Lexical Differences in Autobiographical Narratives from Schizophrenic Patients and Healthy Controls

Kai Hong[1], Christian G. Kohler[2], Mary E. March[2], Amber A. Parker[3], Ani Nenkova[1]
University of Pennsylvania
Philadelphia, PA, 19104, USA
[1]{hongkai1,nenkova}@seas.upenn.edu
[2]{kohler,memarch}@mail.med.upenn.edu
[3]{parker}@sas.upenn.edu

Abstract

We present a system for automatic identification of schizophrenic patients and healthy controls based on narratives the subjects recounted about emotional experiences in their own life. The focus of the study is to identify the lexical features that distinguish the two populations. We report the results of feature selection experiments that demonstrate that the classifier can achieve accuracy on patient level prediction as high as 76.9% with only a small set of features. We provide an in-depth discussion of the lexical features that distinguish the two groups and the unexpected relationship between emotion types of the narratives and the accuracy of patient status prediction.

1 Introduction

Recent studies have shown that automatic language analysis can be successfully applied to detect cognitive impairment and language disorders. Our work further extends this line of investigation with analysis of the lexical differences between patients suffering from schizophrenia and healthy controls.

Prior work has reported on characteristic language peculiarities exhibited by schizophrenia patients. There are more repetitions in speech of patients compared to controls (Manschreck et al., 1985). Patients also tend to repeatedly refer back to themselves (Andreasen., 1986). Deviations from normal language use in patients on different levels, including phonetics and syntax, have been documented (Covington et al., 2005), however lexical differences have not been investigated in detail.

In this paper we introduce a dataset of autobiographical narratives told by schizophrenic patients and by healthy controls. The narratives are related to emotional personal experiences of the subjects for five basic emotions: ANGER, SAD, HAPPY, DISGUST, FEAR. We train an SVM classifier to predict subject status. Our good results on the relatively small dataset indicate the potential of the approach. An automatic system for predicting patient status from autobiographical narratives can aid psychiatrists in tracking patients over time and can serve as an easy way to administer large scale screening. The detailed feature analysis we performed also pinpoints key differences between the two populations.

We study a range of lexical features including individual words, repetitions as well as classes of words defined in specialized dictionaries compiled by psychologists (Section 4). We use several approaches for feature analysis to identify statistically significant differences in the two populations. There are 169 significant features among all of the 6057 features we examined. Through feature selection we are able to obtain a small set of 25 highly predictive features which lead to status classification accuracy significantly better than chance (Section 6.3). We also show that differences between patients and controls are revealed best in stories related to SAD and ANGRY narratives, they are decent in HAPPY stories, and that distinctions are poor for DISGUST and FEAR (Section 6.5).

Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pages 37–47, Jeju Island, Korea, 12–14 July 2012. ©2012 Association for Computational Linguistics

2 Related Work

Research in psychometrics has studied patterns of lexical usage in a large variety of scenarios. A popular tool used for psychometric analysis is Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 2007). One of the most interesting discoveries in that line of research is that people with physical or emotional pain are likely to use first-person singular pronouns more often than the general population (Rude et al., 2004). In the view of therapy, Pennebaker discovered that writing emotional experiences can be helpful in therapeutic process (Pennebaker, 1997). It has also been shown that the usage of pronouns and function words can be indicators of writing styles, physical health and other distinctions (Tausczik and Pennebaker, 2010).

The combination of natural language processing (NLP) and machine learning (ML) has been explored in many psychology related projects, and is gaining popularity. It has been shown that features from language models (LMs) can be used to detect impairment in monolingual and bilingual children (Gabani et al., 2009). Even better results are achieved when features derived from LMs are combined with other surface features to predict language impairment. Similarly, studies on child language development and autism have shown that n-gram cross-entropy from LMs representative of healthy and impaired subjects is a highly significant feature predictive of language impairment (Prud'hommeaux et al., 2011). The feasibility of making use of lexical features to analyze language dominance among bilingual children has also been confirmed (Solorio et al., 2011).

In non-medically related research, LIWC and lexical features have been used to recognize different personalities such as *introvert* vs *extrovert*, *openness* vs *experience*, *conscientiousness* vs *unconscientiousness*, etc. (Mairesse et al., 2007). Similar features have been applied to differentiate author personality of e-mails (Gill et al., 2006), blogs (Gill et al., 2009) and other documents.

Speech-related features and interactional aspects of dialog behavior such as pauses, fillers, etc, have also been found helpful in identifying autistic patients (Heeman et al., 2010).

