ANNUAL INTERNATIONAL CONFERENCE

PROCEEDINGS

06 - 07 March 2017, Singapore

6th Cognitive and Behavioral Psychology (CBP 2017)

PUBLISHED AND ORGANIZED BY
GLOBAL SCIENCE & TECHNOLOGY FORUM (GSTF)

www.globalstf.org

6th Annual International Conference on Cognitive and Behavioral Psychology (CBP 2017)

6 – 7 March 2017
Singapore

Organized & Published By

STEERING INNOVATION, SERVING SOCIETY
www.globalstf.org

Organized, Published and Distributed by
Global Science and Technology Forum (GSTF)
6th Annual International Conference on Cognitive and Behavioral Psychology (CBP 2017)
Tel: +65 6327 0166
Fax: +65 6327 0162
www.globalstf.org | info@globalstf.org

E-mail: secretariat@cognitive-behavior.org
Website: http://cognitive-behavior.org/

6th Annual International Conference on Cognitive and Behavioral Psychology (CBP 2017)
Print ISSN: 2251-1865, E-Periodical ISSN: 2251-1881

Computational Language Analysis for Assessment of Schizophrenia

Lieve Beheydt, Bernard Sabbe, Livia De Picker, Jens Goetschalckx en Walter Daelemans.

Abstract

Computational language analysis can be a useful tool in the assessment of schizophrenia, which is currently for a large part subjective. Given the observed relations of language use with thought disorders, language analysis could play an important role in making the as-sessment more objective. We provide an overview of linguistic features implicated in schizophrenia that can be automatically ana-lyzed in a robust and accurate way given the current state of the art in natural language processing, and describe an explorative pilot study testing these features in the analysis of the language use of one schizophrenic patient.

1 Introduction

Research in schizophrenia is hampered by unclear assessment criteria. The problem is that 'impres-sionistic' assessment, invoked by clinical psy-chologists (and sometimes even made explicit in their interpretation of projective tests where pa-tients can associate on non-structured material like pictures or ink-blots), is not reflected in the pre-vailing biomedical schizophrenia concept or in the widely used structured assessment interviews. The latter focus on scoring the presence of 'nuclear' or 'first rank' symptoms as defined by Schneider, and on evaluating the poor outcome component added by Kraepelin (Ceccherini-Nelli & Crow, 2003). On the other hand, linguistic features of schizophrenic language as attested in literature on the language of schizophrenics (Covington, et al. 2005) and imme-diately recognized by experienced clinicians who are well acquainted with the Bleulerian phenome-nological entity approach of schizophrenia, stand out clearly. Schizophrenic patients associate on sound in normal speech, create neologisms by morphological play and even combine associations on sound primers with associations on semantic primers and even on emotional primers or random incidental primers. The thought process of schizo-phrenic patients is translated in their linguistic be-havior. The original Bleulerian approach, emphasizing this kind of thought disorder or disor-der of association is, unfortunately, not empirically well operationalized and, therefore, the current schizophrenia concept remains limited to the Schneider and Kraepelin conceptualization. Given the lack of validity of the biomedical (Schneiderian and Kraepelinian) schizophrenia concept (Blom, 2007) the present conceptualization constitutes a threat to empirical research of neuropsychological and neurobiological determinants of psychotic dis-orders.

A complementary and reliable contribution to a valid conceptualization of schizophrenia may be expected from natural language processing. Since language is a 'complex dynamic cognitive system which entails integration of multiple levels of lin-guistic and cognitive processing' (Marini et al., 2008), and linguistics has a long tradition in empir-ically analyzing verbal behavior, operationalized in computational methods for automatic language analysis, analysis of natural language of schizo-phrenics could improve and complement the Schneiderian and Kraepelian concept of schizo-phrenia with the empirically based linguistic con-cretization of Bleuler's approach (Ceccherini & Crow, 2003).

