Gender-Based Vocation Identification in Swedish 19th Century Prose Fiction using Linguistic Patterns, NER and CRF Learning

Dimitrios Kokkinakis
Department of Swedish, Språkbanken
University of Gothenburg
Sweden
dimitrios.kokkinakis@gu.se

Ann Ighe
Economic History, School of Business Economics and Law
University of Gothenburg
Sweden
ann.ighe@gu.se

Mats Malm
Department of Literature, History of Ideas and Religion
University of Gothenburg
Sweden
mats.malm@lir.gu.se

Abstract

This paper investigates how literature could be used as a means to expand our understanding of history. By applying macroanalytic techniques we are aiming to investigate how women enter literature and particularly which functions do they assume, their working patterns and if we can spot differences in how often male and female characters are mentioned with various types of occupational titles (vocation) in Swedish literary texts. Modern historiography, and especially feminist and women's history has emphasized a relative invisibility of women's work and women workers. The reasons to this are manifold, and the extent, the margin of error in terms of women's work activities is of course hard to assess. Therefore, vocation identification can be used as an indicator for such exploration and we present a hybrid system for automatic annotation of vocational signals in 19th century Swedish prose fiction. Beside vocations, the system also assigns gender (male, female or unknown) to the vocation words, a prerequisite for the goals of the study and future in-depth explorations of the corpora.

1 Introduction

Can we use literary text as a (valid) source for historical research? Evidence shows that the answer is probably yes and in this paper we investigate how literature can be used as a means to expand our understanding of history; (Rutner & Schonfeld, 2012). This paper presents a system for the automatic annotation of vocational signals in Swedish text, namely 19th century prose fiction. *Vocation* in this context is defined as a single word or a multi word expression intended to capture the (professional) activities with which one occupies oneself, such as employment or other, wider, forms of productive occupations not necessarily paid. Therefore *vocation* is used here in a rather broad sense since we do not want to disallow word candidates that might not fit in a strict definition of the term. Apart from vocation identification, the described system recognizes and assigns gender, i.e. male, female or unknown, to the vocations by using various Natural Language Processing (NLP) technologies.

The purpose of this work is to use literature as means to expand our understanding of history (Pasco, 2014) by applying macroanalytic techniques (Moretti, 2013; Jockers, 2013) in order to start exploring how women enter literature as characters, which functions do they assume and their working patterns. The research questions themselves are not new, but in fact central to the field of gender studies and to a certain extent, economic history. From a historical point of view, the 19th century in Sweden, and several other western countries, is a period with a dramatic restructuring of gender relations in formal institutions such as the civil law, and also a period where the separation of home and workplace came to redefine the spatial arenas for human interaction. Singular works of fiction can be analyzed and interpreted in historical research and current development in digital humanities certainly opens new possibilities in this direction. Therefore, vocation identification can be used as one such indicator for achieving some of the above stated goals. The starting point of this study has been to create an infrastructure of suitable lexical resources and computational tools for empirical NLP in the cultural heritage domain and digital humanities area. Supporting future users of digitized literature collections

Proceedings of NAACL-HLT Fourth Workshop on Computational Linguistics for Literature, pages 89–97,
Denver, Colorado, June 4, 2015.

with tools that enable the discovery and exploration of text patterns and relationships, or even allow them to semantically search and browse (Don et al., 2007; Vuillemot et al., 2009; Oelke et al., 2013), using computer-assisted literary analysis with more semantically oriented techniques, can lay the foundations to more distant reading or macroanalysis of large corpora in novel ways, impossible to investigate using traditional qualitative methods or close reading.

2 Background

Digital humanities is an umbrella term used for any humanities research with potential for real interdisciplinary exchange between various fields and can be seen as an amalgamation of methodologies from traditional humanities disciplines (such as literature and art, corpus linguistics), and social sciences, with computational approaches and tools provided by computer science (such as text and data mining) and digital publishing. During the last couple of decades there has been a lot of research on applying automatic text analytic tools to annotate, enrich, explore and mine historical or other digital collections in various languages and for several reasons (Pennacchiotti & Zanzotto, 2008; Mueller, 2009; Manning, 2011; Piotrowski, 2012; Jockers, 2013; McEnery & Baker, 2014). The focus of such research is to reduce the time consuming, manual work that is often carried out e.g. by historians or other literature scholars, in order to identify valid, useful and meaningful results such as semantic associations, gender patterns and features of human networks (Agarwal et al., 2012). Also, recently, a small number of studies have been published where gender and other biographical characteristics are explored (Hota et al., 2006; Argamon et al., 2007; Garera & Yarowsky, 2009; Bullard & Oveesdotter Alm, 2014). These methods apply various types of classifiers with good performance results.

