Stockholm University Strindberg Corpus: Content and Possibilities

Kristina Nilsson Björkenstam, Sofia Gustafson-Capková & Mats Wirén
Computational Linguistics, Department of Linguistics, Stockholm University

Abstract

We have approached the works of August Strindberg from a computational linguistic point of view, resulting in The Stockholm University Strindberg Corpus, consisting of seven of Strindberg's autobiographical works with linguistic annotation. The corpus is freely available for research.[1]

We use this corpus for three quantitative studies of Strindberg's work: in the first, we describe the novels included in the corpus by keywords; in the second, we compare Strindberg's use of emotionally charged words with selected prose of both his contemporaries and present-day authors; in the third, we explore the semantic prosody of KVINNA ("woman") and MAN ("man").

1 Introduction

The Stockholm University Strindberg Corpus (SUSC) consists of approximately 400 000 tokens and is annotated for parts-of-speech (PoS), including morphological analysis and lemmas. The linguistic annotation follows the Stockholm-Umeå Corpus (Källgren 2006) which is commonly regarded as a reference corpus for Swedish. Furthermore, the annotated texts have been converted to XML which makes the corpus searchable with analysis tools such as Xaira[2] and AntConc.[3] This allows for e.g., searching for keywords, concordances with a specific word form or lemma, for pattern matching (including PoS), and collocation extraction.

In this paper we describe the content of the corpus, how it was constructed, and the linguistic annotation. Furthermore, we discuss how approaches in computational linguistics can be of use within the area of stylistic and literary analysis, and illustrate the possibilities of a corpus-based approach by describing three quantitative studies. We propose the use of computational methods, not as a substitute to other kinds of literary analysis but as a set of tools for exploration and discovery.

2 Content

The current version of the corpus includes seven works by Strindberg which can all be classified as autobiographical (Robinson 2008):[4]

- Tjänstekvinnans son (The son of a servant, 1886-87)
- Han och hon (He and She, 1919)
- Inferno (Inferno, 1897)[5]
- Legender, Jakob brottas (Legends and Jacob Wrestles, 1898)

---

[1] Stockholm University Strindberg Corpus, URL: www.ling.su.se/nlp/susc
[2] Xaira. URL: http://www.oucs.ox.ac.uk/rts/xaira/ Last checked: 2012-10-11
[3] AntConc. URL: http://www.antlab.sci.waseda.ac.jp/index.html Last checked: 2012-10-11
[4] (Robinson 2008) includes En dåres försvarstal (A Madman's Defense), Klostret (The Cloister), and Ockulta dagboken (The Occult Diary) in Strindberg's autobiographical writing. Because we do not at present have access to these works in electronic form, they are not included in the corpus.
[5] Inferno and Legender were written in French and translated to Swedish by Eugene Fahlstedt.

- Fagervik och Skamsund (Fairhaven and Foulstrand, 1902)
- Ensam (Alone, 1903)

We are aware of three other electronic collections of Strindberg's works: Projekt Runeberg,[6] Litteraturbanken,[7] and Språkbanken.[8] The first two collections consist of e-text, whereas the data available through Språkbanken's Korp, a web concordancer interface, is annotated for PoS, lemma, lexico-semantic information, and dependency relations.

While these collections are valuable resources, our corpus is an important addition because, unlike the first two, it is linguistically annotated, and unlike the third, the data is available for download and thus can be processed using the researcher's software of choice. Even more importantly, researchers can add their analyses as new layers of annotation of the corpus.

2.1 Data preprocessing

The starting point was the digitized volumes of *Samlade skrifter av August Strindberg* (Collected works, published 1912-1921), available from Projekt Runeberg, Linköping University.[9] The plain text files for each of the seven works were preprocessed by removing headers, footers, and page numbers. The novels were structured into chapters, paragraphs, sentences, and words. In addition, other segments, such as quotations, poems, and footnotes, were annotated. Preprocessing also included correcting OCR errors.

2.2 Linguistic annotation

The texts are annotated with part-of-speech, morphological analysis and lemma form using STagger,[10] a tagger that has been evaluated against the manually corrected annotation of the SUC corpus with an accuracy of 96.6% (Östling 2012). The results of the tagging were semi-automatically corrected. The most prevalent error types were archaic word forms usch as plural nouns in masculine form, e.g., *gossarne* (“the boys”) and plural forms of verbs, e.g., *gingo* (“went”), foreign words (in English, French, German, Danish, and Latin), and proper names. The PoS tag set is based on the SUC tag set (Källgren 2006). The tags are in PAROLE-format, and a key to the PAROLE codes is provided with the corpus.

