DHQ: Digital Humanities Quarterly

2009
Volume 3 Number 2

Vive la Différence! Text Mining Gender Difference in French Literature

Shlomo Argamon <argamon_at_iit_dot_edu>, Linguistic Cognition Lab, Dept. of Computer Science, Illinois Institute of Technology
Jean-Baptiste Goulain <jibai31_at_gmail_dot_com>, Linguistic Cognition Lab, Dept. of Computer Science, Illinois Institute of Technology
Russell Horton <russ_at_diderot_dot_uchicago_dot_edu>, Digital Library Development Center, University of Chicago
Mark Olsen <markymaypo57_at_gmail_dot_com>, ARTFL Project, University of Chicago

Abstract

In this study, a corpus of 300 male-authored and 300 female-authored French literary and historical texts is classified for author gender using the Support Vector Machine (SVM) implementation SVMLight, achieving up to 90% classification accuracy. The sets of words that were most useful in distinguishing male and female writing are extracted from the support vectors. The results reinforce previous findings from statistical analyses of the same corpus, and exhibit remarkable cross-linguistic parallels with the results garnered from SVM models trained in gender classification on selections from the British National Corpus. It is found that female authors use personal pronouns and negative polarity items at a much higher rate than their male counterparts, and male authors demonstrate a strong preference for determiners and numerical quantifiers. Among the words that characterize male or female writing consistently over the time period spanned by the corpus, a number of cohesive semantic groups are identified. Male authors, for example, use religious terminology rooted in the church, while female authors use secular language to discuss spirituality. Such differences would take an enormous human effort to discover by a close reading of such a large corpus, but once identified through text mining, they frame intriguing questions which scholars may address using traditional critical analysis methods.

Amanda Bonner: What I said was true, there's no difference between the sexes. Men, women, the same.

Adam Bonner: They are?

Amanda Bonner: Well, maybe there is a difference, but it's a little difference.

Adam Bonner: Well, you know as the French say...

Amanda Bonner: What do they say?

Adam Bonner: Vive la difference!

Amanda Bonner: Which means?

Adam Bonner: Which means hurrah for that little difference. (Adam's Rib, 1949)

Introduction

Attempts to identify and characterize differences between male and female discourse have utilized methods such as close reading, sociolinguistic modeling [Tannen 1994], statistical analysis [Olsen 2004], [Olsen 2005], and, more recently, machine learning [Koppel 2002], [Argamon 2003]. The machine learning approach is closely related to purely statistical analysis methods; both approaches exploit differences in aggregate word frequencies to highlight differences between male and female authors in content or style. One advantage of machine learning over simpler forms of statistical analysis lies in its creation of a predictive model of testable accuracy, that can be used to assign gender labels to samples of unknown category, or, as in this study, interrogated to reveal the features most useful in such a classification. The resultant weighted wordlists can be used to support or weaken an existing hypothesis about differences between the corpora, or suggest new directions for investigation, whether by additional machine learning or other, more traditional, critical methods. 1

This study was based on the same male and female corpora used by Olsen in previous statistical analyses [Olsen 2004], [Olsen 2005]. The female corpus was assembled first, due to the more limited digital collection of women's writing at our disposal. 300 texts roughly balanced by genre, collection and time period were chosen, from among texts by French women writers available to us. For each of the 300 texts by 67 female authors (18.5 million words), we selected the chronologically closest male document available in that same genre and, when possible, same collection, leading to a comparison corpus of 300 texts by 170 male authors (27 million words). As noted by Olsen [Olsen 2004], although these texts range from the 12th - 20th centuries, the samples are largely drawn from the 18th-early 20th centuries with strongest representation in the 19th century, owing to the predominance of romantic novelists in the available collections of female writing. The sample is also skewed by a disproportionate number of works by several notable authors, in particular George Sand, with 77 works. Two subsets of the main corpora, each containing 92 documents selected from either the male- or female-authored set, were also selected in an effort to avoid the “Sand Effect.” 2

