Available online at www.sciencedirect.com

ScienceDirect

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

Special issue: Editorial

Language, computers and cognitive neuroscience

Peter Garrard [a,*] and Brita Elvevåg [b,c,**]

[a] Neuroscience Research Centre, Institute of Cardiovascular and Cell Sciences, St George's, University of London, Cranmer Terrace, London, UK

[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

It is not tissue microstructure or functional capability that sets the human brain apart from other organs and systems, but its organisational complexity, and to understand the brain at this level remains one of the great scientific challenges of our age. There is no doubt that computation will prove central to the endeavour, both as a framework for understanding, and a medium for simulating, cognition and its myriad disorders. The power and interconnectedness of modern computing hardware are now being exploited in some of the largest and most ambitious studies of cognition ever undertaken ['Head Start' (Editorial comment) Nature, 2013]. The availability of supercomputing power also opens up the related possibility of exploiting novel information sources that are too large and complex to be captured, organised or analysed using conventional approaches – a resource that, over recent years, has come to be known as 'big data'. The McKinsey Global Institute's 2011 report on this phenomenon is entitled *Big data: The next frontier for innovation, competition, and productivity* (Manyika et al., 2011). The authors show how big data generate value in healthcare, public services, retail and manufacturing. Among our ambitions for this *Cortex* special issue is that it will help to make the case for cognitive neuroscience to be added to the list.

How might big data contribute to the goals of understanding healthy and disordered brains in ways that span Marr's three 'levels of analysis'? (Marr, 1982) (See also Poggio's recent update on this framework, which is available in full at: http://cbcl.mit.edu/publications/ps/MIT-CSAIL-TR-2012-014.pdf.) In a world dominated by digital technology, people contribute to the store of big data simply by going about their daily lives. The focus of interest in the resulting, and constantly growing, body of information will naturally vary: for the business community the behaviour, choices and preferences of users and customers will be critical to the goal of maximising profits, while government and the public sector must aim to formulate indices of economic value and social outcome in order to maximise the efficient use of limited resources. Meanwhile, science has both benefited from and pioneered the understanding of huge datasets and data streams, including those related to particle physics, genomics and climate science – fields that generate quantities of data measured in petabytes ($\times 10^{15}$ bytes) per year (Doctorow, 2008).

It is inevitable that the information people generate as they go about their daily lives will hold some value for cognitive neuroscientists, particularly those who emphasise the importance of 'ecological validity' in the interpretation of behavioural data (Cohen, 1996; Neisser, 1991). We have no interest in reopening any of the wounds inflicted (by both sides) in the debate on the relative merits of everyday memory and traditional laboratory research. Yet few people with a scientific interest in learning and memory would dismiss out of hand a detailed and cumulative record of (for example) all the movements, interactions and web searches carried out by large populations of individuals over a number of years. Although the level of intrusion that would be required to generate such a dataset on private citizens is hardly desirable, there is less cause for squeamishness when one considers the

[*] *Corresponding author*. Neuroscience Research Centre, Institute of Cardiovascular and Cell Sciences, St George's, University of London, Cranmer Terrace, London SW17 0RE, UK.

[**] *Corresponding author*. Psychiatry Research Group, Department of Clinical Medicine, University Hospital of North Norway – Åsgård, Postbox 6124, 9291 Tromsø, Norway.

E-mail addresses: pgarrard@sgul.ac.uk, peter.garrard@gmail.com (P. Garrard), brita@elvevaag.net (B. Elvevåg).

http://dx.doi.org/10.1016/j.cortex.2014.02.012

benefits that could accrue to closed communities, be they real (e.g., work environments or care homes) or virtual (e.g., patient groups with internet connections and/or access to clinical care via a telemedicine programme). Yet there is much practical and ethical ground to move before such data become relevant and usable.

Fewer limitations apply to language data in the form of naturally produced samples of spoken or written language: collection and recording have been taking place for hundreds of years in the form of handwritten and printed texts, audio recording, and most recently the hundreds of millions of digital communications (blogs, tweets, emails and text messages) that are produced each day by an ever more digitally interconnected public. Many of these sources are the product of undirected and spontaneous cognitive activity in single individuals, often with the intention of public communication. In addition, there is a sizeable body of clinical data representing the output of more focused neurocognitive activity in various clinically defined groups (of which, more later). All can be considered in the light of a multitude of dimensions, some of them simple, others reflecting more complex attributes of the symbolic systems in which they are represented. The widespread availability of fast, high capacity, desktop computers means that large volumes can be represented and stored in a digital text format.