2.3 Corpus format

The preprocessed, segmented, and annotated texts were converted to XML format, resulting in corpus data of this format:

```
<s>
<w bf="efter" ps="SPS">Efter</w>
<w bf="tio" ps="MC00N0S">tio</w>
<w bf="år" ps="NCNPG@IS">års</w>
<w bf="vistelse" ps="NCUSN@IS">vistelse</w>
<w bf="i" ps="SPS">i</w>
<w bf="landsort" ps="NCUSN@DS">landsorten</w>
<w bf="vara" ps="V@IPAS">är</w>
<w bf="jag" ps="PF@USS@S">jag</w>
```

[6] Projekt Runeberg, URL: http://runeberg.org/ Last checked: 2011-09-29
[7] Litteraturbanken, URL: http://litteraturbanken.se/ Last checked: 2011-09-29
[8] Språkbanken Korp, URL: http://spraakbanken.gu.se/korp/ Last checked: 2011-09-29
[9] Projekt Runeberg. Strindbergs samlade skrifter. http://runeberg.org/strindbg/
[10] STagger URL: http://www.ling.su.se/nlp/stagger

```
<w bf="åter" ps="RG0S">åter</w>
<w bf="i" ps="SPS">i</w>
<w bf="min" ps="PS@US0@S">min</w>
<w bf="födelsestad" ps="NCUSN@IS">födelsestad</w>
<w bf="och" ps="CCS">och</w>
<w bf="sitta" ps="V@IPAS">sitter</w>
<w bf="nu" ps="RG0S">nu</w>
<w bf="vid" ps="SPS">vid</w>
<w bf="en" ps="DI@NS@S">ett</w>
<w bf="middagsbord" ps="NCNSN@IS">middagsbord</w>
<w bf="bland" ps="SPS">bland</w>
<w bf="den" ps="DF@0P@S">de</w>
<w bf="gammal" ps="AQP0PN0S">gamla</w>
<w bf="vän" ps="NCUPN@DS">vännerna</w>
<w bf="." ps="FE">.</w>
</s>
```

Figure 1: Example XML structure of a sentence from SUSC, with lemma (bf), part of speech and morpho-syntactic analysis in Parole format (ps), and surface form for each word.

The XML-encoding presented in Figure 1, above, is not meant for human readers but makes the data searchable. Corpus search tools, such as Xaira and AntConc, include file view options where the XML data is converted to plain text while still allowing the user to search for combinations of word forms, morpho-syntactic information, and lemmas.

3 Possibilities

In this section we give three examples of corpus based approaches with use of SUSC. Study 1 is based on keywords. Study 2 makes use of word distribution and study 3 exploits semantic prosody. All three studies show how a corpus based approach can be used in order to extend traditional text analysis.

3.1 Study 1: Describing the contents of SUSC by keywords

The content of a text can (to a degree) be described by the words that occur in that text. One way of exploring a collection of texts is to make a word list of surface forms or lemmas ranked by frequency for each text, and compare those lists. (Moon 2007) showed that by dividing word lists into frequency bands identified from reference corpora and examining the most recurrent words within each band, both high-frequency and low-frequency words can be included in the analysis. A commonly used approach to find the most salient words of a text is by comparing the text to a reference corpus of texts written by other authors and extracting the *keywords* of the text (Stubbs 2005, O'Halloran 2007).

We use a reference corpus of texts written by the same author, in order to highlight salient topics and themes for each of the novels, compared to the rest. Using the corpus tool AntConc , we extract the keywords in each work in SUSC by comparing that work to the rest of SUSC; for example, Ensam is compared to Tjänstekvinnans son, Han och hon, Inferno, Legender, and Fagervik och Skamsund.

Below, the ten most salient keywords for each text in SUSC are listed, starting with **Tjänstekvinnans son**. For each keyword, the log-likelihood value and the frequency are listed in brackets. If the log-likelihood value is greater than 6.63, the probability of the result happening by chance is less than 1% ($p < 0.01$), that is, we are 99% certain of the result (Dunning 1993).

Johan (1053.5; 622), han (588.0; 2585), var (577.2; 1741), Han (515.292; 804), hade (173.2; 777), fadren (148.9; 118), skolan (132.7; 71), Hakon (113.3; 50), honom (110.4; 635), Fritz (108.6; 52).

Table 1: Keywords for Tjänstekvinnans son

The ten highest ranking keywords for Tjänstekvinnans son include the names of Strindberg's alter-ego *Johan* and his friend *Fritz*, the noun *fadren* ("the father"), and the masculine third person pronouns *han* ("he") and *honom* ("him"). *Hakon* (*Jarl*) is the subject of a thesis put forth by Johan. This novel describes the childhood and youth of Johan, as evident by the keyword *skolan* ("the school"). Other highly significant keywords signaling a childhood story are *barnen* ("the children"), *modren* ("the mother"), and *gossarne* ("the boys").

Er (1814.9; 533), Ni (1614.0; 516), jag (1562.8; 2446), dig (1039.4; 536), du (762.5; 681), mig (612.6; 1203), skall (454.9; 398), ej (409.1; 518), älskar (349.3; 134), Jag (327.5; 484).

Table 2: Keywords for Han och hon

Han och hon consists of letters, mainly between the two lovers Johan (Strindberg) and Maria (Siri von Essen). The letter format is reflected in the set of the ten highest ranking keywords, consisting of first (*jag*, *mig*, *Jag*) and second person pronouns (*Ni, du, dig*), the verb *skall* ("shall", "will"), the negation *ej* ("not"). The only verb among the ten is *älskar* ("love") in the present tense.

mig (476.0; 946), min (317.8; 462), jag (206.3; 1270), mitt (167.5; 229), av (147.5; 686), mina (146.9; 161), Jag (72.3; 275), har (64.4; 367), ned (61.0; 44), makterna (54.6; 27).

