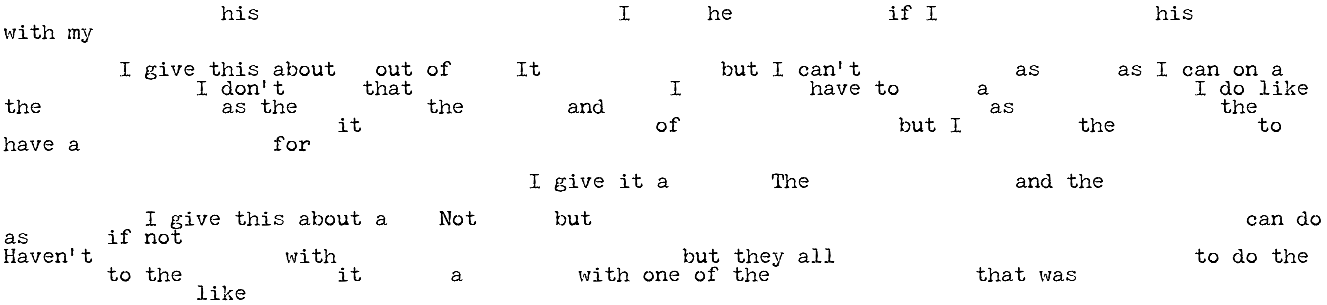
Linguistic Dumpster Diving: Geographical Classification of Arabic Text

Ron Zacharski, Ahmed Abdelali, Stephen Helmreich, and Jim Cowie
New Mexico State University—University of Mary Washington
raz@umw.edu {ahmed, shelmrei,jcowie}@nmsu.edu

1. Introduction

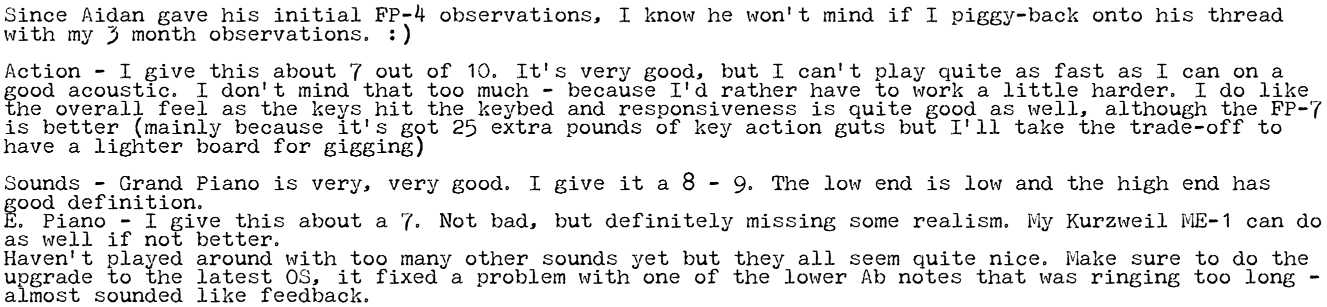
In many applied natural language processing tasks, information is thrown out. For example, in speech recognition systems, prosodic information is commonly discarded; in information retrieval systems, a document is commonly treated as an unordered bag of words and syntactic information is thrown out; and in machine translation systems, pragmatic information (e.g., topic-comment structure and referents of anaphoric expressions) is commonly discarded. Perhaps the most common discarded linguistic forms are the frequent words of a language—words such as those shown in figure 1.[1]



```
                 his                            I      he            if I                 his
with my

         I give this about   out of     It              but I can't            as      as I can on a
               I don't      that                    I          have to      a                I do like
the              as the          the        and                              as                the
                          it                       of                 but I       the           to
have a               for

                                         I give it a        The                and the

           I give this about a   Not      but                                                   can do
as      if not
Haven't               with                           but they all                            to do the
        to the            it       a         with one of the                that was
               like
```

Figure 1: common words in the dumpster

Consider information retrieval systems. These systems enable users to search through a collection of documents. In preparation for the retrieval task, an index of the document collection is created, much like a person would create an index for a book. Like an index for a book, retrieval systems would typically not index words such as *his, I, he, of, a,* and *the*. For example, for the document shown in figure 2 it seems unlikely that a user will search for any of the words shown in figure 1 and these can be safely eliminated prior to indexing. The remaining, less-frequent words (mostly content words), shown in figure 3, are then used for indexing.