Boes (2014) discusses the content of the "Vocations of the Novel Project" which consists of a database of roughly 13,000 German-language prose works, published between 1750-1950, and in which each entry in this database is tagged with vocational metadata identifying occupations that receive extended narrative treatment. Fifteen occupational clusters, such as *agricultural professions*, *health* and *nautical professions*, are used for estimating the proportional distribution of those with the database content, showing for instance that members of the *clergy* diminished after about 1885 or that agricultural professions first declined in importance but then become to rise around the turn of the century, after which they rapidly sank again. However, even closer to our goals is the research by Pettersson & Nivre (2011); Fiebranz et al. (2011) and Pettersson et al. (2012; 2014), who in cooperation with historians, study what men and women did for a living in the early modern Swedish society ("The Gender and Work project", GaW) between 1550-1800. In the context of GaW's verb-orientated studies, historians are building a database with relevant information, by identifying working activities often described in the form of verb phrases, such as *chop wood* or *sell fish*. Automatically extracted verb phrases from historical texts are presented to historians as a list of candidate phrases for revision and database inclusion. Since Swedish NLP tools for that period are very scarce, the historical texts are first normalized to a more modern spelling, before tagging and parsing is applied – the techniques applied in GaW are different and thus complementary to the ones we apply in the research described in this paper.

3 Material

The textual data we use in this work is the content of an 18-19th century Swedish Prose Fiction database – Spf[1]. Spf is comprised by ca 300 prose works that were originally published as books during the years 1800, 1820, 1840, 1860, 1880 and 1900. The material is representational of its times in ways that the canonized literatures are not, in the sense that it contains not only canonized literature but mainstream as well as marginalized treatments of 19th century society. The database makes it possible to examine a particular year span and compare it to the material of other years in order to obtain a comprehensive view of societal development across an entire century. However, the main part of this work deals with the construction and adaptation of several lexical and semantic resources developed for modern Swedish to the language of Spf and algorithmic resources that use those for automatic labeling, and we have left as future work the fine-grained comparison between different time spans.

Furthermore, there are several classificatory systems available where occupations are organized into clearly defined sets of groups according to the tasks and duties undertaken in the job;

[1] <http://spf1800-1900.se/#!/om/inenglish>.

such as the *International Standard Classification of Occupations*[2] or the *National Occupational Classification*[3] which is the nationally accepted reference on occupations in Canada. Such classifications are basically structured by skill type, e.g. *Agricultural, Animal Husbandry and Forestry Workers, Fishermen and Hunters* which can further increase the usability of such resources. However, in this study we did not have the human resources to structure the collected occupations in a finer-grained manner, apart from some very basic and coarse encoding (see further Section 3.1); therefore this task is left as a future work. A large number of vocation and other related terms from three lexically-oriented resources were collected and structured. As our starting point we used ca 3,500 lexical units found in relevant frames, such as *Medical_professionals* and *People_by_origin*, from the Swedish *FrameNet*[4]. Moreover, several other lexical units in related frames, that are slightly more general but still relevant, were used, i.e. entries that belong to other types of both generic and more specific frames that indicate person activities, relationships or qualities of various kinds, such as *Kinship* or *Performer*. Secondly, we used several hundred of vocation names from the Swedish dictionary *Svensk Ordbok*[5] 'Swedish Dictionary' (SO) and finally, several thousand vocation names from the *Alfabetiskt yrkesregister* 'Alphabetically list of professional designators' published by the Statistics Sweden[6].

3.1 Vocation Lexicon Structure

Semi-automatically, all lexicon entries were assigned two features. The first one was *gender* (i.e. Male, Female or Unknown) and the second one *Vocation.* Depending on the content and structure of the three resources we used to extract occupations and similar entries from, we tried to keep and encode any kind of relevant to our goals descriptive information for these entries in tab-separated fields. For our study we only use *Vocation* and combinations with *Vocation* and other labels. For instance, in FrameNet, vocation related frames such as *Medical_professionals* (a label that was transformed to a more generic and shorter one, *Health*, and which consists of single and compound lexical units for health-related occupations) was encoded using both *Vocation* and *Health*, e.g. *patolog* 'pathologist' or *sjuksköterska* 'nurse'. Similarly other combinations of FrameNet-originating labels were extracted and encoded in a similar manner. Since we adopted a broad definition of the term *Vocation* we allow such words to be included in the knowledge base, but not all were used for the study described here if there were not directly vocation-related. For instance, *tjuvpojke* 'thief boy' is coded as *Morality-negative*; here *Morality-negative* is of course not a vocation but rather a general human quality and not used in the study. Also words with the label *Person*, which is the most generic category, including mentions such as *baldrottning* 'prom queen'; *äventyrerska* 'adventuress' or *söndagsskoleelev* 'Sunday school student', are not used in the study presented in this paper.