Table 3: Keywords for Inferno

Inferno, Strindberg's description of a psychological crisis, has the most coherent set of significant keywords of the texts in SUSC: of the ten most significant, six are first person pronouns (*mig, jag, Jag*) and first person possessives (*min, mitt, mina*). The only verb in this list is *har* ("have", present tense). The preposition *ned* ("down") occurs in the context of objects falling down (6 occurrences), somebody sitting down (3), and physically or metaphorically attacking something (1 and 2, respectively).

Important to Inferno are the mysterious forces that guide Strindberg's alter-ego, and the noun *makterna* ("the powers") occurs frequently in contexts where he believes himself saved (examples 1, 2) or condemned by the powers (3, 4), or when he discusses the intentions of the powers (5), for example:

(1) ordningen, förvissad om att vara benådad av **makterna**, vilka tyckas ha uppskjutit straffen
(2) samma väg jag kommit, inom mig tackande **makterna** för att de varnat mig , så viss var jag på
(3) jag befann mig i helvetet! förjagad dit av **makterna**. Vem var då min mästare? Swedenborg
(4) skulle jag ha begripit att jag fann mig av **makterna** dömd till exkrement-helvetet .
(5) den rörelsen, ingenting att ångra, eftersom **makterna** ha så velat, att vår väg skulle gå fram

Among the significant keywords of Inferno are numerous words related to spiritual matters (*Evige* ("Eternal"), *helvetet* ("hell"), *demoner* ("demons"), *försynen* ("providence"), *korset* ("the cross"),

slump (“chance”), *sammanträffande* (“coincidence”)), but also to (pseudo-)science and alchemy (*guld* (“gold”), *experiment* (“experiment”), *svavel* (“sulphur”)).

> jag (223.8; 1297), mig (181.2; 754), la (104.6; 46), av (101.2; 645), har (62.0; 367), rue (57.4; 34), Swedenborg (50.417; 34), natten (46.3; 42), mina (43.1; 112), min (38.3; 284).

Table 4: Keywords for Legender.

Like Inferno, **Legender** centers round a tumultuous period in Strindberg’s life. Among the most significant keywords for Legender are first person personal pronouns and possessives (*jag, mig, mina, min*). Other significant keywords for Legender are French words ”la” and “rue” (referring to addresses in Paris), the preposition *av*, the verb *har* (“have”, present tense), the proper name *Swedenborg*, and the noun *natten* (“the night”).

> han (642.3; 1617), hon (531.1; 485), Torkel (192.8; 53), hennes (128.6; 131), gick (123.1; 131), Och (118.8; 281), hade (114.7; 433), sig (110.7; 633), frun (101.8; 38), Fagervik (98.2; 27).

Table 5: Keywords for Fagervik och Skamsund

In the collection of short stories Fagervik och Skamsund, the most significant keywords are third person personal pronouns (*han, hon*), possessives (*hennes, sig*), the proper names *Torkel* and *Fagervik*, and the noun *frun* (“Madam”). The most high-ranking verbs are *gick* (”went”, “walked”) and *hade* (”had”). Sentence initial *Och* (”And”) is more frequent in Fagervik than in the other novels, although it occurs in all of them.

> jag (243.0; 803), ensamheten (40.4; 19), tjuta (36.2; 8), Vargarne (31.8; 6), förbi (26.7; 19), väggarna (26.2; 6), ty (25.0; 93), människor (24.8; 29), visserligen (23.4; 15), Ahasverus (21.2; 5).

Table 6: Keywords for Ensam.

As can be predicted from the title, among the most significant keywords for Ensam is the noun *ensamheten* (“the loneliness”). Other significant nouns are *Vargarne* (“The wolves”), *väggarna* (“the walls”), and *människor* (“people”). The only verb among the highest ranked words is *tjuta* (“howl”), which occurs in contexts describing what dogs (1 ocurrence) or wolves (7 occurrences, all in a poem) do. The name *Ahasverus* also occurs in a poem.

In Ensam, the adverb *förbi* (“past”) occurs 15 times with the meaning of someone passing something by, and 4 times as in something having ended. The frequent use of the conjunction *ty* (“for”, “because”), with increasing frequency towards the end of Ensam, and the adverb *visserligen* (“indeed”) suggests that this is an argumentative or reasoning text.

We mean that the keyword occurrences to a high degree mirror important aspects of Strindberg’s biography. The keywords reveal relations and emotions that can be seen as representative for the different periods during Strindberg’s life. While such major themes become visible by reading the texts, in addition, the keyword approach offers a clear quantitative foundation for further analysis.