Variables (# Subjects)	Schizophrenia (n=23)	Control (n=16)
Mean age (SD)	33.81 (9.65)	32.29 (6.59)
Mean number of words per story (SD)	192.22 (122.4)	180.79 (95.87)

Table 1: Basic demographic information

Syntax features have been used in approaches of automatic detection of neurological problems. Parsing texts produced by subjects and using bag of rules as features have been applied in analyzing language dominance (Solorio et al., 2011). Methods that quantify syntactic complexity like Yngve score and Fraizer score have been used to analyze autism (Prud'hommeaux et al., 2011). Moreover, there has been research on detecting mild cognitive impairment, which could be an earlier state of Alzheimer's disease: five different ways of evaluating syntactic complexity measures were introduced in their paper (Roark et al., 2011).

In our own work, we focus our analysis exclusively on lexical features. Similarly to prior work, we present the most significant features related to differences between schizophrenic patients and healthy controls. Unlike prior work, instead of doing class ablation studies we perform feature selection from the full set of available features and identify a small set of highly predictive features which are sufficient to achieve the top performance we report. Such targeted analysis is more helpful for medical professionals as they search to develop new therapies and ways to track patient status between visits.

3 Data

For our experiments we collected autobiographical narratives from 39 speakers. The speakers are asked to tell their experience involving the following emotions: HAPPY, ANGER, SAD, FEAR and DISGUST, which comprise the set of the five basic emotions (Cowie, 2000). Most subjects told a single story for each of the emotions, some told two. The total number of stories in the dataset is 201.

The stories were narrated in the doctor's office. The recordings of the narratives were manually transcribed in plain text format. We show age and length in words of the told stories for the two groups in Table 1. There are 23 patients with schizophrenia and 16 healthy controls, telling 120 and 81 stories respectively.

4 Features

Here we introduce the large set of lexical features that we group in three classes: a large class of features computed for individual lexical items, basic features, features derived on the basis of pre-existing dictionaries and language model features. We also detail the way we performed feature normalization and feature selection.

4.1 Surface Features

4.1.1 Basic Features

Basic features include token to type ratio to capture vocabulary diversity, letters per word, words per sentence, sentences per document and words per document. These features describe the general properties of the language used by the subject, without focus on specific words.

Repetitions, revisions, large amount of fillers or disfluencies can be indicators for language impairment. In our basic features we detect the number of repetitions in words, punctuations and sentences for each transcript. Then these three measures are normalized by total number of words or sentences.

We define repetitions as the occurrence of the same token in a sliding window of five items within the same sentence. We count repetitions of words and punctuation separately. The repetition of punctuation, mostly commas and full-stops, are indicative of phrasing in speech which has been indirectly captured in the transcript. Repetition of any word is counted, regardless of which specific word was repeated. For example, for the sentence *I am, am, afraid, that something bad would happen. am* is counted as repeated once, and comma is counted as repeated twice. Finally, sentence repetition captures the amount of overlapping at the beginning of two adjacent sentences, defined as the number of tokens from the beginning of the sentence until the first token where the two sentences differ.

4.1.2 Lexical Features

For words in the vocabulary: we use a real value feature equal to the word frequency for each document. Of particular interest we track the use of pronouns because early research has reported that people with cognitive impairment have a tendency to use subjective words or referring to themselves (Rude et al., 2004).

In addition, for each word in the vocabulary, we apply the presence of the repetition about one particular word.

4.1.3 Perplexity from Language Models

Inspired by the predictive power of language model reported in prior work, we also include several language model features. We build language models on words as well as part-of-speech (POS) tags from Stanford POS-tagger (Toutanova et al., 2003). We tried unigram, bigram and trigram language models by word and POS tag. Experiments showed that bigram performed better than random, and the other two performed below random. Thus in the experiments we report later we train one model for patients and one for controls and use the perplexity of a given text according to the bigram language models on word and POS as features in prediction.

4.2 Dictionaries: LIWC and Diction

Text analysis packages have been widely used in research related to personality analysis, sentimental analysis and psychometric studies. We use two dictionary-based systems, LIWC (Pennebaker et al., 2007)[1] and Diction[2], which both give scores to transcripts based on broad categories.

4.2.1 Linguistic Inquiry&Word Count(LIWC)

LIWC calculates the degree to which people use different categories of words. Several manually compiled dictionaries are at the heart of the application. Each word or word stem could be in one or more word categories or sub-dictionaries. For instance, the word “cried” is part of the following categories: *sadness*, *negative emotion*, *overall affect*, *verb*, and *past tense verb*. When a narrative contains the word “cried”, the scale scores corresponding to these five subcategories are incremented. The final output for each narrative is a real value score for each of the 69 categories.