Comparison With Previous Research

Because we are working with the same corpus previously subjected to a purely statistical analysis [Olsen 2004], [Olsen 2005], we can bring machine learning tools to bear on the questions posed by that work and directly compare our results. Machine learning allows us the possibility of approaching the issue of male and female authorship from a different angle, with a set of metrics for success fundamentally different from those afforded by traditional text analysis methods and statistical inquiries. We ask an SVM model to learn, to the best of its ability, to discriminate between male- and female-authored documents by feeding it labeled examples of each, and applying an algorithm designed to generate predictive models by exploiting generalizable differences in word frequencies between documents in each set. The models give us quantitative feedback regarding their accuracy in their task, and expose their methods by outputting lists of the words which were their input, ranked and weighted as being predictive of one gender or the other. While these metrics do not assure us of an intellectually satisfying outcome from a literary critical viewpoint, they provide a good test of the validity of our process of analysis. 3

Because machine learning algorithms are fundamentally rooted in the exploitation of differential distributions of features (in our case, words), we would expect to see many of the same words appear as highly weighted features in our machine learning results that Olsen found to be significant in his statistical analysis. However, we would not expect the lists to be identical because there are additional factors that influence SVM trained weights that are not captured by differential frequency statistics or other statistical measures such as information gain (IG). Differential frequency and IG are innate properties of an individual word's distribution between sub-corpora, whereas an SVM weight has meaning only within the context of a particular model generated by the learning algorithm, and must be considered in relation to the weights of other features in that model. Differential rates and IG may simply be calculated according to a set formula with unvarying results, whereas SVM weights are heuristically assigned and refined by the learning algorithm in a search for maximum performance on the classification problem. 4

Information gain and other statistical measures of distribution are commonly used as heuristics for reducing feature set dimensionality and for setting initial weights for machine learning algorithms, but there is no guarantee that all words with highly differential frequencies in the corpora will be assigned high weights by the machine learner in the final 5

model. SVM produces two weighted sets of words, male and female, which, taken together, are maximally effective (to the extent of the ability of the algorithm to produce an optimal solution) at discriminating between texts from the two corpora. Words which might exhibit interesting distributions but which do not fit well into a particular model will not be assigned high weights and will escape our notice. Therefore, it is useful to perform a variety of machine learning runs, find what works, and search for common threads in the results. Ultimately, results must find support from a knowledgeable reading of the texts and be fitted with a critical hypothesis to be of great interest from the literary scholar's point of view, although predictive models may have practical uses, such as adding guessed metadata to unclassified documents, independent of their critical value or validity.

Experimental Design

The machine learning algorithm chosen for this classification task is an SVM implementation called SVMLight [Joachims 1999]. SVM has proven to be a model well-suited for text classification, and our initial tests showed that SVMLight achieved the best accuracy in classification among learning algorithm implementations at our disposal, including naive Bayesian and decision tree learners. The SVMLight implementation is freely available and includes key capabilities such as cross-validated accuracy measures via leave-one-out estimation and the ability to extract the weights assigned to each feature. The ability to interrogate the model in this way is essential, because without it we would learn nothing about what word usage patterns distinguish male writing from female writing, merely that such a distinction can be learned with a particular degree of accuracy. A black-box model may be adequate for industrial applications, where the goal is to classify unclassified instances with a certain accuracy, but in this experiment, where the correct classification is already known for all texts, we are far more interested in picking apart the constructed model to determine the orientation and magnitude of the weights of individual words. 6

For our preliminary experiments, we prepared 8 sets of vectors, comprised of the two collections (the full 600 document corpus and 184 document subset) in four versions each: the surface form of the words, the lemmas, the parts of speech (POS) of the words as assigned by TreeTagger, and a simplified part of speech grouping, with broader categories (POSgroup). Each matrix consisted of either 600 or 184 vectors, labeled with 1 for male-authored and -1 for female-authored documents. For a look at the generic data preparation process for text classification, see [ARTFL 2008]. 7

Machine Learning Runs

We then trained SVMLight on each matrix, and obtained the accuracies given in Tables 1 and 2, after cross-validation. Surface form and lemma accuracies cluster around 85%, which means that overall, the models generated by SVMLight can correctly predict the gender of the author about 85% of the time. This is a significant result and indicates that the model has indeed found generalizable differences between the texts in the two corpora. The differences in accuracy between the surface and lemma forms of the words are insignificant, and the POS and POSgroup accuracy differences are generally quite slight as well. The most notable distinction is that POS/POSgroup accuracies are consistently much lower than word/lemma accuracies. The former hover around 70% accuracy, which we have adopted as the borderline for a significant result on a binary classification problem. 70% accuracy is not a particularly compelling result on a “coin-flip” problem, because it shows only 20% improvement over the agreement expected by random chance. Naturally, the more accurate our model is, the more importance we can attach to the words the model weights toward each author gender. 8

