Psychological Medicine (2016), **46**, 2605–2615. © Cambridge University Press 2016
doi:10.1017/S0033291716001215

ORIGINAL ARTICLE

Self-reference in psychosis and depression: a language marker of illness

S. K. Fineberg[1], J. Leavitt[1], S. Deutsch-Link[2], S. Dealy[3], C. D. Landry[1,4], K. Pirruccio[3], S. Shea[1], S. Trent[5], G. Cecchi[6] and P. R. Corlett[1]*

[1] *Department of Psychiatry, Yale University, New Haven, CT, USA*
[2] *School of Medicine, Yale University, New Haven, CT, USA*
[3] *Yale College, New Haven, CT, USA*
[4] *College of Physicians and Surgeons, Columbia University, New York, NY, USA*
[5] *Knox College, Galesburg, IL, USA*
[6] *IBM, Thomas J. Watson Research Center, Yorktown Heights, NY, USA*

Background. Language use is of increasing interest in the study of mental illness. Analytical approaches range from phenomenological and qualitative to formal computational quantitative methods. Practically, the approach may have utility in predicting clinical outcomes. We harnessed a real-world sample (blog entries) from groups with psychosis, strong beliefs, odd beliefs, illness, mental illness and/or social isolation to validate and extend laboratory findings about lexical differences between psychosis and control subjects.

Method. We describe the results of two experiments using Linguistic Inquiry and Word Count software to assess word category frequencies. In experiment 1, we compared word use in psychosis and control subjects in the laboratory (23 per group), and related results to subject symptoms. In experiment 2, we examined lexical patterns in blog entries written by people with psychosis and eight comparison groups. In addition to between-group comparisons, we used factor analysis followed by clustering to discern the contributions of strong belief, odd belief and illness identity to lexical patterns.

Results. Consistent with others' work, we found that first-person pronouns, biological process words and negative emotion words were more frequent in psychosis language. We tested lexical differences between bloggers with psychosis and multiple relevant comparison groups. Clustering analysis revealed that word use frequencies did not group individuals with strong or odd beliefs, but instead grouped individuals with any illness (mental or physical).

Conclusions. Pairing of laboratory and real-world samples reveals that lexical markers previously identified as specific language changes in depression and psychosis are probably markers of illness in general.

Received 10 December 2015; Revised 4 May 2016; Accepted 5 May 2016; First published online 29 June 2016

Key words: Depression, lexical analysis, psychosis, schizophrenia, self-reference.

Introduction

Language has long been studied in schizophrenia. We and others have recently hypothesized that language features may do more than describe experience and mark thought disorder; lexical markers may report on underlying psychopathology. Here, we harnessed the power of 'real-world' writing from public blogs to test the ecological validity and specificity of previously reported differences between psychosis and control subjects in laboratory samples.

* Address for correspondence: P. Corlett, Ph.D., Abraham Ribicoff Research Facilities, Connecticut Mental Health Center, 34 Park Street, New Haven, CT 06511, USA.
(Email: Philip.corlett@yale.edu)

Computational methods for lexical analysis can distinguish language from psychosis and control groups

Recently, computational methods have facilitated the rapid assessments of language features in large text corpora. Several groups have compared the frequencies of words or word categories in psychotic *v.* control subjects, or used language to classify subjects by diagnosis (Junghaenel *et al.* 2008; Strous *et al.* 2009; Hong *et al.* 2015), by degree of anhedonia (Buck *et al.* 2015*a*), and by social functioning and metacognition (Minor *et al.* 2015).

Self-referential words as markers of psychosis

Frequency of self-referential pronouns is a potential marker of psychosis. Three groups have independently identified self-referential pronouns as differing in

https://doi.org/10.1017/S0033291716001215 Published online by Cambridge University Press

Table 1. *Demographic and symptom data for subjects in the interview experiment*

	Setting	Diagnoses, *n*	*n*	Gender, *n*	Mean age, years (S.D.)	Mean duration of education, years (S.D.)	Mean SAPS (S.D.)	Mean SANS (S.D.)
Psychosis	15 hospitalized, 8 stable	Scz 6, sczaff 8, bipolar 7, NOS 1	23	13 male	35.2 (10.4)	12.6 (1.4)	32.1 (28.7)	22.3 (12.2)
Control		None 16, depr 5, anx 1, bipolar 1	23	8 male	28.2 (13.7)	14.8 (1.7)	4.1 (6)	2.3 (4)
p[a]				0.15	0.06	2.61×10^{-5}	4.24×10^{-5}	3.23×10^{-9}

S.D., Standard deviation; SAPS, Scale for the Assessment of Positive Symptoms; SANS, Scale for the Assessment of Negative Symptoms; scz, schizophrenia; sczaff, schizo-affective disorder; bipolar, bipolar disorder; NOS, psychosis not otherwise specified; depr, depression; anx, anxiety

[a] p Values represent significance on the result of Student's t test comparing the two groups.

frequency between psychotic and non-psychiatric control subjects: Strous *et al.* (2009) in Hebrew writing samples, Hong *et al.* (2015) in English speech samples, and Buck & Penn (2015) in English speech samples. Furthermore, two of the three diagnostic classifiers (computational algorithms to sort cases into diagnostic categories, in this case based on word frequencies) which have been reported rely heavily on self-referential pronouns, and both achieve >75% accuracy to distinguish psychosis from control subjects (Strous *et al.* 2009; Hong *et al.* 2015).

Examining lexical changes in patient samples and novel controls opens new windows on etiology

We know that specific word frequencies differ in psychosis *v.* control subjects in the laboratory – we aimed to extend this finding to understand how well these observed differences generalize to language samples produced outside the laboratory or clinic, and what aspects of psychotic illness are associated with these changes.

We tested lexical features in two corpora, one from patient interviews and another from online blog entries. The interview corpus allowed us to replicate others' findings and relate the observed lexical markers to specific symptoms. In particular, we tested three candidate lexical markers of psychosis suggested by others' work: two markers of self-reference: I, biological processes (Strous *et al.* 2009; Buck & Penn, 2015; Hong *et al.* 2015), and negative emotion (Cohen *et al.* 2008, 2009; Buck *et al.* 2015*b*; Minor *et al.* 2015).

