Conference Paper

Digital Humanities: New Tools and New Knowledge

Simon Musgrave

Monash University, Australia

Abstract

A key aspect of the rapidly growing field of Digital Humanities is the application of computational tools to problems in humanistic research, a process which can lead to exciting new knowledge. I will illustrate this development with examples from my own research and from that of other scholars showing how the new tools are applicable across many areas of research in the humanities. In particular, I will discuss how the recent development of machine learning algorithms has made it possible to investigate more fully insights based on a theory of meaning (distributional semantics) which is over 60 years old. Although most of my discussion will focus on the application of new methods for research in the humanities, I will end by switching the perspective and considering how such approaches can enrich education in the humanities and produce graduates equipped with diverse skills which will serve them well in our digital world.

Corresponding Author:
Simon Musgrave

Received: 21 May 2019
Accepted: 26 June 2019
Published: 7 July 2019

Publishing services provided by Knowledge E

Selection and Peer-review under the responsibility of the UICRIC Conference Committee.

1. Introduction

Alfred North Whitehead famously characterised the European philosophical tradition as “a series of footnotes to Plato” (Whitehead 1929). In doing this, Whitehead draws attention to an important aspect of much scholarship in the humanities: the data with which we work, its nature and its extent, is a given. Of course, this is not true for all disciplines or for all research, but it does apply in many cases. And a consequence is that, in many cases, new knowledge in the humanities comes as the result of applying new theories or new methods (or both) to our existing data. In the last two decades, a great deal of new knowledge across the humanities has come as the result of applying methods and tools from computer science to problems in the humanities in the movement now known as digital humanities.

This kind of work has been carried out across the full range of disciplines in the humanities and the social sciences. Early work explored the possibilities of using computers to help elucidate complex texts (Busa 1980; McCarty 2004), but more recently a huge variety of work has been produced. In literary studies, we see numerous

How to cite this article: Simon Musgrave, (2019), “Digital Humanities: New Tools and New Knowledge” in *UNNES International Conference on Research Innovation and Commercialization 2018*, KnE Social Sciences, pages 808–818. DOI 10.18502/kss.v3i18.4770

examples of online archives devoted to the work of particular authors (for example, the Shelley-Godwin archive (http://shelleygodwinarchive.org/ (accessed 02/11/2018))), but we also see research applying machine learning algorithms to literary material (Long & So 2016). In the field of history, we see very large interdisciplinary projects such as the Venice Time Machine (https://vtm.epfl.ch/ (accessed 02/11/2018)) as well as smaller scale projects with a tight focus on a rich body of material, such as the Digital Panopticon project. (https://www.digitalpanopticon.org/ (accessed 02/11/2018)) In the area of art history and cultural heritage, we see projects which make virtual experiences available to new audiences (Kenderdine 2013). Moving away from the core humanities disciplines, we see political scientists employing sophisticated computational techniques to investigate models of democratic process (Gold et al. 2015), and media scholars analysing enormous bodies of data taken from social media (Benkler, Faris & Roberts 2018).

In this essay, I will explore some examples in detail to show the ways that new methods can interact with old data, old theories and new theories to produce new knowledge. These examples will be drawn from my own discipline, linguistics, but I hope that the generality of the argument will be clear. Following discussion of three examples, I will turn briefly to how these new methods can and should inform our teaching.

2. New Methods for Existing Data

I have suggested that, in many cases, the evidence base for research in the humanities is at least partially fixed and that therefore advances in our knowledge depend on the application of new methods or new theories. In this section, I will discuss an extreme example of this type of situation, one where it seems almost impossible that new evidence will ever become available to scholars.

The impact of European settlement on the Indigenous inhabitants of Tasmania was severe, indeed brutal (Ryan 2012). Their social structures were effectively destroyed by the middle of the nineteenth century, and their languages were barely used beyond that date. The last person believed to have command of an Indigenous Tasmanian language (Fanny Cochrane Smith) died in 1905; the only audio record of any Tasmanian language was made by her granddaughters in 1972 (Dixon 1980: 230). The available data for the study of Tasmanian languages is 44 word lists collected by various people between 1777 and 1847 including just over 1000 words (Bowern 2012: 4591). Not surprisingly, given this paucity of evidence, there has been little agreement amongst scholars about even the question of how many distinct languages there might have been on Tasmania.

