Lexical Differences in Autobiographical Narratives from Schizophrenic Patients and Healthy Controls

Ani Nenkova¹, Kai Hong¹, Amber A. Parker², Mary March³, Christian G. Kohler³

Introduction & Purpose of Research
- Lexical differences have not been investigated in detail.
- Evaluate differences in how persons with SZ describe autobiographical experiences of emotional events in standardized setting.

Data: Vignettes from 39 Subjects
- 39 subjects, matched for age, gender and ethnicity.
- SZ: n=23, 115 stories; CO: n=16, 79 stories.
- Five emotions: HAPPY, SAD, ANGER, FEAR and DISGUST.
- Subjects were asked to produce narrative lasting 30-90 sec.
- Narratives transcribed and checked by 2 transcribers.

Demographic information

Gender  Male (19), Female (20)
Age  33.52 (Schizophrenia 32.18, Control 32.19)
Race  Black (23), White (12), Asian (1), Hybrid (3)

Linguistic Inquiry and Word Count & Diction
- Linguistic Inquiry & Word Count (LIWC) (Pennebaker, 2007) calculates the degree to which people use different categories of words. Each word or word stem can be in one or more word categories or sub-dictionaries.

LIWC Significant Features (p-value<0.05)
- SZ: Self(I)**, personal pronoun*, insight^.

LIWC significant categories (# of words included and example)
- insight (195): think, know, consider.
- exclusive words (17): but, without, exclusive.
- inhibition (111): block, constrain, stop.
- complex-words: defined as words >6 letters (sixlfr).

Diction (Hart, 2005) analyzes the lexical characteristics of the transcripts. Similar to LIWC, Diction scores are computed with reference to manually compiled dictionaries.

Diction Significant Features (p-value<0.05)

Some significant Diction categories and definitions
- Cognition: Words referring to cerebral processes.
- Satisfaction: Terms associated with positive affective states.
- Insistence: A measure of code-restriction, indicating a preference for a limited, ordered world.
- Diversity: Words describing individuals or groups of individuals differing from the norm.
- Familiarity: Most common words in English.
- Certainty: Indicating resoluteness in flexibility.

Significance Level: ** [0, 0.01], * (0.001, 0.01], ^ (0.01, 0.05]

Basic Features and Lexical Features
- Basic Features: General properties of language without focus on specific words.
- Lexical Features: Type 1: Analysis of specific words.
- Type 2: Including word analysis and repetition of words.

Type 1 Significant Features (p-value<0.05)
- CO: * ***, sorry*, very*, really*, basically*, relationship*, She’s*, actually*.

Repitition of word: Word has appeared before it with a window of 5.

Significance Level: ** [0, 0.01], * (0.001, 0.01], ^ (0.01, 0.05]

Classification and Feature Selection
- Classification Approach:
  - Using SVM-light as classifier.
  - Applied Leave-one-subject-out (39 times).
  - Voting.

Leave-one-subject-out
- Feature Selection:
  - using Signal-to-Noise (S2N).
  - S2N 2% better than P-value (best performance).

Number of Features for Best Performance

<table>
<thead>
<tr>
<th>emotion</th>
<th>Schizophrenia</th>
<th>Control</th>
<th>General</th>
</tr>
</thead>
<tbody>
<tr>
<td>P(%)</td>
<td>R(%)</td>
<td>F(%)</td>
<td>P(%)</td>
</tr>
<tr>
<td>Story</td>
<td>68.7</td>
<td>75.0</td>
<td>71.7</td>
</tr>
<tr>
<td>Subject</td>
<td>75.0</td>
<td>91.3</td>
<td>82.4</td>
</tr>
</tbody>
</table>

Best Performance Analysis:
- Classification: Specificity: 91.3%, Sensitivity: 56.3%, Accuracy: 76.9% (30 of 39).
- P = Precision, R = Recall
- F = 2P*R / (P+R)=F-measure
- Incorrect Classification: SZ: 2/23, CO 9/16.

- Performance Changing with #Features
  - Best Performance: 25 Features
  - Accuracy dropped: features > 50
  - Less confidence features add noise.

Status Prediction by Emotion

Different Emotions
- HAPPY: SZ talk more about family; higher tendency of ambivalence.
- DISGUST: SZ are more disgusted with dogs, and health. CO has higher communication score, referring to a better social interaction.
- ANGER: SZ show more aggression and cognition.
- FEAR: SZ talk about money. CO use more inhibition words.
- SAD: SZ show more satisfaction & insistence. CO include more working experiences.

Distinguish Power: Anger > Sad > Happy > Disgust

References

Conclusions and future work
- 25 features are enough to get top performance (65%, 77%).
- p-value feature selection.
- Signal-to-noise feature selection.
- Analyze power of different features:
  - Basic features, Lexical features, repetitions.
  - LIWC, Diction.
- Different emotions have different distinguish power.
  - sad, anger > happy > fear, disgust.

Natural Language Engineering Conference, 2010