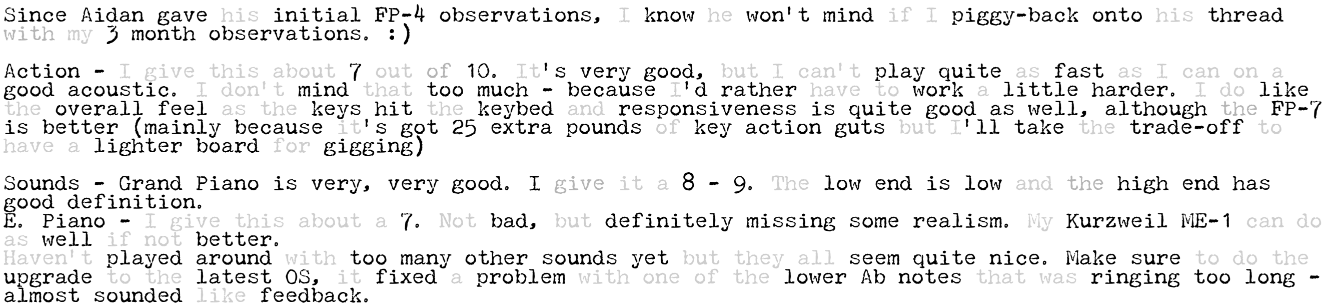
Since Aidan gave his initial FP-4 observations, I know he won't mind if I piggy-back onto his thread with my 3 month observations. :)

Action - I give this about 7 out of 10. It's very good, but I can't play quite as fast as I can on a good acoustic. I don't mind that too much - because I'd rather have to work a little harder. I do like the overall feel as the keys hit the keybed and responsiveness is quite good as well, although the FP-7 is better (mainly because it's got 25 extra pounds of key action guts but I'll take the trade-off to have a lighter board for gigging)

Sounds - Grand Piano is very, very good. I give it a 8 - 9. The low end is low and the high end has good definition.
E. Piano - I give this about a 7. Not bad, but definitely missing some realism. My Kurzweil ME-1 can do as well if not better.
Haven't played around with too many other sounds yet but they all seem quite nice. Make sure to do the upgrade to the latest OS, it fixed a problem with one of the lower Ab notes that was ringing too long - almost sounded like feedback.

Figure 2: a document to index

[1] Frequent words that are routinely removed are commonly called *stop words.*



Since Aidan gave his initial FP-4 observations, I know he won't mind if I piggy-back onto his thread with my 3 month observations. :)

Action - I give this about 7 out of 10. It's very good, but I can't play quite as fast as I can on a good acoustic. I don't mind that too much - because I'd rather have to work a little harder. I do like the overall feel as the keys hit the keybed and responsiveness is quite good as well, although the FP-7 is better (mainly because it's got 25 extra pounds of key action guts but I'll take the trade-off to have a lighter board for gigging)

Sounds - Grand Piano is very, very good. I give it a 8 - 9. The low end is low and the high end has good definition.

E. Piano - I give this about a 7. Not bad, but definitely missing some realism. My Kurzweil ME-1 can do as well if not better.

Haven't played around with too many other sounds yet but they all seem quite nice. Make sure to do the upgrade to the latest OS, it fixed a problem with one of the lower Ab notes that was ringing too long - almost sounded like feedback.

