Recibido / Received: 23/05/2020
Aceptado / Accepted: 27/07/2020

Para enlazar con este artículo / To link to this article:
http://dx.doi.org/10.6035/MonTI.2021.13.10

Para citar este artículo / To cite this article:
Buts, Jan & Henry Jones. (2021) "From text to data: mediality in corpus-based translation studies." En: Calzada, María & Sara Laviosa (eds.) 2021. *Reflexión crítica en los estudios de traducción basados en corpus / CTS spring-cleaning: A critical reflection*. *MonTI* 13, pp. 301-329.

FROM TEXT TO DATA: MEDIALITY IN CORPUS-BASED TRANSLATION STUDIES

JAN BUTS
butsj@tcd.ie
Trinity College Dublin, Ireland

HENRY JONES
h.jones4@aston.ac.uk
Aston University, UK

Abstract

This paper seeks to promote deeper reflection within the field of corpus-based translation studies (CTS) regarding the digital tools by means of which research in this discipline proceeds. It explicates a range of possibilities and constraints brought to the analysis of translated texts by the keyword in context (KWIC) concordancer and other data visualisation applications, paying particular attention to the ways in which these technological affordances have actively shaped central theoretical hypotheses within CTS and related fields, as well as the general principles of corpus construction. This discussion is illustrated through a small case study which applies the suite of corpus analysis tools developed as part of the Genealogies of Knowledge project to the investigation of two English translations of the *Communist Manifesto*.

Keywords: Mediality; Digital technologies; Data visualisation; Corpus construction; KWIC concordancer.

Résumé

Cet article cherche à stimuler une réflexion plus approfondie dans la traductologie de corpus concernant les outils numériques au moyen desquels la recherche est menée

dans cette discipline. L'article explique diverses possibilités et contraintes de l'analyse des textes traduits, assistée par un concordancier « KeyWord In Context » (KWIC) et par d'autres outils de visualisation des données. Une attention particulière sera portée à la manière dont ces affordances technologiques ont façonné des hypothèses théoriques centrales ainsi que les grands principes de la construction de corpus, dans l'approche traductologique basée sur le corpus et dans d'autres domaines proches. Cette discussion est illustrée par une étude de cas appliquant la suite logicielle développée pour l'analyse de corpus dans le cadre du projet Genealogies of Knowledge, à deux traductions en anglais du *Manifeste communiste*.

Mots-clés : Medialité; Technologies numériques; Visualisation des données; Construction de corpus; Concordancier KWIC.

1. Introduction

The use of text corpora for the investigation of language predates the invention of the modern computer (Fenlon 1908; Svartvik 1992: 7). Nevertheless, the success of corpus-based methodologies across the humanities today is primarily associated with the application and assistance of digital technologies in the research process (Luz & Sheehan 2020: 2-3). Beginning with Roberto Busa's work on the *Index Thomisticus* in the 1940s, corpus analysts have exploited the processing power of computers to facilitate the investigation of linguistic patterns repeated across ever larger collections of text. Such techniques rely fundamentally on the core principle of digital media, namely, numerical representation: the ability of the computer to transform any media object into the standardised language of mathematics (Manovich 2001). In order to be interrogated using corpus analysis software, a text must first be digitised; the alphabetic characters or logograms through which its contents are expressed must be converted into a binary code of 1s and 0s, itself an abstract representation of voltage, to be stored and interpreted by the machine. Once the information is stored, each token – a delineated string of characters in the corpus, often corresponding to a word – must be indexed (Luz 2011: 137-139). Tokens are assigned a numerical value that records the items' exact location in the source material. Finally, the researcher's ability to interact with the corpus depends on visualisation tools, such as the classic keyword in context (KWIC) concordance display, which can convert this

mathematical information and reconstitute fragments of the original corpus texts on the screen.