The blog corpus included carefully chosen comparison groups aimed at differentiating the experiences of mental illness, chronic illness, strong belief system, odd beliefs, and social isolation. In addition to validating lexical differences in real-world samples, we aimed to understand the dimensions of psychosis relevant to observed lexical differences.

Method

Speech study

Interview subjects

This protocol was approved by the Yale Institutional Review Board, and consistent with the Declaration of Helsinki. We recruited 23 subjects with psychosis from the in-patient psychiatric hospital and our outpatient clinic. We included people currently treated for psychosis in the context of schizophrenia, schizo-affective disorder or bipolar disorder. Hospitalized patients were admitted for current exacerbation of psychotic symptoms. Subjects with stable psychosis and 23 control subjects without any history of psychosis were identified by posters and word of mouth, and initial inclusion criteria verified by brief telephone screen. Written informed consent for all subjects was obtained by standard procedures. See subject demographics in Table 1 (summary) and online Supplementary Table S1 (details).

Symptom scales

To quantify and describe psychotic symptoms in both groups, trained interviewers met with each subject to discuss psychiatric history and to administer the Scale for the Assessment of Positive Symptoms (SAPS) (Andreasen, 1984) and the Scale for the Assessment of Negative Symptoms (SANS) (Andreasen, 1983) (for results, see Table 1 and online Supplementary Table S1).

https://doi.org/10.1017/S0033291716001215 Published online by Cambridge University Press

Table 2. *Blog groups listed with number of entries in each group and features of interest*

Group	*n*	Chronic illness	Mental illness	Odd belief	Strong belief	Pro-social	Embraces group identity
Psychosis	54	+	+	+	+		
Morgellon's	37	+	+	+	+	+	+
Depression	48	+	+		?		
Aspie	45	+	+				+
Spinal cord injury	41	+					
Cancer	48	+				±	±
Evangelical	48				+	+	+
Anti-religion	49				+	±	+
Conspiracy	44			+	+	±	+

+, Feature present; ?, feature may be present; ±, feature may or may not be present.

Interview speech samples

Interviews were digitally recorded and transcribed. Subjects responded to the prompt 'We like to begin by hearing about you. Would you tell us a little about yourself?' Each was encouraged to continue talking for 10 min. This technique is similar to that in recent reports of lexical analysis in schizophrenia (Cohen *et al.* 2008; St-Hilaire *et al.* 2008; Strous *et al.* 2009; Buck *et al.* 2015*a*, *b*; Minor *et al.* 2015), though others focused more on emotional experience (Junghaenel *et al.* 2008; Cohen *et al.* 2009; Buck & Penn, 2015; Hong *et al.* 2015). This interview question was part of a larger experiment with three subsequent speech prompts and psychological tasks, which will be reported elsewhere.

Blog study

This protocol was determined by the Yale Institutional Review Board to be exempt from oversight.

Blog groups

To investigate the relationship between lexical features and specific aspects of psychosis, we compared word use in psychosis blogs with blogs from eight comparison groups. The comparison groups were blog-writers with chronic illness, mental illness, odd beliefs, strong beliefs and social (dis)engagement (Table 2). Each group was selected to share some, but not others, of these specific features. Each group was self-identified: the blogger declared that they belonged to the group, and there was no evidence that it was a parody. We read each blog in the Psychosis group for further evidence of psychotic illness, such as discussion of hallucinations, delusions, other symptoms, professional diagnosis, medication use, side effects, psychiatric hospitalization, legal problems and other recovery-oriented activities (Fig. 1*b*; for more detail, see online Supplementary Table S2).

Psychosis group. (1) Psychosis ($n = 54$). These bloggers identified as having schizophrenia or schizo-affective disorder. Their heterogeneous experiences included strong and/or odd beliefs, hallucinations, experiences of social stigma and feelings of isolation, as well as physical discomfort and sometimes a sense of pride or community. We aimed to identify language markers of these aspects of the experience of psychosis in the blogs.

Comparison groups

Illness groups. Five groups were formed using blog entries written by people with physical or psychiatric conditions. These groups varied in degree of strong or odd beliefs, physical discomfort, social isolation and sense of community.

(1) Morgellon's ($n = 37$). This group is from a largely Internet-based community of people who rally around physical symptoms (skin lesions, fatigue and overall poor health) that are very bothersome, and that they believe result from infestation or environmental exposure. Most medical experts understand this to be in the psychiatric realm, a kind of delusional parasitosis (Hylwa *et al.* 2012; Pearson *et al.* 2012; Misery, 2013; Mortillaro *et al.* 2013). The bloggers described their distress, their failed efforts to come to common understanding with medical professionals, and relief at finding community. This group had features of strong beliefs, odd beliefs, chronic illness (diagnosis), physical discomfort (itching and pain) and sense of community (group of believers online).
(2) Depression ($n = 48$). People with depression, especially those with suicidality, increase self-reference and bodily focus in their language (Pennebaker,

https://doi.org/10.1017/S0033291716001215 Published online by Cambridge University Press

(a)

Features in psychosis blogs

Gender	**22M, 32F**
Hallucinations	25/54
Delusions	31/54
Other symptoms or side effects	35/54
Taking medications	36/54
Hospitalizations	24/54
Professional diagnosis	43/54
Legal problems	11/54

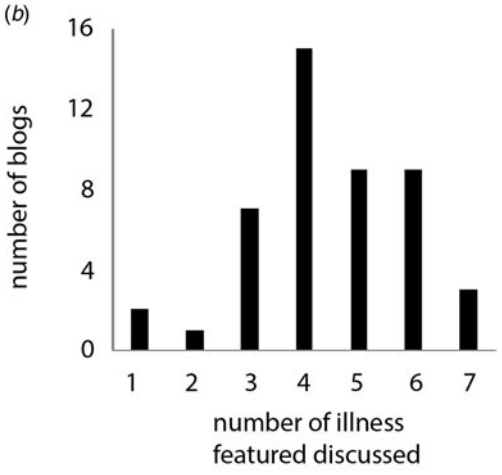


Fig. 1. Blog authors have features of psychosis. (*a*) Psychosis blog sample with regard to gender distribution of writers and what fraction of the sample included mention of each listed feature of psychosis. (*b*) Density of psychosis features in the sample, such that the left-most bar shows that our sample included two blogs mentioning one feature, and the right-most bar shows that our sample included three blogs that mentioned seven features. M, Male; F, female.