Figure 3: words used to index document

Frequent words are removed for two reasons: first, because they are unlikely to contribute in any meaningful way to the results, and, second, removing them can greatly reduce the amount of computation and storage required for the analysis task. For example, the original document shown in figure 2 contains 212 words, while the representation with common words removed in figure 3 contains only128 words. This practice of removing common words in retrieval systems has been known for a long time and is widespread. For example, in 1958 Luhn noted that high frequency words are too common to have the type of significance being sought and would constitute 'noise' in the system (Luhn 1958). Similarly, in his textbook on information retrieval (1979), van Rijsbergen calls high frequency words 'fluff words' and in another place 'non-significant words' and lists 250 such words for English (including *a, about, above, across, after, to, would, yet,* and *you*).

However, frequent words are important for many searches. For example, if I am searching for*flights to Las Cruces* the *to* is critical to my search even though it is a frequent word. In fact, Google holds a patent on a system that detects meaningful frequent words in search queries—non-meaningful frequent words are removed from the query while meaningful ones are not removed (Tong et al. 2008). However, in the vast majority of information retrieval systems, frequent words are removed.

Text classification is another area where the removal of frequent words is common. The task in text classification is to automatically assign a document to a category based on the contents of that document. For example, we may want to categorize texts based on the topic of the text such as digital piano reviews, motorcycle reviews, opinions about the Iraq war and so on. In this case, we would want to categorize the text shown in figure 2 as being about digital pianos and not about motorcycles. The frequent words shown in figure 1 would likely not help us in this classification and can be safely eliminated. Eliminating frequent words is extremely common in classification tasks. For example, in a classic paper on classification by Joachims (1996) he removes 100 of the most common words. Dumais and Chen (2000) in an approach using support vector machines to classify web documents also remove occurrences of frequent words. In discussing the related area of clustering Spangler and Kreulen (2008) write “We don’t need to keep any words that are superfluous because they would simply add noise that obscures the signal we are trying to detect.” Gangolly and Wu (2000) have called such words 'fluff words' in classification tasks. Berry (2004) states that these words “do not bear any content.”

However, there is some evidence that the distribution and use of frequent words is not independent of text categories. For example, the prototypical pronoun in written discourse is one which is interpreted as coreferential with a previous expression in the text. For example, in (1) *it* is coreferential with *a large mansion on Summit Avenue.* However, a certain class of pronouns have no overt direct antecedent in the text as shown in (2) - (4):

(1) Her family lived in a large mansion on Summit Avenue. **It** had been built in 1902. (Gundel et al. 2000)

(2) Seven years of marriage. Yes **we** had our ups and downs, but now **she** says she doesn't love me anymore. [alt.support.divorce] (Gundel et al. 2000)

(3) It is very hard for me to feel supported after recently being discharged from an intensive

treatment program. Today I got weighed and I gained a quarter of a pound and **they** think I water loaded!! ha! [alt.support.eatingdisoders] (Gundel et al. 2000)

(4) I have been tubed a couple of times and **it** is uncomfortable going down. [alt.support.eatingdisoders] (Gundel et al. 2000)

In (2) there is no overt antecedent for *we* and *she*; in (3) there is no overt antecedent for *they*; and in (4) there is no overt antecedent for *it*. There is some evidence that these forms are not independent of topic. For example, in a study of newsgroups Gundel et al. (2000) found that these forms are significantly more frequent in alt.support.eatingdisorders and, to a lesser degree, alt.support.divorce, than in other newsgroups in the study. Nonetheless, it is commonplace to remove frequent words when doing classification tasks.

In sum it is standard practice in a wide range of natural language processing tasks to remove frequent words. In the vast majority of cases this is the correct thing to do. But there is a danger that this practice is so ingrained that it becomes automatic, so that frequent words are removed without thinking.

The usefulness of frequent words.