2 Assessment Difficulties in Schizophre-nia

The DSM-5 (APA, 2013), the fifth edition of the Diagnostical Statistical Manual (or the parallel ICD-10 criteria) is generally used to define schizo-phrenia on the basis of delusions, hallucinations, disorganized speech, grossly disorganized or cata-tonic behavior and negative symptoms, i.e. affec-tive flattening, alogia, or avolition as core symptoms. Apart from the symptoms, schizophre-nia is determined by the dysfunction criterium an

6th Annual International Conference on Cognitive and Behavioral Psychology (CBP 2017)

ISSN 2251-1865
doi: 10.5176/2251-1865_CBP17.37

the duration criterium. It is obvious that symptoms like disorganized speech or disorganized behavior as expressions of disorganized thought (cf. Bleuler), are heavily dependent on subjective interpretation. In systematic structured interviews like SCID I and Mini-plus, the diagnosis is based on self report on questions about unusual experiences like: 'Did you ever have the feeling that there was a conspiracy against you?'. Intelligent patients know it is better to deny such experiences because they sound 'odd'. This is why these questions are followed by three ratings by the clinician evaluating presence of disorganized speed, inability to follow the line of thought, signs of chaotic or catatonic behavior and the presence of negative symptoms, flattening of affect/ and or speech, loss of drive or goal oriented behavior (DSM-5; APA 2013). Such rating, however, makes the assessment of the most stigmatized psychiatric diagnosis very reliant on subjective interpretation. Since schizophrenia is a chronic disorder with psychotic exacerbations, also the interepisodic states without clear first-rank symptoms can be erratic to assess and a more longitudinal approach is essential to produce a trustworthy diagnosis. It is hence customary to ask patients about symptoms in earlier periods in life. But, patients do not always remember accurately enough, as memory problems are common cognitive deficits in schizophrenia (Nielsen, 2011; Kuperberg & Heckers, 2000). Further, lack of insight (in the disorder) is a typical feature of being psychotic, especially a lack of reality testing (DSM-5). Nevertheless, if chances of treatment are missed, it makes prognosis only worse (Hoff et al., 1999). Cognitive and functional deterioration in schizophrenia may be dependent on the length and number of untreated episodes (Nielsen, 2011). Finally, schizophrenia is a highly heterogeneous disorder (Joseph, Narayanaswamy & Venkatasubramanian, 2015), with high comorbidity and high overlap with other disorders like drug-induced psychosis or bipolar disorder, type 1 patients, who may also experience severe psychotic episodes. All this hinders specificity in assessment. In DSM-5, the subtypes of schizophrenia are already left out because of lack of differential validity. A categorical approach of complex disorders like schizophrenia appears not to be feasible, only a multidimensional approach might offer chances to enhance specificity. Factor analysis of psychiatric schizophrenic symptoms produced three independent symptom clusters: positive symptoms (the addition of features in comparison to normal: delusions, hallucinations etc.), negative symptoms (the lack of some abilities in comparison to normal: lack of initiative, lack of drive, lack of expression of emotions, …) and cognitive symptoms (Joseph, Narayanaswamy & Venkatasubramanian, 2015). As positive and negative symptoms are not two extremes of one and the same dimension (Mc Glashan, 1992), it is an interesting question whether the psychiatric symptom triad finds a corresponding or correlative translation in language behavior, and whether general language behavior could add specificity to the diagnosis.

3 Operationalization of Thought Disorders as Language Disorders

In psychiatric literature, language disorders of schizophrenics are often categorized as 'formal thought disorders' (Ketteler et al., 2012). That categorization concerns the form of thoughts more than the quality of content. Typical formal thought disorders are impoverished content, lack of quantty of information in speech, loss of aim (topic maintenance and topic change) and clanging (the sound of vowels being distracting or more important than the content for the association process) (Covington et al., 2005). The consequence of such thought disorder is that communication is disturbed and idiosyncratic. The most important linguistic features according to Andreasen using the TLI-scale (Thought and Language Index) (Andreasen 1986), not specific for schizophrenic patients, were derailment, loss of goal, limited content and tangentiality (partly irrelevant replies), less common are distraction, circumstantiality, neologisms, stilted language, blocking, word approximation (to find substitutes for existing words) and clanging. The first factor analysis yielded one factor only," verbosity" (Andreasen, 1979), characterized by high amount of speech, high level of derailment, illogic language, loss of goal and persistence, incoherence and pressure of speech. A later study (Andreasen, 1986) revealed two factors, "fluent disorganization" and "emptiness", in parallel with positive and negative symptoms. The two main factors, "fluent disorganization" (pressure, distractibility, derailment, loss of goal, and perseveration) and "emptiness" (poverty of speech and content)