2011). In a cognitive–behavioral frame, depression is understood in terms of maladaptive beliefs (Beck, 1992), and people with depression have been shown to exhibit depressive realism – a tendency to make more accurate, less positively biased beliefs than control subjects (Alloy & Abramson, 1979). This group had features of social isolation, chronic illness (diagnosis) and some had physical discomfort.

(3) Aspie ($n = 45$). Some people with high-functioning autism spectrum disorder (previously called Asperger's) self-identify as 'Aspies'. We included this group because they are people with a chronic mental illness who have embraced this identity. We searched for blog authors who self-identified as 'Aspie' – not other autism blogs. These blogs included much speculation about what it is like for non-Aspies to emote and relate, and, for several bloggers, the relief at learning, as adults, that they are Aspies. This group has features of social isolation, chronic illness and strong sense of community

(4) Spinal cord injury (SCI, $n = 41$). We included this group with chronic non-psychiatric illness that may be visible to others (due to gait change or wheelchair use, for example). As the symptoms and behaviors of people with psychosis can at times make them recognizable to others in the community, we expected that many people with SCI may also feel social loss due to the visibility of their condition and attendant social stigma. However, we note that among the bloggers we identified, many had dramatic stories of physical and social feats: climbing mountains, establishing social organizations, etc., after their injury and often found a sense of community online. This group had features of social isolation, chronic illness, physical discomfort, and, for some, a strong sense of community.

(5) Cancer ($n = 48$). We included this group because they have an uncertain prognosis in the non-psychiatric realm: diagnosis may be very upsetting, and could be both isolating (due to not feeling well, others going on with their lives) and socially supportive (strong family support, support groups, gentle care in the hospital). Though we would not suggest that people would be happy to have cancer, many of these bloggers were proud to be survivors. In sum, this group had features of social isolation, chronic illness, physical discomfort and strong sense of community (as above).

No-illness groups. Three groups were formed without illness. They were characterized by strong and/or odd beliefs, but no spoken-of physical or mental health problems.

(6) Evangelical ($n = 48$). We selected this group for their strongly held beliefs that are accepted by

https://doi.org/10.1017/S0033291716001215 Published online by Cambridge University Press

much of society (not 'odd'). They described worries and conflicts with others who do not agree, but more of the focus was on explicating the beliefs, and on the positive affect related to participation in worldwide, local and family groups. This group had features of strong belief (religious doctrine), social engagement and related sense of pride/community (church, missionary work for some).

(7) Anti-religion ($n = 49$). We selected this group for their strongly held beliefs that, although perhaps less popular than religious beliefs, are still broadly accepted. Many of the bloggers in this group identified as having left (mostly Judeo-Christian) groups, and many were angrily separating themselves. This group had features of strong belief, social isolation (many left religious groups) and a sense of community online.

(8) Conspiracy ($n = 44$). We selected this group for their identification with strongly held beliefs that are often considered odd – not as acceptable as Evangelical and Anti-Religion beliefs. They were notable for seeking out others with shared beliefs – many of the blogs included extensive lists of links to other similar blogs – but also for describing the groups they oppose. Many were deeply involved with seeking out evidence to corroborate their beliefs. This group had features of strong belief, odd belief, and both isolation and community due to these beliefs. Though some may have clinical diagnoses, we saw no mention of psychiatric diagnosis, specific symptom-like experiences (other than their strong/odd belief) or psychiatric treatment on the blogs we included.

Blog collection

We collected the most recent 150+ word entry from blogs maintained by single authors (not edited or curated by someone else), written in English, and publicly available on the Internet. One trained study team member read each blog entry for inclusion criteria and processing. Each entry was formatted into a single plain text file using standard procedures (Pennebaker *et al.* 2007). This included correcting spelling errors when the intended word was obvious. We also removed citations and quotations longer than two sentences.

Lexical analysis

Linguistic Inquiry and Word Count 2007 (LIWC) is a software program that counts the frequency of 68 particular word categories in a text corpus (Pennebaker *et al.* 2007) including function words (words that structure the sentence, such as pronouns, prepositions) and content words (words that indicate topic, such as nouns, regular verbs, and some adverbs and adjectives). For between-group comparisons, we compared the frequency of 54 categories (we excluded punctuation and some grammar super-categories, such as 'function words' and 'pronouns').

Statistical analyses

Between-group comparisons

For between-group comparisons, we checked that each word category was used with >1% mean frequency. For the interview sample, we used two-tailed Student's t tests to examine differences between the psychosis and control groups. For the blog sample, we used one-way analysis of variance (ANOVA) to examine differences between the nine groups. For both samples, we set significance at $p < 0.05$ after false discovery rate (FDR) correction (Benjamini & Hochberg, 1995) to reduce the likelihood of type I error from 54 simultaneous word categories tested. For the ANOVA, we used Tukey *post-hoc* tests for pairwise comparisons between the Psychosis group and each of the other groups, with significance set at $p < 0.05$. In the interview corpus, 35 of the 54 categories we tested represented more than 1% of the words in the sample (our cut-off for inclusion in further analysis). Of these, eight categories were significantly different between groups (online Supplementary Table S4).