In addition to the long history of removing frequent words, there is an equally long history that demonstrates the informativeness of frequent words. One compelling example of this is in the area of stylometrics—the analysis of texts to determine the identity of their authors. In stylometrics the task is to find writer invariant features of text–that is, a feature that is similar in all the texts of an author but different in the texts of different authors. A number of writer invariants have been identified including syntax, word length, sentence length, vocabulary, and the frequency of function words. For example, Mosteller and Wallace in their seminal book on stylometrics (1964), noted that the frequencies of various function words could distinguish the writings of Alexander Hamilton and James Madison. They found that Hamilton used the word *upon* far more frequently than Madison did—3.24 times per thousand words versus 0.23. They used the 70 function words shown in figure 4 as part of the feature set they used to classify the documents of the Federalist Papers using a Bayesian approach.

```
a       as      do      has     is      no      or      this
all     at      down    have    it      not     our     to
also    be      even    her     its     now     shall   up
an      been    every   his     may     of      should  upon
and     but     for     if      more    on      so      was
any     by      from    in      must    one     some    were
are     can     had     into    my      only    such    what
```

Figure 4: 70 function words used by Mosteler and Wallace.

Levison et al. (1968) use the distribution of the particle *de*, the conjunction *kai*, as well as as sentence length to argue that the Seventh Letter was not by Plato, but possibly by Speusippus.[2] Hilton (1990) uses frequent words to determine authorship of the Book of Mormon.

2. Task and Method

Our task was to geographically classify Arabic news articles. For example, the task is to categorize the document shown in figure 5 as being from Syria.

لكن بحثه عن معادل الحب والسرور صدعته الخيبة والمرض فأخبر نفسه
الأدب الملحمي الموضوع بين التعدد والتجدد
لاشك ان الألوان الادبية تختلف بين أدب وخر سواء أكان الأدب شعرا أم نثرا فنجد في
أدبنا العربي المديح والفخر والمراسلات والخطابة
بينما نجد في الأدب الأوروبي الملاحم والمسرح والشعر الغنائي والرواية
القضية الفلسطينية في مئة عام إضاءات وراء

[2] This analysis is controversial. See, for example, Deane 1973.

تقيم مؤسسة الشجرة للذاكرة الفلسطينية وتحت رعاية الدكتور محمود السيد وزير
الثقافة المتلقى الفكري حول فلسطين المنعقد بمناسبة يوم التضامن مع الشعب
العربي الفلسطيني تحت عنوان
الإخاء الحقيقي يستلزم دائما نوعا من التواصل بين شخصيات حية يواجه بعضها بعضا
ويتجاوب معظمها مع الخر وينفذ بعضها إلى بعض
ولولا هذا التفتح لكان كل تواصل بين الذوات ضربا من المستحيل
فالحياة الإنسانية لايمكن أن تنمو وتزدهر في محيط قفر من مناجاة الذات لأنها بحاجة
مستمرة إلى التفتح والإشراق في جو دافىء من المحبةوالتبادل والإخاء
شريف محرم الألوان قراءة للروح من زوايا مختلفة
الفنان شريف محرم ينتج لوحته من الملامح العميقة في النفس البشرية ودائما بما
تمليه المناطق الغامضة والمجهولة لتكون اللوحة سرا جميلا يشي بمفاتيحه وأبواب
فتنته
نبيل السمان يعرض تهويمات خاصة في صالة فاتح المدرس
تستضيف صالة فاتح المدرس في هذه الايام معرض التشكيلي نبيل السمان في
مجموعة جديدة من اعماله التي تنزع الى الاشتغال على اللون وقيمته الجمالية الحية

Figure 5: Sample Arabic classification document

The task was not to identify dialects of Arabic—we were not attempting to distinguish the 40 spoken dialects of Arabic from one another—say, Algerian Arabic from Libyan. All the documents are of one dialect—Modern Standard Arabic—and we are attempting to identify regional differences in this one dialect.

2.1 Document Representation

In the classification tasks we described above, we started with the document shown in figure 2, removed the frequent words shown in figure 1, resulting in the document representation shown in figure 3. For geographical classification we do the exact opposite. We start with the document shown in figure 2 keep the frequent words shown in figure 1, discarding the words shown in figure 3. From the resulting representation in figure 1 we process the document further by counting the occurrences of the frequent words. For example, there are 11 occurences of *I*, 8 of *the* and 2 of *his.* We then generate a vector of frequencies—each location in the vector representing a different frequent word. For example: `(0.000462, 0.001865, 0.009324, ...)`

2.2 Corpus

Our corpus consisted of 4,167 articles from 5 different countries as shown in table 1.