The transition from an analogue to a digital work environment and the shift towards binary code as the lingua franca of the twenty-first century has inaugurated a paradigm change within several established scholarly disciplines, but also inspired the creation of hybrid, interdisciplinary approaches to knowledge production. In this respect, *digital humanities* has come to serve as an umbrella term for a variety of practices, including the digitization of texts and artefacts, the study of born-digital material, as well as the development of digital tools and the new methods they facilitate (Sheridan 2016). A key theme within this field of inquiry has been a focus on questions of materiality, in part because technological changes have eroded the seemingly self-evident qualities of previous objects of study and their interpretation: if the digital medium has rendered text more dynamic, its previously more static qualities gain in significance. Similarly, if a hyperlinked environment invites non-linear browsing, reading conventions require renewed attention. In this regard, changes in literacy demands are now central to semiotic debate (Kress 2003). Several scholars have additionally begun to address self-reflexively the emphasis placed on pattern recognition in much digital humanities research, and to critically examine the implications of our focus on patterns – rather than structures or narratives – as the primary objects of study (Dixon 2012; Berry 2011).

In this broad context, the field of translation studies too has shown growing interest in the semiotic and material media in and through which translations are stored, transmitted and – by extension – studied (Armstrong 2020; Pérez-González 2014). There is growing recognition within the discipline that media tools such as books, newspapers, websites and DVDs are not passive conduits for the transmission of information, nor are they inert containers for its storage. Rather, they have their own mediality, they offer their own unique sets of techno-social possibilities and constraints, and they can thus more accurately be considered 'environments' that shape every aspect of our engagement with a text (Jones 2018). Littau (2011, 2016), for example, has explored a series of media-induced transformations in reading, writing and translation practices throughout history, from the oral culture of Ancient Rome through to the network culture of today's digital world. Jones

(2018) has examined how changes in the tools used to produce, distribute and consume audiovisual products during the twentieth century have influenced not only the approaches to translation adopted by subtitlers and their commissioners, but also the ability of ordinary consumers to get involved in this process. Finally, Cronin (2013) has discussed in more general terms the value of recognising the transformative potential of computing as an activity that reconfigures relationships between languages, cultures and texts.

Surprisingly, however, the subdiscipline of corpus-based translation studies (CTS) has remained largely silent on these issues, an omission which may seem particularly striking given the extent to which scholars based in CTS depend upon and interact with technology to enact their research. Reasons for this lacuna may vary. Considerations of mediality in CTS research would have been too distracting if investigated at the outset. When Baker (1993: 243) introduced corpus techniques in translation studies in order to elucidate "the nature of translated text as a mediated communicative event", this left no immediate room to consider the method of study as a mediated communicative event in itself. In addition, technological hesitancy may prompt researchers to turn away from interrogating the tools they use, and to remain silent on the topic. Practitioners of close reading are seldom acquainted with the physics of vision, and even less are they expected to be, yet a lack of digital literacy can inspire scholarly shame and therefore lack of transparency in the dissemination of tools and methodological pathways. Whatever the cause, the convertibility of the sign and its attachment to the binary standard are yet to be consistently questioned.

This paper argues that CTS requires a sustained interrogation of its practices in relation to its conditions of existence: the transformation, by means of code, of text that can be read into data that can be queried. The field of translation studies has successfully demonstrated that expressions across languages are declared rather than found equivalent (Hermans 2007: 6; Tymoczko 2010: 3). We argue that the same holds for expressions across medial environments, and that this includes the setting in which such expressions are studied. Equivalence between a dataset and the objects of study it is made to represent depends upon a specific, situated agreement on perceptual conventions established within a given research culture. Building on these principles, we seek to encourage deeper reflection within CTS and

related fields on the mediality of corpus research. This disciplinary 'dusty corner' is addressed in the following section mainly with reference to the role of the KWIC display in CTS research, which illustrates the convergence of the technological and theoretical boundaries of the field of study. Reflections on the limitations of KWIC analysis and of alternative forms of visualisation are not absent from the field of corpus research (e.g. Anthony 2018). Yet in this article, we seek to shed light not only on the restrictive, but also on the transformative aspects of using a given tool. The logic of the medium is shown to inform not only the mode of analysis, but also various aspects of corpus construction and representation. In the final sections this general account of mediality in CTS is instantiated by means of a discussion of the *Communist Manifesto* as it appears in the Genealogies of Knowledge (GoK) corpora. We illustrate that, just like multiple retranslations of a single text can, even within the space of a phrase, produce widely divergent images that invite a range of different interpretations, varying digital representations of textual material are only equivalent to their source insofar as the medium's mutational qualities are left uninterrogated.