Gender assignment is based on reliable orthographic features of the lexicon entry in question (if available), these include:

- typical gender bearing morphological suffixes, such as –inna [female] *värdinna* 'hostess', *prestinna* 'priestess'or –rska [female] *ångfartygskokerska* 'steamboat's cook', *städerska* 'cleaning woman'.
- gender bearing head noun words that unambiguously can assign gender in compound word forms; for instance *hustru* 'wife' [female]: *soldathustru* 'soldier's wife', *bagarehustru* 'baker's wife'; or *dräng* [male] 'farmhand': *stalledräng* 'stable farmhand', *fiskardräng* 'fisherman farmhand'.

After consulting international efforts in the area, we automatically normalized and attempted to group together and harmonize the labels of all available lexical units (these labels are primarily encoded in the Swedish FrameNet) to the following 14 single types and 4 complex ones, without putting too much effort to introduce finer-grained types for practical reasons (see the previous section for discussion on this issue). Thus, the final set of categories we applied are: *Age* (*yngling* 'youth'), *Expertise-Neg* (*okunnige* 'ignorant'), *Expertise-Pos* (*specialist*), *Jurisdiction* (*invånare* 'resident'), *Kinship* (*dotterdotter* 'granddaughter'), *Morality-Neg* (*tjuv* 'thief'), *Morality-Pos* (*nationalhjälte* 'national hero'), *Origin* (*jugoslav* 'yugoslavian'), *Person* (*vän* 'friend'), *Politics* (*socialdemokrat* 'social-democrat'), *Religion*

[2] <http://www.ilo.org/ public/english/bureau/stat/isco/>.
[3] <http://www5.hrsdc.gc.ca/NOC/Eng lish/NOC/ 2011/OccupationIndex.aspx>.
[4] <http://spraakbanken.gu.se/eng/swefn>.
[5] <http://www.svenskaakademien.se/publikationer/bocker_om_svenska_spraket>.
[6] <http://www.scb.se/>.

(*metodistpredikant* 'methodist preacher'), *Residence* (*granne* 'neighbour'), *Vocation* (*bussförare* 'bus driver'), *Vocation+Health*, *Vocation+Military* (*kavallerilöjtnant* 'cavalry lieutenant'), *Vocation+Performer* (*dragspelare* 'accordionist') and *Person+Disease* (*autistiker* 'autistic person'). The resulting lexicon consists of 19,500 terms, of which over 77% (15,000) are distinct occupational titles (vocations), and used in various ways by the system, mainly as the core lexicon for rule based pattern matching and as a feature for supervised machine learning (see Section 4.4). Moreover, 75% (or 11,000) of all these vocations in the lexical resources have been assigned *Unknown* gender as a default since no classificatory orthographic features, as previously described, could be applied for that purpose.

4 Methods

Figure 1 provides a general outline of the methods applied in the study.

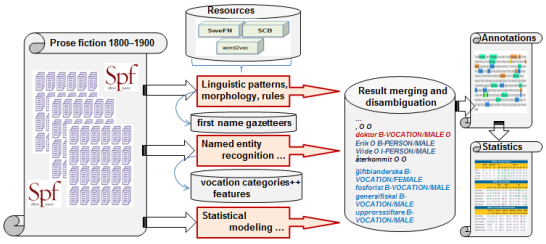

Figure 1. The major steps used to extract gender-bearing vocations from raw text.

4.1 Linguistic Patterns: Morphology, Compounding and Continuous Space Word Representations

The previously outlined lexicon[7] is used for pattern matching based on a number of manually coded rules that explore regularities in the surrounding context of potential vocation words. Inflectional morphology is determined programmatically, while the lexicon's content is also used for discovering "new" terms (not in the static lexicon) using compound segmentation and matching of the head of a candidate vocation to the content of the knowledge base.

For instance, a potential new term candidate, that is a word over 6 characters long not in the lexicon, is de-compounded and its head is matched against the lexicon's content. If a match is found, the new term gets the vocation annotation of the matched head. The length of six characters is determined after testing with other lower values and six is the lowest number that can be safely used and which minimizes the number of false positives returned to a minimum. Suppose *prestinna* 'priestess' is in the lexicon with features *Vocation* and *Female* and a new word, over six characters long, e.g. *öfverprestinna* 'head priestess', is found in a text, *öfverprestinna* will be then decomposed to *öfver+prestinna* and the head *prestinna* will match an existing lexicon entry with the same form; consequently *öfverprestinna* will inherit the annotation *Vocation* and *Female*. Alternative surface text forms, e.g. with hyphenation (*solo-sångare* 'solo singer' versus *solosångare*) are treated in a similar manner, breaking the compound at the hyphen. This processing allows a set of potential new terms to be efficiently recognized, while the results using the lexicon based pattern matching approach show nearly perfect precision scores. The recognition step is made using case insensitive matching against the lexicon's content.

