



Deposited via The University of Leeds.

White Rose Research Online URL for this paper:

<https://eprints.whiterose.ac.uk/id/eprint/195688/>

Version: Accepted Version

Book Section:

Sadler, N (2024) Translation and Datafication. In: Bassnett, S and Johnston, D, (eds.) Debates in Translation Studies. Routledge, Abingdon, Oxon, pp. 106-123. ISBN: 9780367612351.

This item is protected by copyright. This is an author produced version of a chapter published in Debates in Translation Studies. Uploaded in accordance with the publisher's self-archiving policy.

Reuse

Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

Translation and Datafication

We live in a world increasingly saturated with, and shaped by, data. The concept of ‘datafication’, at its broadest, aims to account for this shift, referring to the increasing significance of data in contemporary society. Data is collected, processed and used on a (literally) industrial scale across almost all aspects of daily, and especially economic, activity: data analytics are increasingly important in health care, with companies such as the pharmaceuticals giant Roche offering ‘personalised medicine’ aimed at ‘bringing together a unique understanding of human biology with new ways to analyse health data’ (Roche 2021); in the energy sector, wind turbines are filled with sensors providing real time data about vibration, particulate levels in lubricants, heat, tower sway, wind direction and more.¹ In politics, fine grained demographic data is used to precisely target campaign materials to where they will be most effective (or so campaign managers hope).

Description and analysis of datafication is the subject of a large and growing literature in fields including: education (Jarke and Breiter 2019; Selwyn and Gašević 2020; Williamson, Bayne, and Shay 2020), media studies (Dijck van 2014; Livingstone 2019; Arsenault 2017), urbanism (Iveson and Maalsen 2019; Bibri 2019), health (Kuch, Kearnes, and Gulson 2020; Ruckenstein and Schüll 2017), sport (Fried and Mumcu 2017), epistemology (Bonde Thylstrup, Flyverbom, and Helles 2019) and postcolonialism (Couldry and Mejias 2019; Kotliar 2020). Entire journals, such as *Big Data & Society*, and nascent disciplines, such as Critical Data Studies and Critical Algorithm Studies, have emerged to make sense of it.

Datafication has also transformed, and continues to change, the translation industry. From the development of translation memory (TM) tools in the 1990s to the increasing importance of cloud-based translation management suites today, large scale data collection, processing and use is now a central part of the industry. It is therefore remarkable that it has received so little attention from translation scholars. Despite much, frequently insightful, discussion of translation technology over recent decades, the specific issues pertaining to datafication remain largely untouched. Jiménez-Crespo’s recent review article ‘The “Technological Turn” in Translation Studies: Are we there yet?’ (2020), for example, makes no references to data or datafication at all. Of the 31 entries in *The Routledge Handbook of Translation and Technology* (O’Hagan 2020), none focus specifically on data or datafication. At the time of writing, searching for ‘datafication’ in the leading Translation Studies journals *The Translator*, *Translation Studies*, *Meta* and *Target* does not bring a single result. This suggests a major research gap.

In this context, this chapter is intended to be largely programmatic. It begins with a brief overview of datafication as a phenomenon and its development. This is followed by discussion of three key ideas identified in the existing literature on datafication in terms of their implications for translation as a practice, industry and profession: 1) that it goes beyond digitisation insofar as the shift is not only from the analogue and towards the digital, but also towards the organisation of that digital information in such a way as to render it ‘indexable and thus searchable’ (Mayer-Schönberger and Cukier 2013: 84); 2) that it entails significant quantification of the human; 3) that it is closely connected to behavioural changes. Each is intended to highlight how the concept of datafication can enrich our understanding of translation and raise key questions for research over the coming years. My emphasis throughout will be on commercial translation on the basis that this is the area where the effects of datafication have so far been felt most keenly. Nonetheless, as time goes by, it is likely that its effects will also become increasingly apparent in other areas as well. In the concluding section, I reverse the focus and suggest that the theory of datafication also has much to gain from

¹ See, for example, the Moventas ‘Condition Management System’ (Moventas 2019).

the contribution of translation scholars; it may be a relatively new phenomenon, but many of the issues it raises are closely related to those explored within Translation Studies for many years.

Defining datafication

It is undeniable that there is simply far more quantitative data flying around the world today than there was ten, five, or even two years ago. Nonetheless, a desire to understand the world quantitatively is by no means new. Consider, for example, Heidegger's emphasis in his 1954 essay 'The Question Concerning Technology' on the idea that 'modern science's way of representing pursues and entraps nature as a calculable coherence of forces' (Heidegger 2013: 21); the desire of British colonial officials to produce accurate censuses of the empire's subjects, seen as a necessary part of effective administration (Christopher 2008); or US Secretary of Defense Robert McNamara's attempts in the 1960s to wage a 'scientific' war in Vietnam through the collection and analysis of large volumes of quantitative data (Harrison 1988; Tarpgaard 1995). Even so, the digitisation seen over recent decades was a prerequisite for the pervasive datafication we see today; as more and more of social, cultural and economic life has been carried out digitally, it has become immeasurably easier to collect and process data on a large scale. During the Vietnam war, collecting data about individual military operations was costly and slow, requiring paper forms to be filled in, physically moved from location to location, before being entered into computerised systems for analysis. With digitisation, on the other hand, the data needed for this kind of analysis is often already available in digital form.

