A Quantitative Analysis of Lexical Differences Between Genders in Telephone Conversations

Constantinos Boulis
Department of Electrical Engineering
University of Washington
Seattle, 98195
boulis@ee.washington.edu

Mari Ostendorf
Department of Electrical Engineering
University of Washington
Seattle, 98195
mo@ee.washington.edu

Abstract

In this work, we provide an empirical analysis of differences in word use between genders in telephone conversations, which complements the considerable body of work in sociolinguistics concerned with gender linguistic differences. Experiments are performed on a large speech corpus of roughly 12000 conversations. We employ machine learning techniques to automatically categorize the gender of each speaker given only the transcript of his/her speech, achieving 92% accuracy. An analysis of the most characteristic words for each gender is also presented. Experiments reveal that the gender of one conversation side influences lexical use of the other side. A surprising result is that we were able to classify male-only vs. female-only conversations with almost perfect accuracy.

1 Introduction

Linguistic and prosodic differences between genders in American English have been studied for decades. The interest in analyzing the gender linguistic differences is two-fold. From the scientific perspective, it will increase our understanding of language production. From the engineering perspective, it can help improve the performance of a number of natural language processing tasks, such as text classification, machine translation or automatic speech recognition by training better language models. Traditionally, these differences have been investigated in the fields of sociolinguistics and psycholinguistics, see for example (Coates, 1997), (Eckert and McConnell-Ginet, 2003) or http://www.ling.lancs.ac.uk/groups/gal/genre.htm for a comprehensive bibliography on language and gender. Sociolinguists have approached the issue from a mostly non-computational perspective using relatively small and very focused data collections. Recently, the work of (Koppel et al., 2002) has used computational methods to characterize the differences between genders in written text, such as literary books. A number of monologues have been analyzed in (Singh, 2001) in terms of lexical richness using multivariate analysis techniques. The question of gender linguistic differences shares a number of issues with stylometry and author/speaker attribution research (Stamatatos et al., 2000), (Doddington, 2001), but novel issues emerge with analysis of conversational speech, such as studying the interaction of genders.

In this work, we focus on lexical differences between genders on telephone conversations and use machine learning techniques applied on text categorization and feature selection to characterize these differences. Therefore our conclusions are entirely data-driven. We use a very large corpus created for automatic speech recognition - the Fisher corpus described in (Cieri et al., 2004). The Fisher corpus is annotated with the gender of each speaker making it an ideal resource to study not only the characteristics of individual genders but also of gender pairs in spontaneous, conversational speech. The size and

Proceedings of the 43rd Annual Meeting of the ACL, pages 435–442,
Ann Arbor, June 2005.

scope of the Fisher corpus is such that robust results can be derived for American English. The computational methods we apply can assist us in answering questions, such as *"To which degree are gender-discriminative words content-bearing words?"* or *"Which words are most characteristic for males in general or males talking to females?"*.

In section 2, we describe the corpus we have based our analysis on. In section 3, the machine learning tools are explained, while the experimental results are described in section 4 with a specific research question for each subsection. We conclude in section 5 with a summary and future directions.

2 The Corpus and Data Preparation

The Fisher corpus (Cieri et al., 2004) was used in all our experiments. It consists of telephone conversations between two people, randomly assigned to speak to each other. At the beginning of each conversation a topic is suggested at random from a list of 40. The latest release of the Fisher collection has more than 16 000 telephone conversations averaging 10 minutes each. Each person participates in 1-3 conversations, and each conversation is annotated with a topicality label. The topicality label gives the degree to which the suggested topic was followed and is an integer number from 0 to 4, 0 being the worse. In our site, we had an earlier version of the Fisher corpus with around 12 000 conversations. After removing conversations where at least one of the speakers was non-native[1] and conversations with topicality 0 or 1 we were left with 10 127 conversations. The original transcripts were minimally processed; acronyms were normalized to a sequence of characters with no intervening spaces, e.g. *t. v.* to *tv*; word fragments were converted to the same token *wordfragment*; all words were lowercased; and punctuation marks and special characters were removed. Some non-lexical tokens are maintained such as *laughter* and filled pauses such as *uh, um*. Backchannels and acknowledgments such as *uh-huh, mm-hmm* are also kept. The gender distribution of the Fisher corpus is 53% female and 47% male. Age distribution is 38% 16-29, 45% 30-49% and 17% 50+. Speakers were connected at random from a pool recruited in a national ad campaign. It is unlikely that the speakers knew their conversation partner. All major American English dialects are well represented, see (Cieri et al., 2004) for more details. The Fisher corpus was primarily created to facilitate automatic speech recognition research. The subset we have used has about 17.8M words or about 1 600 hours of speech and it is the largest resource ever used to analyze gender linguistic differences. In comparison, (Singh, 2001) has used about 30 000 words for their analysis.

