Constructing a Corpus of August Strindberg's Collected Works

Mats Wirén, Kristina Nilsson Björkenstam, Gintarė Grigonytė and Sofia Gustafson Capková
Department of Linguistics
Stockholm University
SE-106 91 Stockholm, Sweden
{mats.wiren, kristina.nilsson, gintare, sofia}@ling.su.se

1 Introduction

This abstract outlines our work on constructing a corpus from the National Edition of the collected works of August Strindberg, published 1981–2012 and consisting of 72 printed volumes.[1] The volumes contain edited text by Strindberg, altogether some 20,000 pages or 6 million words, along with critical commentaries. The work reported here is a collaboration with Litteraturbanken[2] at the University of Gothenburg and the Strindberg Project at Stockholm University.[3] The corpus will be based on electronic versions of the printed books provided by Litteraturbanken. The work is a continuation of a previous project on a smaller scale which generated a corpus of Strindberg's autobiographical works, the Stockholm University Strindberg Corpus (SUSC) (Björkenstam et al., 2014).

2 Aim and status

The aim of the project is to distribute the corpus in three versions under a Creative Commons license:

1. A raw-text version without annotation, with the simplest possible structure for representing basic textual units (chapters, paragraphs, headings, etc.) by interspersed blank lines. This version is intended for researchers who want to work with the raw text directly, for example, by using their own scripts or an on-line corpus annotation and lexical analysis tool such as Sketch Engine.[4]
2. A CoNLL version with one word per line and linguistic annotation distributed over additional columns. This version is intended for researchers who want to work with the annotated text without going through the XML version or a search interface such as Språkbanken's Korp (see below).
3. An XML version with an XML schema that encodes the structure and annotation of the text. This version is meant to be bundled with an independent search engine such as The IMS Open Corpus Workbench[5] (CWB) and/or integrated with the Korp infrastructure at Språkbanken.[6] It is primarily intended for literary researchers who want to access the corpus through a search interface or concordancer.

Linguistic annotation of the CoNNL and XML versions will be based on the `efselab`[7] analysis pipeline, which includes tokenisation, part-of-speech tagging and dependency parsing. Further annotation, such as named entities and coreference, may be added at a later stage. In order to handle the archaic features of Strindberg's language, the pipeline will be applied to a version of the text which has undergone spelling normalisation along the lines of Pettersson (2016), though the annotation will be transferred back to the original text.

In addition to the corpus in its different versions, we plan to make available useful corpus statistics involving frequency lists, as well as the literary commentaries.

The project is currently in its initial stages, with most of the work going into extraction of the raw-text version of the corpus.

[1]The work presented here has been supported by an infrastructure grant from the Swedish Research Council (SWE-CLARIN, project 821-2013-2003). We would like to thank Litteraturbanken at Gothenburg University and the Strindberg Project at Stockholm University for support and for providing us with the electronic version of the book material.

[2]http://litteraturbanken.se

[3]http://www.strind.su.se/present.htm

[4]https://www.sketchengine.co.uk/

[5]http://cwb.sourceforge.net/

[6]https://spraakbanken.gu.se/korp

[7]https://github.com/robertostling/efselab

References

Kristina Nilsson Björkenstam, Sofia Gustafson Capková, and Mats Wirén. 2014. The Stockholm University Strindberg Corpus: Content and Possibilities. In Roland Lysell, editor, *Strindberg on International Stages/Strindberg in Translation*. Cambridge Scholars Publishing.

Eva Pettersson. 2016. *Spelling Normalisation and Linguistic Analysis of Historical Text for Information Extraction*. Ph.D. thesis, Uppsala University, Department of Linguistics and Philology.