	Word	Lemma	PoS	PoSgroup
Male	88.3%	87.3%	73.0%	69.7%
Female	83.3%	84.4%	75.7%	78.7%
All	85.7%	85.9%	74.4%	74.2%

Table 1. Preliminary results: 2x300 document sample

	Word	Lemma	PoS	PoSgroup
Male	91.3%	92.4%	73.9%	73.9%
Female	81.5%	81.5%	78.3%	69.6%
All	86.4%	87.0%	76.1%	71.7%

Table 2. Preliminary results: 2x92 document sample

In order to test whether our accuracies were an artifact of the classifier used, rather than demonstrative of true differences between our corpora, we performed the same experiment but with each document randomly labeled as male or female, regardless of true author gender. Over multiple runs, the classifier never achieved more than 50% accuracy in this random falsification experiment, so we can be confident that SVMLight cannot reliably distinguish between and random sub-corpora grouping in this corpus. 9

We can try to learn from our failures here. The fact that SVMLight cannot construct a very accurate prediction model based on POS vectors is a kind of weak evidence against any theory of gendered authorship that holds that men and women speak radically different languages. If, in fact, men and women used the basic building blocks of language in substantially different ways, we might expect to see strong mechanical differences between male and female writing reflected in POS usage rates that the model could exploit to make accurate classifications. That such differences do not widely obtain in this corpus is strongly suggested by the inability of SVMLight to construct a very accurate model to distinguish between the gendered corpora on that basis. Of course, this does not rule out mechanical and stylistic differences that aren't reflected in the simple metric of POS frequencies, but it does suggest a base level of linguistic similarity between the two classes. 10

Based on these initial results, we decided to proceed with further experiments using the surface forms of the words, that being the simplest method and tied for most accurate with the lemmatized forms. All runs cited hereforth were executed within the PhiloMine data mining extensions to the PhiloLogic text search engine [PhiloMine 2007], and are based on vectors of surface forms, and in all cases we achieve an accuracy greater than 70%, most often between 80 and 90%. Now that we were comfortable that the accuracy of our models were significant enough to indicate real differences between our corpora, we investigated the internals of those models to determine where they get their predictive power. We began by extracting the weights assigned to each word in the 2 x 300 surface form features SVMLight model, and sorting them in descending order of magnitude. Words oriented toward male authorship are scored as positive decimals, while those pointing toward female authorship are negative decimals. We obtained the weights of the most influential words in the model, given in Table 5. 11

Our first impulse when examining the feature list was to scan for the presence of “shibboleth” words that trivially identify some subset of works as definitively male- or female-authored, either because they are explicit markers of author gender (such as metadata tags inadvertently retained in the document), or because they are features that occur in only one or a relative handful of works that are homogeneous for author gender. Such terms are gifts to the machine learner, greedily seized upon by our classification model but unlikely to generate any penetrating insight for the scholar. Proper names are the prime example of such features, and we saw several in Table 5, *Consuelo* being the highest-ranked of these. We eliminated terms like *Consuelo* (present in a number of works by Sand) from the input our model receives by stipulating that we will only use words that occur in more than a certain percentage of documents in the corpora. Constructing new vectors using only words that occur in at least 5% of the documents in the combined male and female corpora, we ran the analysis again and extracted the weights for the word given in Table 6. *Consuelo* is gone; a few proper names remain lower on the list, but since they occur in at least 5% of all documents, they may be of broad enough interest to retain. 12

The highest-ranked words in each category are common function words — pronouns, articles, quantifiers, adpositions, common verb forms of *être* and *avoir* — likely to occur frequently in texts of either gender. Several patterns are evident. The female preference for pronouns is quite marked; { *elle, vous, lui, me, ma, moi, mon, il, m', je, toi, tu, votre* } all appear in the top 200 features weighted toward female authors. This is not an unexpected finding given the observation 13

of Olsen [Olsen 2005] of a usage rate for these terms among female authors that is nearly 1.5 times that of male authors. Also of note is the female preference for terms of negative polarity: { *impossible, ne, ni, pas, personne, sans* }. On the male side, we note the preference for determiners such as { *un, le, des, du, les, ce, ces, cette* } and quantifiers such as { *un, deux, une, quelque(s), mille* }.