Furthermore, we also experimented with continuous space word representations (Mikolov et al., 2013) in order to extract, manually review and incorporate lists of near synonyms to vocation-specified words. For instance, the top-10 closest words for the word *soldat* 'soldier' were: *bonde* 'farmer', *officer*, *simple** 'simple', knekt 'foot soldier', *tapper** 'brave', *sjöman* 'seaman', *adelsman* 'nobleman', *munk* 'monk', *duktig** 'capable' and *matros* 'sailor'; only three of these words, marked with '*', are not directly associated with vocations. This experiment resulted into roughly a hundred of new vocation words integrated in the knowledge base.

4.2 Person Entity Recognition

During processing we also use named entity recognition (NER) (Borin et al., 2007), but only a component that deals with person entity identification. Since gender is a prominent attribute for a very large number of *first names*, we apply a NER component that uses a first name gazetteer with 21,000 first names, in which each name has been pre-assigned gender and thus used to assign gender to recognized person first names. For instance the NER processing of the sentence: *Jan och Johan skulle just gå in i stugan, då Maja ropade dem tillbaka* 'Jan and John were about to go into the cottage, when Maja called them back' will recognize three first names (*Jan*, *Johan* and *Maja*) and assign male gender to the first two,

[7] All lexicon entries annotated with the Vocation label are available from: <http://spraakbanken.gu.se/swe/personal/dimitrios#research>.

Jan and *Johan* and female to the last one, *Maja*. Thereafter, the results obtained during the application of the method described in Section 4.1 and the results from the NER will be merged, and a post-NER pattern matching script will try to assign gender to vocation words for which gender is marked as *Unknown* by the process described in 4.1 and there is a first name annotation close by. This is accomplished under the condition that the NER has assigned a gender to a first name in the near context of a "genderless" vocation. For instance, a vocation word, for which its surface characteristics does not reveal any gender affiliation according to the vocation lexicon, can be assigned appropriate gender if a recognized first name appears in its near context; e.g. *bonden Petter* '(the) farmer Petter' or *Gusten är en fiskare* 'Gusten is a fisherman'. In these examples *bonden* and *fiskare* are coded in the knowledge base as vocation words with unknown gender and the process outlined in 4.1 recognizes it as such. *Petter* and *Gusten*, on the other hand, are recognized by the NER as human with *male* gender. The gender attribute will be then propagated to the vocation words *bonden* and *fiskare* which will get the same gender as its appositive *Petter* in the first case and the person entity's gender close by in the second case.

4.3 Local Context Regularities

Since not all vocation annotations get gender assignment during recognition, we use hand-coded rules based on various lexical patterns for that purpose. The heuristics applied to these rules include four major types of reliable information: personal pronouns, gender-bearing adjectives, gender-bearing suffixes and certain forms of local context:

- personal pronouns, e.g. the Swedish *hans* 'his' and *hennes* 'her', are used for gender assignment if they appear in a context very close to a vocation (1 to 5 tokens); e.g. in the text fragment …*fiskaren och hans barn* '… the fisherman and his children', *fiskaren* is identified as a vocation but with unknown gender which at this stage will be assigned male since the pronoun *hans* is male and refers to the *fiskaren* 'fisherman'; while in the text fragment …*hon var en agitator* '… she was an agitator', *agitator* is identified as a vocation but with unknown gender which at this stage will be assigned female since the pronoun *hon* is female referring to *agitator*
- historical forms of Swedish adjectives used to be gender bearing; e.g. the majority of adjectives ending in *-e* designate male gender. For example, *fattige bonden* 'the poor farmer', here *bonden* is identified as a vocation with unknown gender which will be assigned male since the adjective *fattige* is indicating a *male* noun
- similarly to the process described in Section 4.1, we also here take advantage of the fact that many noun suffixes or head words of compounds are also gender bearing; e.g. suffixes *-erska* or *-inna* designate female gender; e.g. *tvätterska* 'laundress' or *värdinna* 'hostess' are assigned female gender because of their gender bearing suffixes. Gender bearing head words are also used for gender assignment; e.g. compounds ending in *-fru* 'wife' such as *bondefru* [bonde+fru] 'peasant wife' will be assigned female gender; while compounds ending in e.g. *karl* 'man' such as *besättningskarl* [besättning+s+karl] 'crew man' will be assigned male gender
- local context is also used to merge two or more consecutive vocation and related annotations into one; typically when a genitive form of a noun precedes another noun. For instance, the text snippet: *ryttmästarns betjent* '[the] rittmeister's servant' will initially receive two vocation annotations (with unknown gender) that will be merged into one [*ryttmästarns+betjent*] which unknown gender; while the snippet *drottningens hofmästarinna* '(the) queen's hofmeisteress' will initially receive two vocation annotations (with female gender) that will be merged into one [*drottningens+hofmästarinna*] with female gender.