Before attempting to analyze the gender differences, there are two main biases that need to be removed. The first bias, which we term the *topic bias* is introduced by not accounting for the fact that the distribution of topics in males and females is uneven, despite the fact that the topic is pre-assigned randomly. For example, if topic A happened to be more common for males than females and we failed to account for that, then we would be implicitly building a topic classifier rather than a gender classifier. Our intention here is to analyze gender linguistic differences controlling for the topic effect as if both genders talk equally about the same topics. The second bias, which we term *speaker bias* is introduced by not accounting for the fact that specific speakers have idiosyncratic expressions. If our training data consisted of a small number of speakers appearing in both training and testing data, then we will be implicitly modeling speaker differences rather than gender differences.

To normalize for these two important biases, we made sure that both genders have the same percent of conversation sides for each topic and there are 8899 speakers in training and 2000 in testing with no overlap between the two sets. After these two steps, there were 14969 conversation sides used for training and 3738 sides for testing. The median length of a conversation side was 954.

3 Machine Learning Methods Used

The methods we have used for characterizing the differences between genders and gender pairs are similar to what has been used for the task of text classification. In text classification, the objective is to classify a document $\vec{d}$ to one (or more) of T predefined topics y. A number of N tuples $(\vec{d}_n, y_n)$

[1] About 10% of speakers are non-native making this corpus suitable for investigating their lexical differences compared to American English speakers.

are provided for training the classifier. A major challenge of text classification is the very high dimensionality for representing each document which brings forward the need for feature selection, i.e. selecting the most discriminative words and discarding all others.

In this study, we chose two ways for characterizing the differences between gender categories. The first, is to classify the transcript of each speaker, i.e. each conversation side, to the appropriate gender category. This approach can show the cumulative effect of all terms on the distinctiveness of gender categories. The second approach is to apply feature selection methods, similar to those used in text categorization, to reveal the most characteristic features for each gender.

Classifying a transcript of speech according to gender can be done with a number of different learning methods. We have compared Support Vector Machines (SVMs), Naive Bayes, Maximum Entropy and the tfidf/Rocchio classifier and found SVMs to be the most successful. A possible difference between text classification and gender classification is that different methods for feature weighting may be appropriate. In text classification, inverse document frequency is applied to the frequency of each term resulting in the deweighting of common terms. This weighting scheme is effective for text classification because common terms do not contribute to the topic of a document. However, the reverse may be true for gender classification, where the common terms may be the ones that mostly contribute to the gender category. This is an issue that we will investigate in section 4 and has implications for the feature weighting scheme that needs to be applied to the vector representation.

In addition to classification, we have applied feature selection techniques to assess the discriminative ability of each individual feature. Information gain has been shown to be one of the most successful feature selection methods for text classification (Forman, 2003). It is given by:

$$IG(w) = H(\mathbf{C}) - p(w)H(\mathbf{C}|w) - p(\bar{w})H(\mathbf{C}|\bar{w}) \quad (1)$$

where $H(\mathbf{C}) = -\sum_{c=1}^{C} p(c) \log p(c)$ denotes the entropy of the discrete gender category random variable $\mathbf{C}$. Each document is represented with the Bernoulli model, i.e. a vector of 1 or 0 depending if the word appears or not in the document. We have also implemented another feature selection mechanism, the KL-divergence, which is given by:

$$KL(w) = D[p(c|w)||p(c)] = \sum_{c=1}^{C} p(c|w) \log \frac{p(c|w)}{p(c)} \quad (2)$$

In the KL-divergence we have used the multinomial model, i.e. each document is represented as a vector of word counts. We smoothed the $p(w|c)$ distributions by assuming that every word in the vocabulary is observed at least 5 times for each class.