being characteristic of mania and schizophrenia, respectively. Liddle et al. (2002) developed a simpler scale on the basis of Andreasen's TLI which yielded three factors : empoverishment, disorganization (e.g. derailment, peculiar syntax, peculiar words) and dysregulation (distraction and perseverance). Remarkably enough, the same features were found in the control group, which is an indication that they are no specific schizophrenic features of speech. Schizophrenic speech is a matter of a combination of frequencies of features. At the same time, the parallel with the psychiatric symptom clusters becomes strikingly obvious, with "impoverishment" being related to negative symptoms/psychomotor impairments, "disorganization" to positive symptoms and "dysregulation" to cognitive symptoms. A very interesting contribution from linguistics was made by the Clang-scale (Chen, 1996) who showed that assessment based on linguistic features had a higher specificity than the prevailing diagnostic system of the ICD-10 (and DSM-IVTR) (Ceccherini & Crow, 2003). The disadvantages of this scale are that it was based on spoken language only, that it was still highly dependent on ratings by an experienced rater and that it was not sensitive to more subtle language symptoms. Ketteler (Ketteler et al., 2012) subsequently developed the HOLT (Higher Order Linguistic Function Test) to overcome some of these problems, but his test was predominantly an experimental language task more than a standardized test. Also, the validation and reliability of this test are as yet not fully explored, as this is a recent test.

4 Objective Language Measures

The contribution of this line of research is to combine the best of two worlds. First of all, since even a severe disorder such as schizophrenia shows language deviations only in terms of fre-quencies, and not in categorical features (all ob-served abnormalities were found, in lower frequencies, in non-patients (Liddle et al. 2002)), it seems to be appropriate to approach language be-havior of schizophrenics with the same existing means of analysis as for standard language. The obvious advantage of such an approach is that it enables comparison with non-patient language and with other psychiatric disorders, too. Different types of language use might show different distributions of specified features. Additionally, a non-categorical approach deserves preference over a categorical one as it allows for a description of a psychiatric disorder in terms of a disease. It might indeed be a sign of stigmatization to categorize properties of psychiatric disease as categorically different, defining abnormality. In clinical practice, to overcome the lack of validity of the prevailing assessment system, many additional neuropsychological tests are administered to the patient to test a whole range of cognitive features. However, an encumbering administration of a vast test battery is not a very elegant and practical way for daily assessment as it is time and energy consuming for the patient and the clinician. Moreover, as is clear from the definition of schizophrenia, negative symptoms make elaborate testing difficult with patients suffering from a lack of drive and initiative, who can react in very unexpected ways, who grieve their loss of intellectual capacity, and who have a very high distractibility. In this regard, it is preferable to introduce objective measures for assessing natural language. But to avoid the just mentioned consequences of negative symptoms, a method should be designed that enables systematic formal analysis. To develop such a method, working with written data has obvious advantages. The benefits are threefold. Using written natural language makes it possible to create a large and hence representative database of language in psychopathology, which can subsequently be further extended and reused for new research purposes. Moreover, scrutinizing large databases is hardly possible with oral language samples. Oral language samples require transcription which cannot be automated with sufficient reliability. On the other hand, computational techniques make it already possible to analyze a considerable number of variables in large data collections of written language. As the standardization of the recording techniques is much easier with written language, and as data collection assignments enable researchers to standardize the subject, the length and the duration of the text samples, analyzability and comparability are highly increased with written data. When also objective computational measures are used to analyze the language samples, the problem of subjective ratings of human interpretation of complex material can be avoided. If, moreover, the same method can be applied to normal samples and to samples of related disorders and to schizophrenic

samples from an acute state, measures of gradation in severity or quality may be developed. This is important in the investigation of markers of psychopathology, where multidimensional graded measures of severity may help to trace the developmental course and specify gradual differences between normality and psychopathology. Different quantitative profiles of combinations of symptoms may possibly be related to basic neuropsychological processes, so that treatment may be better informed and optimized, prevention may be improved, in a way comparable to the approach of complex cardiovascular markers in physical medicine. Subsequently, profiles can be developed for research on endophenotypes, so important in genetic research of psychopathology (Van Der Gaag, Van Wijngaarden-Cremers & Staal, 2012). In short, computational research on natural language in schizophrenic patients may improve the understanding of physiopathology. More specifically, it could test two current competing psycholinguistic theories and foster new insights in neuropsychological functioning in schizophrenia. The first theory is that language dysfunctions in schizophrenia have to do with faster and more profound activation of semantic networks. The typical 'looser' association network could be explained by the heightened automatic or unconscious associative priming effect. In schizophrenia, direct and indirect priming effects are higher, especially in patients with positive symptoms. Apparently, conscious strategic processing is diminished in these patients. The second theory highlights the impairments of working memory and executive function. Specific for schizophrenic patients are impaired retention of information, task related processing of information, replacing or overwriting a dominant reaction in order to adapt to new task demands, thus leading to ambiguous reactions. In short, they experience problems in inhibitory and monitoring executive functions of memory. In linguistic terms, this means that schizophrenic patients are less sensitive to semantic restrictions, have more difficulties understanding lexical ambiguous sentences and suppressing dominant meanings of words in different contexts, and they tend to recognize infringements on logic less. With semantically narrowly related words they are less aware of syntactic improbabilities. The central problem seems to be situated in the combinatory mechanisms to build up a meaningful sentence, their understanding of sentences being too much driven by the dominant semantic content of words and existing semantic associations. The problem is not a purely syntactic deficit, as they are able to evaluate syntactic structure if it does not appeal too much to working memory (Kuperberg, 2010). Perhaps, the deficits in working memory and executive function make them rely more on semantic and associative qualities as a coping mechanism, which may eventually lead to very uncommon or idiosyncratic expressions.