To quote Dixon (1980: 229): “The honest answer to the question ‘how many languages were there in Tasmania?’ is ‘we don’t know’; to say ‘somewhere between eight and twelve’ is to hazard an only slightly informed guess.” The only possibility of adding to the evidence base would be the discovery of as yet unknown records, and this is highly improbable: the available evidence is all that there will ever be as far as we can tell.

I have just described the situation until a few years ago – today we know a good deal more about the linguistic situation of pre-settlement Tasmania as a result of some excellent research which applied new computational methods to the problem. The work of the Australian linguist Claire Bowern (2012) gave us new and exciting insights into this problem, or rather problems, because Bowern’s work applies two different computational methods to two questions: how many languages were there, and is there any evidence for higher order relationships between them (and between them and other languages)?

One of the impediments to studying the Tasmanian materials was that several of the wordlists clearly included data from more than one language, but there was not sufficient evidence to make clear decisions about which words came from a single language (and this difficulty was compounded by the fact that accurate information about the background of the speakers who had been recorded was often lacking). The first stage of Bowern’s research applied the first computational method, a sophisticated clustering algorithm used to identify admixture in the wordlists. The result of this process was a set of vocabularies with no mixing and a preliminary answer to the first question: the best clustering result identified 12 languages. These vocabularies were the input to the second computational method, another clustering algorithm, one developed in genetic biology (Neighbor-Net, see Holland et al. 2004), and this procedure confirmed the model with 12 languages (The criterion for separating languages is quantitative and arbitrary as Bowern acknowledges; as she also points out, the criteria used by linguists to separate languages and dialects are also arbitrary) and added the additional information that the 12 languages could be grouped in five clusters. Perhaps more importantly, Bowern’s results also gave answers to two higher level questions for which no convincing answer had been possible before. She was able to conclude that there is no evidence that the Tasmanian languages were all part of a single language family (that is, they did not share a single ancestor language), nor are the Tasmanian languages related to the Indigenous languages of mainland Australia.

The data on Tasmanian languages is limited and will almost certainly never be extended; it had posed problems for linguists for many years. The application of powerful computational methods by Bowern has significantly increased the state of our linguistic knowledge about Tasmania and has had important implications for historical and anthropological investigation of Indigenous Tasmania also, and these results will stand as our best account of the problem – at least until even more powerful methods become available.

3. New Methods for Existing Theories

Two important linguists of the mid-twentieth century, J.R Firth in the UK and Zellig Harris in the US, both advanced a theory about the relationship between the meanings of words and their distributions (Harris 1954; Firth 1968). The idea, now commonly known as distributional semantics, was pithily summed up by Firth: “You shall know a word by the company it keeps”. The study of collocations, words which occur together in texts, is a part of distributional semantics and has yielded impressive results (see for example Xiao & McEnery 2006; McEnery & Baker 2017), but semantic analysis of this type was restricted to single words or small groups of words until very recently. Work on new algorithms (such as word2vec, Mikolov et al. 2013) made it possible to construct mathematically precise models which locate every word in a text sample in relation to every other word based on their co-occurrence in text. Such models, known as Vector Space Models (VSMs), are multidimensional representations of the association strength between every pair of words in the text sample, given the total frequency of these words. As such, VSMs are distributional semantic models of a sample of a language in use and are rich sources of semantic information; for example, they perform very well on analogy tasks.

One question which I and my colleagues have been investigating using such models is the extent to which the meaning of morphological elements is also recoverable from a VSM, and I will now outline a portion of that work which relates to verbal morphology in Bahasa Indonesia; my collaborators in this research are Gede Primahadi Wijaya Rajeg (Monash University and Universitas Udayana) and Karlina Denistia (Eberhard Karls University, Tübingen) (Rajeg, Denistia & Musgrave 2018). In this work, we have looked at the verbs derived from nouns where suffixation with both –*i* and –*kan* is possible; for example, the noun *akhir* can be the root of three verb forms: *mengakhir, mengakhiri,* and *mengakhirkan*. We looked at such sets of verbs where each member of the set occurred at least 10 times in our corpus, (The model was trained on 184044395 word tokens (184269 word types) drawn from the Leipzig Corpora Collection Indonesian corpus. [http://corpora.uni-leipzig.de/en?corpusId=ind_mixed_2013, accessed -2/11/2018]) and then we analysed how they grouped together. A measure of similarity (cosine similarity

(Cosine similarity is a measure of similarity between two non-zero vectors that measures the cosine of the angle between them. The cosine of 0° is 1 therefore two vectors with the same orientation have a cosine similarity of 1, two vectors oriented at 90° relative to each other have a similarity of 0, and two vectors diametrically opposed have a similarity of -1)) can be calculated between vectors in the model, and such similarity measurements can then be used as the basis for a cluster analysis. The results of this analysis are shown in Figure 1.