[1] See http://www.liwc.net

[2] See http://www.dictionsoftware.com

Because of the elaborate development of dictionaries and categories, LIWC has been used for predicting emotional and cognitive problems from subject's spoken and written samples. Representative applications include studying attention focus through personal pronouns, studying honesty and deception by emotion words and exclusive words and identifying thinking styles (Tausczik and Pennebaker, 2010). Thus it is reasonable to expect that LIWC derived features would be helpful in identifying schizophrenia patients. In Section 6.4 we discuss in more detail the features which turned out to be significantly different between patients and controls within LIWC.

4.2.2 Diction

We also use Diction to analyze the lexical characteristics of the transcripts. Similar to LIWC, Diction scores are computed with reference to manually compiled dictionaries. The master variable scores in Diction include *activity*, *certainty*, *commonality*, *optimism* and *realism*. These five main scores are computed with 33 dictionaries that define pertinent subcategories. The master variable scores are constructed as follows: $S_m = \sum_{i=1}^{n} a_i - \sum_{j=1}^{m} s_j$, where a_i are additive traits, s_j are subtractive traits (giving positive/negative evidence for the presence of the feature, respectively). For example, Certainty and Realism scores are calculated as follows:

Realism = *[Familiarity + Spatial Awareness + Temporal Awareness + Present Concern + Human Interest + Concreteness] - [Past Concern + Complexity]*

Certainty = *[Tenacity + Leveling + Collectives + Insistence] - [Numerical Terms + Ambivalence + Self Reference + Variety]*

We also give definitions for some important categories. The complete description of categories is available in the $Diction$ manual (Hart, 2000).

Cognition: *Words referring to cerebral processes, both functional and imaginative.*

Satisfaction: *Terms associated with positive affective states.*

Insistence: *A measure of code-restriction and contentedness, with the assumption that the repetition of key terms indicates a preference for a limited, ordered world.*

Diversity: *Words describing individuals or groups of individuals differing from the norm.*

Familiarity: *Consisted of the most common words in English.*

Certainty: *Language indicating resoluteness, inflexibility, and completeness and a tendency to speak ex cathedra.*

Realism: *Language describing tangible, immediate, recognizable matters that affect people's everyday lives.*

4.3 Feature normalization

We use two feature normalization approaches: projection normalization and binary normalization. Both of the two approaches are applied to basic features, dictionary features and word features. As for repetition, we don't use normalization, because it is in itself binary. For transcript i, we denote the value of the jth feature as v_{ij}. We denote min_j, max_j, $average_j$ as the minimum, maximum and average value for each feature in the training corpus, respectively. Thus for each feature j, we have: $average_j = \frac{1}{n}\sum_{i=1}^{n} v_{ij}$ $\min_j = \min_i\{v_{ij}\}, \max_j = \max_i\{v_{ij}\}$.

4.3.1 Projection Normalization

Here we simply normalize all feature values to a range of $[0, 1]$, where 0 corresponds to the smallest observed value and 1 to the largest observed value across all transcripts. Then we could have $p_{ij} = \frac{v_{ij} - min_j}{max_j - min_j}$, where p_{ij} is the feature value after normalization.

4.3.2 Binary normalization

Here all features are converted to binary values, reflecting whether the value falls below or above the average value for that feature observed in training. The value p_{ij} of j-th feature for the i-th instance is as below:

$$p_{ij} = \begin{cases} 0 & v_{ij} < \frac{1}{n}\sum_{i=1}^{n} v_{ij} \\ 1 & \text{otherwise} \end{cases}$$

4.3.3 Prediction on the Test Set

All of the previous values, $average_j$, max_j and min_j are derived from the training set. While doing classification, for a new testing instance, we denote the feature vector as $f = (f_1, f_2, \ldots f_n)$.

f_j is then compared with $average_j$ to do binary normalization. We also use $p_j = \frac{f_j - min_j}{max_j - min_j}$ to do projection normalization. If $p_j < 0$, we change p_j into 0; if $p_j > 1$, we change p_j into 1. For the words or features that are not seen in training, we just ignore this dimension.

4.4 Feature selection

All lexically based analysis is plagued by data sparsity problems. In the medical domain this problem is even more acute because collecting patient data is difficult. The number of features we defined outnumbers our samples by orders of magnitude. Therefore, in our classification procedure, we perform feature selection by doing two-sided T-test to compare the values of features in the patient and control groups. The features with p-value ≤ 0.05 are considered as indicative and are selected for later machine learning experiments, in which 169 out of 6057 features have been selected. We discuss the significant features in the full set in Section 6.4 .