Nonetheless, the problem of how to make large datasets tractable, to organise and use them in informative ways remains common to all the enterprises – scientific, technical and commercial – that we have considered so far. Previous attempts to extract meaning from huge datasets have relied on a diverse range of ‘data mining’ techniques, including dimension reduction, information theory, and statistical machine learning – approaches that are represented in a number of the contributions to this special issue. Even simple approaches such as proportional word-counts, however, can produce strikingly informative results, particularly when applied to very large datasets. A leading source of both data and analytical tools is Google: Google Books contains digitally encoded texts of a large (and ever increasing) proportion of all the books ever published; the Google n-gram viewer https://books.google.com/ngrams will plot the change in proportional frequency of any word (unigram) or phrase (n-gram) in books published between the years 1800 and 2000. In a series of fascinating explorations of the data, Michel et al. (2011) reported a selection of instances in which social and cultural evolution and major historical events were reflected in lexical frequency trends. The approach offers limitless possibilities for further exploration, and it is to be hoped that the interdisciplinary nature of cognitive neuroscience will prompt experts from disciplines such as statistics and computer science to modify and add to the analytical armamentarium.

Even if the ‘what?’ and ‘how?’ of large scale language analysis could be fully addressed, we would still be left with the question that even the most rarefied scientific disciplines must nowadays address: ‘to what end?’ We contend that the contributions to this special issue provide a wealth of justifications, predominantly clinical, but also theoretical. Among the latter are the contributions of Montemurro (2014) and Voorspoels et al. (2014). The former advances the idea that the inherent order detectable in the long range co-occurrence of words in texts written in different languages (relative entropy) should be considered a candidate for a quantitative linguistic universal – a bold and testable hypothesis. The latter explores the pitfalls and limitations of the clustering method in arguing for distorted semantic structure in cognitive neuropsychology. Valle-Lisboa, Pomi, Cabana, Elvevåg, and Mizraji (2014) adopt a neurocomputational modelling approach to explore the links between matrix associative memory models and language processing and production, creating a system for exploring how disruptions in connectivity between the underlying representations of concepts can result in various forms of disorganized speech.

Clinically based studies draw on a wide-ranging series of data associated with language change over the course of normal ageing (Ferguson et al., 2014) and tenure of political office (Garrard, Rentoumi, Lambert, & Owen, 2014), as well as linguistic features of cerebral functional disorders including Alzheimer’s disease, primary progressive aphasia (Garrard, Rentoumi, Gesierich, Miller, & Gorno-Tempini, 2014), and schizophrenia. These studies are made possible by the fact that communication is a high-level neurocognitive function providing a rich and extemporaneous dataset that reflects the state of numerous interacting neural and cognitive processes. If assayed appropriately, therefore, communication affords a unique and sensitive window into a person’s state of mental and cognitive health.

As authors, we welcome exposure of our research to the more than usually diverse readership that the interdisciplinary theme of this special issue will attract. As editors, we were struck by the multiplicity of ways in which computer-assisted analysis of large language datasets could contribute to the understanding of brain disorders. Pakhomov and Hemmy (2014) took a large database of verbal fluency responses collected as part of the Wisconsin Nun Study (Snowdon et al., 1996) and interrogated the data for response clusters and switching behaviours using an automated measure of relatedness derived from latent semantic analysis (LSA). Originally conceived as a statistical approach to the acquisition and representation of meaning (Landauer & Dumais, 1997), LSA uses a vector space representation of the words and contexts occurring in large numbers of digitised texts, such that the distance between vectors can be used as a metric of the semantic similarity between the words and/or contexts. This property allows a number of robust measurements to be made in novel text or discourse samples, including those obtained from different patient groups.

Hoffman, Meteyard, and Patterson (2014) use the neighbourhood density of items in a semantic space to derive a measure of ‘semantic diversity’ characterising the vocabulary of patients with conceptual degradation (semantic dementia). Several studies use LSA to examine the properties of discourse samples obtained from patients with schizophrenia. Two papers (those by Holshausen, Harvey, Elvevåg, Foltz, & Bowie, 2014; Tagamets, Cortes, Griego, & Elvevåg, 2014) report correlations between LSA derived measures of patient discourse and other validated functional measures, namely clinical and psychometric indices, and task-related fMRI patterns. A third (Rosenstein, Diaz-Asper, Foltz, & Elvevåg, 2014) examines the effect of a latent semantic variable and a syntactic characteristic to examine the effects of these features on prose recall in patients, their unaffected siblings and healthy unrelated controls. All offer genuine encouragement that statistical features inherent in samples of spoken output could contribute to a quantifiable disease metric for this most elusive of clinical phenotypes. Finally, Nicodemus et al. (2014) report findings that suggest not only clinical but also genetic correlates for LSA derived indices, based on their association with a subset of loci identified by a recent GWAS study of schizophrenia. The potential utility of these simple yet powerful text mining and computational tools for refining the endophenotypic classification of schizophrenia, and with it the significance of the genetic associations, is tantalising (see also Cohen, Blatter, & Patel, 2008; Lyalina et al., 2013).