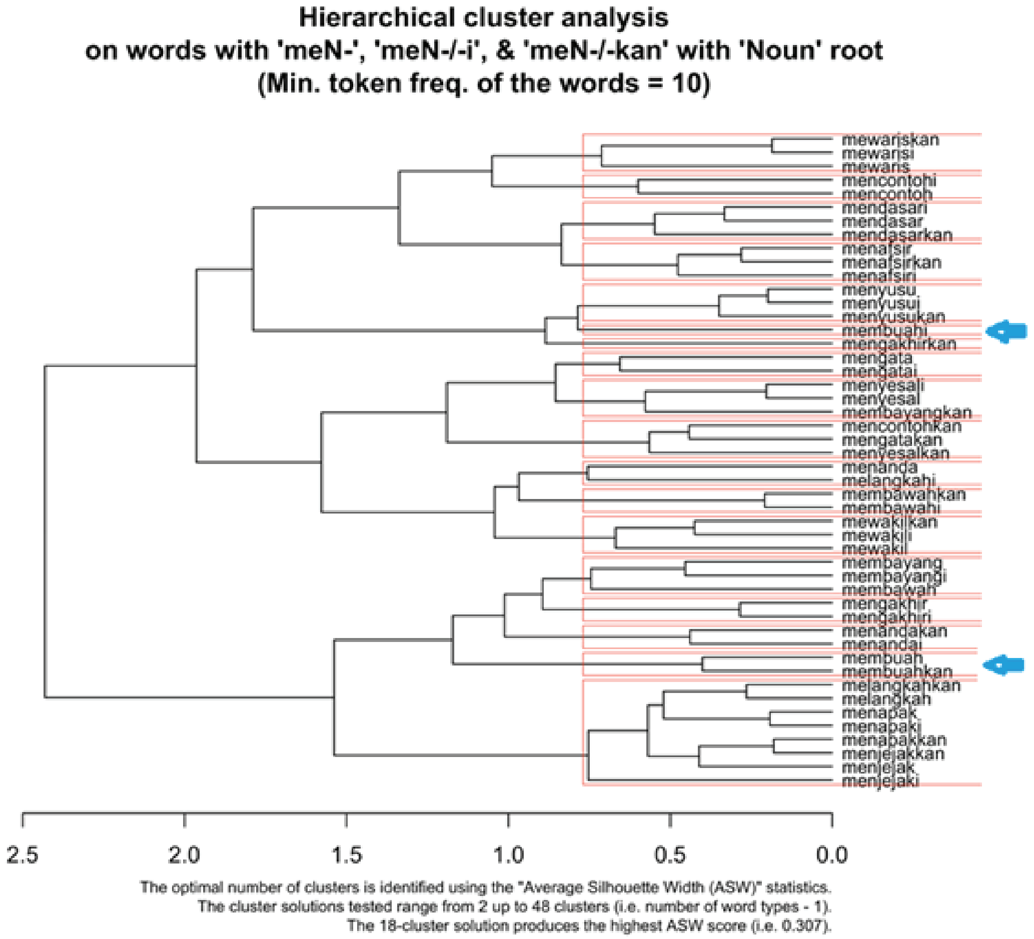


Figure 1

We can see that most sets of verbs cluster together, but not all. The arrows at the right hand side of Figure 1 show the positions of *membuahi* on its own and *membuah* and *membuahkan* clustered together. Other sets which are not immediately adjacent include *mengakhir* and *mengakhiri* which are separated from *mengakhirkan* and *mencontoh* and *mencontohi* separate from *mencontohkan*.

The different clustering possibilities just described can then be the basis for further investigations. Table 1 shows the 10 words closest in the model to each of the three verbs derived from *buah*. (The original word is always returned as its own closest neighbour, so these lists are actually the second to eleventh words most similar) We can see immediately that *membuahi* is located in a very different part of the semantic space compared to the other two verbs. This semantic region is very coherent (all the words are either morphological relatives of the root or are to do with fertilization) and it is dense (the tenth word in this list is only slightly less similar to its root word than the most similar word in either of the other two lists).