Note however that we don't use the features selected on the full dataset for machine learning experiments because when T-tests are applied on the full dataset feature selection decisions would include information about the test set as well. Therefore, we adopt a leave-one-subject-out (LOSO) evaluation approach instead. In each iteration, we set aside one subject as test set. The data from the remaining subjects form the training set. Feature selection is done on the training set only and a model is trained. The predictions are tested on the held out subject. The procedure is repeated for every subject as test set.

The choice of p-value cut-off allows us to relax and tighten the requirement on significance of the features and thus the size of the feature set. We report results with different p-values in Table 3. We also explore alternative feature ranking and feature selection procedures in Section 6.3. In each fold different features may be selected. For ease of discussing feature differences we present a discussion of the 169 significant features on the entire dataset.

5 Our approach

The goal of our system is to classify the person who told a story in one of two categories: Schizophrenia group (SC) and Control group (CO). In order to do this, we give labels to the stories told by each subject. Therefore we could use our model to identify the status of the person who told each individual story, the task is to answer the question "Was the subject who told this story a patient or control?". Then we combine the predictions for stories to predict status of each subject, and the task becomes answering the question "Is this subject a patient or control given that they told these five stories?". Thus in story level prediction we use no information about the fact that subjects told more than one story, while in subject-level prediction we do use this information.

First we present an experiment that relies only on language models for the prediction. Then we present the complete learning-based system that uses the full set of features. Finally, we describe the decision making approach to combine the story level predictions to derive a subject-level prediction.

5.1 Language Model

Language models have been used previously for language impairment on children (Gabani et al., 2009) and language dominance prediction (Solorio et al., 2011). Patients with speaking disorder or cognitive impairment express themselves in atypical ways. Language models (LMs) give a straightforward way of estimating the probability of the productions of a given subject. We expect that the approach would be useful for the study of schizophrenia as well and so start with a description of the LM experiments.

We use LMs on words to recognize the difference between patients and controls in vocabulary use. We also trained a LM on POS tags because it could reduce sparsity and focus more on grammatical patterns. Two separate LMs are trained on transcripts of schizophrenia and controls respectively, using leave-one-subject-out protocol.

Story-level decisions are made by assigning the class whose language model yields lower perplexity:

$$s(t) = \begin{cases} SC & PER_{SC}(t) \leq PER_{CO}(t) \\ CO & \text{otherwise} \end{cases}$$

by Story (%)	SC-F	CO-F	Accuracy	Macro-F
Random	54.4	44.6	50.0	49.5
Majority	74.8	0.0	59.7	37.4
2-gram	62.5	44.4	55.2	53.5
2-gram-Pos	62.2	53.3	58.2	57.8

by Subject (%)	SC-F	CO-F	Accuracy	Macro-F
Random	54.1	45.1	50.0	49.6
Majority	74.2	0.0	59.1	37.1
2-gram	65.2	50.0	58.9	57.6
2-gram-Pos	66.7	54.5	61.5	60.6

Table 2: Language model performance

Here t means a transcript from a subject, while PER_{SC} and PER_{CO} are perplexities for patients and controls, respectively. We experimented with unigram, bigram and trigram LMs on words and POS tags. Laplace smoothing is used when generating word probabilities.

5.2 Classification Phase

Language models are convenient because they summarize information from patterns in lexical and POS use into a single number. However, most of the successful applications of LMs require large amount of training data while our dataset is relatively small. Moreover, we would like to analyze more specific differences between the patient and control group and this would be more appropriately done using a larger set of features.

We have described our features and feature selection process in Section 4. We use SVM-light (Joachims, 1999) for our machine learning algorithm, as its effectiveness has been proved in various learning-based clinical tasks compared to other classifiers (Gabani et al., 2009) .

5.3 Status Decision

Story level predictions are made for each transcript either based on LM perplexity or SVM prediction. The most intuitive way to obtain a subject-level prediction is by voting from story-level predictions between the stories told by the particular subject. The subject-level prediction is simply set to equal the majority prediction from individual stories. On the few occasions where there are equal votes for schizophrenia and control, the system makes a preference towards schizophrenia, because it is more dangerous to omit a potential patient.

P-value cut-off	by Story	by Subject	# Features
0.15	59.0	58.9	450
0.10	61.7	64.1	341
0.05	62.7	**64.1**	169
0.01	57.7	**65.4**	44
0.005	64.2	**71.6**	32
0.001	65.7	**75.6**	18
0.0005	61.7	66.7	14

Table 3: Performance by subject after T-test feature selection in different confidence levels.

6 Experiments and Results

We perform our experiments on the 201 transcripts of the 39 speakers. The two baselines we compare with are doing random assignments and majority class, which for our datasets correspond to predicting all subjects into the Schizophrenia group.