Digitisation was, however, a necessary but insufficient condition for datafication. A move to the digital makes the collection and processing of data much easier, but it does not necessarily demand it. Datafication, on the other hand, constitutes a specific way of taking up and responding to the possibilities offered by digitisation. In terms of translation, digitisation brought a shift from working with paper to electronic texts. Datafication has brought the practice of translators sending segment-level TM data to language service providers (LSPs) to enable aggregation and the creation of large datasets. Not everything that happens digitally is necessarily recorded as data. Even in a world saturated by the digital, data does not occur naturally but must be generated through processes of selection and ordering. As the title of Gitelman's influential edited volume attests, *"Raw Data" is an Oxymoron* (Gitelman 2013). There may be more data than ever but what can be seen in it depends on what is collected and what filtered out as noise – an issue to which we will return later as part of a discussion on the relationship between datafication and contemporary capitalism. As Meijias and Couldry put it, 'in a network, nodes only recognise other nodes, and if something is not represented as a node it does not exist' (Meijias and Couldry 2019: 4). Google researchers Caswell and Liang (2020), for example, emphasise that machine translation 'performance on low-resource languages, like Yoruba or Malayalam, still leaves much to be desired'. That such languages may be used daily by tens of millions of people in digital communication and yet still be considered 'low-resource' emphasises that digitisation and datafication need not run in parallel. The uneven distribution of data, and its capacity to reinforce existing biases, is now the focus of a substantial literature in its own right (Noble 2018; Hamilton 2019; Obermeyer et al. 2019).

Thinking in terms of data, then, is not grounded in a wholly new impulse. Datafication as a contemporary phenomenon, on the other hand, does represent a significant change even from the digital era, at least in terms of scale and tempo. The growth in data is both a requirement for and, as discussed below, a consequence of these changes. It is important to recall that it is not inevitable, all-encompassing, entirely negative in its effects, or immune to resistance. Nonetheless, it seems clear that major changes are underway.

Searchability

Every time I look out of my window, I experience what analytic philosophers such as Bertrand Russell and H.H. Price once referred to as ‘sense-data’. I might refer to a novel as the ‘data’ being analysed by one of my students. Yet neither of these examples constitutes data as understood in the context of datafication; the defining characteristic of data on this view is that it is searchable and indexable. It is difficult to overstate the importance of this distinction. The key requirement for datafication is not an increase in information as such, but in data in formats amenable to automated, computerised analysis. It is an increase in this specific type of data that has driven a significant shift towards quantitative and statistical methods of interpretation. Obviously, neither quantitative nor statistical analysis are new. The change brought with datafication is the extent of their spread. For present purposes, this change has two major implications. The first is a shift, clear-sightedly noted two decades ago by Manovic (2001), from the narrative, temporally-oriented logic which dominated much of human history to one much more strongly grounded in the non-linear, spatially-oriented retrieval based logic of the database. This does not mean that narrative has ceased to be relevant; stories remain important means for making sense of and operationalising non-narrative data, as I and others have previously argued (Sadler 2021; Hayles 2007; Dourish and Gómez Cruz 2018). All the same, it has brought with it a significant increase in the importance of quantified and spatially organised ways of understanding (Halpern 2015; Drucker 2014). We see this in the translation industry, for example, in the charts and tables of now ubiquitous ‘dashboards’ intended to provide ‘insights’ into datasets.²

A second, and comparatively more recent, shift following from the increased availability of searchable data has been an enormous growth in the importance of algorithms in many aspects of life.³ This is both a response to increased data – they help in interpreting otherwise overwhelming quantities of information – as well as a driver of it – an increased emphasis on algorithms creates an incentive to maximise the quantities of data available for them to work with. Initially, their significance was largely focused on facilitating, rather than replacing, human decision making. In 2013, for example, Airbnb produced a basic Japanese version of their platform within a week using crowdsourced translation, using an algorithm to score the importance of different parts of the site to identify which content should be prioritised by translators (Airbnb 2013). In other cases, and with increasing frequency, decision making itself is outsourced to algorithms. In 2020, for example, the school leaving exams taken by 16- and 18-year-olds in the United Kingdom were suspended due to the Coronavirus crisis. In response, the UK government initially used an algorithm to determine results, factoring information such as teachers’ estimates of pupils’ likely grades along with historical data about schools’ past performance. Rather than feeding into a decision-making process, the algorithmically determined result was final.⁴

The increased availability of searchable data has also had profound implications for translation. Most obviously, it has enabled the development of contemporary statistical approaches to machine translation which rely for their training data on large corpora of parallel texts which, beyond merely existing, must also be available in machine readable formats. Newer, and increasingly sophisticated, ‘adaptive’ machine translation builds on this by using complex algorithms to combine information

² Consider, for example, the ‘Taus DQF’ Dashboard which purports to provide insights into the entire translation industry in real time - <https://qd.taus.net/>.

³ Understanding the effects of algorithms is now the focus of a growing body of literature in its own right (e.g. Bucher 2018; Gillespie 2018; Cohn 2019).

⁴ Due to significant discrepancies in the results and public opposition, however, the approach was ultimately changed and the algorithmically determined grades abandoned in most cases (cf BBC 2020).

from very large training sets with smaller datasets produced by individual translators to produce outputs which benefit from the scale of large datasets but also reflect something of the style of individuals. This is only possible when individual translators also produce machine readable data through the computer aided translation (CAT) tools they use. The necessary information may have existed in earlier periods – translation, after all, has happened for millennia. But without datafication it was not effectively useable in this way.

The translation industry is also making increasing use the availability of large, searchable datasets in other ways. Translation quality assurance, for instance, is becoming more datafied. As of the time of writing, the translation technology company memoQ, for example, provides both an 'Automatic Quality Assurance' tool to identify issues such as punctuation errors, missing numbers etc, and a 'Linguistic Quality Assurance' tool, a platform for human reviewers to categorize mistakes and 'create error statistics to provide structured feedback to translators' (memoQ n.d.). Atril's 'Déjà Vu' software enables the creation of 'User Tracking Reports' which give 'the Administrator and project managers... insight into user (translator) performance. Detailed statistics are provided for each user's performance in each project based on the match types' (Atril Solutions n.d.). Straker Translations, on the other hand, claims that their 'X Ray' analytics engine can calculate 'What is the normal rate a translator will work on this type of content and language pair to predict time-frames' and assist in '[f]inding the most suitable translator for the type of content and languages of your project' (Straker Translations n.d.). All, in different ways, rely on the availability of data readily amenable to automated analysis and interpretation by algorithms.

