The Linguistic Annotation System of the Stockholm – Umeå Corpus Project

Gunnel Källgren & Gunnar Eriksson
Institute of Linguistics
Stockholm University
S-106 91 Stockholm
gunnel@ling.su.se, gunnar@ling.su.se

In the Stockholm - Umeå Corpus project, SUC, we have developed and applied a system for representing lexical and morphological information about word forms in unrestricted text. Our poster presents results and experiences from the application of the system to 300,000 word forms, a subpart of a larger corpus.

The application of the system is carried out in two steps, an automatic lexical look up followed by homograph separation, which is done partly automatically, partly manually. Lexical and morphological analysis and disambiguation of Swedish is a rather complicated task, a fact which should hold for several other languages as well. Below a sample text is given, showing both the amount of information that has to be specified for each word form and the degree of ambiguity to be resolved.

```
("<*själv>" <161>
      ("själv" NN NEU SIN IND NOM)
      ("själv" NN NEU PLU IND NOM)
      ("själv" JJ POS UTR SIN IND NOM)
      ("själv" PM NOM))

("<rökar>" <162>
      ("råka" VB PRS AKT)
      ("råk" NN UTR PLU IND NOM))

("<hon>" <163>
      ("hon" PN UTR SIN DEF SUB)
      ("ho" NN UTR SIN DEF NOM))

("<ut>" <164>
      ("ut" AB))

("<för>" <165>
      ("för" PP)
      ("för" AB)
      ("för" SN)
      ("för" KN)
      ("för" NN UTR SIN IND NOM)
      ("för" VB PRS AKT)
      ("för" VB IMP AKT))

("<en>" <166>
      ("en" DT UTR SIN IND)
      ("en" RG UTR SIN IND NOM)
      ("en" PN UTR SIN IND SUB/OBJ)
      ("en" AB)
      ("en" NN UTR SIN IND NOM))

("<kåkfarare>" <167>
      ("kåk_farare" NN UTR SIN IND NOM)
      ("kåk_farare" NN UTR PLU IND NOM))

("<som>" <168>
      ("som" HP - - -)
      ("som" HA)
      ("som" KN))

("<misshandlar>" <169>
      ("miss_handla" VB PRS AKT)
      ("miss_handel" NN UTR PLU IND NOM))

("<och>" <170>
      ("och" KN))
("<förödmjukar>" <171>
      ("förödmjuka" VB PRS AKT))

("<henne>" <172>
      ("hon" PN UTR SIN DEF OBJ))
```