4 Experiments

Having explained the methods and data that we have used, we set forward to investigate a number of research questions concerning the nature of differences between genders. Each subsection is concerned with a single question.

4.1 Given only the transcript of a conversation, is it possible to classify conversation sides according to the gender of the speaker?

The first hypothesis we investigate is whether simple features, such as counts of individual terms (unigrams) or pairs of terms (bigrams) have different distributions between genders. The set of possible terms consists of all words in the Fisher corpus plus some non-lexical tokens such as laughter and filled pauses. One way to assess the difference in their distribution is by attempting to classify conversation sides according to the gender of the speaker. The results are shown in Table 1, where a number of different text classification algorithms were applied to classify conversation sides. 14969 conversation sides are used for training and 3738 sides are used for testing. No feature selection was performed; in all classifiers a vocabulary of all unigrams or bigrams with 5 or more occurrences is used (20513 for unigrams, 306779 for bigrams). For all algorithms, except Naive Bayes, we have used the tf·idf representation. The *Rainbow* toolkit (McCallum, 1996) was used for training the classifiers. Results show that differences between genders are clear and the best results are obtained by using SVMs. The fact that classification performance is significantly above chance for a variety of learning methods shows that

lexical differences between genders are inherent in the data and not in a specific choice of classifier.

From Table 1 we also observe that using bigrams is consistently better than unigrams, despite the fact that the number of unique terms rises from ∼20K to ∼300K. This suggests that gender differences become even more profound for phrases, a finding similar to (Doddington, 2001) for speaker differences.

Table 1: Classification accuracy of different learning methods for the task of classifying the transcript of a conversation side according to the gender - male/female - of the speaker.

	Unigrams	Bigrams
Rocchio	76.3	86.5
Naive Bayes	83.0	89.2
MaxEnt	85.6	90.3
SVM	88.6	92.5

4.2 Does the gender of a conversation side influence lexical usage of the other conversation side?

Each conversation always consists of two people talking to each other. Up to this point, we have only attempted to analyze a conversation side in isolation, i.e. without using transcriptions from the other side. In this subsection, we attempt to assess the degree to which, if any, the gender of one speaker influences the language of the other speaker. In the first experiment, instead of defining two categories we define four; the Cartesian product of the gender of the current speaker and the gender of the other speaker. These categories are symbolized with two letters: the first characterizing the gender of the current speaker and the second the gender of the other speaker, i.e. FF, FM, MF, MM. The task remains the same: given the transcript of a conversation side, classify it according to the appropriate category. This is a task much harder than the binary classification we had in subsection 4.1, because given only the transcript of a conversation side we must make inferences about the gender of the current as well as the other conversation side. We have used SVMs as the learning method. In their basic formulation, SVMs are binary classifiers (although there has been recent work on multi-class SVMs). We followed the original binary formulation and converted the 4-class problem to 6 2-class problems. The final decision is taken by voting of the individual systems. The confusion matrix of the 4-way classification is shown in Table 2.

Table 2: Confusion matrix for 4-way classification of gender of both sides using transcripts from one side. Unigrams are used as features, SVMs as classification method. Each row represents the true category and each column the hypothesized category.

	FF	FM	MF	MM	F-measure
FF	1447	30	40	65	.778
FM	456	27	43	77	.074
MF	167	25	104	281	.214
MM	67	44	210	655	.638

The results show that although two of the four categories, FF and MM, are quite robustly detected the other two, FM and MF, are mostly confused with FF and MM respectively. These results can be mapped to single gender detection, giving accuracy of 85.9% for classifying the gender of the given transcript (as in Table 1) and 68.5% for classifying the gender of the conversational partner. The accuracy of 68.5% is higher than chance (57.8%) showing that genders alter their linguistic patterns depending on the gender of their conversational partner.