We report precision, recall and F-measure for both patient and control groups, as well as overall accuracy and Macro-F value. We get predictions in leave-one-subject-out fashion and compute the results over the complete set of predictions.

6.1 Language Model Performance

Our first experiment relies only on the perplexity from language models to make the prediction. We use the 1,2,3-gram models on word and POS sequences. From the result in Table 2 we can see bigram LM performed better than random baseline for both story and subject level prediction. 3-gram and 1-gram LM did not give a credible performance, with results worse than that of the baselines. Because of space constraints we do not report the specific numbers.

6.2 Classification Result after Feature Selection

Next we evaluate the performance of classification with different number of features from the classes we define in Section 4. As discussed above, we performed feature selection by choosing different levels of significance for the p-value cut-off. Feature selection is performed 39 times for each LOSO training fold. On the standard cut-off p-value ≤ 0.05, our system could achieve 62.7% accuracy on story and 64.1% on patient level prediction. The best performance is achieved when the cut-off p-value is

		Schizophrenia			Control			General	
	Measurement	P (%)	R (%)	F (%)	P (%)	R (%)	F (%)	Accuracy (%)	Macro-F (%)
Story	Random	59.7	50.0	54.4	40.5	50.0	44.6	50.0	49.5
	Majority	59.7	100.0	74.8	NA	0.0	0.0 (NA)	59.7	37.4
	25-Features	68.7	75.0	71.7	57.1	49.4	52.9	64.7	62.3
Subject	Random	59.0	50.0	54.1	41.0	50.0	45.0	50.0	49.6
	Majority	59.0	100.0	74.2	NA	0.0	0.0 (NA)	59.0	37.1
	25-Features	**75.0**	**91.3**	**82.4**	81.8	56.3	66.7	**76.9**	**74.6**

Table 4: Performance on best feature-set by feature ranking using signal to noise

stricter, 0.001, where an accuracy of 75.6% can be reached. In this case only about 18 features are used for the classification. Detailed results are shown in Table 3.

6.3 Performance with Different Feature Size

Next we investigate the relationship between feature set size and accuracy of prediction. We are interested in identifying the smallest possible set of features which gives performance close to the one reported on the full set of significant features. Narrowing the feature set as much as possible will be most useful for clinicians as they understand the differences between the groups and look for indicators of the illness they need to track during regular patient visits. Physicians and psychologists are also interested to know the most significant lexical differences revealed by the stories.

As an alternative to ranking features by p-value, we use the Challenge Learning Object Package (CLOP) [3] (Guyon et al., 2006) . It is a toolkit with a combination of preprocessing and feature selection. We experiment with signal-to-noise (s2n), Gram-Schmidt orthogonalization and Recursive Feature Elimination for finding a subset of indicative features (Guyon and Elisseeff, 2003). The signal-to-noise method gives better results than the other two by at least 6% for the top performance feature set. Thus we pick the best k features according to the s2n result and use only those k features for classification.

Figure 1 shows how prediction accuracy changes with feature sets of different sizes. From the plot we clearly see that our top performance is achieved with 25 to 40 features, after which performance drops. The peak performance is achieved when there are 25 features, where we could reach 75.0% precision, 91.3% recall, 82.4% F-measure for patient, and 76.9% accuracy for overall, as shown in Table 4. Detailed information about the top 30 features can be found in Table 5. '**+**' and '**-**' means more prevalent for patient and control, while '**prj**' and '**01**' correspond to the two normalization approaches in Section 4.3, projection and binary respectively.

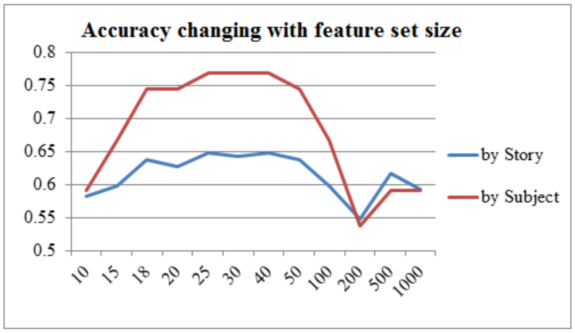

Figure 1: Story and Subject prediction accuracy

6.4 Analysis of Significant Features

In this section we discuss the specific features that were revealed as most predictive by the feature selection methods that we employed. We have seen that it only requires about 25-40 features to obtain peak performance.