3.2 Study 2: Emotionally charged words

Strindberg is said to be an author with an emotionally charged language. For our second study, we have carried out a close study of some emotionally charged words in Strindberg's novels. As a starting point we have selected one of the keywords from study 1, ÄLSKA ("to love") together with the antonym HATA ("to hate") and formed two sets of emotionally charged words. We identify all surface forms of following two sets of positively and negatively charged verbs:

- ÄLSKA ("to love"), DYRKA ("to worship"), BEUNDRA ("to admire")
- HATA ("to hate"), AVSKY ("to abhor"), FÖRAKTA ("to despise")

First, we explore Strindberg's use of emotionally charged words through frequency and distribution over the works included in SUSC. Second, we investigate whether Strindberg used these words more often than other authors, both contemporary and present day.

3.2.1 Word frequency and dispersion over the collection

The most frequent of the positive verbs is ÄLSKA, which occurs 324 times in SUSC. This means that, if we assume that there are about 250 words per printed page, this word occurs on every 49th page. The most frequent negative verb HATA occurs 88 times, or on every 179th page. Table 1, below, shows the frequency of these words in SUSC.

EMOTION	LEMMA	Freq.
POSITIVE	ÄLSKA	324
	DYRKA	14
	BEUNDRA	42
NEGATIVE	HATA	88
	AVSKY	18
	FÖRAKTA	70

Table 7: Frequency of emotionally charged words in SUSC

While the total frequency is interesting, these words are not evenly distributed in SUSC. Figures 1 and 2 (below) show the distribution of positive and negative words over each novel. The most striking observation is that Han och hon is the novel where both the positively and the negatively charged words are used the most frequently (and in this novel *älska* was one of the top 10 keywords). A second observation is that in Ensam, there are only two occurrences of the positive verbs (*beundra* in the context of admiring art, and *älska* in the context of loving one's child), while there are 12 occurrences of the negative verbs: *avsky* (1), *förakta* (2), *hata* (9). The novels Han och hon and Fagervik och Skamsund, with higher frequencies of (mainly positive) emotionally charged words was written during, and centers round, Strindberg's relationships with Siri von Essen and Frida Uhl, respectively, while Ensam describes the author's alter-ego living alone.

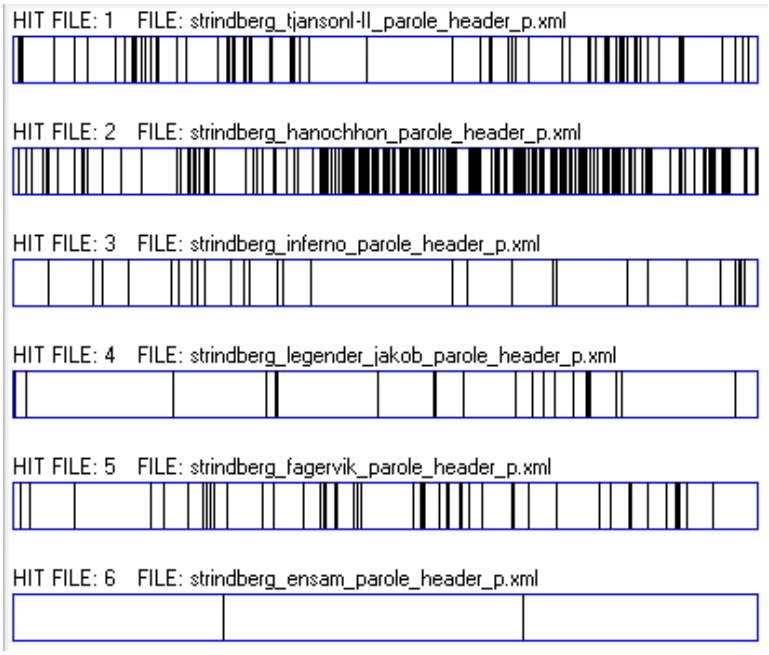

Figure 1: Distribution of ÄLSKA, DYRKA, BEUNDRA over SUSC (plot produced with AntConc).

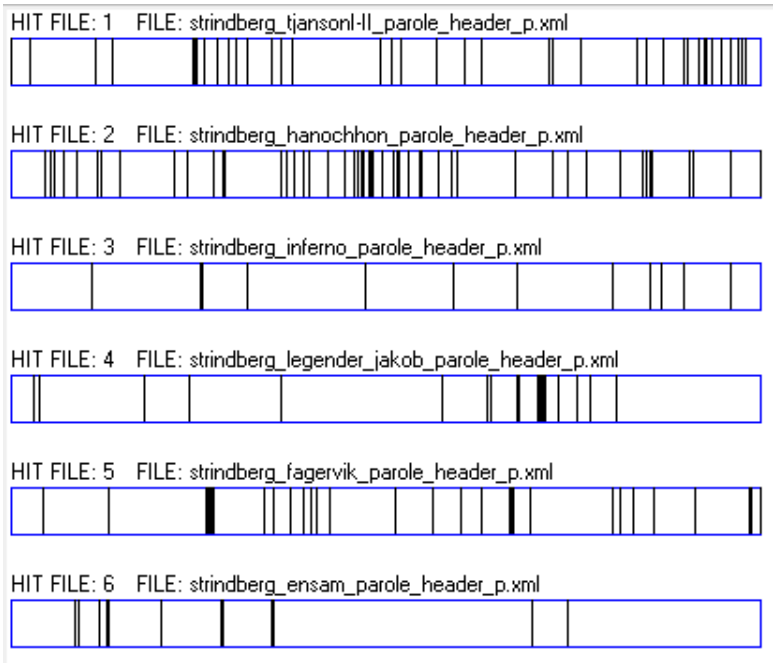

Figure 2: Distribution of HATA, AVSKY, FÖRAKTA over SUSC (plot produced with AntConc).