4.4 Statistical Modeling

Finally, we also use a complementary statistical modelling method, conditional random fields (CRF) for learning gender-assigned vocations in combination with the results of the rule-based system and the NER (the vocation words together with basic features such as n-grams and word shape were used as features for training the learner). For that purpose we use the Stanford CRF off-the-shelf software (Finkel et al., 2005). The purpose of the CRF is to identify vocations and (possibly) correct gender not captured by the previous techniques, in order to increase recall. Training and testing is based on a pre-annotated and manually inspected sample (by the first au-

thor). This sample was randomly selected from Spf and it was first automatically annotated by the rule based and NER components, and then sentences with at least one annotation were selected, manually inspected, corrected and used for training (390,000) and testing (50,000).

5 Results, Evaluation and Analysis

The fact that a large number of vocations in the lexical resources have been assigned *Unknown* gender implies that the computational processing requires to heavily relying on a (wider) context to assign proper gender to these words. This is a serious drawback since there is not always reliable near context, e.g. at the sentence level, that can be used. This fact is mirrored on the results of e.g. the CRF classifier. Different complementary techniques have been tested, but still, a large number of vocations remains with unknown gender. More elaborated ways are probably required to identify gender, perhaps using discourse information, e.g. CRF is used with default features, new features might be necessary to test.

Female	Male	Unknown
hustru [wife] (2016)	kung [king] (976)	präst [priest] (1341)
majorska [majoress] (1054)	prost [provost] (614)	kapten [captain] (532)
grefvinna [countess] (805)	brukspatron [ironmaster] (582)	löjtnant [lieutenant] (487)
jungfru [maid] (698)	patron [land tenure] (430)	major (382)
grevinna [countess] (593)	kyrkoherde [vicar] (325)	bonde [farmer] (343)
överstinna [colonel's wife] (308)	konung [king] (258)	*don (322)
prostinna [minister's wife] (304)	biskop [bishop] (242)	tjänare [servant] (283)
drottning [queen] (274)	grefve [count] (227)	överste [colonel] (241)
hushållerska [housekeeper] (260)	baron (215)	tiggare [beggar] (236)
prästfru [minister's wife] (218)	kejsare [emperor] (208)	husbonde [master] (217)

Table 1. The top-10 most frequent lemmatized vocations (and their occurrences) in the *Selma Lagerlof Archive* ('*': error).

Table 1 above shows the top-20 occurrences of three automatically extracted lists of male, female and unknown gender vocations from the *Selma Lagerlof Archive* (a collection of the author's works of fiction published as books during her lifetime), a completely new corpus, not used for the development of the resources in the study with 3.341.714 tokens.

5.1 Evaluation of the CRF

The results of the classifier's evaluation[8] are given in tables 2 and 3. Low recall scores can be possibly attributed to two facts; one is the amount of test and training texts used for training the classifier and second the use of default features for training. Addition of new features, such as part-of-speech, syntactic and/or co-reference links could have possibly being beneficial, including larger training corpora. Moreover, various types of errors could be identified during all stages of processing. With respect to the CRF component evaluation, most of the errors had to do with the occurrences of e.g. male designated adjectives, such as *svenske* 'Swedish' or *danske* 'Danish'. A number of last names and also common words that were homographic to vocations were also annotated erroneously as such; for instance the last names *Skytte* 'Shooter' (e.g. in the context *Malin Skytte*) and *Snickare* 'carpenter' (e.g. in the context *Gorius Snickare*); and common nouns such as *simmaren* '(the) swimmer' or *vakt* 'guard' in idiomatic contexts such as *hålla vakt* 'be on one's guard'.

	Precision	Recall	f-score
B[9]	96.33%	58.22%	72.57%
I	89.09%	72.06%	79.67%

Table 2. Precision, recall and f-scores for the CRF learner.

	Precision	Recall	f-score
F	97.03%	43.75%	60.31%
M	94.52%	55.44%	69.89%
U	53.25%	45.00%	48.78%

Table 3. Precision, recall and f-scores for the CRF learner on gender (M=male; F=female; U=unknown).

