suggestions' list. A misspelling like *agivoq* (which could be a dialectical hypercorrection for *angivoq*) will thus make the automaton build a list with words like *angivoq*, *anivoq*, *akivoq* etc.

Kukkuníiaat version 1.0 has a coverage and a functionality that by and large compares to the first versions of the Danish spell checker. Coverage on opening day is a little better than 80% of all words in normal text sources. This is entirely satisfactory. We expect very soon to reach a much better coverage since we, by means of a very flexible agreement with Lingsoft, have the possibility to update *Kukkuníiaat*'s lexicons and morpheme lists ad hoc without consulting Lingsoft. This is technically possible because the installation package proper does not include the lexicon material but links to this material on our own server. Lexicons and morphemelists are continuously updated so that we expect to reach a coverage around 90% still under version 1.0 already in half a year or so.

Kukkuníiaat version 1.0 contains some known bugs and shortcomings. Most important is the lack of the so called 'corrections' list. A 'corrections' list is a list of words that need special attention (like recent decisions on new orthography differing with more one character as in old *biili* vs. new *bil* or developments like older *ukallisut* vs. newer *ukalertut*). We still await such a list from the Greenlandic Language Board, but expect to be able to include it in version 2.0. As can be expected we have also found a few plain bugs. As an example the automaton has problems with vowels in nouns undergoing metathesis so that it will accept both correct wordforms like *qeqla* and incorrect ones like *qiqqa*.

The *automaton* is developed by Oqaasileriffik while Lingsoft Ltd. in Helsinki has made it work inside *Office* (technically speaking 'compiled' it). Lingsoft is the software house that has made the writing tools for all the Nordic languages plus a number of European and Asian languages.

The *Automaton* takes its data from manually coded lexicons with about 12.000 proper nouns, about 40.000 nouns, about 35.000 verbs, and a little less than 1000 particles. The automaton furthermore contains a number generator and a generator crunching acronyms to handle the many 'words' like *SIK-mi*, *2006-imi*, and *KNAPK*. The lexicon files

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

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

are coded using *Erik Fleischer's* compilation of 350.000 different Greenlandic words which Erik donated to Oqaasileriffik in 2002. Future updates will more and more be made by the system itself since it now has the capacity to automatically extract unresolved words from unedited texts like *Sermitsiaq* or *Atuagagdliutit* and present it as a wordlist for the staff to evaluate.

The *Automaton* first evaluate a given stem taken from the lexicons and pass it on to the lexicons holding the derivational morphemes that can be added to the stem in question. The newly generated "words" will then be passed on to the "grammar" where proper flexives will be added. Finally the "word" will go the final lexicon to fetch the clitic particles needed. Finally the "word" will run through a system of ordered phonologic and orthographic rules that will handle the alternations taking place on morpheme borders etc. *qimmeqaravinnqooq* ('It is because he has dogs they say') is one example:

The word is made up of four morphemes *qimmeq+qar+gavit+gooq* that after being processed in the rules will come out as the expected *qimmeqaravinnqooq* and so be a Greenlandic word which *Kukkuniiaat* will accept and accordingly not underline.

The *Automaton* is developed by Oqaasileriffik over a period of almost two years. This is a very limited time consumption which not least can be explained by the fact that the project has had much attention and much help from fellow linguists outside Greenland. First and foremost has ass. professor *Trond Trosterud* from the University of Tromsø played a prominent role in the process. *Trond* is a computer linguist at a very advanced level. He has donated several hundred hours of work to us and has opened a number of important doors. He introduced us to the Saamic Divvun project from which we could copy the overall framework and thus save hundreds of hours of trial-and-error in the initial phases. It is also a server in Tromsø that holds our back-up copies and handle the complicated CVS (ConcurrentVersioncontrolSystem). *Trond* has even actively partaken in concrete programming for instance by providing us with the complicated *numbercruncher*.

We have also been lucky enough to have a computer scientist high above average to help us, namely *Tero Avellan*. *Tero* is a Finnish programmer who though still only a student of computer engineering is already headhunted to *Nokia's* inner circles (the prototype development center in Tampere). He has been our guru on the rather heavy Linux platform needed for this kind of development and he is the one to take over whenever development in Perl or PHP exceeded Oqaasileriffik's own staff's computer know-how.

Senior advisor *Per Langgård* has made the program and developed the grammar and the rules without which the automaton could not work. Coding of *Erik Flescher's* wordlist has been taken care of by research assistants *Marianne Hansen* and *Elisa Isaksen* with *Erik* as an always available oracle on the telephone from Paamiut. The technical duties with updates will in the years to come be handled by research assistant *Aviaq Tobiassen* with researchers *Nuka Møller* and *Lisathe Møller Kruse* being linguistic tutors.

Development of *Kukkuniiaat* has been supported from *Nunafonden* in the very earliest stage and from *Nordens Sprogråd* for development proper. Apart from that the work has been kept inside Oqaasileriffik's annuum with a special grant of 380.000 kroner from Home Rule in 2005.