5 Pilot Study

In an explorative pilot study, we implemented the methodology introduced above and tested it in a case study.

5.1 Approach

A computational analysis approach to written texts by schizophrenic patients will currently have to focus on lexical and syntactic aspects. It is well known that the major linguistic impairment in schizophrenic language, especially in schizophrenic speech, is formal thought disorder. Yet, formal thought disorder is particularly hard to operationalize, as it is dealing with logical sequencing on the discourse level, for which reliable analysis tools are lacking. Moreover, for diagnostic purposes, detecting formal thought disorder is less interesting as this symptom is unequally frequent in acute and chronic cases.

A lexical analysis, on the other hand, may prove highly rewarding, as a typical symptom of schizophrenic language is impairment of lexical access. Mistaken lexical choices appear to be frequent. A notorious example of such errors in lexical retrieval is: "Oh, it [life in hospital] was superb, you know, the trains broke, and the pond fell in the front doorway" (Oh et al., 2002, p. 235). Mistaken lexical choices also occur in normal speech, but not in the heightened frequency found in schizophrenic speech. Hence, if erroneous lexical choices abound in written schizophrenic language, a significant difference in frequency could be a reliable index of schizophrenia. Of course, operationalizing 'erroneous lexical choice' in a formal computational way is not a straightforward procedure.

Also typical for the lexical usage of schizophrenics is the use of "unusual words". Unusual words is, of course, not a standard linguistic term.

With schizophrenics it may, for instance, refer to low frequency words, typical for stilted speech. Schizophrenics tend to use rare words (Pinard & Lecours, 1983), so a relatively high rate of low frequency words (compared to frequency distributions of texts in a reference corpus) may be one of the indices of 'unusual word usage', as is the use of neologisms, personally made-up words (such as handshoe for 'glove'). This can be computed by matching against the word list of a control corpus. Unusual words may also make use of morphological processes: opaque compounds (which may also be a lexical choice problem) or unusual derivations and inflections. It appears that all these types of unusual words occur in schizophrenic speech. They might also turn up in written language.

Syntax and morphology are most often not impaired, and if impaired, only occasionally. This means that grammatical wellformedness would not constitute a reliable index of schizophrenia. However, syntactic simplification does seem to be a reliable index. It appears that syntactic complexity is diminished with increasing deterioration in schizophrenia (Covington et al. 2005: 91). Syntactic simplification can be measured using standard readability measures like mean utterance length, as utterance length correlated with syntactic complexity. This measure is usually defined as the total number of words divided by the total number of sentences. As this is a simple and stable measure, which is in general independent of sample length, it is a first safe approximation to syntactic complexity. It varies with text type (genre, register), however. Taking into account variation around the mean provides a more reliable measure. A more varied distribution (e.g. using the standard deviation) around the means generally indicates a greater syntactic complexity. Significant differences between these measures with those in reference texts may indicate differences in syntactic complexity.

One other possible measure of syntactic complexity that has been proposed to be a reliable measure of reduced syntactic complexity in schizophrenics, is that of 'phrasal complexity' as measured by the total number of sentences, clause conjunctions and clause embeddings (Bickerton 1990). This measure, in turn, can be automatically computed after using existing parsers.