Another type of important omission has been the fact that a large number of vocations are assigned

[8] For the evaluation we used the *conlleval* script, vers. 2004-01-26 by Erik Tjong Kim Sang.

[9] We use the IOB tags for file representation: I (inside), O (outside), or B (begin). A token is tagged as B if it marks the beginning of a chunk. Subsequent tokens within the chunk are tagged I. All other tokens are tagged O. E.g. the context *Fru majorskan skrek till majorn…* 'Mrs. major's wife shouted to the major ...' is represented as:

Fru B-Female
majorskan I-Female
skrek O
till O
majorn B-Male ...

unknown gender since there is no reliable context (at the sentence level) that could be used. Moreover, we have been restrictive to the gender assignment of certain vocations in the resources, although, in principle, and considering the nature and publication time of the texts, we could by default assign gender to a large number of these vocations. For instance, a large number of military-related vocations, such as *löjtnant* 'lieutenant' or *generalmajor* 'major general' are assigned unknown gender, although these, predominantly, refer to males in the novels. Moreover, identical singular and plural forms of vocation terms are yet another difficult problem, e.g. *politiker* 'politician' or 'politicians' or *spritfabrikarbetare* 'distillery worker' or 'distillery workers'. Some sort of linguistic pre-processing, such as idiom identification and part of speech annotation, could probably exclude word tokens in plural form (or verbs in that matter, but such cases were extremely rare in our data), nevertheless part of speech tagging is not used at the moment. Also, more elaborative models could be used to first determine who the personal pronouns refer to before an attempt could be made to assign the pronoun's gender to a vocation word with unknown one.

5.2 Evaluation of the Knowledge-based Components

Besides the evaluation of the CRF learner, we also conducted an analysis on a small random sample of similar text from different, but comparable, corpora, in order to investigate the contribution of the different components for the vocation identification. A selection of a randomized subset of 1000 sentences from two sources was conducted from the *August Strindberg's Collected Works* ("August Strindbergs Samlade Verk") and the *Selma Lagerlof Archive* ("Selma Lagerlöf-arkivet"), both parts of the Swedish Literature Bank[10]. These 1000 sentences were automatically annotated by: i) the rule-based system without any sort of disambiguation or other processing only lexicon look-up; this can be considered as a baseline system where only inflectional morphology is considered; and ii) all the rest without the CRF. That included the rule-based system *with* the use of lexical patterns for disambiguation, compound segmentation and the named entity recognition.

A total of 341 of vocation identifiers could be manually recognised and confirmed the assumption that best results are produced by using all available resources at hand. Moreover, the precision of the rule-based system (i.e. the lexicon lookup) is very high. Out of the 341 possible vocations, the baseline, i.e. the rule-based system without any sort of disambiguation, compound analysis etc., identified 329 vocations (46% with the correct gender and the rest with unknown gender); 12 (most of them compounds such as *klädessömmerska* 'clothing dressmaker') and a few others such as *penitentiarius* 'confessor', could not be found and 15 tokens were annotated as vocations but were wrong. These 15 wrong ones originate from (possibly) inappropriate lexicon vocation entries, entries that shouldn't have entered the lexicon as *vocations*, such *Sachsare* 'a person from Saxony' *befrämjare* 'promoter' (a borderline case) and homographs with proper names, such as *Jarl* (which is a title given to members of medieval royal families before their accession to the throne, but also used as a last name).

The combination of all available tools and resources improved these figures; marginally on the gender but substantially on the recognition of the compounds. All vocation compounds were recognized (10) and also four more that were wrong, such as *Nekropolis* 'Nekro+polis' (since *polis* 'police' is in the lexicon) and Notre-Dame 'Notre-+Dame' (since *dame* 'female equivalent of the honour of knighthood' is in the lexicon as well). These compounds could be identified because of the compound decomposition step and matching of the compounds' heads to the lexicon content. Compared to the baseline results the percentage of vocations with correct gender raised to 49.6%.