First we briefly review the features that turned out to be statistically significant (for 0.05 p-value cut-off). Table 7 provides a list of the features with higher values for Schizophrenia and Control respectively. [4] We group the significant features according to the feature classes we introduced in

[3] See http://clopinet.com/CLOP/

[4] LM1 is defined as the ratio of CO perplexity and SC perplexity from LMs, LM7 comes from projection normalization of LM1. If LM perplexity for CO is smaller than that of SC, then we set LM3 as 1; otherwise we set LM4 as 1.

Rank	Feature	Category	P-value
1	Prj-Self +	Diction	5.33E-06
2	01-Self +	Diction	7.34E-06
3	Prj-punctuation -	Basic	1.33E-05
4	01-I +	LIWC	2.73E-05
5	01-sorry -	Lexical	0.007
6	01-money +	Lexical	6.95E-05
7	01-punctuation -	Basic	4.88E-05
8	prj-I +	LIWC	5.12E-05
9	01-extremely +	Lexical	5.10E-05
10	prj-mildly +	Lexical	0.0006
11	prj-sorry -	Lexical	0.011
12	prj-I +	Lexical	0.0002
13	LM1 +	LM	0.0002
14	LM7 +	LM	0.0002
15	I +	Repeat	0.0003
16	and +	Repeat	0.0002
17	01-mildly +	Lexical	0.0004
18	prj-adverb -	LIWC	0.0006
19	01-relationship -	Lexical	0.024
20	01-late -	Lexical	0.024
21	prj-comma -	Lexical	0.001
22	Repeat word -	Basic	0.001
23	prj-late -	Lexical	0.034
24	prj-very -	Lexical	0.007
25	prj-extremely +	Lexical	0.001
26	01-couldn't +	Lexical	0.001
27	prj-relationship -	Lexical	0.037
28	very -	Repeat	0.007
29	prj-? +	Lexical	0.002
30	prj-moderately +	Lexical	0.006

Table 5: Table of the top 30 features by signal-to-noise ranking

Section 4. Of the 169 significant features, 111 are more prevalent in patients, 58 are more prevalent among the controls. If a feature was significant with both normalizations we use, we list it only once in Table 7.

Among the words indicative of schizophrenia, subjective words such as *I* and LIWC category *self* are among the most significant. This finding conforms with prior research that patients with mental disorders refer to themselves more often than regular people. Patients produce more questions (as indicated by the significance of the question mark as a feature). It is possible that this indicates a disruption in their thought process and they forget what they are talking about. Further work will be needed to understand this difference better.

In terms of words, patients talked more about *money*, *trouble*, and used adverbs like *moderately* and *basically*. Repetition in language is also a revealing characteristic of the patient narratives. There is a substantial difference in the appearance of repetitions between the two groups, as well as repetition of specific words: *I*, *and*, and repetition of filled pauses *um*. As patients focus more on their own feelings, they talked a lot about their family, using words such as *son*, *grandfather* and even *dogs*.

Diction features revealed some unexpected differences. The schizophrenia group scores higher in the *Self*, *Cognition*, *Past*, *Insistence* and *Satisfaction* categories. This indicates that they are more likely to talk about past experience, using cognitive terms and having a repetition of key terms. We were particularly curious to understand why patients score higher on *Satisfaction* ratings. On closer inspection we discovered that patients' stories were rated higher in *Satisfaction* when they were telling SAD stories. This finding has important clinical implications because one of the diagnostic elements for the disease is inappropriate emotion expression. Our study is the first to apply an automatic measure to detect such anomaly in patients' emotional narratives. Prompted by this discovery, we take a closer look at the interaction between the emotion expressed in a story and the accuracy of status prediction in the next section.

The control group exhibited more word complexity, sentence complexity and thoughtfulness in their stories. They use more adverbs and exclusive words (e.g. but, without, exclude) on general trend. They use the word *sorry* significantly more often than patients.

6.5 Status Prediction by Emotion

We also investigate if classification accuracy differs depending on the type of conveyed emotion. Accuracy per emotion with three feature selection methods is shown in Table 6. When using signal-to-noise, we can see that on SAD stories the two groups can be distinguished better. Story-level accuracies on HAPPY stories reach 72.5%, and that the accuracy on HAPPY stories is the next highest one. When applying the 0.05 p-value cut-off to select significant features, ANGER stories become the ones for which the status of a subject

Accuracy (%)	s2n (25)	T-test (0.05)	T-test (0.001)
Happy	66.7	59.0	**71.8**
Disgust	63.4	61.0	51.2
Anger	61.0	**70.7**	**70.7**
Fear	60.0	55.0	67.5
Sad	**72.5**	60.0	67.5
Story	64.7	62.9	65.7
Patient	76.9	64.1	74.4
Majority	59.0	59.0	59.0

Table 6: Accuracy per emotion by different feature-sets

can be predicted most accurately. Using the threshold of 0.001 for selection gives the best overall prediction. In that case, HAPPY and ANGER are the emotions for which recognition is best. The changes in the recognition accuracy depending on feature selection suggests that in future studies it may be more beneficial to perform feature selection only on stories from a given type because obviously indicative features exist at least for the SAD, ANGER and HAPPY stories.