A final major methodological theme of the special issue is the application of information theory and machine learning to the clinical classification of patients, using written texts, sets of neurolinguistic features, or transcribed speech samples as raw data. van Velzen, Nanetti, and de Deyn (2014) use the type-token ratio to plot changes in lexical richness over the careers of a selection of prolific English and Dutch authors, and apply models of increasing complexity to describe the resulting time series. They go on to select mathematically the most parsimonious models, suggesting that the selected models may map on to different patterns of cognitive ageing. Wilson et al. (2010) have previously described the wealth of information that can be extracted from samples of connected speech in primary progressive aphasia. The methods used by Fraser et al. (2014), overcome the time-consuming disadvantage of the hand-scoring process by applying machine learning classification to sets of features that can be automatically extracted from digital transcripts. Meteyard, Quain, and Patterson (2014) employ similar automated methodologies in pursuit of evidence from patients with semantic dementia that both lexical retrieval and grammatical encoding can be incorporated within a common constraint-satisfaction model. Garrard, Rentoumi, Gesierich, et al. (2014) show that machine learning algorithms can make at least one of these classifications (that of semantic dementia with high reliability on the strength of no more than the vocabulary of the speech sample, even when no information about word-order is available (the so-called 'bag of words' assumption)). The same approach appeared also to have some traction on the more difficult clinical distinction between right and left temporal lobe predominant semantic dementia. Finally, Clark et al. (2014) show how the performance of a machine learning classifier in predicting cognitive decline can be enhanced by using novel statistical methods to extract information from verbal fluency task responses.

For all the analytical sophistication and volumes of available data, we must acknowledge that none of the studies presented in this special issue moves beyond the representation of language as text. Prosody, emotional and sociolinguistic connotation, and other 'paralinguistic' elements that play such a critical role in verbal communication, are not considered. Finding stable and reliable ways of incorporating these features into data representations remains a major challenge for the future. Even taking account of these limitations, however, the contents of this special issue illustrate the challenges of applying computational linguistics to the cognitive neuroscience field, as well as the power of these techniques to frame questions of theoretical interest and define clinical groups of practical importance. The future of digital written language sampling is inexorably in the direction of rapid growth in data, a movement that will obviate many of the laborious acquisition steps. Similar progress in the automated transcription of spoken language has been slower, but the potential richness of recorded speech data continues to justify the investment of pre-processing time, with novel biological and clinical insights into neurological and psychiatric illness as the ultimate payoff. Deployment of clinically and biologically relevant assays on a large scale, during the evolution of neurodegenerative and neuropsychiatric conditions, can only enhance our ability to quantify such elusive entities as disease risk, rate of progression, prognosis and, in the case of psychiatric illness, relapse.

REFERENCES

Clark, D. G., Kapur, P., Geldmacher, D. S., Brockington, J. C., Harrell, L., & Marson, D. C. (2014). Latent information in verbal fluency lists predicts functional decline in persons at risk for Alzheimer disease. *Cortex*. http://dx.doi.org/10.1016/j.cortex.2013.12.013.

Cohen, G. (1996). *Memory in the real world*. Hove, E. Sussex: Psychology Press.

Cohen, T., Blatter, B., & Patel, V. (2008). Simulating expert clinical comprehension: adapting latent semantic analysis to accurately extract clinical concepts from psychiatric narrative. *Journal of Biomedical Informatics, 41*(6), 1070–1087. http://dx.doi.org/10.1016/j.jbi.2008.03.008.

Doctorow, C. (2008). Big data: welcome to the petacentre. *Nature,* 455(7209), 16–21. http://dx.doi.org/10.1038/455016a.

Ferguson, A., Spencer, E., Craig, H., & Colyvas, K. (2014). Propositional idea density in women's written language over the lifespan: computerized analysis. *Cortex*. http://dx.doi.org/10.1016/j.cortex.2013.05.012.

Fraser, K. C., Meltzer, J. A., Graham, N. L., Leonard, C., Hirst, G., Black, S. E., et al. (2014). Automated classification of primary progressive aphasia subtypes from narrative speech transcripts. *Cortex*. http://dx.doi.org/10.1016/j.cortex.2012.12.006.

Garrard, P., Rentoumi, V., Gesierich, B., Miller, B., & Gorno-Tempini, M. L. (2014). Machine learning approaches to diagnosis and laterality effects in semantic dementia discourse. *Cortex*. http://dx.doi.org/10.1016/j.cortex.2013.05.008.