6 Conclusions

Women's history has emphasized a relative invisibility of women's work and women workers. The reasons to this are manifold, and the extent, the margin of error in terms of women's work activities is of course hard to assess, e.g. work wasn't a tax base also work wasn't the point of departure for a collective political interest (i.e. labour) as it later developed into; while the political organisation also excluded women from some certain areas and particularly from formal authority. This means that women were to a lesser extent mentioned, authorised, appointed or nominated in formal sources. This is obviously

[10] Information about the Literature Bank is here: <http://litteraturbanken.se/#!/om/inenglish>.

the case for many but not all traditional sources, population registers, cameral, and fiscal sources. However, we still don't have good reasons to believe that this means *either* that women didn't work *or* that other types of material couldn't be more rewarding. And the suggestion of this paper is that prose fiction is still possible to utilise further and that the methodological developments in digital humanities should be tried in this endeavour, e.g. by investigating women's work and economic activities represented longitudinally in prose fiction and the differences between the texts of female, male (and unknown) authors, in all those aspects. In this work we have applied automatic text analytic techniques in order to identify vocation signals in 19th century Swedish prose fiction. Our goal has been to reduce the time consuming, manual work that is usually carried out by historians, literature scholars etc., in order to e.g. identify and extract semantically meaningful information such as gender patterns and semantic associations. Literature is a comprehensive source for data on employment and occupation, economy and society, and to e.g. an economic historian or gender researcher such data can be of immense value, particularly for the period between 1800-1900 since gender relations in and through work is a long-standing problem due to repeated underestimation of women's work attributed to among other compelling reasons, the systematic under-reporting of women's work in the used sources (Humphries & Sarasua, 2012; Ighe & Wiechel, 2012). Prose fiction does not necessarily have the same limitations and can be utilized as a fruitful point of departure.

For future work we would like to explore even in more detail the variation of both the performance of the processing steps and also compare the results across time periods and authors' gender. Deeper analysis could provide interesting insights on the nature of which types of person activities are used by different authors or compare and explore other types of collections[11] from the same period, and thus confirm or reject established hypotheses about the kind of vocabulary used; e.g. do male authors use more vocation or kinship labels?

[11] Such collections could be the *Dramawebben* (<http://www.dramawebben.se/>), i.e. digital versions of over 500 plays of Swedish drama from the 1600s to modern times; or the *Digidaily*, i.e. digitized Swedish newspapers from the 18-19th century (<https://riksarkivet.se/digidaily>).

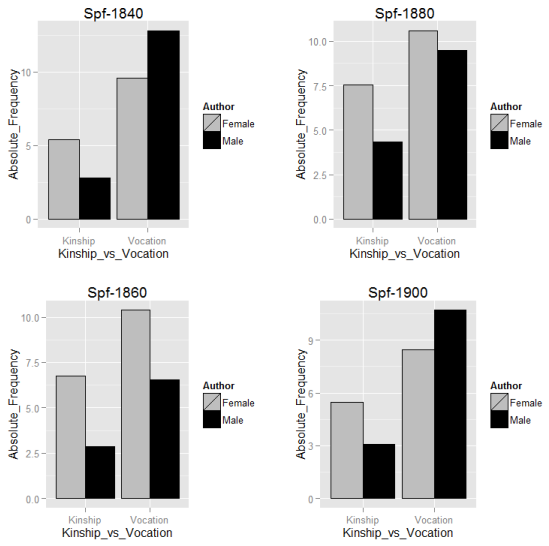

Figure 2. Comparison male and female authors (based on absolute frequencies) of the *Vocation* and *Kinship* categories during the period 1840-1860-1880-1900.

As Fig. 2 shows, other types of investigations are possible and the analysis can provide information about the kind of vocabulary used by various authors during different periods; e.g. do male authors use more vocation than kinship labels and of which type? Kinship (section 3.1) is one of the categories already encoded in the knowledge base and can be easily used to compare the style of authors diachronically.

Acknowledgements

This work is partially supported by the Swedish Research Council's framework grant "Towards a knowledge-based culturomics" dnr 2012-5738.

References

Apoorv Agarwal, Augusto Corvalan, Jacob Jensen and Owen Rambow. 2012. Social Network Analysis of Alice in Wonderland. *Workshop on Computational Linguistics for Literature*. Pp 88–96, Montréal, Canada.

Shlomo Argamon, Russell Horton, Mark Olsen and Sterling Stuart Stein. 2007. Gender, Race, and Nationality in Black Drama, 1850-2000: Mining Differences in Language Use in Authors and their Characters. *Digital Hum*. Pp. 8-10. U. of Illinois.

Tobias Boes. 2014. The Vocations of the Novel: Distant Reading Occupational Change in 19th Century German Literature. *Distant Readings – Topolo-*

gies of German Culture in the Long 19th Century. Erlin&Tatlock (eds). Pp. 259-283. Camden House.

Lars Borin, Dimitrios Kokkinakis and Leif-Jöran Olsson. 2007. Naming the past: Named entity and animacy recognition in 19th century Swedish literature. *1st LaTeCh*. Pp. 1-8. Prague.