14 These results are striking in that they replicate almost exactly those of a similar analysis of female- and male- authored texts in the British National Corpus (BNC) [Argamon 2003]. The primary findings of that study were that females tended to use both more personal pronouns such as {*I, you, she, her, their, myself, yourself, herself*} and negative particles such as {*not, no, never*}, and that males used more determiners such as {*a, the, that, these*} and quantifiers such as {*one, two, more, some*}. Although reflexive pronouns are not expressed by a single word in French as they are in English, and hence do not show up distinctly in our analysis, the rest of the findings match almost exactly. The issue of reflexive pronouns might be investigated in subsequent tests by using word bigrams as features rather than, or in addition to, single words. The strong agreement between these two experiments is all the more remarkable for the very different texts involved in these two studies. Argamon et al. [Argamon 2003] analyzed 604 documents from the BNC spanning an array of fiction and non-fiction genres from a variety of sources, all in Modern British English (post-1960), whereas the current study looks at predominantly fictional French works from the 12th - 20th centuries. This cross-linguistic similarity could be supported with further research in additional languages.

15 Somewhat lower down the list than the function words, we start to encounter content words, and some of the same phenomena noted by Olsen in his statistical analyses are apparent. { *aime, aimer, aimable* } all show up on the female list, which squares with Olsen's observation of a use rate of *aim** by females roughly at roughly 1.5 times that of males across all genres. In noting the female preference for personal pronouns and emotional language, Olsen argues “[female] space may be characterized by a more personal, emotive and interactive frame that is not explained by differences in genre or period” [Olsen 2005], and we can support this hypothesis with our machine learning analysis.

16 Having found support for previous findings in Argamon [Argamon 2003] and Olsen [Olsen 2005], [Olsen 2004], we looked for additional patterns in the heavily weighted terms for each gender. Our corpus spans a wide time range, and we are most interested in discovering patterns that persist across that span. To that end, we split our 600 document combined male- and female-authored corpus into two time range sub-corpora, one comprised of all documents from 1100-1799 (244 documents) and one for all other documents, spanning 1800-2000 (356 documents). Separate SVMLight training runs were performed on each time range corpus using those words that appeared in at least 20% of all documents in that corpus, and the 500 highest-weighted features for male- and female-authored documents from each period were extracted. Taking the union of the two male lists and the two female lists, we found 153 male and 192 female features that are among the top 500 features for both time period runs. No single text or group of contemporary texts can force the inclusion of any word into these merged lists because each text occurs in only one time range sub-corpus, so inclusion on both lists indicates a widespread and enduring trend in usage. The relatively common words in Table 3 are consistently useful in distinguishing male and female French writing over a wide time range, and must reflect real differences in style or content between the genders in the corpora.

