Sexual Self-Schemas in the Real World: Investigating the Ecological Validity of Language-Based Markers of Childhood Sexual Abuse

Article in Cyberpsychology, Behavior, and Social Networking · June 2017
DOI: 10.1089/cyber.2016.0657

3 authors, including:

Amelia M. Stanton
University of Texas at Austin
16 PUBLICATIONS 12 CITATIONS

Ryan L. Boyd
University of Texas at Austin
34 PUBLICATIONS 170 CITATIONS
Sexual Self-Schemas in the Real World: Investigating the Ecological Validity of Language-Based Markers of Childhood Sexual Abuse

Amelia M. Stanton, BA, Cindy M. Meston, PhD, and Ryan L. Boyd, PhD

Abstract

This is the first study to examine language use and sexual self-schemas in natural language data extracted from posts to a large online forum. Recently, two studies applied advanced text analysis techniques to examine differences in language use and sexual self-schemas between women with and without a history of childhood sexual abuse. The aim of the current study was to test the ecological validity of the differences in language use and sexual self-schema themes that emerged between these two groups of women in the laboratory. Archival natural language data were extracted from a social media website and analyzed using LIWC2015, a computerized text analysis program, and other word counting approaches. The differences in both language use and sexual self-schema themes that manifested in recent laboratory research were replicated and validated in the large online sample. To our knowledge, these results provide the first empirical examination of sexual cognitions as they occur in the real world. These results also suggest that natural language analysis of text extracted from social media sites may be a potentially viable precursor or alternative to laboratory measurement of sexual trauma phenomena, as well as clinical phenomena, more generally.

Keywords: social media, childhood sexual abuse, language, meaning extraction method, methodology

Introduction

Social media provides unique advantages for psychologists seeking to answer complex research questions. In particular, the quantity of natural language data that can be gleaned from social media is enormous. According to data from 2013, almost 75 percent of Americans reported household Internet use, and 71 percent of Americans with Internet access in their homes use social media. Much online data also come at either low or no cost and can be accessed in a relatively short period of time. Finally, naturally occurring language data gathered from social media sites are free from biases that are introduced by experimenters in formal laboratory studies, thereby reflecting ecologically valid social and psychological phenomena.

Recent developments in natural language processing techniques have enabled researchers to better understand the nuances of complex psychological processes. Over the past 10 years, studies have examined natural language data to draw conclusions on personality measurement, online search behaviors, core values, and sexual cognition, to name just a few domains. In some of these studies, such as that of Boyd et al., researchers collected natural language data from social media sites, which are primarily characterized by the fact that they are composed almost exclusively of user-generated content. The increasing availability of social media data offers researchers the opportunity to study psychological traits and cognitions, and their links to behaviors and psychopathology, on a larger scale than ever before.

Despite concerns of self-selection biases and data quality in online settings, a recent systematic review of social network sites as a source of health data found that social media can be used to effectively collect data on health outcomes and other health phenomena. Analytic advances, combined with the wide availability of social media data, may be particularly beneficial for the study of certain clinical phenomena that exist in difficult-to-access populations. Given that the extraction of language data from social media sources bypasses the need for subject recruitment, the anonymity of the Internet may help facilitate discussion of shameful experiences and topics considered to be taboo in these populations.

A highly relevant area of research that benefits from the use of social media data is that of childhood sexual abuse (CSA) which, according to recent estimates, affects 8–31 percent of girls and 3–17 percent of boys worldwide. CSA is associated with numerous detrimental psychological outcomes in both childhood and adulthood, including...
decreased quality of life, increased prevalence of mental health problems, and sexual dysfunction.

Importantly, the psychological impact of CSA has been found to be imprinted in a person’s language as well. In previous laboratory samples, Lorenz and Meston examined linguistic differences between women with and without CSA histories when writing about their sexual experiences and beliefs about sexuality. In line with previous research on language and depression, Lorenzo and Meston demonstrated that women with CSA histories, who are more likely than nonabused women to show symptoms of depression and suicidality, exhibited higher rates of first person singular pronoun use than did nonabused women when writing about their sexual experiences and beliefs. Lorenz and Meston also noted key linguistic differences between women with and without CSA histories, including the use of differing rates of affect words and words related to genuineness.