5.2 Data

The pilot study was executed in 2014. A longitudinal approach was set up with the diary (2009-2011) of a schizophrenic patient, 50 years old, 45-47 years when keeping the diary. She was diagnosed with schizoaffective disorder, so there were also episodes of manic behavior. She had a first psychotic episode when she was 26, and five years later she had a second psychosis followed by a first residential stay in psychiatry. In 2005 she was admitted in a psychiatric ward with a third psychotic episode with observed delusions. Further, she had restrained intellectual functioning and fierce mood changes. In January 2010, a new hospitalization started which was preceded by three manic episodes in 2009. At the time of her admission she was manic again. The mania episode had started on the first of March 2010 and was over by March 15th. She had no structure and experienced severe problems with circadiurnal rythm. She was dismissed from the psychiatric ward on May 12th, 2010. The last available written text is dated December 27th 2011. In this pilot study, a lot of variables are not controlled. The coexisting mood pathology in schizoaffective disorder may, for instance, have had a strong impact on the results found, but the primary purpose of the present study was to investigate whether the designed methodology was efficient, whether any significant differences would appear, whether schizophrenic features could be recognized and operationalized, and whether a relationship with hospitalization or psychotic episodes would be found.

The text data for the longitudinal analysis is a time annotated digital written text in diary-style. It covers the period between July 14th 2009 and December 27th 2011. Because she did not write every day, analyses were made for each day sample and for each month. She wrote on 48 days and in 10 of the studied range of 30 months. Information for this period could be compared to the medical file for 2010 and 2011. This is the only period for which the presence of psychosis could be ascertained. It is indeed not known when the episodes of mania appeared in 2009 and whether they coincided with any writing of the patient. Comparisons with the control time segments were, however, used to make predictions about the mental state episodes in 2009. For statistical analysis, only 8433 tokens were analyzed on 18 days, distributed

over 4 months. 2080 tokens were written in known psychotic episodes of the patient, written on 8 days in one month. In 2010 and 2011, 6353 tokens were found of which it is certain that they were produced at a time she was not psychotic, written on 10 days, distributed over 3 months. For 2009 16,187 tokens are available, without information whether she was psychotic at the time of writing. These were used to make predictions about possible psychotic episodes. A reference language corpus was used as a comparative basis. This corpus was the only reference for the 2009 time segment because no medical file is available for this time segment.

The control corpus used for comparison was the STEVIN Dutch written corpus (SoNaR corpus) with more than 500 million words from different sources (Oostdijk et al. 2013). Here the Dutch language subcorpus for blogtext was used as a reference, because this is stylistically the most appropriate basis for comparison with diary fragments. It contains 188,233 words collected in 773 blogs, with a mean of 153 tokens for every section. In the text of the patient, titles were left out because they were no complete sentences and that would have biased the averages of length and syntax and it could have produced unreliable standard deviations. The minimum length of a blog was set to 50 tokens. The preprocessing of the text in this way produced 620 files with a minimum of 51 tokens and a maximum of 853 tokens.

5.3 Linguistic Analysis

We analyzed general text features, and selected features typical for schizophrenia as described above.

General linguistic analysis of thematic words

The LIWC (Linguistic Inquiry and Word Count) is a software application for text analysis developed by Pennebaker, Booth and Francis. It measures for different thematic categories of words (collected in a dictionary) the frequency of words associated with each category in a text. We only used the 66 dictionaries of the Dutch version of the system (Zijlstra et al. 2004) and did our own frequency analysis. Categories contain words reflecting a psychological state (like positive emotions), but also pronouns which may be linked to demographic or psychological style features. Some words may belong to different categories, because of different meanings of the word or because of partial overlap of the categories. Although LIWC was not developed specifically for schizophrenia, it gives representations of psychological processes that may reflect changes in mental state by changes in relative frequencies of certain categories of words that could contribute to the detection of schizophrenic episodes.

Analysis of linguistic features related to Schizophrenia

For evaluating patterns of schizophrenic language the following text features were investigated.

Lexical diversity was measured using Type Token Ratio (TTR), the ratio of different words (types) to the total number of words (tokens) in the text sample. TTR is expected to be significantly lower for schizophrenics (Vetter 1970:9).

Richness of the semantic content was investigated by computing specificity as measured by position in the hypernymy hierarchy of the Cornetto semantic wordnet for Dutch. The lower in the genericity hierarchy a noun or verb is, the more specific. Text specificity was measured as the mean specificity of all content words of the text. It was expected to be lower in psychotic episodes. Poverty in content and vagueness should also result in higher relative frequencies of indefinite nouns (such as thing, something, nothing, ...) and low specificity verbs (such as be, go, get, ...). To compute this index a list of 5 indefinite nouns and 25 low specificity verbs was selected and their frequencies computed. The frequency of these words was expected to increase in episodes with more negative symptoms.