2. A Medium Shaping a Discipline: CTS and the KWIC concordance

The term *medium* typically denotes a channel of communication and does so in abstract fashion. The medium is 'television' rather than the television set. In this sense, any medium is intimately connected to a technology, of which particular tools are instantiations. Different compounds such as *print media* and *social media* highlight different characteristics of a medial environment – among other aspects, one can focus on material conditions (as in 'print') or pragmatics (as in 'social'). The divide between these main aspects of media technologies often gives rise to different perceptions of how a medium takes shape and influences its users. On the one hand, a focus on the constraints enforced by new media technologies may lead to technological determinism: the tools you use condition the actions you undertake, and ultimately, the thoughts you have. This line of thinking has come to be associated with McLuhan's (1964: 7) mantra that "the medium is the message". On the other hand, a focus on the way new media technologies are received and put to particular uses by their human users may foreground the social construction

of technology (Klein & Kleinman 2002): the nature of our adoption of and engagement with technological innovations is the result of choices made at the individual as well as group level. The distinction between the material and the pragmatic perspective in the study of media thus runs along the classic divide between agency and structure. As several commentators since McLuhan have noted (Cronin 2013; Littau 2006), the most productive accounts of our ultimately social reality will take both perspectives into account, thus treating a medium as a compromise: a consensus arrived at by agents to act in accordance with a constraint imposed by a tool that, on the whole, facilitates a common goal.

Within CTS, this common goal is fundamentally the ability to identify and interrogate recurring patterns within large collections of written or transcribed text, selected according to a specific set of criteria, held in machine-readable form, and compiled in order to investigate a hypothesis about the process or products of translation (Baker 1995: 225). Such a collection of texts is itself little more than a mute database. It is the mode of access that turns a collection into a corpus in the modern sense: collected text is to be "analysed automatically or semi-automatically using different types of software specifically created for linguistic research" (Malamatidou 2018: 43). A certain vagueness about the type of software to be used has been part of discipline-defining statements in CTS since its early days, but in actual fact the predominant mode of display throughout the last three decades has been the KWIC concordance, identified by Baker (1995: 226) as "the corpus analyst's stock in trade". Indeed, the concordance line is the commodity the analyst deals in, as well as the currency that guarantees, through the provision of textual evidence, trust and recognizability within a larger research community.

Thus, in the case of CTS, one should not just consider 'the computer' or 'the screen' as the medium through which research proceeds. The KWIC concordance interface, a specific application of the computer as environment, is the prime medial display. In the following paragraphs we argue that the concordancer as medium is not a neutral tool of representation: not only does its design reflect a specific set of concerns among researchers interested in language use, but its affordances have also actively shaped central theoretical hypotheses within CTS and related fields, as well as the general principles

of corpus construction. None of what follows is intended to suggest that the principles of corpus research are defective, or that KWIC analysis is not informative. Rather, the examples we offer serve to indicate that there is a myriad of ways to approach language, and the tools we use put in place specific constraints that narrow down the possibilities productively, but in the process also transform the qualities of the textual material in view as well as our intuitions about it.

2.1 Theory and Technology

Electronic KWIC retrieval predates CTS by quite a stretch and is typically traced back to the work of Luhn (1960). The prior development of this technique can itself be situated within a much longer history of producing analogue concordances recording every instance of an alphabetical list of keywords along with a snippet of the immediate co-text for each occurrence. In addition to providing a solution to specific problems (for example, the need to identify suitable passages from the Bible with which to illustrate a sermon – Fenlon 1908), such endeavours reflect a fundamental interest – shared among theologians, philologists and ultimately linguists – in the iterability of the sign, or the ways in which lexical repetition and contextual variation constitute meaning.