Furthermore, linguistic patterns resulting from CSA are often highly specific and nuanced. Building on findings that women with a history of CSA report less positive sexual self-schemas than nonabused women, Stanton et al. conducted a language analysis of sexual self-schema themes from the same language samples used by Lorenz and Meston. This study identified seven overarching schema themes: family and development, virginity, abuse, relationship, sexual activity, attraction, and existentialism. In addition to highlighting these sexual self-schema themes, Stanton et al. found differences in the prominence of each theme as a function of sexual abuse history. Specifically, women with CSA histories showed greater invocation of the abuse and attraction themes, whereas nonsexually abused (NSA) women had markedly higher invocation of the virginity and the relationship themes relative to abused women.

Importantly, all previous research on language-CSA links has been conducted using highly controlled laboratory samples. The degree to which previous work extends into real-world samples and, more importantly, the degree to which the above-described patterns are manifest in daily life remain hitherto unknown. It is critical, then, to begin investigating the psychological patterns that have been associated with CSA in daily life, including patterns embedded within trace behavioral data such as Internet activity. By extending the findings of previous work to unconstrained online samples, two primary benefits may arise. First, by establishing generalizability, we are able to better understand the impact of CSA on daily functioning and cognition. Second, should the previously discovered findings hold true in a messy, demographic-ambiguous sample, we propose that online samples may be gainfully put to use in the context of pilot research programs. When examining the psychological sequelae of difficult-to-study topics, researchers may consider conducting initial investigations using natural language data that are already available online; downstream recruitment for controlled laboratory samples may be put to better use once a clear direction of research is established using exploratory accessible samples.

The aim of the current study was to investigate the two goals described above. This work sought to expand upon previous work by examining differences in general language use and sexual self-schema themes in an unconstrained online sample. In accordance with previous research, we predicted that the same linguistic differences between CSA and NSA groups noted by Lorenz and Meston and by Stanton et al. would emerge. In addition, we sought to determine if the linguistic patterns that were established in the laboratory generalize to a large, unconstrained real-world sample that lacks careful laboratory control. Should the findings of previous work extend to the current context, we are able to provide support for the notion that passive online sampling is a valid and justifiable method for researching difficult clinical phenomena in a “flipped” paradigm.

### Methods

#### Participants

We collected archival data from the Reddit website that were directly relevant to previous studies on language use and sexual self-schemas. Reddit is a large social media website that is subdivided into topic-specific discussion boards dedicated to the sharing of information, hyperlinks, and personal narratives related to various topics. For the current study, we collected user posts from two boards that were parallel to the comparison groups of previous research. The first board was r/AdultSurvivors, a board described as “a community for adults who experienced sexual abuse as children.”—This group was selected as a parallel to the CSA group in previous laboratory studies. The second board was r/Sex, a general discussion board “about all facets of sexuality and sexual relationships.”—This group was selected to parallel the NSA group in past studies.

On each board, forums are structured in such a way that users can either (a) make an original post (OP) to which others can reply or (b) leave comments in response to the posts of others. We were primarily interested in psychological information embedded in the language of OPs. Because OPs are unprompted and spontaneous, they necessarily reflect the psychological state of users before shifts that may occur during subsequent interactions; by considering only the OPs, we excluded data that are confounded by such social processes.

Data were collected using a custom script that acquired archival data from the Reddit website’s application programming interface (API). Collection occurred in

*www.reddit.com

A website API is a programming interface that allows coders to directly interact with the data without having to do so through the typical web interface. The Reddit API is open to the public, and users can opt out of allowing their data to be accessed in this manner. Data from Reddit have been used in multiple other studies to study social and psychological phenomena, such as altruism and support-seeking styles for individuals diagnosed with cancer.