Table 1: Ten words closest to each verb derived from *buah*.

membuah	Cosine Sim.	**membuahkan**	Cosine Sim..	**membuahi**	Cosine Sim.
membuahkan	0.5991543	berbuah	0.6716316	dibuahi	0.8372309
barreto	0.5974444	mem-buahkan	0.6296921	ovum	0.8330801
memperdaya	0.5889142	tercipta	0.6214626	sperma	0.7965994
bobol	0.5863349	membuah	0.5991543	gamet	0.7326975
tandukan	0.5792586	menuai	0.5729809	pembuahan	0.7300101
tandukannya	0.5776066	tendangannya	0.5533927	terbuahi	0.7016097
bareto	0.5766841	ditepis	0.5530693	spermatozoid	0.6901719
blunder	0.5719315	kerasnya	0.5528045	spermatozoa	0.6794663
menjebol	0.5712235	pinalti	0.5482891	parthenogenesis	0.6775866
baretto	0.5712207	dimentahkan	0.5478375	zigot	0.6712285

A more conventional analysis of the collocates of the different derived verbs confirms this picture, details are presented in an Appendix.

Our research into these questions is at a very preliminary stage, but I hope that what I present here already shows the fascinating further questions which have appeared. For example, are the semantic patterns different when it is the –*kan* derived verb which does not cluster with the others? Is it possible to distinguish the semantic contribution of the two functions of the –*kan* suffix. What is exciting about this new method here is that it provides a way of approaching such questions in a very precise way.

4. New Methods and New Theories

In this section, I will briefly discuss some very recent research which uses VSMs to provide representation of meaning which is precise enough to be used as part of a machine learning process. This work uses a combination of methodologies, all of which are recent in origin: VSMs, functional magnetic resonance imaging (fMRI) and supervised machine learning. One might argue that this work is outside of the humanities; however, I believe that the research is linguistic in important and challenging ways and further that it raises profound questions about meaning and mental operations which certainly should be of interest to humanistic scholars.

Pereira and his colleagues (Pereira et al. 2018) acquired fMRI images of the brains of subjects reading words and sentences. They also constructed mathematical representations of the meanings of those words and sentences using a VSM. In each case, the representation was a vector, that is, a numerical value for each dimension in the model. The model consists of such vectors for each word in the input text, so vectors for individual words are easily recovered. Vectors representing sentences were constructed by taking the mean of the vectors of the words in the sentence. The decoder was then trained on pairings of fMRI scan and meaning vector. After training, the decoder was tested on unseen scans and produced meaning representations which corresponded to the stimulus at levels well above chance; for example, average accuracy on identifying single words was 0.74 ($p < 0.01$). Similar results were obtained from sentence discrimination tasks, and even relatively open-ended decoding showed the decoder was identifying content words at better than chance performance.

I find these results astonishing (And rather disturbing – it is not hard to imagine ways in which people might attempt to use such techniques for bad purposes) and their implications are far-reaching. I do not have space to follow them here; the point that I do want to make is that this cutting-edge research crucially depends on work in digital humanities as I conceive that endeavour. A computational method (VSMs) has been applied to a classic problem in the humanities (how to represent meaning). Using a radical abstraction of the input data, a prototypical computational approach, we can reach one solution to that problem (certainly not the only one) which in turn can contribute to addressing another profoundly complex problem, how meaning might be represented in the human brain.

