CORTEX XXX (2014) 1–10

Available online at www.sciencedirect.com

ScienceDirect

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

Special issue: Research report

Category fluency, latent semantic analysis and schizophrenia: a candidate gene approach

Kristin K. Nicodemus [a,*], *Brita Elvevåg* [b,c], *Peter W. Foltz* [d,e], *Mark Rosenstein* [d], *Catherine Diaz-Asper* [f] *and Daniel R. Weinberger* [f,g,h]

[a] *Neuropsychiatric Genetics Group, Department of Psychiatry, Trinity Centre for Health Sciences, Trinity College Dublin, St James Hospital, Dublin, Ireland*
[b] *Psychiatry Research Group, Department of Clinical Medicine, University of Tromsø, Norway*
[c] *Norwegian Centre for Integrated Care and Telemedicine (NST), University Hospital of North Norway, Tromsø, Norway*
[d] *Pearson Knowledge Technologies, Boulder, CO, USA*
[e] *Institute for Cognitive Science, University of Colorado, Boulder, CO, USA*
[f] *Clinical Brain Disorders Branch, National Institute of Mental Health/NIH, Bethesda, MD, USA*
[g] *Lieber Institute for Brain Development, Baltimore, MD, USA*
[h] *Departments of Psychiatry, Neurology, Neuroscience and The Institute of Genomic Medicine, Johns Hopkins School of Medicine, Baltimore, MD, USA*

ARTICLE INFO

Article history:
Received 15 July 2013
Reviewed 01 October 2013
Revised 14 November 2013
Accepted 11 December 2013
Published online xxx

Keywords:
Cognition
Verbal learning and recall
Gene
Latent semantic analysis
Schizophrenia

ABSTRACT

Background: Category fluency is a widely used task that relies on multiple neurocognitive processes and is a sensitive assay of cortical dysfunction, including in schizophrenia. The test requires naming of as many words belonging to a certain category (e.g., animals) as possible within a short period of time. The core metrics are the overall number of words produced and the number of errors, namely non-members generated for a target category. We combine a computational linguistic approach with a candidate gene approach to examine the genetic architecture of this traditional fluency measure.

Methods: In addition to the standard metric of overall word count, we applied a computational approach to semantics, Latent Semantic Analysis (LSA), to analyse the clustering pattern of the categories generated, as it likely reflects the search in memory for meanings. Also, since fluency performance probably also recruits verbal learning and recall processes, we included two standard measures of this cognitive process: the Wechsler Memory Scale and California Verbal Learning Test (CVLT). To explore the genetic architecture of traditional and LSA-derived fluency measures we employed a candidate gene approach focused on SNPs with known function that were available from a recent genome-wide association study (GWAS) of schizophrenia. The selected candidate genes were associated with language and speech, verbal learning and recall processes, and processing speed. A total of 39 coding SNPs were included for analysis in 665 subjects.

Results and discussion: Given the modest sample size, the results should be regarded as exploratory and preliminary. Nevertheless, the data clearly illustrate how extracting the

* *Corresponding author*. Neuropsychiatric Genetics Research Group, Department of Psychiatry, Trinity Centre For Health Sciences, Trinity College Dublin, St James's Hospital, Dublin 8, Ireland.
E-mail address: nicodemk@tcd.ie (K.K. Nicodemus).
0010-9452/$ – see front matter
http://dx.doi.org/10.1016/j.cortex.2013.12.004

meaning from participants' responses, by analysing the actual content of words, generates useful and neurocognitively viable metrics. We discuss three replicated SNPs in the genes ZNF804A, DISC1 and KIAA0319, as well as the potential for computational analyses of linguistic and textual data in other genomics tasks.