Garrard, P., Rentoumi, V., Lambert, C., & Owen, D. (2014). Linguistic biomarkers of Hubris syndrome. *Cortex*. http://dx.doi.org/10.1016/j.cortex.2013.08.014.

Head Start. (2013). *Nature, 503*(7474), 5.

Hoffman, P., Meteyard, L., & Patterson, K. (2014). Broadly speaking: vocabulary in semantic dementia shifts towards general, semantically diverse words. *Cortex*. http://dx.doi.org/10.1016/j.cortex.2012.11.004.

Holshausen, K., Harvey, P. D., Elvevåg, B., Foltz, P. W., & Bowie, C. R. (2014). Latent semantic variables are associated with formal thought disorder and adaptive behavior in older inpatients with schizophrenia. *Cortex*. http://dx.doi.org/10.1016/j.cortex.2013.02.006.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*(2), 211–240. http://dx.doi.org/10.1037//0033-295x.104.2.211.

Lyalina, S., Percha, B., Lependu, P., Iyer, S. V., Altman, R. B., & Shah, N. H. (2013). Identifying phenotypic signatures of neuropsychiatric disorders from electronic medical records. *Journal of the American Medical Informatics Association: JAMIA*. http://dx.doi.org/10.1136/amiajnl-2013–001933.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., et al. (2011). *Big Data: The next frontier for innovation, competition, and productivity*. McKinsey Global Institute.

Marr, D. (1982). *Vision*. New York: W.H. Freeman and Company.

Meteyard, L., Quain, E., & Patterson, K. (2014). Ever decreasing circles: speech production in semantic dementia. *Cortex*. http://dx.doi.org/10.1016/j.cortex.2013.02.013.

Michel, J. B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., The Google Books Team, et al. (2011). Quantitative analysis of culture using millions of digitized books. *Science, 331*(6014), 176–182. http://dx.doi.org/10.1126/science.1199644.

Montemurro, M. A. (2014). Quantifying the information in the long-range order of words: semantic structures and universal linguistic constraints. *Cortex*. http://dx.doi.org/10.1016/j.cortex.2013.08.008.

Neisser, U. (1991). A case of misplaced nostalgia. *American Psychologist, 46*, 34–36.

Nicodemus, K. K., Elvevåg, B., Foltz, P. W., Rosenstein, M., Diaz-Asper, C., & Weinberger, D. R. (2014). Category fluency, latent semantic analysis and schizophrenia: a candidate gene approach. *Cortex*. http://dx.doi.org/10.1016/j.cortex.2013.12.004.

Pakhomov, S. V., & Hemmy, L. S. (2014). A computational linguistic measure of clustering behavior on semantic verbal fluency task predicts risk of future dementia in the Nun Study. *Cortex*. http://dx.doi.org/10.1016/j.cortex.2013.05.009.

Rosenstein, M., Diaz-Asper, C., Foltz, P. W., & Elvevåg, B. (2014). A computational language approach to modeling prose recall in schizophrenia. *Cortex*. http://dx.doi.org/10.1016/j.cortex.2014.01.021.

Snowdon, D. A., Kemper, S. J., Mortimer, J. A., Greiner, L. H., Wekstein, D. R., & Markesbery, W. R. (1996). Linguistic ability in early life and cognitive function and Alzheimer's disease in late life. Findings from the Nun Study. *JAMA, 275*(7), 528–532.

Tagamets, M. A., Cortes, C. R., Griego, J. A., & Elvevåg, B. (2014). Neural correlates of the relationship between discourse coherence and sensory monitoring in schizophrenia. *Cortex*. http://dx.doi.org/10.1016/j.cortex.2013.06.011.

Valle-Lisboa, J. C., Pomi, A., Cabana, Á., Elvevåg, B., & Mizraji, E. (2014). A modular approach to language production: models and facts. *Cortex*. http://dx.doi.org/10.1016/j.cortex.2013.02.005.

van Velzen, M. H., Nanetti, L., & de Deyn, P. P. (2014). Data modelling in corpus linguistics: how low may we go? *Cortex*. http://dx.doi.org/10.1016/j.cortex.2013.10.010.

Voorspoels, W., Storms, G., Longenecker, J., Verheyen, S., Weinberger, D. R., & Elvevåg, B. (2014). Deriving semantic structure from category fluency: clustering techniques and their pitfalls. *Cortex*. http://dx.doi.org/10.1016/j.cortex.2013.09.006.

Wilson, S. M., Henry, M. L., Besbris, M., Ogar, J. M., Dronkers, N. F., Jarrold, W., et al. (2010). Connected speech production in three variants of primary progressive aphasia. *Brain, 133*(Pt 7), 2069–2088. http://dx.doi.org/10.1093/brain/awq129.