Semantic cohesion is reported to be compromised in schizophrenic patients as they show a high tendency to revert to semantic associations between more unrelated words. The latter tendency could be expected to be reflected in the semantic cohesion of words near to each other in a text. For every ten nouns of the target text semantic relatedness was measured for all pairs of words by computing the length of the shortest pathway in the Cornetto semantic network (following relations of hypernymy, causality, involvedness etc.). The average of the mean similarity of each of the ten

words to the nine others is used as a measure of semantic relatedness. The hypothesis was that cohesion would be decreased in psychotic periods.

Syntactic complexity. Schizophrenic language is supposed to be syntactically less complex; there could be syntactic simplification in periods of negative symptoms of the disorder (Covington et al. 2005:91). Syntactic complexity is computed using the dependency analysis present in the Frog text analysis software (Bosch et al., 2011). The average syntactic parse tree complexity for all sentences in the text is used as a measure of syntactic complexity. This measure is expected to be lower in the case of syntactic simplification. As a second measure of syntactic complexity, the proportion of syntactic function words (function words contributing to structural cohesion of a sentence such as pronouns, prepositions, articles, numerals and conjunctions) to the total number of words is used. This ratio is hypothesized to be lower when the disorder becomes more severe, due to syntactic simplification and raised semantic association, leading to a more complex content with a less complex syntax, and consequently fewer syntactic function words. Conversely, in case of more negative symptoms, a more vague content could yield a higher proportion of functional words. As a third feature, the average number of clauses per sentence is computed. Low values of this index point to simplification. A fourth feature of syntactic complexity is the average number of constituents, measured on the basis of the phrase chunks that the Frog tool delivers.

Mean Length of Utterance (MLU). MLU is measured by dividing the number of words in all sentences by the total number of sentences in the sample. In syntactic simplification due to cognitive impairment or deterioration, the working memory is less efficient leading to shorter sentences. Mean word length may also be an indication of grammatical simplification.

Proportion of connectors. Connectors are words primarily used for the connection of phrases, sentences and paragraphs. They are indications of the attempt to construe logical coherence in a goal-oriented text. Coherence of expository discourse is a notorious problem in schizophrenic patients. Earlier findings indicated that they have the greatest trouble with pragmatics and discourse planning. Therefore, a list of 40 connectors was collected and their relative frequencies were counted in the corpuses. This measure was predicted to be lower in more severe psychotic episodes.

Distribution of parts of speech. For every time segment, the relative frequencies of different parts of speech was computed using Frog. Prominent negative symptoms can have an impact on the ratio of nouns and verbs versus other less necessary informative words like adjectives. So this could be a sign of limited content and lowered specificity.

Pronoun usage. There have been studies of pronoun usage to discern among schizophrenics and non-patients. These studies "involved both counts of general pronoun usage and the incidence of first person pronouns and self-reference statements". Although for schizophrenic language use, the relevance of pronoun distributions is debated, their importance is stressed in computational stylometry research, including personality profiling (e.g. Pennebaker, 2011). It could be expected that schizophrenia has indeed an impact on psychological processes expressed in pronoun usage. Note that pronoun measurement (here on the basis of Frog output) partially overlaps with some of the dictionaries of the LIWC system.

5.4 Results and Discussion

Using the features described in the previous section, we (i) compare schizophrenic text with control text, (ii) compare schizophrenic text written in manic psychotic episodes with text written in residual episodes with more negative symptoms and cognitive impairment, according to the medical recordings, and (iii) try to predict psychotic episodes using Machine Learning techniques.

Schizophrenic texts were compared to standard text applying a one-sample t-test for every feature. Standards were established by computing the means of the features in the control texts. For the comparison between psychotic/manic episodes and residual texts, an independent samples t-test was performed. For the prediction of psychosis on the basis of texts, a 10-fold-cross-validation was performed.

LIWC. Forty-three out of 66 LIWC features showed a significant difference in frequency when

comparing schizophrenic and control texts. In the schizophrenic text, there was more talking about people, about self and others and relations and there were more negative emotions and fewer features related to 'inhuman' issues like money, leisure, achievements, etc. Remarkably, there were also higher frequencies in features like insight and certainty. The latter is in contrast with the classical known feature of disappearance of insight in schizophrenics due to thought disorders or wrong attribution and lack of reflectiveness (tolerating doubt). LIWC features were better able to differentiate between schizophrenic text and normal text than between psychotic and residual episodes (only 15 significantly different LIWC features were found here). Maybe this was to be expected as the thematic features of LIWC predominantly measure psychological factors such as personality, which is known to represent a trait and not a mental status. Using LIWC features, it was impossible to predict psychotic episodes (with clustering methods).