Joseph Bullard and Cecilia Oveesdotter Alm. 2014. Computational analysis to explore authors' depiction of characters. *Proc. of the 3rd Workshop on Computational Linguistics for Literature.* Pp. 11–16. Gothenburg, Sweden.

Anthony Don et al. 2007. Discovering interesting usage patterns in text collections: Integrating text mining with visualization. *16th ACM Conf. on info & knowledge management (CIKM).* Pp. 213-222.

Rosemarie Fiebranz, Erik Lindberg, Jonas Lindström and Maria Ågren. 2011. Making verbs count: the research project 'Gender and Work' and its methodology. *Scan Econ Hist Rev*. 59:3, pp. 273-293.

Jenny Rose Finkel, Trond Grenager and Christopher Manning. 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. *43rd ACL*. Pp. 363-370.

Nikesh Garera and David Yarowsky. 2009. Modeling latent biographic attributes in conversational genres. *47th Annual Meeting of the ACL and 4th IJCNLP*. Pp. 719–718. Singapore.

Sobhan Raj Hota, Shlomo Argamon and Rebecca Chung. 2006. Gender in Shakespeare: Automatic stylistics gender character classification using syntactic, lexical and lemma features. *Chicago Colloq. on Digital Humanities and Computer Science* (DHCS). Pp. 100-106. U. of Chicago, USA.

Jane Humphries and Carmen Sarasua. 2012. Off the Record: Reconstructing Women's Labor Force Participation in the European Past. *Feminist Economics*. Vol. 18:4. Taylor and Francis.

Ann Ighe and Anna-Helena Wiechel. 2012. Without title. The dynamics of status, gender and occupation in Gothenburg, Sweden, 1815-45. *Eur. Social Science History Conf.*. Glasgow, Scotland.

Matthew L. Jockers. 2013. *Macroanalysis - Digital Methods and Literary History. Topics in the Digital Humanities*. University of Illinois Press.

Christopher Manning. 2011. Natural Language Processing Tools for the Digital Humanities. Available at <http://nlp.stanford.edu/~manning/courses/DigitalHumanities/> (Visited 20150105).

Tony McEnery and Helen Baker. 2014. The Corpus as Social History - Prostitution in the 17th Century. *Exploring Historical Sources with LT: Results / Perspectives CLARIN Workshop*. Den Haag, The Netherlands. <https://www.clarin.eu/sites/default/files/Mcenery_DenHaag.pdf> (Visited 20150222)

Tomas Mikolov, Wen-tau Yih and Geoffrey Zweig. 2013. Linguistic Regularities in Continuous Space Word Representations. *Proc. of NAACL HLT*. Pp. 746–751. Atlanta, USA.

Franco Moretti. 2013. *Distant Reading*. Verso.

Martin Mueller. 2009. Digital Shakespeare, or towards a literary informatics. *Shakespeare*. 4:3, 284-301, Routledge.

Daniela Oelke, Dimitrios Kokkinakis and Daniel Keim. 2013. Visual Literature Analysis: Uncovering the dynamics of social networks in prose literature. *15th Eurographics Conf. on Visualization (EuroVis)*. Pp. 371-380. Leipzig, Germany.

Allan H. Pasco. 2004. Literature as Historical Archive. *New Literary History*. Vol. 35:3, pp 373-394. John Hopkins University Press.

Marco Pennacchiotti and Fabio M. Zanzotto. 2008. Natural Language Processing across time: an empirical investigation on Italian. *Proc. of GoTAL*. LNAI. Vol. 5221. Pp 371-382. Springer.

Eva Pettersson and Joakim Nivre. 2011. Automatic Verb Extraction from Historical Swedish Texts. *5th LaTeCH*. Pp 87-95. Oregon, USA.

Eva Pettersson, Beáta Megyesi and Joakim Nivre. 2012. Parsing the Past – Identification of Verb Constructions in Historical Text. *6th LaTeCH*. Pp 65-74. Avignon, France.

Eva Pettersson, Beáta Megyesi and Joakim Nivre. 2014. Verb Phrase Extraction in a Historical Context. *Proc. of the first Swedish national SWE-CLARIN workshop*. Uppsala, Sweden.

Michael Piotrowski. 2012. *Natural Language Processing for Historical Texts*. Synthesis Lectures on HLT. 5(2):1. Morgan & Claypool Publ.

Jennifer Rutner and Roger C. Schonfeld. 2012. *Supporting the Changing Research Practices of Historians*. National Endowment for the Humanities. Ithaca S+R. New York.

Romain Vuillemot, Tanya Clement, Catherine Plaisant and Amit Kumar. 2009. What's being said near "Martha"? Exploring NEs in Literary Text Collections. *VAST*. Pp. 107-114. NJ, USA.