It is not difficult to imagine how this might expand in the future. If data is collected continuously as translators work, it would be possible, for instance, for LSPs to see that a particular translator is typically more productive in the mornings than the afternoons or that their output tends to dip in the days following a period of intense work. This could be factored into decisions about which translators are commissioned and which passed over. As in other areas, algorithms could be used more extensively for automated, real-time surveillance of translators, continuing the gradual move seen over recent decades towards 'disciplinary' (Foucault 1977) modes of control in which employers specify not only the final product required, but also precisely control the manner of its production. All the data needed to make these possibilities a reality is already routinely collected – it would simply be a case of developing new ways of processing and utilising it.

The quantification of the human

Datafication has brought ever greater quantification of aspects of human existence which might, in the past, have been understood as beyond quantification. This is, again, not a new phenomenon in and of itself. The essays written by students taking degree programmes in the humanities, for example, are emphatically qualitative in nature and yet are typically assigned a single numerical grade to facilitate their being directly comparable with one another. Lyotard's *The Postmodern Condition: A report on knowledge* (1984), first published in 1979, discussed at length the increasing externalisation, commodification and quantification of knowledge provoked by the, then still relatively new, technology of computers. He suggested it was greatly accelerating a long-standing process of separating knowledge from individual knowers which began with the development of writing and was at that time gaining pace through 'microstorage, computerized data, and data banks of hitherto unimaginable proportions' (Lyotard 1984: xii). This, for Lyotard, had the consequence that:

The old principle that the acquisition of knowledge is indissociable from the training (*Bildung*) of minds, or even of individuals, is becoming obsolete and will become ever

more so... Knowledge is and will be produced in order to be sold, it is and will be consumed in order to be valorized in a new production' (Lyotard 1984: 4)

As Lyotard has it, the separation of knowledge from knowers through quantification not only enables but is driven by a desire to commoditise and objectify knowledge, enabling it to be bought and sold.

In the context of translation, this is exactly what we see with TMs. On an individual level, they allow for individual translators to externalise much of the factual information to which they refer in their daily work, storing it as electronic data rather than as biological memory. Once quantified and externalised as data, it can be assigned a price and traded, for example on the Taus 'Data Marketplace'⁵ which offers translators the opportunity to 'monetize data you collected overtime'. The ways in which translation data can be monetized also emphasise a need to rethink where the value of translated material lies. In isolation, a translated text functions as a non-fungible commodity – it is not directly substitutable in the sense that one translation cannot be simply swapped with another in the way that one barrel of oil can be substituted with another. It has little or no value beyond the specific context of its commissioning and use – a translation of a legal document pertaining to the divorce of Mr and Mrs Smith is of no direct value in the context of the separation of Mr and Mr Klein. As Lyotard suggests, and Sadowski (2019) addresses directly, datafication changes this. Once seemingly qualitative texts are quantified as indexed and searchable data, and aggregated with data generated from other texts, they are transformed into something more like a kind of capital which can be 'valorized in a new production' (Lyotard 1984: 4) which need not be directly connected to the circumstances of its original creation. As Sadowski (2019: 4), drawing on Marx, puts it, 'the capitalist is not concerned with the immediate use of a data point or with any single collection, but rather the unceasing flow of data-creating'.

Quantification of the human also has profound existential implications. In his early works, Heidegger (2010; 1982) demonstrates the inadequacy of thinking of human existence in terms of autonomous individuals.⁶ He emphasises instead the central importance of what he calls 'Dasein' – 'being-there' in the sense that all being happens in particular contexts and through its interconnections with the environment. The salient part of Heidegger's approach for the present discussion is the central role he assigns to 'equipment' understood as 'the *nearest things* that surround us' (Heidegger 1982: 163). He argues that in dealing with equipment, we do not 'grope forward from one to the next, progressively taking the single things together' but proposes that rather there is 'always already a manifold of equipment' that 'environs us'. This 'contexture' shapes what it means to exist in the particular place and manner of our existing: it is 'the prius, within which specific beings, as beings of this or that character, are as they are and exhibit themselves correspondingly' (Heidegger 1982: 164). From this basic account, we begin to see that the existence of a translator is grounded not in an autonomous Cartesian individual, but in the ensemble of the translator along with the equipment that they use. In a very real sense, a translator who routinely works with CAT tools as a core part of their practice cannot be understood as the same translator without that equipment.⁷

⁵ <https://datamarketplace.taus.net/>

⁶ From a Translation Studies perspectives, there are some parallels between Heidegger's approach and Blumczynski's (2016) emphasis on thinking about translation, and ontology more broadly, in terms of 'how-ness' rather than 'what-ness'.

⁷ There are parallels between this idea and work on translation as 'extended cognition' (Risku and Windhager 2013; Risku 2014) and 'human-computer interaction' (O'Brien 2012; Paulsen Christensen and Schjoldager 2010; Bundgaard, Paulsen Christensen, and Schjoldager 2016). Heidegger's ontological perspective,

Datafication does not undermine the centrality of equipment to human existence. Nonetheless, it does further extend it. In thinking about what makes a given translator the translator that they are, it is increasingly necessary to consider the data they use as a basic part of the existential ensemble. In more concrete terms, when a client hires a translator, they hire not just a person but the whole ensemble of the translator and their data. Indeed, the suitability of a translator for a given task may lie as much in the data at their disposal as in their skills. The TM is, therefore, not something simply bolted on to an otherwise autonomous translator – the TM is a key part of what constitutes their being as a translator at its most basic. For a translator to lose access to their TM data could be as drastic as a woodworker losing a hand or a musician having their instrument destroyed. In each case, the translator, woodworker or pianist would not simply cease to exist. Nonetheless, their way of being a translator, pianist or woodworker would significantly change. Conversely, this also suggests limits to the commodification and sale of translation data. If a translator is only the translator that they are with their data, the data is only the data that it is along with the individual translator. Elements cannot simply be removed from the ensemble and continue to function in precisely the same way.