1. Introduction

1.1. *Category fluency and schizophrenia: the role of the intermediate phenotype*

A complex combination of susceptibility genes and environmental factors is assumed to contribute to the overall clinical presentation of psychiatric disorders. Applying a reductionist approach to the diverse presenting phenomenology is not only daunting, but likely overlooks much of the associated deficits in the case of schizophrenia (but see Morar et al., 2011) where cognitive deficits are quite central to the neurodevelopmental course of the illness (Elvevåg & Weinberger, 2001). With such complex medical disorders one way to reduce the complexity of genetic effects is the 'intermediate phenotype' approach where it is argued that the putative risk genes should show greater effects at the intermediate level. Applied to psychiatry, this research strategy argues for bridging the gap between the emergent psychosis and the effects of genes on cells that directly modulate neurocognition (Goldberg & Weinberger, 2004; Meyer-Lindenberg & Weinberger, 2006; Tan, Callicott, & Weinberger, 2008). Such a research framework is appealing (but see Flint and Munafo (2007) for a different opinion), as the resulting intermediate phenotypes (e.g., working memory, episodic memory, semantic memory) are more amenable to systematic neurobiological research than the transient phenomenology (Elvevåg & Weinberger, 2009). Crucially, in psychiatric disorders it is at this intermediate phenotype level that genetic associations often show both stronger penetrance (Tan et al., 2008) and inheritance (Snitz, MacDonald, & Carter, 2006) than at the level of clinical diagnosis. Consequently, several major challenges emerge, namely the unavoidable required refinements to the intermediate phenotype and the management of the huge amount of data resulting from investigations of intermediate phenotypes.

Given the increasing importance of genome-wide association studies (GWAS) in neuropsychiatric research, it is increasingly apparent that intermediate phenotypes are potentially the means with which genomic discoveries will be made, but also may be limiting factors. Indeed, this new approach is magnitudes more complex than any enterprise embarked on hitherto in psychiatric genetics and arguably requires sophisticated phenotypes in order to unravel the complexities and thus eventually the pathologies within neural functional systems. Bilder and colleagues argue that cognitive ontologies need to be developed and refined to not only enable greater consistence and collaboration in research, but also to facilitate connections between intermediate phenotypes and genes (Bilder et al., 2009).

One crucial part of this puzzle is a modern cognitive neuroscientific re-operationalization of common psychometric concepts and terms. Here we focus on one of the most widely used neuropsychological tests – the category fluency task – to illustrate the current limitations of the 'verbal descriptions' of the underlying cognitive constructs and the issues that emerge when trying to explore the genetic architecture of the associated constructs. Specifically, the recall process likely involves a search for meanings as reflected in the 'clustering' of words in the output. Many approaches have been employed to examine the structure of the clustering, but are often problematic given the subjective judgements of cluster boundaries or have turned out to be simply unreliable (Voorspoels et al., 2013). We have previously adopted Latent Semantic Analysis (LSA) as an objective and reliable methodology to chart the flow of meaning in words and discourse (Elvevåg, Foltz, Weinberger, & Goldberg, 2007), and briefly describe this technique below. Our current motivation is that the 'content' of words has rarely been considered a useful candidate in investigations concerning genomics. This absence may be partially due to the notoriously subjective and labour intense efforts required in quantifying the content of words. However, advances in computational linguistics provide a viable framework within which the meanings of words can be rigorously investigated.

1.2. *Latent semantic analysis: building a semantic space*

Latent Semantic Analysis (LSA) is a statistical approach to the acquisition and representation of meaning, which allows similarities among the elements of a language (e.g., words, sentences, or passages) to be computed based on word co-occurrence patterns in large corpora of naturally produced discourse. LSA is a computational model of meaning that closely mimics human understanding of the contextual use of language, which has been widely used for information retrieval, machine understanding of text, and applications such as automated essay scoring (for an overview, see Landauer, Kintsch, McNamara, & Dennis, 2007). Unlike standard keyword-based methods, LSA can detect subtle aspects of semantic content. LSA has been widely used for cognitive modelling of learning and memory processes as well as for computing coherence in language and thought processes. The reduced dimension semantic representation from LSA allows comparison by computing the semantic similarity between individual terms or groups of terms (see Supplementary Methods for further details and an example).

In the case of the category fluency task, the total number of words produced has been shown to be an important metric and poor performance (i.e., production of substantially fewer words than expected based on demographically based

Please cite this article in press as: Nicodemus, K. K., et al., Category fluency, latent semantic analysis and schizophrenia: a candidate gene approach, Cortex (2014), http://dx.doi.org/10.1016/j.cortex.2013.12.004