153 persistent features in Male-authored documents: 1, a, abord, action, affaire, ajouta, amie, article, au, aura, auteur, autour, autre, aux, avons, bas, bouche, bras, c, capitaine, cent, chacun, chair, champ, charles, chez, christ, ciel, cinq, comment, comtesse, contre, corps, coup, coups, crime, côté, d', des, deux, diable, dis, docteur, doigts, dont, doute, droite, du, entre, est, face, fait, façon, femme, feu, fin, fit, fois, foule, gens, gros, haut, histoire, homme, hé, hôtel, ils, in, jacques, jean, juge, jusqu', la, laquelle, le, les, leurs, ligne, long, lorsque, main, mains, maîtresse, messieurs, mis, mit, moins, monseigneur, monsieur, montre, mot, même, nez, nom, nombre, nos, oeil, oeuvres, ordre, oreille, ou, oui, où, par, passage, pied, pieds, présente, président, prêtre, quatre, quelqu', quelque, quelques, question, qui, quoi, reprit, reste, rue, récit, saint, saints, salut, sang, second, seconde, selon, ses, seulement, simple, sire, soit, sous, sur, table, tirer, tour, toute, trente, trois, un, v, ventre, vers, vieux, village, vin, vingt, voici, y, yeux, à
192 persistent features in Female-authored documents: 192 persistent features in Female-authored documents: absence, admiration, afin, agréable, ai, aimable, aime, aimer, aller, amitié, amour, anglais, angleterre, auguste, auprès, aurais, avais, avait, avec, avez, avoir, beaucoup, belle, bien, bonheur, bonne, brillante, but, cacher, car, caractère, celle, chagrin, chercher, chère, coeur, comprendre, compte, comte, confiance, conserver, cour, crois, destinée, disant, donner, douceur, douleur, doux, elle, elles, empêcher, encore, enfance, enfant, enfants, entièrement, envie, esprit, espérance, estime, eût, faisait, fallait, faut, fièvre, fleurs, france, frère, fût, gloire, goût, grande, grandes, généreux, henri, hiver, ici, il, imagination, impossible, inquiétude, inspire, inspirer, instant, intérêt, jamais, jardin, jours, liberté, lui, lumières, m, ma, mais, malgré, manière, manières, me, moi, mon, montrer, mère, ne, ni, nécessaire, opinion, parce, parler, parlez, passion, pauvre, pays, personne, personnes, petite, peut, peuvent, plaire, plaisir, pleurs, plusieurs, possible, pourquoi, pourrais, pouvait, prince, princes, princesse, pu, puisque, puissance, père, quand, que, quitter, regarder, reine, repos, retrouver, revenir, roi, sais, sait, sans, savoir, secret, sentiment, sentir, seule, si, son, souffrir, souvenir, souvent, soyez, suis, supporter, surprise, tant, toi, toujours, tous, toutes, trop, trouva, trouver, très, tu, utile, veux, vie, vit, vivre, voir, vois, vos, votre, voulait, voulut, vous, voyage, voyant, véritable, âme, éducation, égard, égards, émotion, épouser, était, êtes

Table 3. Features appearing in the top 500 highest-weighted in both time range models

Within the male and female lists, it is possible to identify a number of interesting semantic groupings of words. Reassuringly, the female pronouns and negative polarity items and male quantifiers discussed earlier are still present. In addition, there are a number of other semantic categories of words that appear to cohere:

Enduring Male Terms	Enduring Female Terms
Quantifiers: quelqu', quelque(s) *Religiosity: christ, ciel, corps, diable, saint(s), saints, sang(?)* *Numericality: 1, cinq, cent, deux, nombre, quatre, second(e), trois, trente, un, vingt* *Anatomy: bouche, bras, chair, corps, doigts, face(?), main, nez, pied(s), oeil, oreille, sang, yeux, ventre* *Authority: capitaine, docteur, juge, président, sire* *Other notables: action, amie, femme, feu, histoire, homme, maîtresse, rue, salut, vieux, village, vin*	*Pronouns: me, moi, mon, vos, votre, vous* *Spirituality: âme, chercher, coeur, destinée, espérance, esprit, imagination, inspire, inspirer, passion* *Quantifiers: tous, toutes, (toujours)* *Emotion: agréable, aimable, aime, aimer, amitié, amour, bonheur, douceur, douleur, doux, émotion, envie, espérance, plaire, plaisir, pleurs, sentiment, sentir, seule* *Family: enfant(s), épouser, frère, mère, père* *Nobility: prince(s), princesse, reine, roi* *Negatives: impossible, ne, ni, pas, personne, sans* *Other notables: éducation, impossible, inquiétude, gloire, liberté, lumières, opinion, pauvre, possible, puissance, quitter, sais, sait, savoir, secret, seule, souffrir, souvenir, supporter, surprise, vivre, voyage, voyant, voulait, voulut*