The quantification of the human also contributes to changing how the world is understood on a basic level. In his later work, Heidegger argues that the ‘essence of [modern] technology is by no means anything technological’ but rather lies in a particular way of encountering the world.⁸ He proposes that with the gradual shift from artisanal craftwork to the era of ‘machine-powered technology’ (Heidegger 2013: 13), the distinctly technological way of encountering the world became to cease to see things as objects and begin to engage with them as ‘*Bestand*’ – a standing reserve of resources to be exploited. He proposes, then, that woodland comes to be seen solely as a source of timber, a river as a source of hydroelectric power and so on. As he puts it in relation to passenger aeroplanes:

Yet an airliner that stands on the runway is surely an object. Certainly. We can represent the machine so. But then it conceals itself as to what and how it is. Revealed, it stands on the taxi strip only as standing-reserve, inasmuch as it is ordered to ensure the possibility of transportation. For this it must be in its whole structure and in every one of its constituent parts, on call for duty, i.e., ready for takeoff (Heidegger 2013: 17)

Its meaning comes to lie solely in the uses to which it can be put and it no longer comes into view as an object in and of itself.

Translation remains, to a significant degree, an industry of craftwork rather than mass production; translators continue to produce individual translations in response to individual source materials. The transformation of translations into capital wrought by datafication, however, allows translated material to subsequently come into view as *Bestand* in Heidegger’s sense for perhaps the first time. Once broken into segments, quantified and indexed within databases, the notion of the text dissolves. They cease to exist as objects understood in the sense of their being defined through their differentiation from other objects. Instead, they are absorbed within a larger standing reserve of data, ordered to ensure the possibility of the extraction of future value. They are no longer valuable in themselves, but in terms of the use that might later be gleaned from them. We might therefore

nonetheless, leads us to situate tool-use as a basic component of existence itself, assigning it a far more fundamental role than that typically seen in Translation Studies.

⁸ As this suggests, Heidegger’s understanding of ‘technology’ is very different to how it is typically understood in Translation Studies.

say that datafication is not itself something technological in that it does not refer to any particular technology. Nonetheless, it arises from the essence of technology as Heidegger describes it, bringing with it equally profound implications.

The modification of human activity

It will be clear from the preceding sections that datafication is associated with changes, of varying importance, to human behaviour. In some cases, they follow datafication that has already taken place. In others, changes are engineered specifically to enable the generation of data to facilitate the advancement of datafication – they are a requirement *for* datafication, rather than a simple consequence of it. Social media platforms, for example, not only enable human sociality in a such a way that it can be recorded and become productive of valuable data but are designed to encourage it to take place in this way. One purpose of Google Maps is to make it easier to find your way around; another is to collect data about its users' movements to facilitate the development of individualised advertising profiles. One purpose of CAT tools is to make translation faster for translators; another is to generate data useable by agents other than the translator.

This pattern of changing existing behaviours, or adding new ones, to either generate new data or capitalise on existing data is a recurrent one. Face-to-face conversations create no usable data; Facebook messenger conversations do. Paper shopping lists create no usable data; shopping list apps such as 'Out of Milk' and 'Mealime' do. Going for a run, in itself, produces no data; running while having location, heartrate, duration, speed etc tracked by an app such as Apple Health does. Visiting a beach generates no data, 'checking in' to the beach with Instagram does. In each of these instances, seemingly small changes of behaviour are required to transform activities which do not produce data into those that do. This can, in turn, lead to larger behavioural changes – rather than simply using Google Maps to find my way to restaurants, I may start following its recommendations about which restaurants to go to in the first place; rather than simply checking in with Instagram at the beach, I may start visiting places specifically so that I can check in there.⁹ Rather than simply taking advantage of the data generated when translators use CAT tools, an LSP may start requiring them to use them specifically so that more data can be collected.

Several important points arise. First, this is an ambivalent process. The datafication involved in each of these examples is valued, albeit to varying degrees, by their users.¹⁰ The behavioural changes wrought by datafication are not necessarily for the worse. Google Maps' data-driven service, for example, makes finding a hotel in an unfamiliar city hugely easier than it used to be without it. In other cases, the benefits of behavioural changes accrue predominantly to the institutions encouraging the changes. There are few obvious benefits to Facebook users, for instance, of the company's practice of semi-covertly collecting data about its users' location either from smartphones' own location services or via the EXIF data attached to photos uploaded to the site or sent through its Messenger app (Doffman 2020).

Second, in many cases, datafication is intimately connected to contemporary forms of capitalism which emphasise generating monetary value and exerting disciplinary control through the collection

⁹ There is growing evidence, for example, that social media has a significant impact on choices of travel destination, both in terms of it shaping perceptions of destinations and people traveling specifically to create social media content (Smith 2018; 2019; The Economist 2019).

¹⁰ Just how valuable, however, remains a subject of debate. Recent studies have suggested, for example, that the amount the average Facebook user would have to be paid to deactivate their account for a year ranged from \$453 (Brynjolfsson, Eggers, and Gannamaneni 2018) to over \$1000 (Corrigan et al. 2018). When asked the same question about search engines, on the other hand, far higher figures were given – a median of \$17,530 in Brynjolfsson, Eggers and Gannamaneni (2018: 30)

and manipulation of data.¹¹ The behavioural changes described in the previous paragraph are all intended to generate data which is commercially useful and can be monetized. Google, for example, uses data to build detailed profiles of its users: their likes and dislikes, their patterns of daily movements, health, shopping habits etc. This information, as well as access to users, can then be sold at a premium to advertisers. This model has been remarkably successful – Google Ads generated \$147 billion in 2020 (Graham and Elias 2021). Not all datafication is motivated by capitalism. Neither the datafication of health in the UK National Health Service or of authoritarian surveillance in China, for example, are aimed at serving the needs of capital. But the technology firms which have played such a central role in driving datafication do so squarely in the pursuit of economic capital, even as they espouse techno-utopian discourses of positive data-driven transformation. This emphasises a need to be critical with the behavioural changes associated with datafication.