In the next experiment we design two binary classifiers. In the first classifier, the task is to correctly classify FF vs. MM transcripts, and in the second classifier the task is to classify FM vs. MF transcripts. Therefore, we attempt to classify the gender of a speaker given knowledge of whether the conversation is same-gender or cross-gender. For both classifiers 4526 sides were used for training equally divided among each class. 2558 sides were used for testing of the FF-MM classifier and 1180 sides for the FM-MF classifier. The results are shown in Table 3.

It is clear from Table 3 that there is a significant difference in performance between the FF-MM and FM-MF classifiers, suggesting that people alter their linguistic patterns depending on the gender of the person they are talking to. In same-gender conversations, almost perfect accuracy is reached, indicating that the linguistic patterns of the two genders be-

Table 3: Classification accuracies in same-gender and cross-gender conversations. SVMs are used as the classification method; no feature selection is applied.

	Unigrams	Bigrams
FF-MM	98.91	99.49
FM-MF	69.15	78.90

come very distinct. In cross-gender conversations the differences become less prominent since classification accuracy drops compared to same-gender conversations. This result, however, does not reveal how this convergence of linguistic patterns is achieved. Is it the case that the convergence is attributed to one of the genders, for example males attempting to match the patterns of females, or is it collectively constructed? To answer this question, we can examine the classification performance of two other binary classifiers FF vs. FM and MM vs. MF. The results are shown in Table 4. In both classifiers 4608 conversation sides are used for training, equally divided in each class. The number of sides used for testing is 989 and 689 for the FF-FM and MM-MF classifier respectively.

Table 4: Classifying the gender of speaker B given only the transcript of speaker A. SVMs are used as the classification method; no feature selection is applied.

	Unigrams	Bigrams
FF-FM	57.94	59.66
MM-MF	60.38	59.80

The results in Table 4 suggest that both genders equally alter their linguistic patterns to match the opposite gender. It is interesting to see that the gender of speaker B can be detected better than chance given only the transcript and gender of speaker A. The results are better than chance at the 0.0005 significance level.

4.3 Are some features more indicative of gender than other?

Having shown that gender lexical differences are prominent enough to classify each speaker according to gender quite robustly, another question is whether the high classification accuracies can be attributed to a small number of features or are rather the cumulative effect of a high number of them. In Table 5 we apply the two feature selection criteria that were described in 3.

Table 5: Effect of feature selection criteria on gender classification using SVM as the learning method. Horizontal axis refers to the fraction of the original vocabulary size ($\sim$20K for unigrams, $\sim$300K for bigrams) that was used.

		1.0	0.7	0.4	0.1	0.03
KL	**1-gram**	88.6	88.8	87.8	86.3	85.6
	2-gram	92.5	92.6	92.2	91.9	90.3
IG	**1-gram**	88.6	88.5	88.9	87.6	87.0
	2-gram	92.5	92.4	92.6	91.8	90.8

The results of Table 5 show that lexical differences between genders are not isolated in a small set of words. The best results are achieved with 40% (IG) and 70% (KL) of the features, using fewer features steadily degrades the performance. Using the 5000 least discriminative unigrams and Naive Bayes as the classification method resulted in 58.4% classification accuracy which is not statistically better than chance (this is the test set of Tables 1 and 2 not of Table 4) . Using the 15000 least useful unigrams resulted in a classification accuracy of 66.4%, which shows that the number of irrelevant features is rather small, about 5K features.

It is also instructive to see which features are most discriminative for each gender. The features that when present are most indicative of each gender (positive features) are shown in Table 6. They are sorted using the KL distance and dropping the summation over both genders in equation (2). Looking at the top 2000 features for each number we observed that a number of swear words appear as most discriminative for males and family-relation terms are often associated with females. For example the following words are in the top 2000 (out of 20513) most useful features for males *shit, bullshit, shitty, fuck, fucking, fucked, bitching, bastards, ass, asshole, sucks, sucked, suck, sucker, damn, goddamn, damned*. The following words are in the top 2000 features for females *children, grandchild,*

Table 6: The 10 most discriminative features for each gender according to KL distance. Words higher in the list are more discriminative.