Correlation analyses

Pearson's correlations were used to test the relationships between the three LIWC categories that are candidate lexical markers of psychosis based on the previous work described in the Introduction of this paper (I, Biological Processes, and Negative Emotions) and interview subject symptom scores (SAPS and SANS) with significance set at $p < 0.05$ (for results, see online Supplementary Table S3). We included positive symptom subscales for hallucinations, delusions, thought disorder and bizarre delusions. We also included negative symptom subscales for anhedonia, inattention, avolition, alogia and affective flattening.

Clustering

We used an unbiased clustering approach to parse the blog entries into groups based on LIWC results. We entered the frequencies of the 37 LIWC content word categories that were not umbrella categories, e.g. 'see', but not 'perceptions', into a principal components analysis to reduce them to five factors. Single blog entries were clustered based on these factor

https://doi.org/10.1017/S0033291716001215 Published online by Cambridge University Press

values using the 'mclust' algorithm in the statistical program 'R'. Mclust is a package for model-based clustering which determines the optimum number of clusters based on the Bayesian information criterion (Fraley *et al*. 2012).

Results

Two language samples: a laboratory sample and a real-world sample

To validate and extend the above-described previous findings about word use in the language of people with psychosis, we collected speech samples from patient and control subject interviews (for details, see Fig. 1*a* and online Supplementary Table S1) and writing samples from blogs by people with psychosis and eight comparison groups (for details, see Fig. 1*b* and online Supplementary Table S2).

Interview corpus

We compared demographics in the two interview subject groups (Fig. 1*a* and online Supplementary Table S1). Psychosis and control subject groups were age and gender matched. The control group had on average 2 more years of education than the psychosis group. Psychosis group diagnoses (made by their clinicians prior to the study) included schizophrenia, schizo-affective disorder and bipolar disorder. As expected, our interview elicited more positive and negative symptoms from the psychosis group.

Blog corpus

We collected blog entries from writers with psychosis and each of eight other groups designed to test aspects of psychosis that might be associated with language changes (Fig. 1*b*). Each writer contributed one entry to the corpus. We examined each psychosis group blog (multiple entries) for mentions of the writer having experiences consistent with psychosis. These features were present in many of the blogs (Fig. 1*a*) and many of the blogs included multiple features (Fig. 1*b*).

Three candidate lexical markers of psychosis

We analysed our interview sample, in which subjects spoke for 10 min on the prompt 'Tell me about yourself', for putative lexical markers of psychosis (Cohen *et al.* 2008, 2009; Junghaenel *et al*. 2008; St-Hilaire *et al.* 2008; Strous *et al.* 2009; Hong *et al.* 2015; Buck *et al.* 2015*b*; Buck & Penn, 2015; Minor *et al.* 2015) (Fig. 2, black bars). First-person pronouns (Fig. 2*a*, black bars, $t = 3.66$, FDR adjusted $p = 7.00 \times 10^{-3}$) and biological process words (Fig. 2*b*, black bars, $t = -4.34$, FDR adjusted $p = 1.67 \times 10^{-3}$) were more frequent in Psychosis than Control. Negative emotion words (Fig. 2*c*, black bars, $t = 3.66$, FDR adjusted $p = 0.02$) were also more frequent in Psychosis than Control.

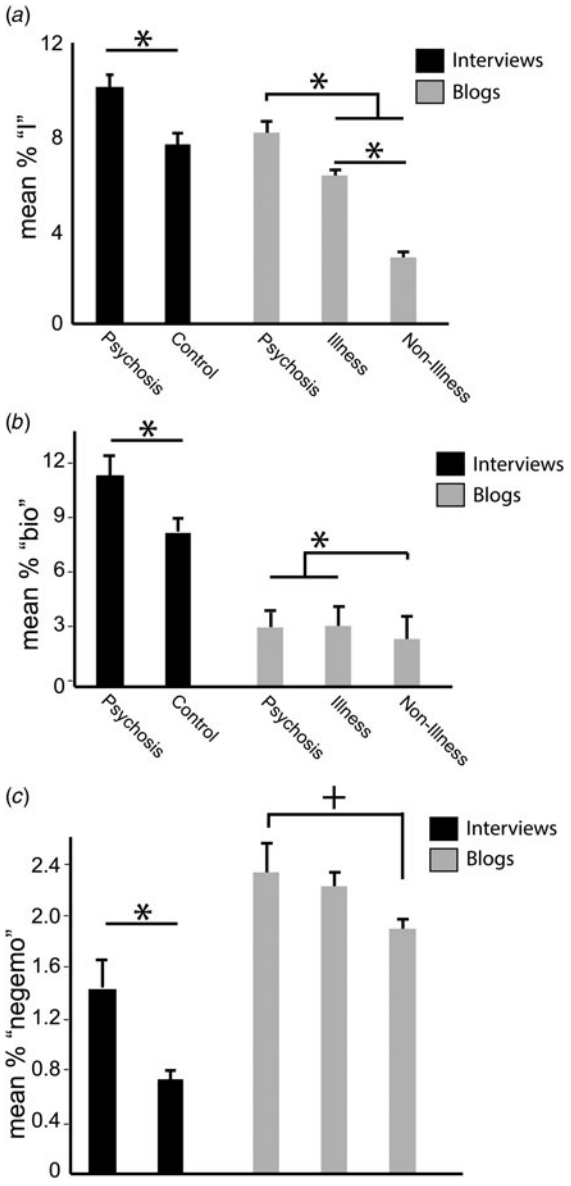


Fig. 2. Candidate lexical marker frequencies in interview and blog corpora. Frequencies of candidate lexical markers (*I*, *bio*, and *negemo*) in the interview (■) and blog (■) corpora. Values are means, with standard deviations represented by vertical bars. * Mean frequencies that significantly differ ($p < 0.05$). $^{+}$ Mean frequencies that trend toward a significant difference ($p < 0.1$).