Country	**Website**	**Number of documents**
Egypt	ahram.or.eg	1146
Sudan	almshaheer.com	749
Libya	akhbar-libya.com	999
Syria	thawra.com	263
UK	asharqalawsat.com	1010

Table 1: Distribution of documents in the corpus

The average size of an article was 15 kilobytes or roughly 7,500 characters.

We varied the size of the frequent word list from 58 to 1000 words.[3] The reason for this variation was to determine if classification accuracy would improve with the size of the list. As in the above example, each Arabic document is represented as a vector of common word frequencies. A subset of these words with their translations is shown below.

[3] The use of a list of 58 words instead of using a seemingly more reasonable number like 50 was because the 58 word list was a pre-existing one used in other Arabic analysis tasks. The list was hand edited to remove frequent content words (e.g., the names of newspapers).

حول	around	امس	yesterday
أي	any	ان	that
اعلان	announce	انه	that he
حيث	where	او	or
الاول	first	اي	any
التي	which	ايضا	also
الذين	which	قبل	before
الى	to	بعد	after
اليوم	today	بين	between
امام	in front of	حتى	until

Each of the 4,167 documents of the corpus were converted to this vector format. We then trained on this data using a support vector machine approach to build a classifier. The basic approach of such training algorithms is as follows. Suppose we plot out documents written by Hamilton and Madison in two dimensional space as shown in figure 6a. The x-axis represents the frequency of the word *enough* in the documents and the y-axis represents the frequency of *upon*. In (6a) the documents written by Hamilton are indicated by *h* and those of Madison by *m*. As you can see from (6a) there were more occurrences of *upon* and *enough* in documents written by Hamilton than in documents written by Madison.

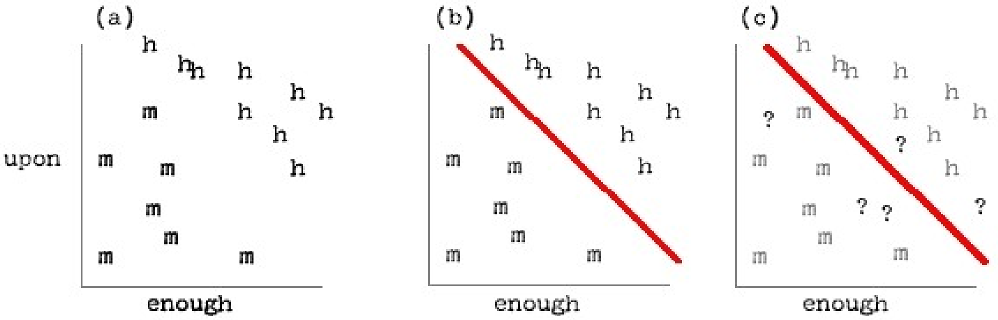


Figure 6: distribution of documents

What the training algorithm does is find a line that best separates the two classes as shown in (6b). Once we have this line we can use it to classify new documents. In (6c) new documents are indicated by question marks. Our classifier will classify the new documents below the line as being authored by Madison and those above as belonging to Hamilton. In the Arabic task the dimensions match the size of the word lists, which ranges from 58 to 1,000. So minimally we have a 58 dimensional space and instead of a line separating the classes we have a hyperplane. Regardless of the number of dimensions, the approach is the same as the two dimensional one.

The specific support vector machine algorithm we used is the sequential minimal optimization algorithm (Platt 1998). We evaluated the algorithm using 10-fold cross-validation. We compared the accuracy of using 5 word lists differing in how many words they contained: 58, 100, 250, 500, 1000. The 58 word list was a pre-existing one. The remaining lists were constructed by combining the 58 word list with a list of frequent words in the Arabic newspaper corpus.