Male	Female
dude	*husband*
shit	*husband's*
fucking	*refunding*
wife	*goodness*
wife's	*boyfriend*
matt	*coupons*
steve	*crafts*
bass	*linda*
ben	*gosh*
fuck	*cute*

child, grandchildren, childhood, childbirth, kids, grandkids, son, grandson, daughter, granddaughter, boyfriend, marriage, mother, grandmother. It is also interesting to note that a number of non-lexical tokens are strongly associated with a certain gender. For example, *[laughter]* and acknowledgments/backchannels such as *uh-huh,uhuh* were in the top 2000 features for females. On the other hand, filled pauses such as *uh* were strong male indicators. Our analysis also reveals that a high number of useful features are names. A possible explanation is that people usually introduce themselves at the beginning of the conversation. In the top 30 words per gender, names represent over half of the words for males and nearly a quarter for females. Nearly a third were family-relations words for females, and 17

When examining cross-gender conversations, the discriminative words were quite substantially different. We can quantify the degree of change by measuring $KL_{SG}(w) - KL_{CG}(w)$ where $KL_{SG}(w)$ is the KL measure of word w for same-gender conversations. The analysis reveals that swear terms are highly associated with male-only conversations, while family-relation words are highly associated with female-only conversations.

From the traditional sociolinguistic perspective, these methods offer a way of discovering rather than testing words or phrases that have distinct usage between genders. For example, in a recent paper (Kiesling, in press) the word *dude* is analyzed as a male-to-male indicator. In our work, the word *dude* emerged as a male feature. As another example, our observation that some acknowledgments and backchannels (*uh-huh*) are more common for females than males while the reverse is true for filled pauses asserts a popular theory in sociolinguistics that males assume a more dominant role than females in conversations (Coates, 1997). Males tend to hold the floor more than women (more filled pauses) and females tend to be more responsive (more acknowledgments/backchannels).

4.4 Are gender-discriminative features content-bearing words?

Do the most gender-discriminative words contribute to the topic of the conversation, or are they simple fill-in words with no content? Since each conversation is labeled with one of 40 possible topics, we can rank features with IG or KL using topics instead of genders as categories. In fact, this is the standard way of performing feature selection for text classification. We can then compare the performance of classifying conversations to topics using the top-N features according to the gender or topic ranking. The results are shown in Table 7.

Table 7: Classification accuracies using topic- and gender-discriminative words, sorted using the information gain criterion. When randomly selecting 5000 features, 10 independent runs were performed and numbers reported are mean and standard deviation. Using the bottom 5000 topic words resulted in chance performance ($\sim$5.0)

	Top 5K	Bottom 5K	Random 5K
Gender ranking	78.51	66.72	74.99$\pm$2.2
Topic ranking	87.72	-	74.99$\pm$2.2

From Table 7 we can observe that gender-discriminative words are clearly not the most relevant nor the most irrelevant features for topic classification. They are slightly more topic-relevant features than topic-irrelevant but not by a significant margin. The bottom 5000 features for gender discrimination are more strongly topic-irrelevant words.

These results show that gender linguistic differences are not merely isolated in a set of words that

would function as markers of gender identity but are rather closely intertwined with semantics. We attempted to improve topic classification by training gender-dependent topic models but we did not observe any gains.

4.5 Can gender lexical differences be exploited to improve automatic speech recognition?

Are the observed gender linguistic differences valuable from an engineering perspective as well? In other words, can a natural language processing task benefit from modeling these differences? In this subsection, we train gender-dependent language models and compare their perplexities with standard baselines. An advantage of using gender information for automatic speech recognition is that it can be robustly detected using acoustic features. In Tables 8 and 9 the perplexities of different gender-dependent language models are shown. The SRILM toolkit (Stolcke, 2002) was used for training the language models using Kneser-Ney smoothing (Kneser and Ney, 1987). The perplexities reported include the end-of-turn as a separate token. 2300 conversation sides are used for training each one of {FF,FM,MF,MM} models of Table 8, while 7670 conversation sides are used for training each one of {F,M} models of Table 9. In both tables, the same 1678 sides are used for testing.