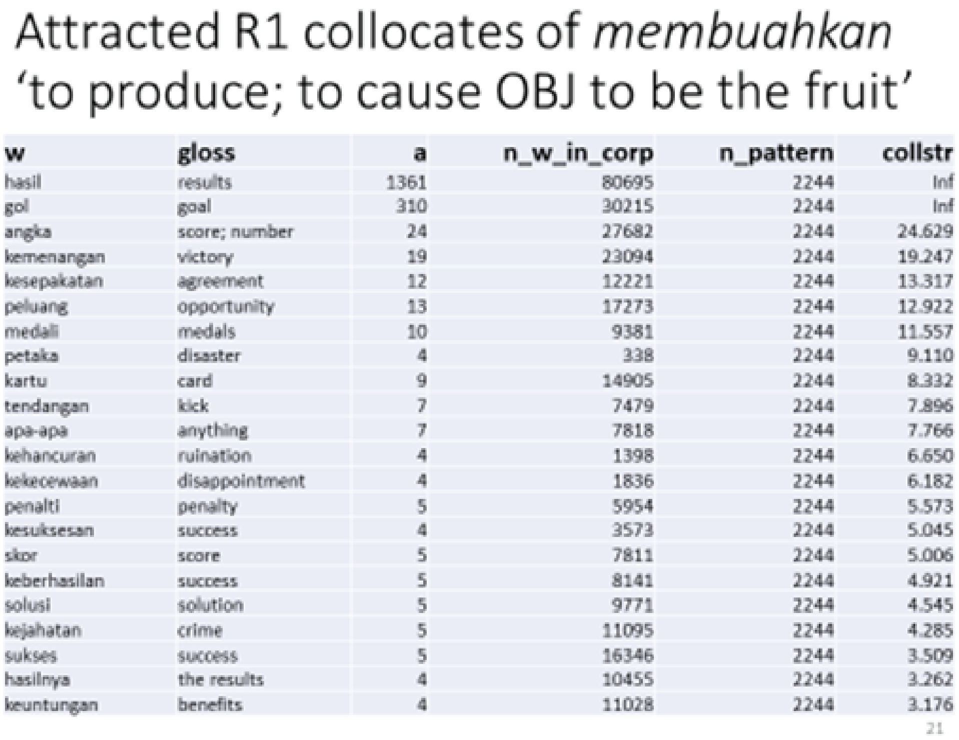
5. New Methods and Education

The humanities, both in research and in teaching, have not ignored the technological change happening to our world. There is by now a huge literature exploring and critiquing what it means to be human in a digital age (to cite only two examples: Dourish & Bell 2011; Tredinnick 2008), but this is only a part of what I would include in digital humanities and the development of that strand of research does not mean that the methodological aspect of digital humanities has been adopted everywhere. In fact, it is only very recently that digital humanities could be reasonably spoken of as ‘widespread’ and even then with the qualification that the spread is mainly restricted to research and graduate education. The impact of digital humanities on undergraduate teaching is still limited, but I would like to finish by suggesting that there are good reasons why we should be working to change this situation.

Firstly, we all (I hope) can subscribe to the idea that our teaching should reflect the best available scholarship but also should cover the most recent work in our fields of expertise. That recent work is increasingly likely to involve some aspect of digital scholarship and we therefore have to be prepared for students to be interested to learn about the methods which are used to produce such research, perhaps even to learn the methods themselves.

Secondly, we have a responsibility to equip our students for the world in which they will live and work. The concept of digital literacy (or literacies) can be approached in a variety of ways (e.g. Eshet-Alkalai 2004), but however we understand the concept, we must aim to train students in digital competencies where they are relevant to our disciplines. Even beyond this, acquiring skills is an important part of staying on the right side of the so-called digital divide (Van Dijk & Hacker 2003); the base level of skill may be higher for many of our students today than it was a few years ago, but the skill level needed for full engagement has also lifted. The kind of jobs which have been thought of as career paths for humanities graduates, such as school teaching or policy work for government agencies, are increasingly dependent on digital skills such as knowing how to access reliable sources of data and knowing how to apply computational tools to analyse and present that data. These are the skills used in digital humanities and it is therefore, I believe, entirely sensible for us to teach digital humanities to all our students.