Furthermore, frequency of these candidate lexical markers correlated with use of the other candidate markers (online Supplementary Table S3A), positive symptom scores (online Supplementary Table S3B) and negative symptom scores (online Supplementary Table S3C).

https://doi.org/10.1017/S0033291716001215 Published online by Cambridge University Press

Lexical patterns united bloggers with illness more than bloggers with strong or odd beliefs

In light of previous work identifying self-referential and negative emotion words as more frequent in both psychosis and depression (Junghaenel *et al.* 2008; Strous *et al.* 2009; Bedi *et al.* 2015; Buck *et al.* 2015*a*; Buck & Penn, 2015; Hong *et al.* 2015; Minor *et al.* 2015), we thought that these words could be markers of mental illness. However, we predicted that overall word use pattern in Psychosis would most reflect the strong odd beliefs (delusions) that often pervade the language of people with psychotic disorders. We used factor analysis followed by clustering to describe the patterns of word frequencies in the 37 thematic (non-grammar) categories. We predicted that these methods would reveal similarities between psychosis and other writers with strong odd beliefs (Conspiracy theorists, people with Morgellon's syndrome) and to a lesser degree, writers with strong beliefs (Evangelicals, Anti-religion writers). Contrary to our prediction, we found more word categories significantly differing from Psychosis in Conspiracy than any other group (Fig. 3*a*). Factor analysis revealed that word use in the Psychosis group was more similar to that of other groups with illness than to that of non-illness groups (Fig. 3*b* and *c*). To verify these groupings, we applied an unbiased clustering algorithm to the factor data. Each blog entry (without consideration of group) was assigned to a cluster: a three-cluster solution emerged. Psychosis (and other illness group) blog entries were predominantly assigned to cluster 2, whereas Conspiracy (and other strong belief) blogs were predominantly assigned to cluster 3 (Fig. 3*d*).

The three candidate markers were increased across illness groups

To test the revised hypothesis that the lexical markers that are increased in psychosis and depression are actually markers of illness in general rather than of having strong odd beliefs or mental illness in particular, we checked the frequency of self-referential and negative emotion words in the blog corpus. They did follow the predicted pattern (Fig. 2, gray bars). '*I*' was most frequent in psychosis, less in the Ill group, and even less in the Non-ill group (Fig. 2*a*, gray bars. Means: 8.16, 6.33, 2.84%, ANOVA $F = 82.71$ and $p = 7.61 \times 10^{-31}$, and $p < 0.05$ for psychosis *v.* ill, psychosis *v.* non-ill, and ill *v.* non-ill). '*Bio*' was more frequent in psychosis and the Ill group than in the Non-ill group (Fig. 2*b*, gray bars. Means: 3.18, 3.33, 2.31%, ANOVA $F = 10.71$ and $p = 2.94 \times 10^{-5}$, and $p < 0.05$ for psychosis *v.* non-ill and ill *v.* non-ill). '*Negemo*' was most frequent in psychosis, less in the Ill group, and even less in the Non-ill group (Fig. 2*c*, gray bars. Means: 2.39, 2.25, 1.91%, ANOVA $F = 3.65$ and $p = 0.03$, and $p < 0.1$ for psychosis *v.* non-ill, and ill *v.* non-ill).

Furthermore, in the interview corpus, these three categories correlated with each other and with symptom burden (online Supplementary Table S3).

Discussion

We replicated the finding that first-person pronouns are increased in psychosis (Junghaenel *et al.* 2008; Strous *et al.* 2009; Bedi *et al.* 2015; Buck *et al.* 2015*a*; Buck & Penn, 2015; Hong *et al.* 2015; Minor *et al.* 2015). We found that another self-referential category, biological processes, which includes words such as body and health, was more frequent in psychosis. Self-reference word categories such as 'I', 'ingestion,' and 'sex' have been previously identified as markers of depression (Oxman *et al.* 1988; Rude *et al.* 2004) and suicidality (Stirman & Pennebaker, 2001; Lester, 2009; Baddeley *et al.* 2011). In a corpus of published essays, we have previously found more self-referential words and fewer third-person plural words ('they') in depression than schizophrenia (Fineberg *et al.* 2015).

We replicated these findings in the more ecologically valid setting of blog writing: candidate markers were increased in psychosis *v.* eight comparator groups (for *I, negemo*). However, using factor analysis and clustering, we found that word use patterns across a large set of categories do not align groups that share the presence of strongly held or odd beliefs. Rather, patterns are most similar across groups with illness. Furthermore, in the interview sample, the three candidate lexical markers correlated with symptom intensity, and each marker was more frequent in the illness than non-illness blogs. Future work will determine if more symptomatic patients with other illnesses also use more of these lexical features.

Our finding that frequencies of these three lexical markers are positively correlated with one another in our interview subjects could suggest that these lexical markers relate to some common illness-associated process in these subjects.

Self-reference and illness experience

Illness is attended by shifts in self-concept and discomforts that pull one's attention inward. A functional magnetic resonance imaging marker of psychosis, dorsolateral prefrontal cortex activation, also correlates with distress (Corlett & Fletcher, 2012). Furthermore, belief-associated distress separates people with schizophrenia from others with high magical ideation, e.g. conspiracy theorists (Swami *et al.* 2010), and predicts delusion-motivated violence (Coid *et al.* 2013).

https://doi.org/10.1017/S0033291716001215 Published online by Cambridge University Press