Table 8: Perplexity of gender-dependent bigram language models. Four gender categories are used. Each column has the perplexities for a given test set, each row for a train set.

	FF	**FM**	**MF**	**MM**
FF	**85.3**	91.1	96.5	99.9
FM	85.7	**90.0**	94.5	97.5
MF	87.8	91.4	**93.3**	95.4
MM	89.9	93.1	94.1	**95.2**
ALL	**82.1**	**86.3**	**89.8**	**91.7**

In Tables 8 and 9 we observe that we get lower perplexities in matched than mismatched conditions in training and testing. This is another way to show that different data do exhibit different properties. However, the best results are obtained by pooling all the data and training a single language model. Therefore, despite the fact there are different modes, the benefit of more training data outweighs the benefit of gender-dependent models. Interpolating ALL with F and ALL with M resulted in insignificant improvements (81.6 for F and 89.3 for M).

Table 9: Perplexity of gender-dependent bigram language models. Two gender categories are used. Each column has the perplexities for a given test set, each row for a train set.

	F	**M**
F	**82.8**	94.2
M	86.0	**90.6**
ALL	**81.8**	**89.5**

5 Conclusions

We have presented evidence of linguistic differences between genders using a large corpus of telephone conversations. We have approached the issue from a purely computational perspective and have shown that differences are profound enough that we can classify the transcript of a conversation side according to the gender of the speaker with accuracy close to 93%. Our computational tools have allowed us to quantitatively show that the gender of one speaker influences the linguistic patterns of the other speaker. Specifically, classifying same-gender conversations can be done with almost perfect accuracy, while evidence of some convergence of male and female linguistic patterns in cross-gender conversations was observed. An analysis of the features revealed that the most characteristic features for males are swear words while for females are family-relation words. Leveraging these differences in simple gender-dependent language models is not a win, but this does not imply that more sophisticated language model training methods cannot help. For example, instead of conditioning every word in the vocabulary on gender we can choose to do so only for the top-N, determined by KL or IG. The probability estimates for the rest of the words will be tied for both genders. Future work will examine empirical differences in other features such as dialog acts or turntaking.

References

C. Cieri, D. Miller, and K. Walker. 2004. The Fisher corpus: a resource for the next generations of speech-to-text. In *4th International Conference on Language Resources and Evaluation, LREC*, pages 69–71.

J. Coates, editor. 1997. *Language and Gender: A Reader*. Blackwell Publishers.

G. Doddington. 2001. Speaker recognition based on idiolectal differences between speakers. In *Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech 2001)*, pages 2251–2254.

P. Eckert and S. McConnell-Ginet, editors. 2003. *Language and Gender*. Cambridge University Press.

G. Forman. 2003. An extensive empirical study of feature selection metrics for text classification. *Machine Learning Research*, 3:1289–1305.

S. Kiesling. in press. Dude. *American Speech*.

R. Kneser and H. Ney. 1987. Improved backing-off for m-gram language modeling. In *Proc. Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 181–184.

M. Koppel, S. Argamon, and A.R. Shimoni. 2002. Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, 17(4):401–412.

A. McCallum. 1996. Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. http://www.cs.cmu.edu/ mccallum/bow.

S. Singh. 2001. A pilot study on gender differences in conversational speech on lexical richness measures. *Literary and Linguistic Computing*, 16(3):251–264.

E. Stamatatos, N. Fakotakis, and G. Kokkinakis. 2000. Automatic text categorization in terms of genre and author. *Computational Linguistics*, 26:471–495.

A. Stolcke. 2002. An extensible language modeling toolkit. In *Proc. Intl. Conf. on Spoken Language Processing (ICSLP)*, pages 901–904.