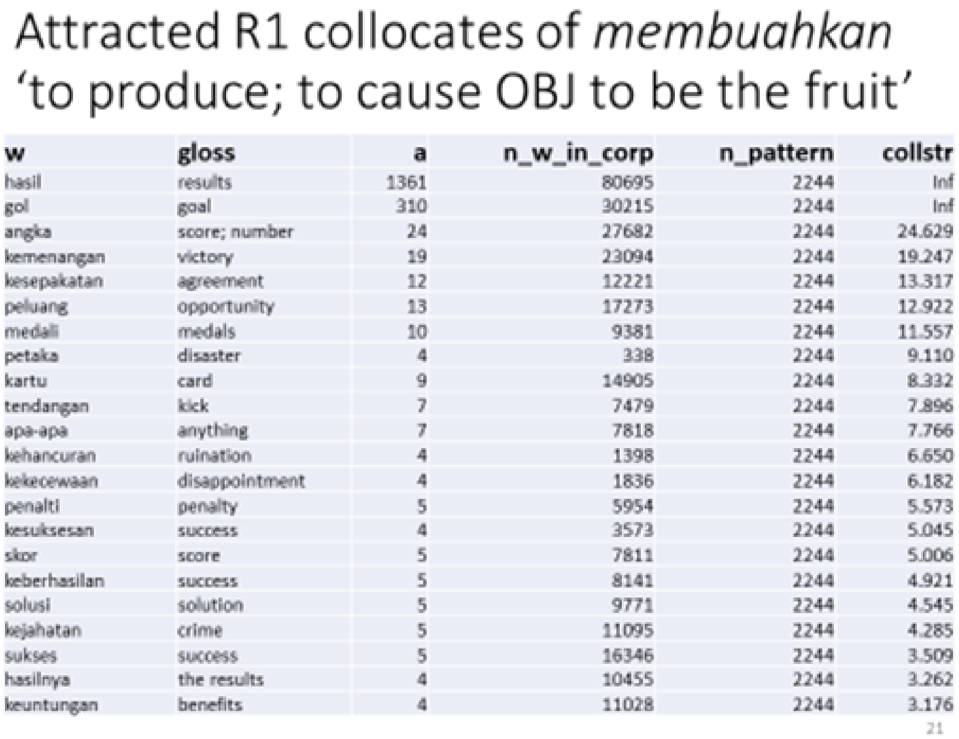
Appendix



Attracted R1 collocates of *membuahkan* 'to produce; to cause OBJ to be the fruit'

w	gloss	a	n_w_in_corp	n_pattern	collstr
hasil	results	1361	80695	2244	Inf
gol	goal	310	30215	2244	Inf
angka	score; number	24	27682	2244	24.629
kemenangan	victory	19	23094	2244	19.247
kesepakatan	agreement	12	12221	2244	13.317
peluang	opportunity	13	17273	2244	12.922
medali	medals	10	9381	2244	11.557
petaka	disaster	4	338	2244	9.110
kartu	card	9	14905	2244	8.332
tendangan	kick	7	7479	2244	7.896
apa-apa	anything	7	7818	2244	7.766
kehancuran	ruination	4	1398	2244	6.650
kekecewaan	disappointment	4	1836	2244	6.182
penalti	penalty	5	5954	2244	5.573
kesuksesan	success	4	3573	2244	5.045
skor	score	5	7811	2244	5.006
keberhasilan	success	5	8141	2244	4.921
solusi	solution	5	9771	2244	4.545
kejahatan	crime	5	11095	2244	4.285
sukses	success	5	16346	2244	3.509
hasilnya	the results	4	10455	2244	3.262
keuntungan	benefits	4	11028	2244	3.176

21



Attracted R1 collocates of *membuahkan* 'to produce; to cause OBJ to be the fruit'

w	gloss	a	n_w_in_corp	n_pattern	collstr
hasil	results	1361	80695	2244	Inf
gol	goal	310	30215	2244	Inf
angka	score; number	24	27682	2244	24.629
kemenangan	victory	19	23094	2244	19.247
kesepakatan	agreement	12	12221	2244	13.317
peluang	opportunity	13	17273	2244	12.922
medali	medals	10	9381	2244	11.557
petaka	disaster	4	338	2244	9.110
kartu	card	9	14905	2244	8.332
tendangan	kick	7	7479	2244	7.896
apa-apa	anything	7	7818	2244	7.766
kehancuran	ruination	4	1398	2244	6.650
kekecewaan	disappointment	4	1836	2244	6.182
penalti	penalty	5	5954	2244	5.573
kesuksesan	success	4	3573	2244	5.045
skor	score	5	7811	2244	5.006
keberhasilan	success	5	8141	2244	4.921
solusi	solution	5	9771	2244	4.545
kejahatan	crime	5	11095	2244	4.285
sukses	success	5	16346	2244	3.509
hasilnya	the results	4	10455	2244	3.262
keuntungan	benefits	4	11028	2244	3.176

21

References

[1] Benkler, Yochai, Rob Faris & Hal Roberts. 2018. *Network propaganda: manipulation, disinformation, and radicalization in American politics*. New York, NY: Oxford University Press.

[2] Bowern, Claire. 2012. The riddle of Tasmanian languages. *Proceedings of the Royal Society of London B: Biological Sciences* 279(1747). 4590–4595. doi:10.1098/rspb.2012.1842.

[3] Busa, Roberto. 1980. The annals of humanities computing: The index thomisticus. *Computers and the Humanities* 14(2). 83–90.

[4] Dixon, Robert M. W. 1980. *The languages of Australia* (Cambridge Language Surveys). Cambridge, Eng; New York: Cambridge University Press.

