CORTEX XXX (2014) 1–19

Available online at www.sciencedirect.com

ScienceDirect

Journal homepage: www.elsevier.com/locate/cortex

Special issue: Research report

A computational language approach to modeling prose recall in schizophrenia

Mark Rosenstein [a,*], *Catherine Diaz-Asper* [b], *Peter W. Foltz* [a,c] *and Brita Elvevåg* [d,e]

[a] *Pearson Knowledge Technologies, Boulder, CO, USA*
[b] *Clinical Brain Disorders Branch, National Institute of Mental Health/NIH, Bethesda, MD, USA*
[c] *University of Colorado, Institute of Cognitive Science, Boulder, CO, USA*
[d] *Psychiatry Research Group, Department of Clinical Medicine, University of Tromsø, Norway*
[e] *Norwegian Centre for Integrated Care and Telemedicine (NST), University Hospital of North Norway, Tromsø, Norway*

ARTICLE INFO

Article history:
Received 4 June 2013
Reviewed 28 August 2013
Revised 11 December 2013
Accepted 22 January 2014
Published online xxx

Keywords:
Episodic memory
Wechsler memory scale
Text analysis
Latent semantic analysis

ABSTRACT

Many cortical disorders are associated with memory problems. In schizophrenia, verbal memory deficits are a hallmark feature. However, the exact nature of this deficit remains elusive. Modeling aspects of language features used in memory recall have the potential to provide means for measuring these verbal processes. We employ computational language approaches to assess time-varying semantic and sequential properties of prose recall at various retrieval intervals (immediate, 30 min and 24 h later) in patients with schizophrenia, unaffected siblings and healthy unrelated control participants. First, we model the recall data to quantify the degradation of performance with increasing retrieval interval and the effect of diagnosis (i.e., group membership) on performance. Next we model the human scoring of recall performance using an *n*-gram language sequence technique, and then with a semantic feature based on Latent Semantic Analysis. These models show that automated analyses of the recalls can produce scores that accurately mimic human scoring. The final analysis addresses the validity of this approach by ascertaining the ability to predict group membership from models built on the two classes of language features. Taken individually, the semantic feature is most predictive, while a model combining the features improves accuracy of group membership prediction slightly above the semantic feature alone as well as over the human rating approach. We discuss the implications for cognitive neuroscience of such a computational approach in exploring the mechanisms of prose recall.

1. Introduction

Human memory is to a large extent genetically controlled, and thus it is considered to be a heritable, polygenic trait. In schizophrenia impaired cognitive function is a core feature of the illness (Elvevåg & Goldberg, 2000) and some of the most prominent deficits are in verbal episodic memory (Aleman, Hijman, de Haan, & Kahn, 1999; Barch, 2005; Cirillo & Seidman,

* *Corresponding author*. Pearson, 4940 Pearl East Circle, Suite 200, Boulder, CO 80301, USA.
E-mail address: mark.rosenstein@pearson.com (M. Rosenstein).
http://dx.doi.org/10.1016/j.cortex.2014.01.021

2003; Kalkstein, Hurford, & Gur, 2010). The disproportionate impairment in verbal episodic memory relative to visual episodic memory may suggest that a useful endophenotype is a deficit in verbal processing, rather than memory impairment *per se* (Skelley, Goldberg, Egan, Weinberger, & Gold, 2008). In this paper, we use recalls from a widely used prose recall test to explore the usefulness of an automated scoring methodology that has the potential to provide equivalent or more sensitive scoring metrics to that of human raters, as well as a more detailed characterization of recall performance over time.[1]

Measures of verbal episodic memory typically include the learning and subsequent recall of word lists or prose passages (stories), and one of the most comprehensive, popular and enduring scales is the Wechsler Memory Scale (WMS; Wechsler, 1945, 1987, 1997, 2009), currently in its 69th year and fourth revision. With minor modifications over time, the Logical Memory subtest has remained a core component of the battery, and is one of the most widely-used measures of prose recall in the research literature (Rabin, Barr, & Burton, 2005).