### Table 1. Examples of Words Associated with Each Sexual Self-Schema Theme

<table>
<thead>
<tr>
<th>Sexual self-schema themes</th>
<th>Example words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Family and development</td>
<td>House, father, mom, mother, sister</td>
</tr>
<tr>
<td>Virginity</td>
<td>High, school, virginity, lost, guy</td>
</tr>
<tr>
<td>Abuse</td>
<td>Abuse, happen, ashamed, disgust, molest</td>
</tr>
<tr>
<td>Relationship</td>
<td>Important, person, intimacy, enjoy, monogamous</td>
</tr>
<tr>
<td>Sexual activity</td>
<td>Hand, hold, penis, kiss, intercourse</td>
</tr>
<tr>
<td>Attraction</td>
<td>Bring, attractive, physically, arouse, lesbian</td>
</tr>
<tr>
<td>Existentialism</td>
<td>Thought, read, understand, learn, question</td>
</tr>
</tbody>
</table>

From Stanton et al.
September 2015, starting from the oldest post threads and working forward in time. The entire post history of the /r/AdultSurvivors board was collected in short time given the relatively low amount of activity compared to other boards on Reddit. Conversely, the /r/Sex board is a high-activity discussion forum on the website, and comprehensive data collection was unnecessary for the purpose of collecting an adequate sample; streamed archival data were collected from this board for 3 days.

Data collection resulted in a sizeable sample for analysis, including 163 OPs to the /r/AdultSurvivors board and 8,898 OPs to the /r/Sex board. After excluding content that contained an insufficient quantity of language data (e.g., posts that linked to other websites or images; posts with <50 words), our final sample included 132 OPs to the /r/AdultSurvivors board and 3,359 OPs to the /r/Sex board.

Analysis

First, we conducted language analyses parallel to those of Lorenz and Meston. All OPs were analyzed with LIWC2015, a text analysis program that scores texts along ~80 well-validated psychological/cognitive measures, we investigated only those LIWC dimensions examined in previous work.

Second, we quantified the seven sexual self-schema themes identified in previous work using the Meaning Extraction Method (MEM) previously described, examples of

FIG. 1. (a) “I” use. (b) Affect word use. (c) Positive emotion word use. (d) Negative emotion word use. (e) Psychological distancing/authenticity word use.
each theme are presented in Table 1. Using a word-counting approach, which has been demonstrated to be a powerful method for detecting cognitive patterns in past work, texts were analyzed for specific words previously shown to capture distinct cognitive facets of the individual’s sexual self-schema in laboratory contexts.

Finally, in some cases, users made several OPs to their respective Reddit board. Given the unit of interest for this study was the individual, all linguistic measures were averaged by person to establish the average psychological states encoded in each person’s natural language. This resulted in a final sample of $N = 100$ for the /r/AdultSurvivors board and $N = 2,959$ for the /r/Sex board.

Results

For each analysis, we conducted an independent samples $t$-test to assess differences in the use of both specific linguistic categories and sexual self-schema themes by group (/r/AdultSurvivors, /r/Sex). Tests were conducted without equality of variances assumptions, a method well-suited to imbalanced sample sizes.\(^28\) Note that, for our analyses, the statistical degrees of freedom are yoked to the smaller of the two groups (i.e., the /r/AdultSurvivors group). Results and case-wise comparisons pertaining to general language use are presented in Figure 1. Results pertaining to sexual self-schema themes are presented in Figure 2. Descriptive and

<table>
<thead>
<tr>
<th>Category</th>
<th>Previous research(^a)</th>
<th>Current study</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NSA($N = 99$) Mean (SD)</td>
<td>CSA($N = 128$) Mean (SD)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>/r/Sex board($N = 2,959$) Mean (SD)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>/r/AdultSurvivors($N = 100$) Mean (SD)</td>
</tr>
<tr>
<td>Personal pronouns</td>
<td>NA(^b)</td>
<td>NA(^b)</td>
</tr>
<tr>
<td>‘I’ use</td>
<td>10.24 (2.01)</td>
<td>11.52 (2.53)</td>
</tr>
<tr>
<td>Affect words</td>
<td>5.80 (1.29)</td>
<td>6.40 (1.70)</td>
</tr>
<tr>
<td>Positive emotion words</td>
<td>2.90 (0.97)</td>
<td>2.52 (0.87)</td>
</tr>
<tr>
<td>Negative emotion words</td>
<td>1.31 (0.75)</td>
<td>1.83 (0.67)</td>
</tr>
<tr>
<td>Distancing or authenticity words</td>
<td>28.30 (4.93)</td>
<td>29.75 (4.50)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\(^a\)Lorenz and Meston.\(^15\)

\(^b\)Data unavailable.