3.2.2 Frequency compared to other authors

Next, we will compare Strindberg's use of these words to other authors, both his contemporaries and ours. For this comparison, we use resources provided by Språkbanken[11] as reference corpora, grouped as follows:

- **Strindberg's contemporaries:** novels published between 1830 and 1930 by both male and female authors, e.g., C.J.L Almqvist, Fredrika Bremer, Victoria Benedictson, Hjalmar Bergman, Selma Lagerlöf, Hjalmar Söderberg, and Viktor Rydberg. Total: 4.3 million words.

[11] Språkbanken. URL: http://spraakbanken.gu.se/swe/resurser

- **Our contemporaries:** four sets of novels published between 1976 and 1999**:** Bonniersromaner I (1976-1977), Bonniersromaner II (1980-1981), SUC-romaner (1992), Norstedtsromaner (1999). Total: 18 million words.

Because the data sets are of different sizes we calculate the normalized frequency (as per million words) for each word in each data set. This allows for comparisons between the data sets. However, a difference in frequency between two data sets may be due to chance, and therefore we test the statistical significance of the difference by calculating the log-likelihood value for each result. A log-likelihood value of 3.84 or more means that the probability of the result being due to chance is less than 5% ($p < 0.05$), that is, the result is statistically significant (Dunning 1993). All results presented below are highly significant (with a log-likelihood value over 6.63; $p < 0.01$).

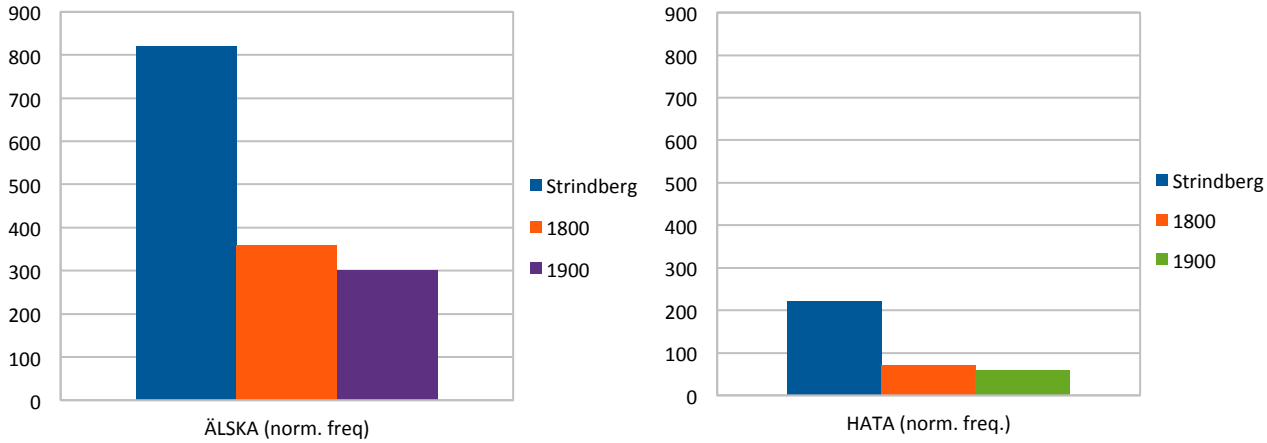

Figure 3: The use of ÄLSKA and HATA (normalized frequencies: per million words) by Strindberg, compared to his contemporaries (1800) and present-day authors (1900).

The diagram on the left in figure 3 shows the normalized frequency of all surface forms of ÄLSKA in the works of Strindberg, the set of works by 19th century authors (labeled 1800) and 20th century authors (1900). The differences are highly significant ($p < 0.0001$): ÄLSKA occurs more often in SUSC than in the two reference corpora. The diagram on the right shows that the surface forms of HATA occurs more often in SUSC than in the reference corpora. Again, this result is highly significant ($p < 0.0001$).