For early adopters such as the linguist and lexicographer John Sinclair, whose work provided the main inspiration for the use of corpora in translation studies, the assistance of the machine offered a powerful means of developing more objective and productive methodologies for linguistic analysis, in marked contrast with previously dominant introspective approaches to the study of language (McIntyre & Walker 2019: 6; Stubbs 1996: 24). In particular, the use of an electronic KWIC concordancer made vastly more accurate and efficient the study of collocation, understood as the frequent co-occurrence of linguistic elements (Firth 1968: 14). It is important to note, however, that in the Firthian tradition out of which Sinclair's work emerged collocation had been understood to depend upon a broad, situational understanding of context. For Firth collocational patterns could also consist of sonorous markers such as alliteration, or organisational features such as paragraph structure (Partington 1998: 16-17). Once the KWIC display became

the dominant interface for accessing corpora, however, this tool came to strongly inform the theoretical priorities of linguistic research, as can be illustrated with reference to hypotheses about extended units of meaning.

One of the major arguments of corpus linguistics is that it can offer proof that individual words cannot be considered clear and distinct units of meaning (Hunston 2007: 250; Sinclair 2008: 409). Notably, the alternative proposals for extended units of meaning have been made conveniently in tune with the constraints of the medium. Semantic prosody, for instance, is a corpus-based theoretical innovation that refers to the evaluative or attitudinal function shared by a sequence of frequently co-occurring lexical items (Stubbs 2009: 124-125). The first elaborate illustration of the phenomenon can be found in Louw (1993), and despite some valid criticism regarding the coherence of the concept (Hunston 2007; Stewart 2010; Whitsitt 2005), semantic prosody has received sustained attention in corpus-based translation studies (Munday 2011; Stewart 2009). Louw's (1993: 170) original argument was rich in examples, such as the attested use of *symptomatic of*, a phrase which tends to be followed by a lexical item with a negative overtone (e.g. *tensions, inadequacies*), and which is thus argued to be itself indicative of a negative attitude towards a given topic. At times, however, examples are found where speakers use *symptomatic of* followed by a supposedly positive reference, as in *symptomatic of our good reputation*. According to Louw, this suggests that the speaker might be ironic, insincere, or perhaps even unaware of their own attitude.

Throughout Louw's article one finds a depreciation of introspective linguistics, and a plea for the corpus-based method to be adopted (Louw 1993: 173). Louw provides the reader with sets of concordance lines at several stages, because the crucial factor in identifying a semantic prosody, identifying deviations from it, as well as coming to grips with the reason for these deviations, is the perusal of a sequence of similar phrases. If some theoretical innovations can only be revealed as well as illustrated through concordances, the medium becomes indispensable to linguistic research, and ultimately determines the discipline's view of language. Indeed, in response to the received lexicographical paradigm and its focus on the confines of the individual word as the primary unit of meaning in any given language, Sinclair (2004: 34) would ultimately come to argue that semantic prosody

was a more likely candidate to constitute "the boundary of the lexical item". The markers of semantic prosody are typically situated within a very limited span on either side of the keyword, a span that has more in common with the KWIC concordance than with sentence structure or features of text organization. Once lexis is severed from the writing system's main word divisions, there is no obvious reason to situate the boundaries elsewhere, whether at two, three, or twenty positions from the keyword, yet the concordance view strongly nudges one first to restrict collocational relevance to a limited span, and secondly to interpret this restriction as meaningful.

CTS research has examined whether there is a connection between the idea of an extended unit of meaning and the idea of the translation unit, or "the smallest segment of the utterance whose signs are linked in such a way that they should not be translated individually" (Kenny 2011; Vinay & Darbelnet 1995: 21). Tognini-Bonelli (2001: 133) argues that the two notions are different: the unit of meaning is a mere linguistic convention, while a translation unit is strategic, and "the result of explicit balancing decisions", including broad contextual considerations. Linguistic conventions, however, are equally context-dependent formalizations of speakers' strategies, and it is highly likely that the difference posited is an effect of the medium: looking at text in a different manner may suggest that it has been produced according to different motivations, but what suggests itself as a theoretical discrepancy is in fact a technological one. Every utterance expresses concern for macro-structural features, even though the decontextualised lines of a concordance may temporarily suggest otherwise.