Regardless of the feature selection approach, it is more difficult to tell the two groups apart when they tell DISGUST and FEAR stories. These results seem to indicate that when talking about certain emotions patients and controls look much more alike than when other emotions are concerned. Future data acquisition efforts can focus only on collecting autobiographical narratives relevant to the emotions for which patients and controls differ most.

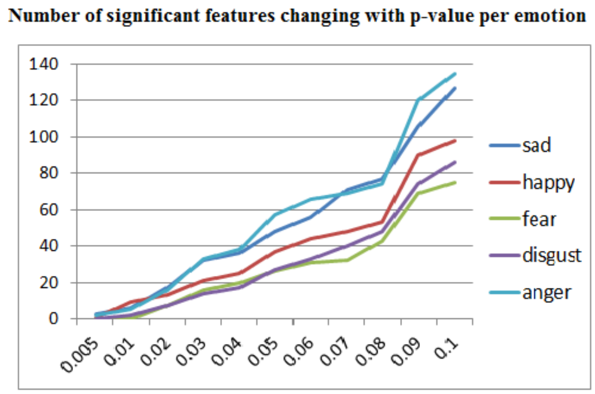

Figure 2: Number of significant features by P-value selection on different thresholds (per emotion)

In future work we would like to use only stories from a given emotion to classify between patients and controls. Doing this with our current dataset is not feasible because there are only about 40 transcripts per emotion. Therefore, we use our data to identify significant features that distinguish patients from controls only on narratives from a particular emotion. For example, we compare the differences of SAD stories told by patients and controls. We count the number of significant features between patients and controls with 11 different p-value cut-offs, and provide a plot that visualizes the results in Figure 2. From the graph, it is clear that there are many more differences between the two groups in ANGER and SAD narratives. HAPPY comes next, then DISGUST and FEAR. However, at lower confidence levels, HAPPY has equal number of significant features as ANGER and SAD, which is in line with the result in Table 6.

The feature analysis performed by emotion reveals more differences between patients and controls, beyond common features such as *self*, *I*, etc. For HAPPY stories, patients talk more about their *friends* and *relatives*; they also have a

Types	Significant features more common in **SCH**
Basic	repeat-word, sentence/document
LIWC	I, insight, personal-pronoun
Diction	self, cognition, past, insistence, satisfaction
Lexical	?, ain't, alone, at, aw, become, before, behind care, chance, confused, couldn't, December, dog dogs, extreme, extremely, feeling, forty, friends god, got, grandfather, guess, guy, hand, hanging hearing, hundred, increased, looking, loved mental, met, mild, mildly, moderate, moderately money, my, myself, outside, paper, passed, piece remember, sister, son, stand, step, story, take taken, throwing, took, trouble, use, wake wanna, way
Repeat	a, and, I, um, was
LM	LM1, LM4, LM7

Types	Significant features more common in **CO**
Basic	length/word, words/sentence
LIWC	$\geq$6-letters, adverb, exclusive words, inhibitive
Diction	certainty, cooperation, diversity familiarity, realism
Lexical	",", able, actually, are, basically, be, being, get's in, late, not, really, relationship, result, she's sleep, sorry, tell, their, there's, very, weeks
Repeat	very, ","
LM	LM3

Table 7: Significant features (p-value $\leq$ 0.05)

higher tendency of being *ambivalent*. For DISGUST stories, patients are more disgusted with *dogs*, and they talk more about *health*. The control group shows a higher *communication* score, referring to a better social interaction. ANGER is one of the emotions that best reveals the differences between groups, and schizophrenia patients show more *aggression* and *cognition* while talking, according to features derived from Diction. The control group sometimes talks more about *praise*. In FEAR stories patients talk about *money* more often than controls. Meanwhile, the control group uses more *inhibition* words, for instance: block, constrain and stop. An interesting phenomenon happens in SAD narratives. When talking about sad experiences, patients sometimes show *satisfaction* and *insistence*, while the controls talked more about *working* experiences.

7 Conclusion

In this paper, we analyzed the predictive power of different kinds of features for distinguishing schizophrenia patients from healthy controls. We provided an in-depth analysis of features that distinguish patients from controls and showed that the type of emotion conveyed by the personal narratives is important for the distinction and that stories for different emotions give different sets indicators for subject status. We report classification results as high as 76.9% on the subject level, with 75.0% precision and 91.3% on recall for schizophrenia patients.