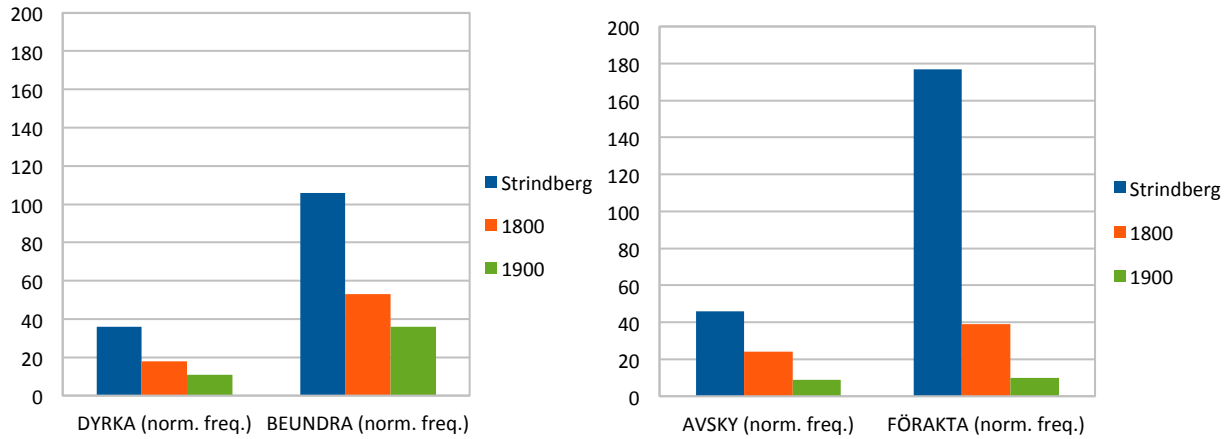

Figure 4: Positive verbs DYRKA and BEUNDRA (left). Negative verbs AVSKY and FÖRAKTA (right). Normalized frequencies (per million words) by Strindberg, his contemporaries (labeled 1800) and present-day authors (1900).

The normalized frequencies of all surface forms of DYRKA ("to worship") and BEUNDRA ("to admire") are shown in the diagram to the left in figure 4. Both words occur more frequently in SUSC than in the reference corpora, written by Strindberg's contemporaries and our present-day authors. In the diagram to the right, we see the same pattern for the negatively charged words AVSKY ("to detest") and FÖRAKTA ("to despise").

3.2.3 Contexts of ÄLSKA

The high frequency of ÄLSKA in SUSC prompts the question in which types of contexts this word is used. We first attempt to answer this question by looking at the sequential word patterns (n-grams) in which ÄLSKA occurs. Such patterns are a means of approximating phrases (c.f., Starcke 2006). We extracted word patterns (consisting of two to five words) from SUSC with the AntConc n-gram tool.

Pronouns occur in four out of the five most frequent two-word patterns with ÄLSKA (ranked by frequency). The only non-pronoun pattern is *att älska* ("to love").

PATTERN	Freq.
Jag/jag/Du/du älskar	110
älskar dig/mig/er	101
älska dig/mig	26
han/jag älskade	18
att älska	18

Table 8: The most frequent two-word patterns of ÄLSKA in SUSC.

If we look for longer patterns, of three or four words, we still find mostly patterns with pronouns ("*jag älskar dig*", "*att jag älskar Er*") with decreasing frequency. If we look for patterns of five words we find only 11 patterns in all of SUSC (each pattern occurring only twice):

1. du att jag älskar dig (2)
2. jag älskar dig! Och du (2)
3. jag älskar Er så att (2)
4. jag älskar honom, men jag (2)
5. måste säga att jag älskar (2)
6. O, vad jag älskar Er (2)
7. sant att du älskar mig (2)
8. skall jag älska dig så (2)
9. älska mig! Ja älskade! älska (2)
10. älskar Er, älskar Er, älskar (2)
11. är sant att du älskar (2)

These sequential patterns show a strong relationship between ÄLSKA and personal pronouns, but we are also interested in non-sequential relationships between ÄLSKA and other lexical items. To this end, we use statistical measures to find *collocations* of ÄLSKA, that is, frequently co-occurring words

within a window of *n* words.

We used the AntConc collocation tool to extract collocations with two measures of collocation strength, Mutual Information (MI) and t-score, on a window of four words to the left, and four words to the right of ÄLSKA. These two measures typically result in somewhat different sets of collocations for the same word on the same data: MI tends to find collocations with rare words, while t-score finds collocations with frequent words. Therefore, following (Church et al., 1994), we select the overlap between the sets of MI and t-score collocations for ÄLSKA (with threshold values of more than 3 for the MI value, and more than 2 for the t-score value). The result is listed below, grouped by word class:

- NOUNS *guds* (”god’s”), *gud* (”god”), *kvinna* (”woman”), *barnet* (”the child”), *barn* (”child”), *kärlek* (”love”)
- PRONOUNS *min* (”my”), *du* (”you”+subj), *dig* (”you”+obj), *din* (”your”+utr), *ditt* (”your”+neutr), *henne* (”her”), *ni* (”you”+subj plur), *er* (”you”+obj plur), *varandra* (”each other”)
- VERBS *tro* (”to believe”), *tror* (”believe”), *vet* (”know”)
- ADVERB *varför* (”why”)
- ADJECTIVES *älskade* (”beloved”), *egen* (”own”), *lilla* (”little”)
- INTERJECTIONS *farväl* (”goodbye”), *tack* (”thank you”), *förlåt* (”forgive”), *ja* (“yes”)

The sequential patterns showed a strong relationship between personal pronouns and ÄLSKA. In addition, the collocations show that objects (denoted by nouns) related to ÄLSKA includes “woman”, “child”, and “god”, and that related actions (denoted by verbs) are “believe” and “know”.