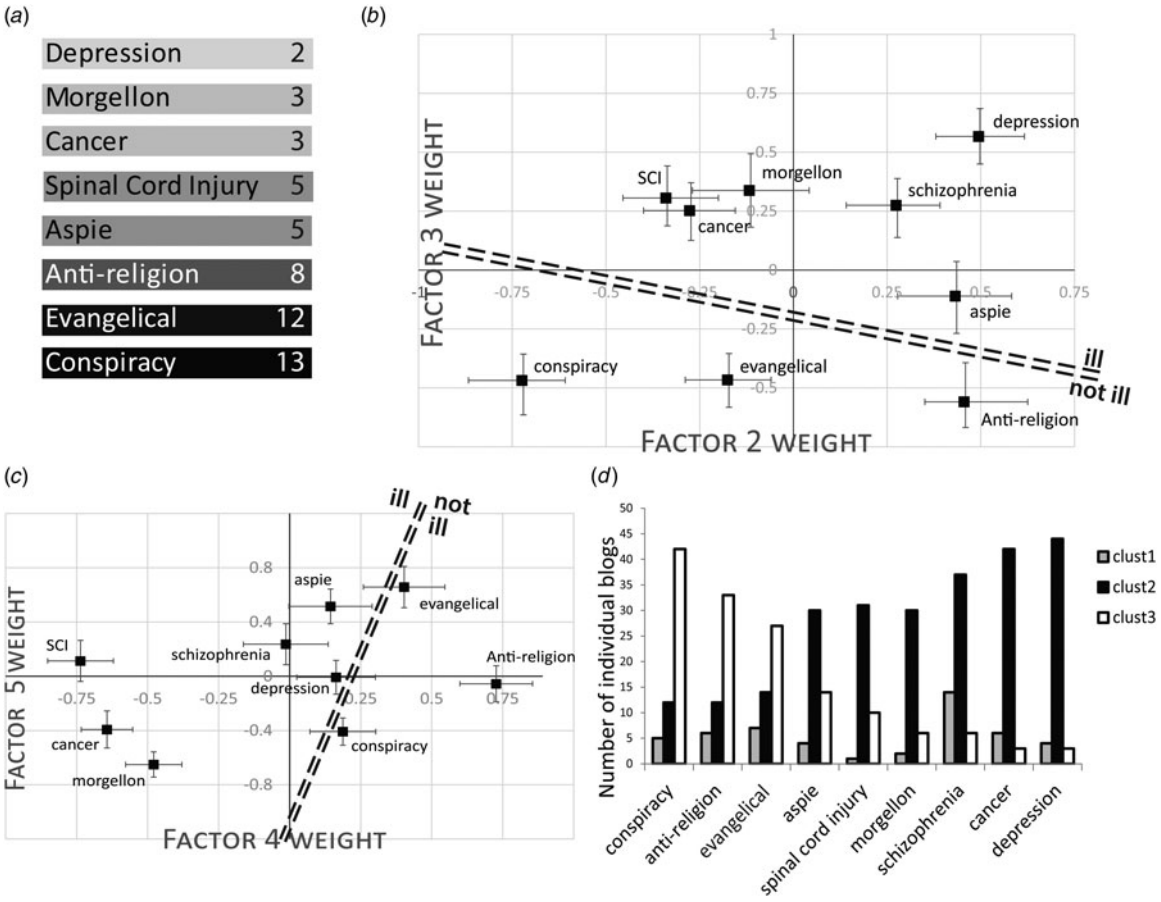


Fig. 3. Blog corpus thematic word frequencies and clustering results. (*a*) Numbers of word categories significantly different from Psychosis are listed for each group. (*b, c*) Factor weights for each group are plotted. (*d*) Number of blogs assigned to each cluster (clust) is marked for each group. SCI, Spinal cord injuries.

The similarity in word use among the ill groups in our sample also accords with research in neuro-immunology. Circulating inflammatory markers can promote 'sickness behaviors', a syndrome of low energy and depressed mood quite similar to major depression (Dantzer *et al.* 2008). Psychiatric symptoms are also linked to interferon-α treatment of hepatitis C (Smith *et al.* 2011), post-coronary recovery period (Lafitte *et al.* 2015) and chronic gut inflammation (Zonis *et al.* 2015). Inflammation has also been implicated in post-traumatic stress disorder (O'Donovan *et al.* 2015) and schizophrenia (Minor *et al.* 2015). We are not aware of any studies directly testing the relationship between inflammatory markers and word use patterns with LIWC, although there are some studies considering the health effects of sharing one's illness narrative (Pennebaker *et al.* 1988; Petrie *et al.* 1995, 2004). Perhaps the most parsimonious explanation of our present data is that the psychological state of being distressed is common to both physical and mental illness. This state may have immune consequences. Future studies relating language markers to inflammatory processes in psychiatric syndromes may help discern this effect and perhaps even target therapy.

Affective words

Previous data on affect word use in psychosis are mixed. Some have found more negative emotion words in psychosis *v.* control, specifically in discussion of neutral/unpleasant *v.* pleasant topics (Cohen *et al.* 2009). In psychosis with anhedonia, both decreased (Cohen *et al.* 2008) and increased use have been documented (Cohen *et al.* 2009). One study found that patients with anhedonia use more negative emotion words when prompted to talk about pleasant experiences (Cohen *et al.* 2009). However, Junghaenel *et al.* found no relationship between affect word use and anhedonia (Buck *et al.* 2015*b*), and they and others have reported no significant difference in negative emotion words in schizophrenia *v.* control subjects (Junghaenel *et al.* 2008; St-Hilaire *et al.* 2008). Two groups have reported no differences between groups in positive emotion word use (Cohen *et al.* 2008, 2009), but

https://doi.org/10.1017/S0033291716001215 Published online by Cambridge University Press

Junghaenel *et al.* found fewer positive emotion words in schizophrenia *v.* control subjects ($F = 4.9$, $p < 0.05$) (Junghaenel *et al.* 2008). In our speech samples, negative emotion words were more frequent in Psychosis than Control, and positive emotion words did not differ.

As our sample seems to have included more acutely psychotic subjects (many of our subjects were interviewed during psychiatric hospital admissions, whereas other studies included mostly stable out-patients), we wondered whether negative emotion words might correlate with symptom burden, perhaps explaining differences between reported results. Others have found that emotion word use in schizophrenia but not control subjects was positively correlated with trait anxiety (St-Hilaire *et al.* 2008). In another study, negative emotion word frequency correlated with PANSS score (Minor *et al.* 2015), and with the global affective flattening subscale of the SANS (not the overall SANS score) (Cohen *et al.* 2008). Positive emotion words did not correlate with SAPS or SANS (Cohen *et al.* 2008) or PANSS (Minor *et al.* 2015). Positive but not negative emotion word frequency correlated with social functioning (Cohen *et al.* 2008). Here, we found that frequency of negative emotion words was positively correlated with positive and negative symptom burden (online Supplementary Table S3). This, together with the positive correlations between self-referential words and symptom scores, suggests that these word categories could be markers of the intensity of symptoms.

