PIP THORNTON

WORDS AS DATA: THE VULNERABILITY OF LANGUAGE IN AN AGE OF DIGITAL CAPITALISM

The security of the data that circulates the internet is dependent on much more than cryptographic key exchange. Data can represent all manner of information that might threaten personal and national securities and safety, be it through the misuse of social media or mapping data, the tracking of personal information for advertising, or the state-led gathering of financial or communications data.

Some of these data (mis)uses can of course be avoided, mitigated or challenged, but there is one type of data that underpins almost every aspect of our digital lives, regardless of who we are, which is much harder to shield from forces of commercialisation, surveillance and the systemic biases of technology – and that is linguistic data.

The language that flows through the platforms and portals of the Web is increasingly mediated and manipulated by large technology companies that dominate the internet, and in particular for the purpose of advertising by companies such as Google and Facebook. Whether through keyword targeting, email, search engine optimisation techniques, or the dissemination of news or status updates, the words that circulate through digital space are increasingly laden with economic value.

In this respect, words-as-data become detached from their original function as a means of human communication, and instead become vessels for the flow of advertising and cultural capital around the online and offline world. This has significant consequences.

We all need to communicate, access information and keep up in the modern marketplace, but in today's digitally networked society, the words we enter into Web-based platforms such as search engines and social media have themselves been turned into valuable pieces of data. And when words are digitised for transmission and processing through the Web, they lose their original context. Just like any other type of data, linguistic data becomes vulnerable to manipulation and monetisation.

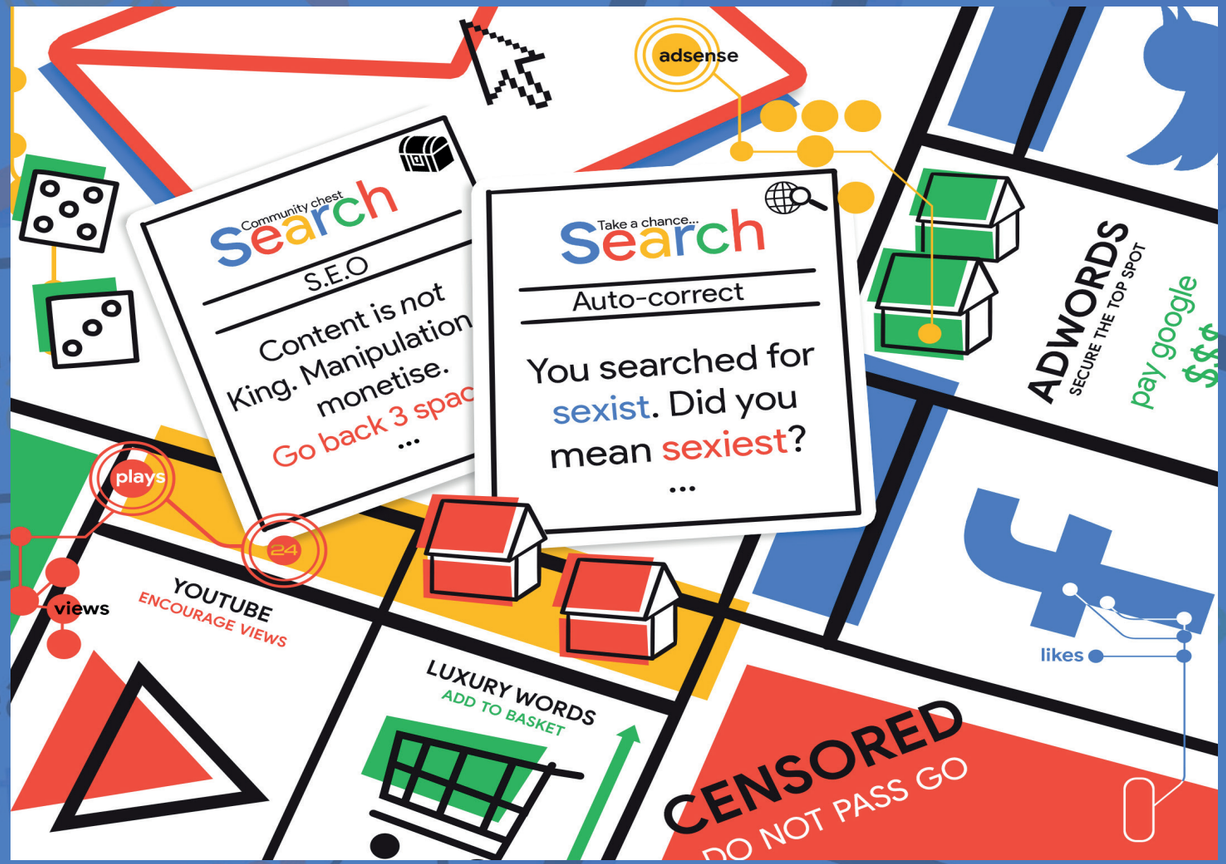
The computational manner in which linguistic data is processed is responsible for the sometimes amusing, but also sometimes dangerously stereotypical and controversial auto-predictions that appear when you start typing in the Google search bar. Auto-predictions are based on a mixture of aggregated previous searches, and the existing data available on the Web. Words and phrases that appear more frequently next to each other in this 'searchable database' will therefore be more likely to complete your search query.