We find that the figures from our study of emotionally charged words in Strindberg’s autobiographical works strongly support the view of Strindberg as an author with an emotionally charged language. In addition, our results also mirror tendencies over time in Strindberg’s emotional life as communicated by the autobiographical works. Findings like this are interesting complements to a closer reading of the texts.

3.3 Study 3: Semantic prosody of KVINNA and MAN

Inspired by the findings in study 1 and 2, which indicate the presence of emotional relations in Strindberg’s autobiographical works, we conducted a study on how Strindberg describes women and men in his texts.

Semantic prosody, first proposed by Louw (1993), can be described as the semantic coloring a word gets by frequent co-occurrence with other words of a specific semantic category (Wynne 2006). For our final study, we explore the semantic prosody of KVINNA (“woman”) and MAN (“man”) in SUSC. First, we extract and analyze compounds with KVINNA and MAN, and second, we extract and categorize the most significant collocations of these words.

3.3.1 Compounds

We used the corpus search tool Xaira to search for compounds with KVINNA (”woman”) or MAN (“man”) as either a prefix or a suffix. To this end, we searched for words tagged as noun and matching one of the following regular expressions (where \b means word boundary):

- KVINNA
  - Prefix: /\b(kvinna|kvinno)/
  - Suffix: /(kvinna|kvinnor|kvinnan|kvinnorna)s?\b/
- MAN

- Prefix: /\b(man|män)/
- Suffix: /(man|män|mannen|männen)s?\b/

This query resulted in lists of concordances, which were further analyzed manually. The frequencies are low, both of each compound and of such compounds as a set, but the tendencies are clear.

Compounds with MAN are almost exclusively lexicalized forms, while there are few lexicalized compounds with KVINNA. The list of lexicalized compounds with KVINNA includes *kvinnofrågan* (lit. ”the woman-question”), *kvinnokroppen* (“the female body”), and *tjänstekvinna* (“female servant”).[12] The list of lexicalized compounds with MAN is longer, and includes e.g., *gentleman*, *hedersman* (lit. “honour-man”, an honorable man), *herreman* (”gentleman”), *ämbetsman* (”official”), *uppsyningsman* (”inspector”), *ålderman* (”elder”), *vetenskapsman* (”scientist”), and *överman* (”master”).

Few of the compounds with MAN are negative; one such example is *avundsman* (lit. “envy-man”, an envious man). In contrast, compounds with KVINNA are emotionally charged and mostly negative, for example:

- *kvinnodyrkan* (lit. ”woman-worship”); *kvinnohat* (”woman-hate”)
- *kvinnodjävul* (lit. ”woman-devil”, ”she-devil”)
- *primitivkvinnans* (lit. ”primitive-woman’s”), *vildkvinnans* (lit. ”wild-woman’s”)
- *kvinnodömet* (lit. ”woman-dome”, the reign of women), *kvinnosidan* (lit. “woman-side”, the party of women)
- *kvinnotungor* (lit. “woman-tounges”)

3.3.2 Collocations and semantic categories of KVINNA and MAN

We extracted collocations of all surface forms of KVINNA and MAN using the AntConc collocation tool (measure MI, a window of four words to the left and right of the node word). Below, the collocations with a collocation measure above 3 are divided into a combination of semantic and grammatical categories.

Collocations of MAN:

- **Personal traits:** *halvung* (“youngish”), *rättfärdige* (“righteous”), *unge* (“young”), *egyptisk* (“Egyptian”), *handlingens* (“the action’s”, as in “man of action”), *vise* (“wise”), *store* (“great”), *hygglig* (“good”), *framstående* (“prominent”), *rike* (“rich”+masc), *lille* (“little”), *hederlig* (“honourable”), *unga* (“young”+pl), *rik* (“rich”), *ung* (“young”), *intelligens* (“intelligence”), *ära* (“honour”)
- **Personal attributes:** *utseendet* (“the appearance”)
- **Feelings:** *vrede* (”wrath”)
- **Family:** *familjeförsörjare* (”family provider”), *äkta* (”wedded”), *kvinnan* (”woman”)
- **Spiritual:** *andans* (as in “man of the cloth”), *Evige* (“Eternal”)
- **Other:** *Plutark*, *brutet* (”broken”), *egenskaper* (”traits”), *osynlige* (”invisible”), *skulden* (”guilt”), *misstankar* (”suspicions”), *tidens* (”time’s”), *kronor* (”money”),
- **Verb:** *gissa* (”guess”), *kurtiserade* (”courted”), *erfarit* (”experienced”), *erfar* (”experience”), *slutade* (”stopped”, ”ended”)
- **Funktion words:** *denne, Denne* (”this”), *per* (as in ”per man”)

Collocations of KVINNA:

- **Personal traits:** *skönaste* (“the most beautiful”), *ful* (“ugly”), *rå* (“raw”)
- **Personal attributes:** *barhuvad* (”bareheaded”), *hår* (”hair”)

[12] *Tjänstekvinna* only occurs in the phrase *tjänstekvinnans son* (“the son of a servant”).