We consider the results presented here to be a pilot study. We are currently collecting and transcribing additional stories from the two groups which we would like to use as a definitive test set to verify the stability of our findings. We plan to explore syntactic and coherence models to analyze the stories, as well as emotion analysis of the narratives.

References

Nancy C. Andreasen. 1986. Scale for the assessment of thought, language, and communication (TLC). *Schizophrenia Bulletin*, 12:473 – 482.

Michael A. Covington, Congzhou He, Cati Brown, Lorina Naci, Jonathan T. McClain, Bess Sirmon Fjordbak, James Semple, and John Brown. 2005. Schizophrenia and the structure of language: The linguist's view. *Schizophrenia Research*, 77(1):85 – 98.

Roddy Cowie. 2000. Describing the emotional states expressed in speech. In *Proceedings of the ISCA Workshop on Speech and Emotion*.

Keyur Gabani, Melissa Sherman, Thamar Solorio, Yang Liu, Lisa Bedore, and Elizabeth Peña. 2009. A corpus-based approach for the prediction of language impairment in monolingual english and spanish-english bilingual children. In *Proceedings of HLT-NAACL*, pages 46–55.

Alastair J. Gill, Jon Oberlander, and Elizabeth Austin. 2006. Rating e-mail personality at zero acquaintance. *Personality and Individual Differences*, 40(3):497 – 507.

Alastair J. Gill, Scott Nowson, and Jon Oberlander. 2009. What are they blogging about? personality, topic and motivation in blogs. In *Proceedings of the AAAI ICWSM'09*.

Isabelle Guyon and André Elisseeff. 2003. An introduction to variable and feature selection. *J. Mach. Learn. Res.*, 3:1157–1182, March.

Isabelle Guyon, Jiwen Li, Theodor Mader, Patrick A. Pletscher, Georg Schneider, and Markus Uhr. 2006. Feature selection with the CLOP package. Technical report, http://clopinet.com/isabelle/Projects/ETH/TM-fextract-class.pdf.

Rodrick Hart. 2000. Diction 5.0, the text-analysis program user's manual, Scolari Software, Sage Press. http://www.dictionsoftware.com/.

Peter A. Heeman, Rebecca Lunsford, Ethan Selfridge, Lois M. Black, and Jan P. H. van Santen. 2010. Autism and interactional aspects of dialogue. In *Proceedings of the SIGDIAL 2010 Conference*, pages 249–252.

T. Joachims. 1999. Making large–scale SVM learning practical. In B. Schölkopf, C. J. C. Burges, and A. J. Smola, editors, *Advances in Kernel Methods — Support Vector Learning*, pages 169–184, Cambridge, MA. MIT Press.

F. Mairesse, M. A. Walker, M. R. Mehl, and R. K. Moore. 2007. Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text. *Journal of Artificial Intelligence Research*, 30:457–500.

Theo C. Manschreck, Brendan A. Maher, Toni M. Hoover, and Donna Ames. 1985. Repetition in schizophrenic speech. *Language & Speech*, 28(3):255 – 268.

J.W. Pennebaker, R.J. Booth, and Francis. 2007. Linguistic inquiry and word count (LIWC

2007): A text analysis program. Austin, Texas. http://www.liwc.net/.

James W. Pennebaker. 1997. Writing about Emotional Experiences as a Therapeutic Process. *Psychological Science*, 8(3):162–166.

Emily T. Prud'hommeaux, Brian Roark, Lois M. Black, and Jan van Santen. 2011. Classification of atypical language in autism. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, CMCL'11, pages 88–96.

Brian Roark, Margaret Mitchell, John-Paul Hosom, Kristy Hollingshead, and Jeffrey Kaye. 2011. Spoken language derived measures for detecting mild cognitive impairment. *IEEE Transactions on Audio, Speech & Language Processing*, 19(7):2081–2090.

Stephanie Rude, Eva-Maria Gortner, and James Pennebaker. 2004. Language use of depressed and depression-vulnerable college students. *Cognition & Emotion*, 18(8):1121–1133.

Thamar Solorio, Melissa Sherman, Y. Liu, Lisa Bedore, Elizabeth Peña, and A. Iglesias. 2011. Analyzing language samples of spanish-english bilingual children for the automated prediction of language dominance. *Natural Language Engineering*, 17(3):367–395.

Yla R. Tausczik and James W. Pennebaker. 2010. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, 29(1):24–54, March.

Kristina Toutanova, Dan Klein, and Christopher D. Manning. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of HLT-NAACL 03*.