CSA, childhood sexual abuse; NSA, nonsexually abused.
statistical analyses on language use in the current sample and in previous research\textsuperscript{15} are presented in Tables 2 and 3, respectively. Tables 4 and 5 provide descriptive and statistical analyses on sexual self-schema themes from both the current sample and from previous research.\textsuperscript{6}

Across all analyses, our results perfectly replicated the patterns found in the work of previous laboratory studies. For the LIWC-based analyses of psychological processes, we found that the \textit{/r/AdultSurvivors} group used more personal pronouns overall and used the word “I” significantly more often than did posters to the \textit{/r/Sex} board. Posters to the \textit{/r/AdultSurvivors} board also used significantly more affect words, negative emotion words, and authenticity words, but significantly less positive emotion words than \textit{/r/Sex} board posters. For the MEM-based analyses of sexual self-schema themes, our results revealed that posters to the \textit{/r/AdultSurvivors} board used significantly more words pertaining to the abuse theme and the attraction theme relative to posters to the \textit{/r/Sex} board. Conversely, posters on the \textit{/r/Sex} board were significantly more likely to invoke the relationship theme and the virginity theme than were posters to the \textit{/r/AdultSurvivors} board.

**Discussion**

We examined language use and sexual self-schema themes in a large online sample. Our results using this real-world sample replicated those found in a carefully selected laboratory sample of women with and without a history of CSA. To our knowledge, this study offers the first empirical examination of sexual cognitions as they occur in the real world.

The findings from this research provide us with valuable information about the ecological patterning of maladaptive sexual cognitions. While past research on language–sexual schema links has been conducted in a highly artificial manner in the context of laboratory research and clinical interventions, our current findings demonstrate that the same patterns of language that are diagnostic of a history of CSA are manifest in unsolicited social media language as well. In addition, given the well-established links between cognitive and behavioral phenomena, our results may be considered a step forward in understanding how core sexual cognitions (typically measured under artificial conditions) may be related to real-world behaviors that can have dire physical and mental health consequences, such as impulsive or hypersexual behaviors and substance abuse.\textsuperscript{29} Furthermore, for many of the group differences found in the present study, the magnitude of several effects was particularly pronounced relative to previous research in this area. While the greater magnitude of several differences should be considered tentatively, the current results can be considered preliminary information for better understanding just how impactful and disruptive phenomena such as CSA may be upon cognitive patterns in the real world.

Our results may be particularly relevant for clinicians who treat patients with a history of CSA. Particularly during intake sessions, a rapid analysis of a client’s past social media language will be useful for identifying possible CSA history, as well as the severity of maladaptive cognitive patterns. Following intake, repeated analysis of social media language will likely also prove highly informative. Past research has found that changes in sexual self-schemas and cognitions over the course of a clinical intervention are an important marker of treatment success when working with this

---

**Table 3. Analytic Statistics of Linguistic Categories**

<table>
<thead>
<tr>
<th>Category</th>
<th>Previous research: comparison of NSA group to CSA group\textsuperscript{a}</th>
<th>Current study: comparison of \textit{/r/Sex} board to \textit{/r/AdultSurvivors} board</th>
<th>Online sample replication?</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(F) (\text{df})</td>
<td>(t) (\text{df})</td>
<td></td>
</tr>
<tr>
<td>Personal pronouns</td>
<td>16.63*** 210</td>
<td>9.691*** 109.49</td>
<td>Yes</td>
</tr>
<tr>
<td>“I”</td>
<td>6.43* 212</td>
<td>10.20*** 106.57</td>
<td>Yes</td>
</tr>
<tr>
<td>Affect words</td>
<td>5.07* 213</td>
<td>3.74*** 107.59</td>
<td>Yes</td>
</tr>
<tr>
<td>Positive emotion words</td>
<td>9.66** 213</td>
<td>6.54*** 114.68</td>
<td>Yes</td>
</tr>
<tr>
<td>Negative emotion words</td>
<td>24.18*** 213</td>
<td>9.40*** 104.13</td>
<td>Yes</td>
</tr>
<tr>
<td>Distancing or authenticity words</td>
<td>4.68* 207</td>
<td>4.78*** 108.44</td>
<td>Yes</td>
</tr>
</tbody>
</table>