- **Feelings:** *älskade* (”loved”, ”beloved”), *älska* (”to love”), *älskat* (”loved”), *föraktar* (”despise”), *kärleken* (”love”, noun), *känslor* (”feelings”)
- **Family:** *förbindelse* (“relationship”), *maka* (“wife”), *moder* (“mother”), *mannens* (“the man’s”), *skild* (“divorced”, “separated”), *gift* (“married”), *mannen* (“the man”)
- **Spiritual**: *familjehelgd* (“the sanctity of the family”), *korsfästa* (“crucify”), *korsfäst*(“crucified”) , *gud* (“god”), *vördnad* (“awe”)
- **Emancipation:** *oavhängiga* (”independent”), *arbetsmarknad* (”labor market”), *fria* (”free”)
- **Other:** *dubbla* (”double”), *tavla* (”painting”), *sidor* (”sides”), *värd* (”entitled to”), *talan* (”say”), *blod* (”blood”)
- **Verb:** *ingick* (“entered into”), *funnits* (“has been”), *gällde* (“related to”), *född* (“born”) , *tillhörde* (“belonged to”), *stiger* (“to rise”)
- **Funktion words:** *allmänhet* (part of idiom: ”in general”), *vissa* (”some”)

The semantic categorization of the collocations for MAN shows that there are many, mainly positive, collocations describing the personal traits of MAN: e.g., *rättfärdige* (“righteous”), *vise* (“wise”), *store* (“great”), *framstående* (“prominent”). The verbs indicate that MAN is an agent of courtship and experiences. A number of collocations of MAN are associated with negative feelings, e.g., *vrede* (“wrath”), *skulden* (“guilt”), *misstankar* (“suspicions”). The collocations associated with family describe the role of the man as family provider and husband.

In contrast, there are few and mainly negative collocations describing the personal traits of KVINNA, related to appearance (*skönaste*, “the most beautiful”; *ful*, “ugly”)), and behavior (*rå*, “raw”). A long list of collocations describing feelings are related to KVINNA, mostly positive (forms of ÄLSKA) but also negative (DESPISE). Strongly associated with KVINNA are words describing family and relationships, both nouns (*förbindelse, maka, moder, skild, gift, mannen*) and verbs (*ingick, född, tillhörde*).

In summary, collocations of MAN shows the man as an active experiencer defined through personal traits, while the collocations of KVINNA shows the woman as defined through relationships and family, with personal traits relating to beauty (or the lack thereof). These are the semantic prosodies of a small set of words in a selection of auto-biographical works by Strindberg; a fuller understanding of what they mean requires further analysis into both the life and times of Strindberg.

4 Concluding remarks

In this paper we have presented the Stockholm University Strindberg Corpus, a new resource consisting of seven of Strindberg’s autobiographical works. This is our main contribution. In addition, we have carried out three corpus studies on SUSC using keywords, word distribution, and semantic prosody. Our findings support a prevalent view of Strindberg as an author using an emotionally charged language, with intense relationships to both men and women. A close reading of the texts reveals the emotions expressed in Strindberg’s texts; by complementing with corpus techniques we can also say something about the extent and degree of such expressions.

References

Kenneth Ward Church, William Gale, Patrick Hanks, Donald Hindle & Rosamund Moon. 1994. Lexical substitutability. In: *Computational approaches to the lexicon*, 153-177. Oxford: Oxford University Press.

Ted Dunning. 1993. Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, 19(1):61-74.

Gunnel Källgren. 2006. *SUC 2.0* (eds.) Sofia Gustafson-Capková & Britt Hartmann, Department of Linguistics, Stockholm University.

Bill Louw. 1993. *Irony in the Text or Insincerity in the Writer? The Diagnostic Potential of Semantic Prosodies.* In Baker, M., Francis, G. & Tognini-Bonelli, E. (eds): *Text and Technology*. Philadelphia/ Amsterdam: John Benjamins.

Rosamund Moon. 2007. Words, frequencies, and texts (particularly Conrad): A stratified approach. Journal of Literary Semantics*, 34(1)*:87-104.

Kieran O'Halloran. 2007. The subconscious in James Joyces 'Eveline': a corpus stylistic analysis that chews on the `Fish hook'. Language and literature, 16(3):227-244.

Michael Robinson. 2008. *An International Annotated Bibliography of Strindberg Studies 1870-2005*. (Vol. 1 General Studies, Vol. 2 The Plays, Vol. 3 Autobiographies, Novels, Poetry, Letters, Historical Works, Natural History and Science, Linguistics, Painting and the Other Arts, Politics, Psychopathology, Biography, Miscellaneous, Dissertations), MHRA.

Bettina Starcke. 2006. The phraseology of Jane Austen's Persuasion: Phraseological units as carriers of meaning. ICAME Journal. Computers in English Linguistics, 30:87-104.

Michael Stubbs. 2005. Conrad in the computer: examples of quantitative stylistic methods. Language and Literature, 14(1):5-24.

Martin Wynne. 2006. Stylistics: Corpus Approaches. In: *Encyclopedia of Language and Linguistics*, pp. 223–226. Oxford: Elsevier, 2nd edition.

Robert Östling. 2012. Stagger: A modern POS tagger for Swedish. In: Proceedings of the Fourth Swedish Language Technology Conference, October 24-26 2012, Lund, Sweden.