3. Results

The results are shown in the following tables. They range from 92% accurate in classifying documents to over 99%[4] The rows of the tables represent the actual country the documents were from; the columns represent how the document was classified by our algorithm. For example, in table 2, 1,145 documents from Egypt were correctly classified as being from Egypt; 1 document from Egypt was incorrectly classified as being from Libya. In the next row, 713 of the documents from Sudan were correctly classified as being from Sudan; 1 was incorrectly classified as being from Egypt, 2 from Libya, and 33 from the UK.

	Egypt	Sudan	Libya	Syria	UK
Egypt	1145	0	1	0	0
Sudan	1	713	2	0	33
Libya	21	0	895	0	83
Syria	0	0	13	195	55
UK	1	7	77	30	895

Table 2: 58 word vector: 92.23% accuracy

	Egypt	Sudan	Libya	Syria	UK
Egypt	1144	1	0	1	0
Sudan	0	733	0	0	16
Libya	4	0	978	0	17
Syria	0	1	3	227	32
UK	0	3	5	25	977

Table 3: 100 word vector: 97.41% accuracy

	Egypt	Sudan	Libya	Syria	UK
Egypt	1145	0	0	1	0
Sudan	0	746	0	0	3
Libya	4	0	989	0	6
Syria	0	0	0	252	11
UK	0	0	3	10	997

Table 4: 250 word vector: 99.09% accuracy

	Egypt	Sudan	Libya	Syria	UK
Egypt	1145	0	1	0	0
Sudan	0	748	0	0	1
Libya	4	0	992	0	3
Syria	0	0	0	260	3
UK	0	0	1	7	1001

Table 5: 500 word vector: 99.5% accuracy

	Egypt	Sudan	Libya	Syria	UK
Egypt	1145	0	1	0	0
Sudan	0	748	0	0	1
Libya	4	0	993	0	2
Syria	0	0	0	263	0
UK	0	0	0	1	1009

Table 6: 1000 word vector: 99.78% accuracy

As the tables show, accuracy improves as the size of the vector increases. In addition we evaluated the performance on 249 blog entries using the same 100 word list as used in the newspaper task. The results are shown in table 7 and table 8. When we trained on these blog entries and tested using 10-fold cross validation the accuracy was 75.9%. When we used the classifier trained on newspapers to classify these blog entries our accuracy was 43.78 %.

[4] We also did preliminary investigation of this method on English text using the ICE-SIN Corpus (the Department of English Language & Literature, The National University of Singapore), the ICE-IND Corpus (Shivaji University, Kolhapur, and the Freie Universität Berlin), and the ICE-PHI Corpus (the College of Liberal Arts, De La Salle University, Manila, The Philippines). Using the same method outlined above with 100 common English words taken from the Brown Corpus, 86.79% of the instances were classified correctly (identifying whether a document was from India, Singapore, or the Philippines). The difference in accuracy between Arabic and English will be investigated in future work.

	Egypt	Sudan	Libya	Syria	UK
Egypt	31	0	0	0	19
Sudan	1	43	0	2	4
Libya	4	5	23	4	12
Syria	0	1	0	48	2
UK	4	1	0	1	44

Table 7: forum entries: 75.9% accuracy

	Egypt	Sudan	Libya	Syria	UK
Egypt	42	0	2	0	6
Sudan	9	6	3	2	30
Libya	19	1	6	1	21
Syria	0	0	1	28	22
UK	23	0	0	0	27

Table 8: forum entries (newspapers) 43.78%

4. Discussion

This work suggests that newspaper articles can be geographically classified with high accuracy using a support vector machine approach. However, when using this approach with blog entries the accuracy is significantly lower. There could be several reasons for this difference. One likely reason for part of this difference is that the size of the training set is substantially smaller for the blog data set than for the newspaper data set (249 documents compared to over 4,000). In our future work, we plan on increasing the size of our blog training corpus.

Another reason for the poor performance with blogs is that while the blog itself is situated in a country, the blog contributors can be geographically dispersed. So a particular blog may have blog entries that are authored by people in different countries. We plan on performing a more detailed by-hand analysis of the blog data to determine if this explanation can be supported by the data.