Third, a major question remains as to the extent of choice available. On the one hand, runners need not log their runs with Strava or communicate using Facebook Messenger. Nonetheless, the datafication of society means that social life increasingly happens in and through tools like these – not merely recording activities which would have otherwise taken place in the same way, but rather modifying the activities themselves and identities which go with them. For club cyclists, for instance, group rides may be organised on Facebook and recording and uploading data to tracking services can be an important part of what it means to be part of the club. While it is therefore not impossible to opt out of at least some aspects of datafication, it can be difficult. As Lanier (2013: 207) puts it, ‘what might have started out as a choice is no longer a choice after a network effect causes a phase change. After that point we effectively have less choice. It’s no longer commerce, but soft blackmail’.

These issues also arise with, and influence, translation. As already discussed, the practice of translation, especially in its commercial forms, has been greatly influenced by the emergence and now dominance of CAT tools. Consequently, translation practices have changed. It has been routine for many years for LSPs to require translators to submit TM data at the end of each task they complete so that it can be monetised. It is now routine for LSPs and TTPs to collect data on anything and everything possible, even if ‘the amounts collected may vastly exceed a firm’s imaginative reach or analytic grasp’ (Fourcade and Healy 2017: 13). In the context of post-editing, for example, the availability of data on editing time and edit distance per segment has allowed for new types of benchmarking and comparison between translators. Quoting for jobs may now involve using services such as ‘memoQ project manager’ which can produce a preliminary ‘analysis report’ which:

Shows the size of the project and the savings from translation memories, LiveDocs corpora, and homogeneity. This is where you get the word counts that you use to create a quote for your client and estimate the costs of running the project.¹²

Which tasks an LSP does or does not accept along with the prices they charge, then, comes to depend not only the intrinsic complexity of the text, turnaround time, availability of translators, for example, but also on the data that they currently hold. Likewise, a difficult text which otherwise might seem to offer little prospect of turning a profit taken in isolation may become more appealing if completing it will generate data which can be exploited in the future. The shift to cloud-based translation management means that very large volumes of aggregate data are now routinely supplied to TTPs. The opportunities such platforms offer for real-time monitoring of individual

¹¹ Data’s role in capitalism is the focus of growing literature. For example: Chandler and Fuchs (2019); Iveson and Maalsen (2019); Zuboff (2015); Sadowski (2019).

¹² <https://www.memoq.com/products/memoq-project-manager>

translator contributions allows larger tasks to be shared between many different translators without specific parts being allocated to individuals; since the contributions of individual translators are logged as they are made, payment can be made according to individual contributions once the task is completed.

Individual translators' behaviours have also changed. Most obviously, their decision to use CAT tools will be influenced by the datafication of the industry more widely which has changed expectations as to the volume a translator can be expected to translate each day and the potential benefits of datafying their own work, in terms of time saved in the future. A translator deciding to adopt cloud-based working practices may do so because of a perceived need to obtain a ranking useful for getting future work (cf Garcia 2017: 67/68).¹³ The shift to thinking in terms of segments rather than texts is both a response to existing datafication (because it is at the level of the segment that suggestions from TMs can be most effectively made) and driven by its future needs (it is also the level at which the most useful data is generated). The operative unit of translation as a practice is changed from the chunks most amenable to human interpretation to the units which are most useful for serving the needs of data.

As with datafication more broadly, these changes are not amenable to easy appraisals as positive or negative. The datafication of the translation industry is clearly grounded in contemporary capitalism. As seen in the discussion so far, and noted by Garcia (2017), the benefits accruing to LSPs are clear in terms of the surplus value that can be generated: datafying translators' working practices facilitates control and maximises revenue.¹⁴ The situation for individual translators is more complex. TMs reduce the time translators must spend on dull, repetitive work and, in building their own TMs as they work, allow them to also generate surplus value in each task they complete. They also accrue data capital over time and there is little doubt that datafication enables translators to handle larger volumes of words per day than they otherwise could, even if there is some evidence that it comes at a cost to creativity (Kenny and Winters 2020). Data on translator incomes is notoriously scarce, but increased productivity has been matched by changed perceptions as to what rate of work can reasonably be expected of translators, exerting downward pressure on per-word payment.

It is more difficult to see benefits for translators themselves from the more recent shift to cloud-based platforms. These changes enable the breaking down of complex tasks into many small tasks, distributed among many providers. Moorkens' (2020) associates this kind of 'digital taylorism' with reduced job satisfaction for translators. The data tracked by CAT tools are those of value to capital rather than necessarily those that matter most to human translators: words, productivity and edit distance rather than weekends lost to unreasonable client demands, or support labour provided by partners and families. Constant monitoring costs translators autonomy and privacy but gains them little. The extra surplus value produced by metadata accrues almost entirely to LSPs since it is most useful when aggregated with data from other translators. This may partly explain why the parts of translation technology company websites promoting cloud-based services to translators frequently make little or no mention of monitoring while they often feature prominently in materials targeted to project managers.

There are also limits to the extent to which translators can opt out of datafication, as with the phenomenon more broadly. A translator might choose not to use any CAT tools at all, therefore not

¹³ This also emphasises the possibility of datafication to lead to 'stratifying effects' as data contributes to the establishment and maintenance of new hierarchies (Fourcade and Healy 2017).