The Logical Memory task requires participants to repeat back two orally-presented short stories, both immediately after presentation, and following a 30 min delay.[2] The scoring criteria, or rubric, generally specifies that one point is awarded for each key word or narrowly defined concept correctly recalled, with a maximum of 25 points per story, summed for a total raw score out of 50. A measure of forgetting ["percent retained" (Russell, 1988) or "saving score"[3] (Munro Cullum, Butters, Tröster, & Salmon, 1990)] can also be calculated as the total number of items recalled following the delay interval, divided by the total number recalled immediately after initial presentation. Prose recall tasks such as Logical Memory likely rely heavily upon multiple cognitive and memory systems, including language comprehension, conceptual organization, schema formation, working memory, and episodic and semantic memory (Baddeley & Wilson, 2002; Dunn, Almeida, Barclay, Waterreus, & Flicker, 2002). Since performance on this task relies upon hippocampal memory systems (Ho et al., 2006; Lim et al., 2006; O'Driscoll et al., 2001), it is a sensitive assay of verbal episodic memory dysfunction in a variety of neuropsychiatric conditions, including schizophrenia and Alzheimer's disease (Egan et al., 2003; Matsui et al., 2007; Vassos et al., 2010). Importantly, it demonstrates a genetic load effect in schizophrenia, with unaffected siblings typically performing intermediary between patients with schizophrenia and healthy controls (Goldberg et al., 1995; Skelley et al., 2008; Toulopoulou, Rabe-Hesketh, King, Murray, & Morris, 2003). While Logical Memory has proven a useful clinical measure of verbal episodic memory, there are several limitations. Early versions (WMS, Wechsler, 1945; Wechsler Memory Scale-Revised (WMS-R), Wechsler, 1987) relied heavily upon the recall of salient words from the story, known as "story units", yet prose recall is rarely verbatim (e.g., Kintsch, 1998). Rather, it is filled with approximate renderings of the passage that may include substitutions, omissions, additions and elaborations, and shifts in the story's sequence (Lezak, Howieson, & Loring, 2004). These common deviations in recall are not adequately captured by the relatively simplistic "story units" measurement. More recent revisions of the test (e.g., WMS-III, Wechsler, 1997) have introduced "thematic" scoring units in addition to story units, wherein larger chunks of discourse pertaining to a theme are sought rather than the verbatim recall of select key words, presumably to better capture gist recollection. However, Dunn et al. (2002) contend this measure is merely a subset of story units and adds no additional information. The approach further relies on the subjective judgment by the scorer about the degree of match of recall to themes. For these reasons, in this study only the story unit rubric was used.

A few studies illustrate how departing from the constraints of standard administration and scoring can provide complimentary information on verbal episodic memory function. For example, when Skelley et al. (2008) examined episodic memory function in patients with schizophrenia, their unaffected siblings and healthy unrelated controls, they utilized the "savings score" calculation on total raw scores on Logical Memory at three different time points (immediate, 30 min, and 24 h). They reported that both patients and siblings displayed the greatest impairment in initial learning (from immediate to 30 min) and little impairment in long-delay savings (from 30 min to 24 h).

An alternative approach to obtain further information from prose recalls is to assay the effect of the underlying cognitive processes integral to prose recall – the sequential construction of the words and semantic processes – but this approach may introduce a level of subjectivity potentially compromising reliability and validity (Dunn et al., 2002). However, a way to obviate this concern is to employ automated language analysis methods. The first question we address is whether automated methods can perform as well as humans in the scoring task, and then having established a baseline performance, whether features arising from automated analysis might actually outperform the existing rubric in predicting group membership (i.e., diagnosis). We have previously shown that departing from global scoring techniques and employing a data-driven methodology can provide useful information concerning cognitive strategies that individuals use in order to remember lists of words (Longenecker, Kohn, et al., 2010). In the case of prose recall, given the "story unit" rubric's strong emphasis on capturing exact words and phrases,[4] a language sequential categorization algorithm based on natural language processing techniques may be able to capture much of how humans

[1] Although we illustrate this method with a test from the WMS-R (Wechsler, 1987), the techniques can naturally be applied to other verbal memory tests.

[2] A third recall at 24 h was added to the protocol for this study.

[3] For clinical purposes, the raw score may be converted to a standardized scaled score (0–19) based on the normative tables published in the test manual. The concept of "saving score" has a long history (e.g., Ebbinghaus, 1885/1913). Robinson and Heron (1922) define "saving score" in the context of memorizing lists, though in their case practice over time improved performance, so instead of directly reporting the fraction presented here, the fraction was first subtracted from 100. This metric is reportedly less vulnerable than standardized scaled scores to the well-documented declines in performance on the Logical Memory test with advancing age (Lezak et al., 2004), and also differentiates cortical from subcortical dementias (Tröster et al., 1993).

[4] Especially in the WMS-R which is the test version we employed.

Please cite this article in press as: Rosenstein, M., et al., A computational language approach to modeling prose recall in schizophrenia, Cortex (2014), http://dx.doi.org/10.1016/j.cortex.2014.01.021