Schizophrenia related features. The following Table shows which features were significantly higher or lower in the schizophrenic text compared to the control texts.

HIGHER frequencies (p-value)
Low specific words (<0.001)
Ratio syntactic words (<0.001)
Mean length of sentences (0.003)
Proportion of connectors (<0.001)
Mean number of constituents (<0.001)
Mean number of clauses (<0.001)
Lexical cohesion (0.007)
LOWER frequencies (p-value)
Lexical diversity (0.04)
Mean length of words (<0.001)
Specificity (0.015)

For the differences between psychotic episodes versus residual text, the results of statistical analysis were less clear and often went against the expected trend. Strikingly, the syntactic complexity measures were in general higher instead of lower in a psychotic episode. It should be remembered, however, that three independent factors are at work in schizophrenia, the positive, the negative and the cognitive. In this patient, medical records reported delusions, thought disorders and some negative symptoms, but not hallucinations or total disorganization. Specificity and lexical cohesion were significantly lower both in schizophrenic language and in psychotic state, as expected.

Despite the less clear discrimination of these features between psychotic and residual states, psychotic episodes could be predicted with up to 80% accuracy using these features in a machine learning set-up (logistic regression), showing that in combination, they provide considerable predictive (and therefore potentially diagnostic) power. Further analysis of the interaction of the different features explaining this high predictive power is needed.

6 Conclusion

The results of the pilot study, notwithstanding the severe limitations (especially the fact that we have little data of only one schizophrenic patient), showed that the features we extracted allow accurate distinction between schizophrenic text and control text, but less between psychotic states and residual states within the same patient. Predictability of psychotic states in this patient is however highly accurate with combinations of these features. At a theoretical level, the results seem to provide evidence for some hypotheses in schizophrenia research, but currently our analysis only provides indications that the methodology we propose could be used for investigating such hypotheses when a sufficient number of cases can be investigated. In future research we want to investigate for example the hypothesis of Kuperberg that thought disorder reflects a deficit in semantic processing that can be dissociated from deficits in semantic and working memory (Kuperberg & Heckers, 2000), for which we found indications in our case study analysis. Moreover, this study also suggested evidence that cognitive function is partly independent from positive symptoms as was put forward by Szoke et al. (2008). The apparently contradictory results for 'local' semantic coherence and higher order structural incoherence ('thought disorder') can be explained by the opposite tendencies between the semantic associative inclination, on the one hand, and the incapacity to monitor higher order structural planning, on the other. We will also investigate further use of the developed method for assessment.

References

American Psychiatric Association (2000). *Diagnostic and statistical manual of mental disorders, Text Revision* (DSM-IV-TR) (4th ed.). Washington, DC: American Psychiatric Association.

American Psychiatric Association (2013). *Diagnostic and Statistical Manual of Mental Disorders, 5th edition* (DSM-5). Arlington : American Psychiatric Association. ISBN-10 : 0890425558.

Andreasen, N.C. (1979). Thought, language, and communication disorders : clinical assessment, definition of terms, and assessment of their reliability. Arch. Gen. Psychiatry, 36, 1351-1321.

Andreasen, N.C. (1986). Scale for the assessment of thought, language, and communication (TCL). Schizophr. Bull., 12, 473-482.

Bickerton, D. (1992). Language and species. University of Chicago Press.

Bleuler, E. (1911). Dementia Praecox or the group of schizophrenias. New York : International University Press.

Blom, J.D. (2007). Honderd jaar schizofrenie. Van Bleuler naar de DSM-5. Tijdschrift voor psychiatrie, 49 (12), 887-895.

Bosch, Antal van den, Bertjan Busser, Sander Canisius, and Walter Daelemans (2007) "An efficient memory-based morphosyntactic tagger and parser for Dutch." LOT Occasional Series 7 (2007): 191-206.

Carroll, J.B. (1964). Language and thought. Englewood Cliffs.

Ceccherini-Nelli, A. & Crow, T.J. (2003). Disintegration of the components of language as the path to a revision of Bleuler's and Schneider's concepts of schizophrenia : Linguistic disturbances compared with first-rank symptoms in acute psychosis. The British Journal of Psychiatry, 182, 233-240.