\(\text{From Lorenz and Meston.}^{15}\)  
\(*p<0.05; **p<0.01; ***p<0.001.\)

---

**Table 4. Descriptive Statistics for Sexual Self-Schema Categories**

<table>
<thead>
<tr>
<th>Theme</th>
<th>Previous research\textsuperscript{a}</th>
<th>Current study</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NSA ((N=101)) (\text{Mean (SD)})</td>
<td>(\text{/r/Sex board (N=2,959)}) (\text{Mean (SD)})</td>
<td>(\text{/r/AdultSurvivors (N=100)}) (\text{Mean (SD)})</td>
</tr>
<tr>
<td>Virginity</td>
<td>2.70 (1.52)</td>
<td>2.01 (1.54)</td>
<td>1.76 (1.06)</td>
</tr>
<tr>
<td>Abuse</td>
<td>0.90 (0.82)</td>
<td>0.85 (0.88)</td>
<td>2.15 (1.24)</td>
</tr>
<tr>
<td>Relationship</td>
<td>2.03 (1.34)</td>
<td>1.11 (1.10)</td>
<td>0.80 (0.86)</td>
</tr>
<tr>
<td>Attraction</td>
<td>0.58 (0.57)</td>
<td>0.35 (0.64)</td>
<td>0.70 (0.66)</td>
</tr>
</tbody>
</table>

\(\text{From Stanton et al.}^{6}\)
population. Following a targeted intervention, the sexual self-schemas of women with a history of CSA have been shown to become more similar to those of nonabused women, and these changes were associated with decreases in depression and post traumatic stress disorder symptoms. Indeed, altering maladaptive schemas is a core element of cognitive behavioral therapy. Our replication of laboratory findings in a large online sample provides further support for the importance of attending to sexual self-schemas during treatment.

Our current research has one primary limitation: given the nature of public archival data from online forums like Reddit, demographic data on the posters in our sample were not available. For example, it is possible that individuals posting to the r/AdultSurvivors board may not have experienced sexual abuse during childhood; they may instead have been seeking information on behalf of family or friends, or they may simply be asking questions about the topic. However, this feature of our study may also be considered one of its greatest strengths. We found that previous laboratory findings accurately extended into a sample over which we had no control or condition verification. While previous studies have included only females who were heavily screened on factors like sexual activity, suicidality, and recent traumatic event histories, our online sample is likely composed of a far less homogeneous group of individuals. Nevertheless, the hypothesized psychological patterns in the current sample’s language provided perfect conceptual replication of previous work.

Finally, our current study provides strong support for the notion of a “flipped” paradigm to conducting initial clinical research. While it is true that our current study shows that carefully-elicited language patterns from laboratory studies are also diagnostic in messy real-world contexts, the opposite logically holds true as well. The ability to search for broad group differences in online samples may facilitate initial coarse understandings of difficult-to-study psychological phenomena, allowing researchers to dedicate limited “up-front” resources to an area of study, then honing in on promising research areas downstream using more sophisticated laboratory paradigms. Such an approach would allow research to gain an initial understanding of psychological phenomena by first discovering patterns “in the wild,” thereafter honing in on specific processes using more traditional, empirical laboratory techniques.

Acknowledgments

This research was supported, in part, by grants from the National Institutes of Health (5R01GM112697-02), the John Templeton Foundation (#48503), and the National Science Foundation (IIS-1344257). The views, opinions, and findings contained in this chapter are those of the authors and should not be construed as position, policy, or decision of the aforementioned agencies, unless so designated by other documents.

Author Disclosure Statement

No competing financial interests exist.

References


Address correspondence to:
Ryan L. Boyd, PhD
Department of Psychology
The University of Texas at Austin
108 E. Dean Keeton, Stop A8000
Austin, TX 78712
E-mail: ryanboyd@utexas.edu