¹⁴ Although it is important to note, as Sadowski (2019: 4) highlights, much data collection happens 'without specific uses in mind', with collection becoming an end in itself in a context of intense datafication.

enabling the collection of segment-level data on their translation process. This approach remains viable in some case – perhaps in literary and academic contexts. In others, such as legal and commercial translation, however, commissioners frequently stipulate the use of specific CAT tools when advertising for translators. In keeping with the dominance of capital over labour in contemporary capitalism, a power dynamic clearly visible in the translation industry, LSPs are far more strongly placed to dictate terms than individual translators. Even where translators are not explicitly required to use CAT tools, a translator opting not to use them would likely struggle to compete with competitors using them. An alternative might be to use CAT tools but refuse to use cloud-based platforms collecting real-time data or refuse to send exported TM data at the conclusion of jobs. Such a stance would nonetheless be difficult to maintain without the cooperation of LSPs, given the significant power imbalances at play; they may determine that it is easier to simply find another translator willing to hand over TM data and permit remote surveillance than to negotiate with one who is not.

Conclusions and what translation has to offer

Datafication, then, brings a host of issues for Translation Studies. Changes in the industry have been well researched over the years but the central role of data in driving them, as part of a much broader societal change, requires greater attention than it has so far received. In this chapter, I have sought to raise a number of questions to be explored in depth in future work. An emphasis on searchability introduces subtle, but significant, changes to what data is understood to be, and to the kinds of analysis and uses to which it can be put. At the same time, a massive emphasis on quantifying human activity hugely increases the amount of information that can be understood, and exploited, as data in this way. Taken together, these factors have led to significant behavioural changes, both in response to the analysis of data and to enable the generation of data in the first place. All three trends reach far beyond translation. Yet their effects have been, and continue to be, keenly felt by translators, even if they have not been extensively researched by translation scholars.

In the final paragraphs, nonetheless, I want to emphasise that translational thinking has a great deal to offer thinking on datafication: first, datafication involves constant movement between signifying systems. This does not mean between different natural languages as emphasised in much traditional translation research but rather across ‘membranes’ in the broader sense proposed by Marais (2019). This kind of translation itself happens in two major ways: 1) from ‘non-data’ to ‘data’ – e.g. in the quantification of biological and physical processes such as human experience, the growth of a plant, or the functioning of a weather system. 2) Between datatypes: formats differ in how they work and what they enable to be done; it is possible to convert between them, but, as anyone who has tried to convert a PDF into a Word document will know, this is not necessarily a straightforward process. In both cases, the conceptual tools and experience available to Translation scholars mean that they are strongly positioned to investigate these transitions, along with the losses and opportunities which result.

Second, datafication in no way diminishes the importance of culture. Rather, it has led to the emergence of ‘data cultures’ (Albury et al. 2017) regulating how data is produced, cultivated and used. Furthermore, data does not necessarily remain within a single data culture but frequently moves between them. This can lead to complex intercultural interactions of a kind familiar to translation scholars. With regard to data collection and handling, for instance, the practices of major American technology companies such as Facebook, Uber, and Google are tightly connected to their technocapitalist cultures which prize the acquisition of the maximum possible volumes of data and

place very little emphasis on privacy and transparency.¹⁵ As their global influence has spread, tensions have arisen as they have attempted to apply that model within different data cultures – for example, in the European Union, their approach has clashed with legal structures and social norms strongly emphasising individual privacy, leading to the enactment of the sweeping General Data Protection Regulation in 2018 to control data storage, transfer, and collection. Authoritarian regimes in countries such as Turkey, Egypt and China, have strongly opposed American technology companies’ reluctance to share user data with state authorities. Again, these are issues that translation scholars are very well placed to address given their expertise in thinking cross-culturally.

Third, datafication does not reduce the importance of human interpretation. Databases and algorithms are built by humans. All databases are grounded in human decisions about what is and is not important, what should and should not be acknowledged as salient. The creation of databases is a translational process, to transform otherwise unmanageable volumes of information into useable data. Many algorithms also serve a similarly translational function: in processing data, they render comprehensible what is otherwise largely incomprehensible through a process of mediation. Even so, further acts of translation are often needed. At the turn of the millennium, Manovich lamented a situation in which there was ‘too much information and too few narratives to tie it all together’ (Manovich 2001: 217). Datafication is a way of interpreting the world but, all too often, its outputs are difficult for non-specialists to interpret and act upon. This establishes a need for what a 2018 paper by the management consultancy McKinsey describe as ‘analytics translators’ as the ‘new must-have role’ to mediate between technical specialists and other parts of businesses (Henke, Levine, and McInerney 2018). Translation scholars are ideally positioned to theorise such acts of translation and, perhaps in the future, also to prepare students for this type of work in translator training programmes.

¹⁵ Indeed, in commercial contexts, the term ‘data culture’ is frequently conflated with ‘data-driven culture’ and understood in terms of maximising the monetary value generated from data (e.g. Kalb 2021; Díaz, Rowshankish, and Saleh 2018; Rollings, Duncan, and Logan 2020; Cuelogic 2021).