We received a number of good comments during the question component of our presentation at the workshop. Patrick Juola suggested we look at non-linear methods. He said he has had good luck with nearest neighbor classifiers. It seems worthwhile performing this comparison. Another person questioned how we knew the classifier was classifying geographically rather than picking up on individual authors of the newspapers. There is some evidence from our English analysis that we are picking up geographical classes rather than individual writers. In the English corpus we were 87% accurate in categorization English from India, the Philippines, and Singapore. This English corpus was carefully constructed and represents a wide range of writers.

5. Conclusion

In this paper, we have shown that frequent words in a corpus, that are often ignored in such tasks as information retrieval, text classification, and data mining, are useful in distinguishing geographical provenance of newspaper articles in Modern Standard Arabic. Accuracy of up to 99.8% was achieved on a corpus of over 4000 documents from five different locations. Initial results are also promising in using this technique on more colloquial Arabic texts (blogs) and also on distinguishing geographical varieties of English (Singapore/India/UK), so continued research in this area is well-warranted.

References

Berry, Michael W. 2004. *Survey of Text Mining: Clustering, Classification, and Retrieval*. Springer.

Deane, Philip. 1973. Stylometrics do not Exclude the Seventh Letter. *Mind* 82.325: 113-117.

Dumais, Susan, and Hao Chen. 2000. Hierarchical classification of web content. *Proceedings of SIGIR-00, 23rd ACM International Conference on Research and Development in Information Retrieval* edited by Nicholas J. Belkin, Peter Ingwersen, and Mun-Kew Leong, 256-263. New York: ACM Press.

Gangolly, Jagdish, and Wu Yi-Fang. 2000. On the Automatic Classification of Accounting concepts: Preliminary Results of the Statistical Analysis of Term-Document Frequencies. *New Review of Applied Exert Systems and Emerging Technologies*, 6: 81-88.

Gundel, Jeanette; Nancy Hedberg; and Ron Zacharski. 2000. Statut cognitif et forme des anaphoriques indirects. *Verbum* 22: 79-102.

Hilton, John L. 1990. On verifying wordprint studies: Book of Mormon authorship.*Book of Mormon Authorship Revisited: The evidence for ancient origins*, edited by Noel Reynolds, 225-253. Provo Utah: Foundation for Ancient Research and Mormon Studies.

Joachims, T. 1996. A probabilistic analysis of the Rocchio Algorithm with TFIDF for text categorization. *Proceedings of the Fourteenth International Conference on Machine Learning*, 143-151.

Levison, M., A. Q. Morton, and A. D. Winspear. 1968. The Seventh Letter of Plato.*Mind* 77: 309-325.

Luhn, H. P. 1958. The automatic creation of literature abstracts.*IBM Journal of Research and Development* 2: 159-165.

Mosteller, Frederick, and David K. Wallace. 1964.*Inference and Disputed Authorship: The Federalist*. Reading, MA: Addison-Wesley Publishing.

Platt, John C. 1998. Fast Training of Support Vector Machines using Sequential Minimal Optimization. *Advances in Kernel Methods - Support Vector Learning* edited by Bernhard Schölkopf, Christopher J. C. Burges, and Alexander J. Smola, 185-208. Cambridge: MIT Press.

Spangler, Scott, and Jeffrey Kreulen. 2008.*Mining the Talk: Unlocking the Business Value in Unstructured Information*. Upper Saddle River, New Jersey: IBM Press.

Tong, Simon; Uri Lerner; Amit Singhal; Paul Haahr; and Stephen Baker. 2008. Locating meaningful stopwords or stop-phrases in keyword-based retrieval systems. United States Patent http://patft.uspto.gov/netacgi/nph-Parser?Sect1=PTO2&Sect2=HITOFF&u=%2Fnetahtml%2FPTO%2Fsearch-adv.htm&r=1&p=1&f=G&l=50&d=PTXT&S1=7,409,383.PN.&OS=pn/7,409,383&RS=PN/7,409,383

van Rijsbergen, C.J. 1979. *Information Retrieval*. London: Butterworths.