The problem with this is that any omission, manipulation or bias in the searchable database is therefore reproduced and compounded. So the word 'man' or 'male' might be more often associated with nouns like 'doctor', 'boss' or 'CEO', and this will be reflected in search results and auto-completions. It is also the reason why online translation services like Google Translate are often so bad.

Google can at any time also interfere, censoring certain keywords so that they won't be included in the construction of search results. This might be for political, commercial, legal or ethical reasons. Google is not a neutral and democratic gatekeeper of the world's information, and it is crucially important not to treat what comes out of the search engine as unmediated truth.

The way digitised language is structured is also dependant on the monetary value of words in the online advertising industry. Google is one of the main players in this marketplace – its commodification and exploitation of language has been described as a form of 'linguistic capitalism'. Google has around a 95% market share of internet searches in the UK, and its advertising platforms AdWords and AdSense have an ever increasingly significant impact on how all kinds of information circulates on the Web.

AdWords is the system by which advertisers bid and pay for keywords and phrases in order to secure the top spots on Google's search engine results page. Each time somebody searches for a word on Google, a mini auction takes place, and the advertiser with the highest bid for that particular word at that time wins, and as long as their advert is considered worthy by the algorithmic ranking system, their advert will appear at the top of the search results page, above the 'organic', non-paid results. The information appearing before our eyes is therefore mediated by the vagaries and complexities of a linguistic market. Even the so-called 'organic' search results are significantly affected by the forces of linguistic capitalism. A whole Search Engine Optimisation industry has grown out of identifying and valuing keywords to make online text more attractive to search algorithms.



And this is a really important point. Much of the text that exists online is structured and restricted by digital processing systems, and/or created or optimised not for human readers, but for the algorithms that scrape text for the purposes of targeted advertising. The information we receive through search engines is therefore susceptible and vulnerable to the fluctuations and restrictions of an algorithmic marketplace. The value – and therefore the reliability – of language has become destabilised by digital capitalism.

Digital capitalism also has a huge role to play in the rise of fake news. While propaganda and subversive advertising are nothing new, many of the 'fake news' stories that circulated the Web in the run up to the 2016 US Presidential election were written not for any particular political motive, but because Google pays website owners to host adverts through its AdSense platform. The more views a website (and the adverts served on it) has, the more money the owner makes, regardless of its content.

A politically controversial story, spread virally through media such as Facebook 'likes', 'shares' and 'comments', can generate thousands of dollars in advertising revenue. What is important to remember here, is that the stories being generated, while often completely made up – as in the case of many of the anti-Clinton stories in 2016 – become embedded into the fabric of the Web, their linguistic data contributing to future searches, translations, and other informational systems.

The influence and control of language on the Web therefore translates into a frightening power over the generation and dissemination of information. As a result, we need to be asking what narratives are we creating when our online discourse is optimised for the spread of capital rather than for narrative communication? What does it mean that every query we make of a search engine is influenced by (often opaque) algorithmic 'market forces', or that YouTube videos aimed at children contain sexual or violent material to encourage more views and therefore more advertising revenue? As we have seen in the revelations about Cambridge Analytica, the spread of fake news through digital advertising is perhaps the tip of the iceberg.

The systemic manipulation and monetisation of digitised language is a threat to the security and stability of modern society. The very words we use to communicate, learn, debate, and critique have become compromised by opaque algorithmic organisation and optimisation, and the market-driven profits of private companies such as Google. We might therefore ask ourselves, just how resilient and secure is language in the digital age? Indeed, how can we even talk about security when we cannot talk securely?

Pip Thornton is a Post-Doctoral Research Associate in Creative Informatics at the University of Edinburgh. The material in this essay is based on her recently published articles 'A Critique of Linguistic Capitalism: Provocation/Intervention' *(2018) and* 'Geographies of (con)text: Language and Structure in a Digital Age' *(2017), and on her research blog www.linguisticgeographies.com*