[5] Dourish, Paul & Genevieve Bell. 2011. *Divining a digital future: Mess and mythology in ubiquitous computing*. Cambridge MA: MIT Press.

[6] Eshet-Alkalai, Yoram. 2004. Digital literacy: A conceptual framework for survival skills in the digital era. *Journal of Educational Multimedia and Hypermedia* 13(1). 93.

[7] Firth, J.R. 1968. A synopsis of linguistic theory 1930-1955. In F. R. Palmer (ed.), *Selected Papers of J.R. Firth 1952-1959*, 168–205. London: Longman.

[8] Gold, Valentin, Mennatallah El-Assady, Tina Bögel, Christian Rohrdantz, Miriam Butt, Katharina Holzinger & Daniel Keim. 2015. Visual Linguistic Analysis of Political Discussions: Measuring Deliberative Quality. *Digital Scholarship in the Humanities* fqv033. doi:10.1093/llc/fqv033.

[9] Harris, Zellig S. 1954. Distributional Structure. WORD 10(2–3). 146–162. doi:10.1080/00437956.1954.11659520.

[10] Holland, Barbara R., Katharina T. Huber, Vincent Moulton & Peter J. Lockhart. 2004. Using Consensus Networks to Visualize Contradictory Evidence for Species Phylogeny. *Molecular Biology and Evolution* 21(7). 1459–1461. doi:10.1093/molbev/msh145.

[11] Kenderdine, Sarah. 2013. “Pure Land”: Inhabiting the Mogao Caves at Dunhuang. *Curator: The Museum Journal* 56(2). 199–218. doi:10.1111/cura.12020.

[12] Long, Hoyt & Richard Jean So. 2016. Literary pattern recognition: Modernism between close reading and machine learning. *Critical Inquiry* 42(2). 235–267.

[13] McCarty, Willard. 2004. The Analytical Onomasticon. *An Analytical Onomasticon to the Metamorphoses of Ovid*. http://www.mccarty.org.uk/analyticalonomasticon/ (2 November, 2018).

[14] McEnery, Tony & Helen Baker. 2017. *Corpus linguistics and 17th-century prostitution: computational linguistics and history* (Corpus and Discourse. Research in Corpus and Discourse). London□; New York, NY: Bloomsbury Academic. http://ezproxy.lib.monash.edu.au/login?url=http://oapen.org/download?type=document&docid=625761 (9 July, 2018).

[15] Mikolov, Tomas, Kai Chen, Greg Corrado & Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

[16] Pereira, Francisco, Bin Lou, Brianna Pritchett, Samuel Ritter, Samuel J. Gershman, Nancy Kanwisher, Matthew Botvinick & Evelina Fedorenko. 2018. Toward a universal decoder of linguistic meaning from brain activation. *Nature Communications* 9(1). doi:10.1038/s41467-018-03068-4. http://www.nature.com/articles/s41467-018-03068-4 (7 March, 2018).

[17] Rajeg, Gede Primahadi Wijaya, Karlina Denistia & Simon Musgrave. 2018. Semantic Vector Space Model and the usage patterns of Indonesian denominal verbs. *The Twenty-Second International Symposium On Malay/Indonesian Linguistics (ISMIL 22), UCLA (11-12 May 2018)*. doi:doi.org/10.4225/03/5acffc60eb649.

[18] Ryan, Lyndall. 2012. *Tasmanian Aborigines: a history since 1803*. Crows Nest, N.S.W: Allen & Unwin.

[19] Tredinnick, Luke. 2008. *Digital information culture: the individual and society in the digital age*. Elsevier.

[20] Van Dijk, Jan & Kenneth Hacker. 2003. The digital divide as a complex and dynamic phenomenon. *The information society* 19(4). 315–326.

[21] Whitehead, Alfred North. 1929. *Process and Reality, an Essay in Cosmology, by Alfred North Whitehead,...* The University Press.

[22] Xiao, Richard & Tony McEnery. 2006. Collocation, Semantic Prosody, and Near Synonymy: A Cross-Linguistic Perspective. *Applied Linguistics* 27(1). 103–129. doi:10.1093/applin/ami045.

Simon Musgrave is a lecturer in the School of Languages, Literatures, Cultures and Linguistics at Monash University, where he also co-ordinates the undergraduate major in Digital Humanities. His research interests include the use of computational tools in linguistic research, Austronesian languages, and the influence of communities of practice in language description.