Study limitations

Though our interview sample is matched on age and gender, we note that the discrepancy in education may reflect not only differences in years in the classroom, but also the vulnerable age when many of our psychosis subjects may have had education interrupted by symptom onset. By contrast, many of our control subjects are students with trajectory toward yet more education.

Language collected from blogs also has important limitations: most critically, we can not verify the identity or affiliations of the authors. Without meeting these writers or having access to their psychiatric histories, we cannot be certain of how well each blogger fits into their assigned group, and how many might possibly fit into another group as well. This does increase the potential for greater heterogeneity in the samples. However, we did find mention of experiences consistent with having a psychotic disorder in most of the psychosis blogs (Fig. 1). Furthermore, we did see similar word categories differ in our Interview sample (including many acutely ill patients). Also, writers with diagnosed conditions may be a subset of those with illness: those doing well enough to engage in the blogging project. Our correlation results from the Interview experiment do, however, suggest that self-referential and negative emotion words actually correlate with more not fewer symptoms.

Lastly, Conspiracy, Anti-religion and Evangelical bloggers may write more about strong belief than about themselves, which could account for some of the observed differences. However, the consideration of these data in the context of interview studies (ours and others') suggests that the differences we observed are in fact quite similar to those seen in Psychosis and Control subjects responding to the same prompts. Nonetheless, future studies to examine language from people talking about strong beliefs or other kinds of prompts will be important.

Conclusions

The similarities we have observed in lexical markers across mental and physical illnesses reinforce the importance of considering common mechanisms as we describe the pathophysiology of mental illness. This work does not allow us to define causality (this would be better understood in a prospective repeated-measures design). Future work will help us to understand the relative contributions of the social and biological aspects of illness to changes in lexical markers. In particular, lexical analyses may be of interest as novel therapies are introduced, word use metrics can be simply (and cheaply) tracked through illness onset, recovery and relapse across diagnoses. Based on our present findings, word use may represent an assay of current clinical status, distress and discomfort, regardless of specific symptoms and diagnosis.

Supplementary material

The supplementary material for this article can be found at http://dx.doi.org/10.1017/S0033291716001215

Acknowledgements

This work was supported by the Connecticut Mental Health Center (CMHC) and Connecticut State Department of Mental Health and Addiction Services (DMHAS). P.R.C. was funded by an IMHRO/Janssen Rising Star Translational Research Award and CTSA Grant Number UL1 TR000142 from the National Center for Research Resources (NCRR) and the National Center for Advancing Translational Science (NCATS), components of the National Institutes of Health (NIH), and NIH roadmap for Medical Research. S.K.F. was supported by National Institute of Mental Health (NIMH) grant no. 5T32MH019961,

https://doi.org/10.1017/S0033291716001215 Published online by Cambridge University Press

'Clinical Neuroscience Research Training in Psychiatry,' and by a NARSAD Young Investigator Grant. The contents of this work are solely the responsibility of the authors and do not necessarily represent the official view of the NIH or the CMHC/DMHAS.

References

Alloy LB, Abramson LY (1979). Judgment of contingency in depressed and nondepressed students: sadder but wiser? *Journal of Experimental Psychology-General* **108**, 441–485.

Andreasen NC (1983). *The Scale for the Assessment of Negative Symptoms* (*SANS*). The University of Iowa: Iowa City, IA.

Andreasen NC (1984). *The Scale for the Assessment of Positive Symptoms* (*SAPS*). The University of Iowa: Iowa City, IA.

Baddeley JL, Daniel GR, Pennebaker JW (2011). How Henry Hellyer's use of language foretold his suicide. *Crisis* **32**, 288–292.

Beck AT (1992). Cognitive therapy – a 30 year retrospective. *Which Psychotherapies in Year 2000?* **6**, 13–28.

Bedi G, Carrillo F, Cecchi GA, Slezak DF, Sigman M, Mota NB, Ribeiro S, Javitt DC, Copelli M, Corcoran CM (2015). Automated analysis of free speech predicts psychosis onset in high-risk youths. *npj Schizophrenia* **1**, 15030.

Benjamini Y, Hochberg Y (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society* **57**, 289–300.

Buck B, Minor KS, Lysaker PH (2015*a*). Differential lexical correlates of social cognition and metacognition in schizophrenia; a study of spontaneously-generated life narratives. *Comprehensive Psychiatry* **58**, 138–145.

Buck B, Minor KS, Lysaker PH (2015*b*). Lexical characteristics of anticipatory and consummatory anhedonia in schizophrenia: a study of language in spontaneous life narratives. *Journal of Clinical Psychology* **71**, 696–706.

Buck B, Penn DL (2015). Lexical characteristics of emotional narratives in schizophrenia: relationships with symptoms, functioning, and social cognition. *Journal of Nervous and Mental Disease* **203**, 702–708.

Cohen AS, Alpert M, Nienow TM, Dinzeo TJ, Docherty NM (2008). Computerized measurement of negative symptoms in schizophrenia. *Journal of Psychiatric Research* **42**, 827–836.

Cohen AS, St-Hilaire A, Aakre JM, Docherty NM (2009). Understanding anhedonia in schizophrenia through lexical analysis of natural speech. *Cognition and Emotion* **23**, 569–586.

Coid JW, Ullrich S, Kallis C, Keers R, Barker D, Cowden F, Stamps R (2013). The relationship between delusions and violence: findings from the East London First Episode Psychosis Study. *JAMA Psychiatry* **70**, 465–471.

Corlett PR, Fletcher PC (2012). The neurobiology of schizotypy: fronto-striatal prediction error signal correlates with delusion-like beliefs in healthy people. *Neuropsychologia* **50**, 3612–3620.

Dantzer R, O'Connor JC, Freund GG, Johnson RW, Kelley KW (2008). From inflammation to sickness and depression: when the immune system subjugates the brain. *Nature Reviews Neuroscience* **9**, 46–56.