Works cited

- Airbnb. 2013. "Launching Airbnb.Jp in Record Time." Medium. September 2013. <https://medium.com/airbnb-engineering/launching-airbnb-jp-in-record-time-52f8b0af965d>.
- Albury, Kath, Jean Burgess, Ben Light, Kane Race, and Rowan Wilken. 2017. "Data Cultures of Mobile Dating and Hook-up Apps: Emerging Issues for Critical Social Science Research." *Big Data and Society* 4 (2): 1–11.
- Arsenault, Amelia. 2017. "The Datafication of Media: Big Data and the Media Industries." *International Journal of Media & Cultural Politics* 13 (1–2): 7–24.
- Atril Solutions. n.d. "Companies and LSPs." Accessed July 28, 2021. <https://atril.com/companies-and-lsps/>.
- BBC. 2020. "A-Levels and GCSEs: How Did the Exam Algorithm Work?" BBC Explainers. August 20, 2020. <https://www.bbc.co.uk/news/explainers-53807730>.
- Bibri, Simon Elias. 2019. "The Anatomy of the Data-Driven Smart Sustainable City Instrumentation, Datafication, Computerization and Related Applications." *Journal of Big Data* 6 (59).
- Blumczynski, Piotr. 2016. *Ubiquitous Translation*. London & New York: Routledge.
- Bonde Thylstrup, Nanna, Mikkel Flyverbom, and Rasmus Helles. 2019. "Datafied Knowledge Production: Introduction to the Special Theme." *Big Data and Society* 6 (2): 1–5.
- Brynjolfsson, Erik, Felix Eggers, and Avinash Gannamaneni. 2018. "Using Massive Online Choice Experiments to Measure Changes in Well-Being." 24514. *National Bureau of Economic Research*. NBER Working Paper Series. Cambridge, MA.
- Bucher, Taina. 2018. *If...Then: Algorithmic Power and Politics*. Oxford & New York: Oxford University Press.
- Bundgaard, Kristine, Tina Paulsen Christensen, and Anne Schjoldager. 2016. "Translator-Computer Interaction in Action — an Observational Process Study of Computer-Aided Translation." *The Journal of Specialised Translation* 25: 106–30.
- Caswell, Isaac, and Bowen Liang. 2020. "Recent Advances in Google Translate." Google AI Blog. 2020. <https://ai.googleblog.com/2020/06/recent-advances-in-google-translate.html>.
- Chandler, David, and Christian Fuchs, eds. 2019. *Digital Objects, Digital Subjects*. London: University of Westminster Press.
- Christopher, A.J. 2008. "The Quest for a Census of the British Empire c.1840-1940." *Journal of Historical Geography* 34 (2): 268–85.
- Cohn, Jonathon. 2019. *The Burden of Choice: Recommendations, Subversion and Algorithmic Culture*. New Brunswick & London: Rutgers University Press.
- Corrigan, Jay R., Saleem Alhabash, Matthew Rousu, and Sean B. Cash. 2018. "How Much Is Social Media Worth? Estimating the Value of Facebook by Paying Users to Stop Using It." *PLoS ONE* 13 (12): 1–11.
- Couldry, Nick, and Ulises A. Mejias. 2019. "Data Colonialism: Rethinking Big Data's Relation to the Contemporary Subject." *Television and New Media* 20 (4): 336–49.
- Cuelogic. 2021. "What Is Data Culture? Why Is It Essential for Solving Organizational Problems?" 2021. <https://www.cuelogic.com/blog/data-culture>.

- Díaz, Alejandro, Kayvaun Rowshankish, and Tamim Saleh. 2018. "Why Data Culture Matters." *McKinsey Quarterly* September.
- Dijck van, José. 2014. "Datafication, Dataism and Dataveillance: Big Data between Scientific Paradigm and Ideology." *Surveillance and Society* 12 (2): 197–208.
- Doffman, Zak. 2020. "Facebook 'Secretly' Tracks Your iPhone Location—This Is How To Stop It." *Forbes*. December 2020. <https://www.forbes.com/sites/zakdoffman/2020/12/13/facebook-tracks-apple-iphone-locations-even-ios-14-and-iphone-12-12-pro-and-pro-max/?sh=3e4decb4156b>.
- Dourish, Paul, and Edgar Gómez Cruz. 2018. "Datafication and Data Fiction: Narrating Data and Narrating with Data." *Big Data and Society* 5 (2): 1–10.
- Drucker, Johanna. 2014. *Graphesis: Visual Forms of Knowledge Production*. Cambridge, MA & London: Harvard University Press.
- Foucault, Michel. 1977. *Discipline and Punish: The Birth of the Prison*. Translated by Alan Sheridan. London: Allen Lane.
- Fourcade, Marion, and Kieran Healy. 2017. "Seeing like a Market." *Socio-Economic Review* 15 (1): 9–29.
- Fried, Gil, and Ceyda Mumcu, eds. 2017. *Sport Analytics: A Data-Driven Approach to Sport Business and Management*. London & New York: Routledge.
- Garcia, Ignacio. 2017. "Translating in the Cloud Age: Online Marketplaces." *Hermes*, no. 56: 59–70.
- Gillespie, Tarleton. 2018. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. New Haven & London: Yale University Press.
- Gitelman, Lisa, ed. 2013. *"Raw Data" Is an Oxymoron*. Cambridge, MA & London: MIT Press.
- Graham, Megan, and Jennifer Elias. 2021. "How Google's \$150 Billion Advertising Business Works." *CNBC*. 2021. <https://www.cnn.com/2021/05/18/how-does-google-make-money-advertising-business-breakdown-.html>.
- Halpern, Orit. 2015. *Beautiful Data*. Durham, NC: Duke University Press.
- Hamilton, Melissa. 2019. "The Sexist Algorithm." *Behavioral Sciences and the Law* 37 (2): 145–57.
- Harrison, Fisher. 1988. "Computers, Electronic Data, and the Vietnam War." *Archivaria* 26: 18–32.
- Hayles, Katherine. 2007. "Narrative and Database: Natural Symbionts." *PMLA* 122 (5): 1603–8.
- Heidegger, Martin. 1982. *Basic Problems of Phenomenology*. Translated by Albert Hofstadter. Bloomington: Indiana University Press.
- . 2010. *Being and Time*. Translated by Joan Stambaugh. Albany, NY: State University of New York Press.
- . 2013. *The Question Concerning Technology and Other Essays*. New York, London, Toronto, Sydney, New Delhi, Auckland: Harper Perennial.
- Henke, Nicolaus, Jordan Levine, and Paul McNerney. 2018. "Analytics Translator: The New Must-Have Role." 2018. <https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/analytics-translator>.
- Iveson, Kurt, and Sophia Maalsen. 2019. "Social Control in the Networked City: Datafied Dividuals, Disciplined Individuals and Powers of Assembly." *EPD: Society and Space* 37 (2): 331–49.