Table 4. Subjective thematic groups among the persistent features

The number of strongly cohesive thematic groupings that can be constructed from the highly-weighted features that obtain in both time periods suggest that male and female writers in the corpus exercise markedly different topic selection. Although the identification of these persistent themes marks the endpoint of this machine learning analysis of the corpus, the themes themselves form a natural starting point for a scholar interested in pursuing the differences between male and female writing from a traditional literary critical viewpoint. It would be quite interesting, for example, to explore why male authors favor religious terminology rooted within the church, whereas female authors spend more time discussing spirituality in a personal, more secular language. Similarly, why should so many anatomical terms rank in the very top of male-weighted features, and are they literal expressions of physicality, or rooted in metaphorical usage? Clearly, these thematic groupings cannot be taken as definitive, universal statements about gendered authorship, but they are clearly identifiable trends that provide a neat snapshot of some basic differences between male and female authors, while suggesting potentially fruitful areas for further analysis, either computer-assisted or using traditional methods. Scholars intrigued by these questions could narrow the context for a close reading by refining the text mining analysis, focusing on questions such as which authors and works best exemplify the discovered trends, and which provide exceptions and counter-examples. 18

Conclusion

Our research demonstrates the utility of using support vector machine models to find contrasting features of male and female writing by interrogating the trained models to identify patterns of word usage that distinguish the gendered corpora. We found little advantage to using lemmatized forms of words as our features and a significant disadvantage to using parts of speech, and therefore used the surface forms of the words for the bulk of our research, achieving 19

accuracies in classification between 80% and 90%. Of the words found to be most useful in distinguishing male and female writing, several distinct functional and semantic groupings were identified. The more personal and emotional frame of reference found in female authors' writing by Olsen in his statistical analysis of the same corpus was supported by our machine learning models. The marked male preference for determiners and female preference for personal pronouns and negative polarity items was a particularly promising finding, as it echoes very closely previous work by Argamon et al. [Argamon 2003] on a different corpus in a different language (excerpts from the English-language British National Corpus). Among the other patterns we identified were a number of cohesive semantic groupings of words that were consistently highly weighted towards males or females across the wide time range of the corpus, such as anatomical and religious terms favored by males, and familial and emotional vocabulary favored by females. The close, contextual reading of a corpus of this magnitude could be the life's work or more of a dedicated scholar, with no guarantee that such trends would be salient enough to be noticed. Through the use of machine learning techniques, we can efficiently analyze vast swathes of texts and achieve results that are interesting and enlightening both in and of themselves, and as a spur to further research using other critical methods.

Male Features		**Female Features**	
Word	**Weight**	**Word**	**Weight**
qui	3.032	elle	-4.270
un	2.706	ne	-2.768
à	2.568	vous	-2.256
le	2.512	pas	-1.812
des	2.392	et	-1.594
du	1.993	avec	-1.435
les	1.847	mais	-1.433
au	1.598	lui	-1.365
monsieur	1.396	était	-1.346
est	1.302	si	-1.245
deux	1.264	avait	-1.178
de	1.250	me	-1.127
sur	1.033	ma	-1.069
a	0.953	pour	-0.952
homme	0.884	sans	-0.811
par	0.867	moi	-0.794
ce	0.746	consuelo	-0.779
madame	0.690	quand	-0.779
d'	0.656	bien	-0.702
une	0.594	roi	-0.676
ces	0.590	l'	-0.666
ses	0.586	il	-0.614
dont	0.566	beaucoup	-0.570
quelque	0.554	n'	-0.560

femme	0.535	henri	-0.543
ils	0.528	m'	-0.535
où	0.511	jamais	-0.523
tems	0.496	reine	-0.513
charles	0.493	je	-0.482
ou	0.487	princesse	-0.479
autre	0.451	toujours	-0.470
aux	0.449	car	-0.465
yeux	0.429	ai	-0.462
main	0.417	votre	-0.459
fit	0.392	esprit	-0.453
leurs	0.386	avais	-0.447
quelques	0.384	m	-0.444
leur	0.380	personne	-0.430
cette	0.379	albert	-0.419
fait	0.379	temps	-0.400
après	0.374	mon	-0.393
avois	0.374	bonne	-0.383
reste	0.363	être	-0.381
mille	0.355	dans	-0.379
même	0.327	ça	-0.371
saint	0.326	se	-0.365
fille	0.324	liberté	-0.364
francs	0.309	la	-0.360
tout	0.307	âme	-0.356
lettre	0.299	très	-0.356
étoit	0.298	enfants	-0.349
entre	0.287	peut	-0.347