Fineberg SK, Deutsch-Link S, Ichinose M, McGuinness T, Bessette AJ, Chung CK, Corlett PR (2015). Word use in first-person accounts of schizophrenia. *British Journal of Psychiatry* **206**, 32–38.

Fraley C, Raftery AE, Murphy TB, Scrucca L (2012). *mclust Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation*. Department of Statistics, University of Washington: Seattle, WA.

Hong K, Nenkova A, March ME, Parker AP, Verma R, Kohler CG (2015). Lexical use in emotional autobiographical narratives of persons with schizophrenia and healthy controls. *Psychiatry Research* **225**, 40–49.

Hylwa SA, Foster AA, Bury JE, Davis MD, Pittelkow MR, Bostwick JM (2012). Delusional infestation is typically comorbid with other psychiatric diagnoses: review of 54 patients receiving psychiatric evaluation at Mayo Clinic. *Psychosomatics* **53**, 258–265.

Junghaenel DU, Smyth JM, Santner L (2008). Linguistic dimensions of psychopathology: a quantitative analysis. *Journal of Social and Clinical Psychology* **27**, 36–55.

Lafitte M, Tastet S, Perez P, Serise MA, Grandoulier AS, Aouizerate B, Sibon I, Capuron L, Couffinhal T (2015). High sensitivity C reactive protein, fibrinogen levels and the onset of major depressive disorder in post-acute coronary syndrome. *BMC Cardiovascular Disorders* **15**, 23.

Lester D (2009). Learning about suicide from the diary of Cesare Pavese. *Crisis* **30**, 222–224.

Minor KS, Bonfils KA, Luther L, Firmin RL, Kukla M, MacLain VR, Buck B, Lysaker PH, Salyers MP (2015). Lexical analysis in schizophrenia: how emotion and social word use informs our understanding of clinical presentation. *Journal of Psychiatric Research* **64**, 74–78.

Misery L (2013). Morgellons syndrome: a disease transmitted via the media. *Annals of Dermatology and Venereology* **140**, 59–62.

Mortillaro G, Rodgman C, Kinzie E, Ryals S (2013). A case report highlighting the growing trend of Internet-based self-diagnosis of 'Morgellon's disease'. *Journal of Louisiana State Medical Society* **165**, 334–336.

O'Donovan A, Chao LL, Paulson J, Samuelson KW, Shigenaga JK, Grunfeld C, Weiner MW, Neylan TC (2015). Altered inflammatory activity associated with reduced hippocampal volume and more severe posttraumatic stress symptoms in Gulf War veterans. *Psychoneuroendocrinology* **51**, 557–566.

Oxman TE, Rosenberg SD, Schnurr PP, Tucker GJ (1988). Diagnostic classification through content analysis of patients' speech. *American Journal of Psychiatry* **145**, 464–468.

Pearson ML, Selby JV, Katz KA, Cantrell V, Braden CR, Parise ME, Paddock CD, Lewin-Smith MR, Kalasinsky VF, Goldstein FC, Hightower AW, Papier A, Lewis B, Motipara S, Eberhard ML, Unexplained Dermopathy Study Team (2012). Clinical, epidemiologic, histopathologic and molecular features of an unexplained dermopathy. *PLOS ONE* **7**, e29908.

https://doi.org/10.1017/S0033291716001215 Published online by Cambridge University Press

Pennebaker JW (2011). *The Secret Life of Pronouns: What our Words Say About Us*. Bloomsbury Press: New York.

Pennebaker JW, Chung CK, Ireland M, Gonzales A, Booth RJ (2007). The Development and Psychometric Properties of LIWC2007 (software manual). University of Austin: Austin, TX (http://www.liwc.net/LIWC2007LanguageManual.pdf). Accessed June 2016.

Pennebaker JW, Kiecolt-Glaser JK, Glaser R (1988). Disclosure of traumas and immune function: health implications for psychotherapy. *Journal of Consulting and Clinical Psychology* **56**, 239–245.

Petrie KJ, Booth RJ, Pennebaker JW, Davison KP, Thomas MG (1995). Disclosure of trauma and immune response to a hepatitis B vaccination program. *Journal of Consulting and Clinical Psychology* **63**, 787–792.

Petrie KJ, Fontanilla I, Thomas MG, Booth RJ, Pennebaker JW (2004). Effect of written emotional expression on immune function in patients with human immunodeficiency virus infection: a randomized trial. *Psychosomatic Medicine* **66**, 272–275.

Rude SS, Gortner EM, Pennebaker JW (2004). Language use of depressed and depression-vulnerable college students. *Cognition and Emotion* **18**, 1121–1133.

Smith KJ, Norris S, O'Farrelly C, O'Mara SM (2011). Risk factors for the development of depression in patients with hepatitis C taking interferon-α. *Neuropsychiatric Disease and Treatment* **7**, 275–292.

St-Hilaire A, Cohen AS, Docherty NM (2008). Emotion word use in the conversational speech of schizophrenia patients. *Cognitive Neuropsychiatry* **13**, 343–356.

Stirman SW, Pennebaker JW (2001). Word use in the poetry of suicidal and nonsuicidal poets. *Psychosomatic Medicine* **63**, 517–522.

Strous RD, Koppel M, Fine J, Nachliel S, Shaked G, Zivotofsky AZ (2009). Automated characterization and identification of schizophrenia in writing. *Journal of Nervous and Mental Disease* **197**, 585–588.

Swami V, Chamorro-Premuzic T, Furnham A (2010). Unanswered questions: a preliminary investigation of personality and individual difference predictors of 9/11 conspiracist beliefs. *Applied Cognitive Psychology* **24**, 749–761.

Zonis S, Pechnick RN, Ljubimov VA, Mahgerefteh M, Wawrowsky K, Michelsen KS, Chesnokova V (2015). Chronic intestinal inflammation alters hippocampal neurogenesis. *Journal of Neuroinflammation* **12**, 65.

https://doi.org/10.1017/S0033291716001215 Published online by Cambridge University Press