- Jarke, Juliane, and Andreas Breiter. 2019. "Editorial: The Datafication of Education." *Learning, Media and Technology* 44 (1): 1–6.
- Jiménez-Crespo, Miguel A. 2020. "The 'Technological Turn' in Translation Studies: Are We There yet? A Transversal Cross-Disciplinary Approach." *Translation Spaces* 9 (2): 314–41.
- Kalb, Aaron. 2021. "What Is Data Culture and Why Do You Want One?" 2021. <https://www.alation.com/blog/what-is-data-culture/>.
- Kenny, Dorothy, and Marion Winters. 2020. "Machine Translation, Ethics and the Literary Translator's Voice." *Translation Spaces* 9 (1): 123–49.
- Kotliar, Dan M. 2020. "Data Orientalism: On the Algorithmic Construction of the Non-Western Other." *Theory and Society* 49 (5–6): 919–39.
- Kuch, Declan, M. Kearnes, and K. Gulson. 2020. "The Promise of Precision: Datafication in Medicine, Agriculture and Education." *Policy Studies* 41 (5): 527–46.
- Lanier, Jaron. 2013. *Who Owns the Future?* New York, London, Toronto, Sydney, New Delhi, Auckland: Simon and Schuster.
- Livingstone, Sonia. 2019. "Audiences in an Age of Datafication: Critical Questions for Media Research Published in Television and New Media." *Television & New Media* 20 (2): 170–83.
- Lyotard, Jean-François. 1984. *The Postmodern Condition: A Report on Knowledge*. Translated by Geoff Bennington and Brian Massumi. Manchester: University of Manchester Press.
- Manovich, Lev. 2001. *The Language of New Media*. Cambridge, MA & London: The MIT Press.
- Maraís, Kobus. 2019. *A (Bio)Semiotic Theory of Translation: The Emergence of Social-Cultural Reality*. New York & London: Routledge.
- Mayer-Schönberger, Viktor, and Kenneth Cukier. 2013. *Big Data: A Revolution That Will Transform How We Live, Work and Think*. London: John Murray.
- Meijas, Ulises, and Nick Couldry. 2019. "Datafication." *Internet Policy Review* 8 (4).
- memoQ. n.d. "What Is Translation Quality Assurance." Accessed July 22, 2021. <https://www.memoq.com/tools/translation-quality-assurance>.
- Moorkens, Joss. 2020. "'A Tiny Cog in a Large Machine': Digital Taylorism in the Translation Industry." *Translation Spaces* 9 (1): 12–34.
- Moventas. 2019. "Moventas CMaS." 2019. <https://www.moventas.com/moventas-condition-management-system-cmas/>.
- Noble, Safiya Umoja. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: New York University Press.
- O'Brien, Sharon. 2012. "Translation as Human–Computer Interaction." *Translation Spaces* 1: 101–22.
- O'Hagan, Minako, ed. 2020. *Routledge Handbook of Translation and Technology*. London & New York: Routledge.
- Obermeyer, Ziad, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations." *Science* 366 (6464): 447–53.
- Paulsen Christensen, Tina, and Anne Schjoldager. 2010. "Translation-Memory (TM) Research: What Do We Know and How Do We Know It?" *Hermes—Journal of Language and Communication Studies* 44: 89–101.

- Risku, Hanna. 2014. "Translation Process Research as Interaction Research: From Mental to Socio-Cognitive Processes." *MonTI. Monografías de Traducción e Interpretación*, 331–53.
- Risku, Hanna, and Florian Windhager. 2013. "Extended Translation." *Target. International Journal of Translation Studies* 25 (1): 33–45.
- Roche. 2021. "This Is Personalised Healthcare: Driving a Better Future for Patients." 2021. https://www.roche.com/about/priorities/personalised_healthcare.htm.
- Rollings, Mike, Alan D. Duncan, and Valerie Logan. 2020. "10 Ways CDOs Can Succeed in Forging a Data-Driven Organization." 2020. <https://www.gartner.com/doc/reprints?id=1-26OCBF3S&ct=210629&st=sb>.
- Ruckenstein, Minna, and Natasha Dow Schüll. 2017. "The Datafication of Health." *Annual Review of Anthropology* 46: 261–78.
- Sadler, Neil. 2021. *Fragmented Narrative: Telling and Interpreting Stories in the Twitter Age*. London & New York: Routledge.
- Sadowski, Jathan. 2019. "When Data Is Capital: Datafication, Accumulation, and Extraction." *Big Data and Society* 6 (1): 1–12.
- Selwyn, Neil, and Dragan Gašević. 2020. "The Datafication of Higher Education: Discussing the Promises and Problems." *Teaching in Higher Education* 25 (4): 527–40.
- Smith, Sean P. 2018. "Instagram Abroad: Performance, Consumption and Colonial Narrative in Tourism." *Postcolonial Studies* 21 (2): 172–91.
- . 2019. "Landscapes for 'Likes': Capitalizing on Travel with Instagram." *Social Semiotics* 31 (4): 604–24.
- Straker Translations. n.d. "XRAY Data Analytics Engine - Lowering Your Costs with Data Analytics." Straker Translations. Accessed July 27, 2021. <https://www.strakertranslations.com/technology/ai-translation-technology/xray/>.
- Tarpgaard, Peter. 1995. "McNamara and the Rise of Analysis in Defense Planning: A Retrospective." *Naval War College Review* 48 (4): 8.
- The Economist. 2019. "Daka Destinations." *The Economist*, August 17, 2019.
- Williamson, Ben, Sian Bayne, and Suellen Shay. 2020. "The Datafication of Teaching in Higher Education: Critical Issues and Perspectives." *Teaching in Higher Education* 25 (4): 351–65.