Table 5. Weights have been scaled to 10,000 times their original values for ease of reading

Male Features		**Female Features**	
Word	**Weight**	**Word**	**Weight**
qui	3.043	elle	-4.291
un	2.716	ne	-2.780
à	2.578	vous	-2.265
le	2.522	pas	-1.820

des	2.400	et	-1.599
du	2.000	avec	-1.441
les	1.856	mais	-1.439
au	1.603	lui	-1.366
monsieur	1.400	était	-1.348
est	1.305	si	-1.250
deux	1.269	avait	-1.179
de	1.252	me	-1.127
sur	1.037	ma	-1.072
a	0.956	pour	-0.956
homme	0.888	sans	-0.814
par	0.870	moi	-0.795
ce	0.749	quand	-0.782
madame	0.690	bien	-0.706
d'	0.657	roi	-0.679
une	0.597	l'	-0.668
ces	0.592	il	-0.621
ses	0.587	beaucoup	-0.572
dont	0.568	n'	-0.564
quelque	0.555	henri	-0.549
femme	0.537	m'	-0.536
ils	0.530	jamais	-0.526
où	0.513	reine	-0.515
tems	0.498	je	-0.483
charles	0.495	princesse	-0.481
ou	0.488	toujours	-0.471
autre	0.452	car	-0.466
aux	0.450	ai	-0.462
yeux	0.430	votre	-0.460
main	0.418	esprit	-0.455
fit	0.394	avais	-0.447
leurs	0.387	m	-0.445
quelques	0.386	personne	-0.431
cette	0.381	albert	-0.420
leur	0.381	temps	-0.402
fait	0.380	mon	-0.392

après	0.375	bonne	-0.385
avois	0.375	être	-0.380
reste	0.364	dans	-0.378
mille	0.356	ça	-0.375
même	0.329	se	-0.366
saint	0.327	liberté	-0.365
fille	0.325	la	-0.358
francs	0.311	très	-0.358

Table 6. Weights have been scaled to 10,000 times their original values for ease of reading.

Works Cited

ARTFL 2008 Warning: Biblio formatting not applied. ARTFL. *ARTFL Technical Report: Creating Vectors for Text Classification Machine Learning*. ARTFL. 2008. http://artfl.uchicago.edu/TechReports/VectorsForTextClassification.

Argamon 2003 Argamon, Shlomo, Moshe Koppel, Jonathan Fine and Anat Rachel Shimoni. “Gender, Genre, and Writing Style in Formal Written Texts”. *Text* 23: 3 (August 2003).

Joachims 1999 Joachims, Thorsten. “Making Large-Scale SVM Learning Practical”. In Bernhard Schölkopf Christopher J.C. Burges and Alexander J. Smola, eds., *Advances in Kernel Methods -- Support Vector Learning*. Cambridge, MA: MIT Press, 1999.

Joachims 2008 Joachims, Thorsten. *SVMLight*. University of Dortmund, 2008. http://svmlight.joachims.org/.

Koppel 2002 Koppel, Moshe, Shlomo Argamon and Anat Rachel Shimoni. “Automatically Categorizing Written Texts by Author Gender”. *Literary and Linguistic Computing* 17: 4 (2002), pp. 401-412.

Olsen 2004 Olsen, Mark. “Making Space: Women's Writing in France, 1600-1950”. Presented at *ALLC/ACH 2004*. *The Joint International Conference of the Association for Computers and the Humanities and the Association for Literary and Linguistic Computing* (2004).

Olsen 2005 Olsen, Mark. “Écriture Féminine: Searching For an Indefinable Practice?”. *Literary and Linguistic Computing* 20 (2005), pp. 147-164.

PhiloMine 2007 *Philomine*. The ARTFL Project. http://philologic.uchicago.edu/philomine/rationale.html.

Schmid 2006 Schmid, Helmut. *TreeTagger: a language independent part-of-speech tagger*. Institute for Natural Language Processing, University of Stuttgart, 2006. http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html.

Stein Achim Stein, The University of Stuttgart. http://www.uni-stuttgart.de/lingrom/stein/forschung/resource.html

Tannen 1994 Tannen, Deborah. *Gender and Discourse*. New York: Oxford University Press, 1994.