Chen, E.Y.H., Lam, L.C.W., Kan, C.S., Chan, C.K.Y., Kwok, C.L., Nguygen, D.G.H., Chen, R.Y.L. (1996). Language disorganization in schizophrenia : validation and assessment with a new clinical rating instrument. Hong Kong, J. Psychiatry, 6(1), 4-13.

Covington, M.A., He, C., Brown, C., Naçi, L., Mc Clain, J.T., Fjordbak, B.S., Semple, J., Brown, J. (2005). Schizophrenia and the structure of language : The linguist's view. Schizophrenia Research, 77, 85-98.

Hoff, A. L., Sakuma, M., Wieneke, M., Horon, R., Kushner, M., & DeLisi, L. E. (1999) . Longitudinal neuropsychological follow-up study of patients with first-episode schizophrenia. American Journal of Psychiatry, 156(9), 1336-1341.

Joseph,B. , Narayanaswamy,J. C. , Venkatasubramanian,G. (2015). Insight in schizophrenia : relationship to positive, negative and neurocognitive dimensions. Indian J Psychol Med, 37(1), 5-11.

Ketteler, D., Theodoridou, A., Ketteler, S., Jäger, M. (2012). High Order Linguistic Features such as ambiguity processing as relevant diagnostic markers for schizophrenia. Schizophrenia Research and Treat-ment. Doi : 10.1155/2012/825050.

Kuperberg, G. & Heckers, S. (2000). Schizophrenia and cognitive function. Current Opinion in Neurobiolo-gy, 10, 205-210.

Kuperberg, G.R. (2010). Language in schizophrenia Part 1 : an introduction. Lang Linguist Compass., 4(8), 576-589.

Kuperberg, G.R. (2010). Language in Schizophrenia Part 2 : what can psycholinguistics bring to the study of schizophrenia and vice versa? Lang Linguist Compass, 4(8), 590-604.

Liddle, P.F. et al. (2002). Thought and Language Index: an instrument for assessing thought and language in schizophrenia. Br J Psychiatry, 181, 326-330.

Marini, A., Spoletini, I., Rubino, I.A., Ciuffa, M., Bria, P., Martinotti, G., Banfi, G., Boccascino, R., Strom, P., Siracusano, A., Caltaginone, C., Spalletta, G. (2008). The language of schizophrenia : an analysis of micro and macrolinguistic abilities and their neuropsychological correlates. Schizophrenia Research, 105, 144-155.

Mc Glashan, T.H., fenton, W.S. (1992). The positive-negative distinction in schizophrenia. Review of natural history validators. Arch gen Psychiatry, 49(1), 63-72.

Nielsen, R.E. (2011). Cognition in schizophrenia, a systematic review. Drug Discovery Today : Therapeutic Strategies, 8, 1(2), 43-48.

Oh, T.M., Mc Carthy, R.A. & Mc Kenna, P.J. (2002). Is there such thing as a schizophasia? A study applying a single case approach to formal thought disorder in schizophrenia. Neurocase, 8, 233-244.

Oostdijk, Nelleke, Martin Reynaert, Véronique Hoste, and Ineke Schuurman. (2013) "The construction of a 500-million-word reference corpus of contemporary written Dutch." In Essential speech and language technology for Dutch, pp. 219-247. Springer Berlin Heidelberg, 2013.

Pennebaker, J.W. (2011). The secret life of pronouns. What our words say about us. New York : Blooms-berry Press.

Pinard, G., & Lecours, A. R. (1983). The language of psychotics and neurotics. Aphasiology. Balliere Tin-dall, London, 313-335.

Stevin (2009). Cornetto Lexical Database, Versie 7. Nederlandse Taalunie.Harold J.

Szöke, A., Trandafir, A., Dupont, M. E., Méary, A., Schürhoff, F., & Leboyer, M. (2008). Longitudinal studies of cognition in schizophrenia: meta-analysis. The British Journal of Psychiatry, 192(4), 248-257.

Van der Gaag, R.J., Van Wijngaarden-Cremers, P.J.M., Staal, W.G. (2012). Stagering : een ontwikkeling-spuzzel. Tijdschrift voor psychiatrie, 11, 965-972.
Vetter (1970) language behavior and psychopathology. Rand McNally & Company: Chicago.
Zijlstra, H. et al. (2004). De Nederlandse versie van de 'Linguistic Inquiry and Word Count' (LIWC). Een gecomputeriseerd tekstanalyseprogramma.