Apparent disregard for integral texts as communicative units relates not just to corpus analysis, but in some cases extends to corpus construction, as is evident from the use of sample corpora, which are made up of parts of originally longer texts, and which reveal a belief in the primacy of patterns over narrative or text structure as objects of study in CTS and across the broader field of the digital humanities. Sampling is a method developed to ensure representativeness, a factor that became important in relation to the claim that corpora provide a view of natural language in use. Early lexicographical work dictated that potential variety must be investigated, meaning that corpus representativeness was equated with a search for either comprehensiveness (the corpus should contain as many different texts as possible)

or balance (the corpus should contain a comparable number of tokens for each kind of text selected). However, neither balance nor comprehensiveness necessarily produce good intimations of linguistic and ultimately cognitive reality. Effective propaganda or advertising, for instance, typically produces limited unique textual output. In any given language, Coca-Cola slogans throughout history would barely fill a page, but they are read and heard many times over, with significant psychological and economic consequences (see also Baker *et al*. 2008: 283). Despite many corpora today focusing on matters of ideology and influence, textual balance still overrides repetition and textual impact in the selection of corpus materials. Thus, despite aiming to represent language in use, no corpora take actual use into account as a construction principle.

Research into the connection between speech and cognition, such as Hoey's (2007, 2011) work on lexical priming, attaches great importance to repetition. Words are encountered in a certain context, come to be associated with it, and thus are produced again when a similar context is encountered. The field of corpus-assisted discourse studies (e.g. Baker 2006) presents repetition as a means of persuasion and a hallmark of ideology. Lexical priming and discourse analysis have both found ample application in CTS, and generally, identifying repeated collocations or other lexical patterns can be seen as the central objective of corpus research. However, while repetition is sought after in corpus analysis, variation is central to corpus construction. If texts were represented multiple times within a corpus, the concordance would not be helpful, as it would return mantras of identical lines, meaningless without the immediate provision of adequate context outside of the medium. When Luhn introduced the electronic KWIC concordance, he devised it as a tool to query indexes of technical literature in order to find material relevant to one's study. In this context, duplicating entries would not have made sense. The expansion of the use of the concordance in CTS, as in other disciplines, to include matters of social and political impact means that the absence of duplicate texts is now a mere convention, partly sustained by the affordances of the tool.

2.2 Description and Representation

It is not only the theoretical priorities of linguistic analysis that are shaped by the mediality of the CTS research environment. The processes of preparing, describing and representing corpus texts are also influenced to a large extent by the affordances of the digital medium, beginning, for instance, with the use of a document type definition (DTD). The DTD is a common means of determining how documents stored in a corpus database are to be interpreted by an extensible markup language (XML) application such as a concordance browser (Luz 2011: 133; Zanettin 2011: 112). The DTD explicates the guidelines for producing a valid XML document for a given document type. Typically, a DTD will resemble a minimal grammar: it consists of a skeletal set of elements complemented with a list of potential attributes. XML is used for markup, meaning that its tags are normally not displayed in the concordance output. Nonetheless, all concordance lines returned in a corpus search will have been matched to a defined category, and potentially have undergone structural alterations for this purpose. These are made in addition to the inevitable changes in material texture, font, size, imaging, colour, location and so on that already affect each element of a publication prepared for inclusion in a corpus. Consequently, a concordance line is always a back-translation. The text it contains has first been translated into a form that conforms to a markup syntax that the software can interpret, and then is subsequently returned to the concordance user as a representation of the original text. This representation is declared equivalent to its source but at least implicitly operates along different linguistic as well as material constraints.

The conventional separation between text and paratext may serve as an example. Typically, the bottom of a page is reserved, where relevant, for footnotes in small font. Footnotes do not belong to the main argument and are therefore placed outside the main verbal sequence. As per publishing conventions, footnotes may be provided either by authors or by additional contributors such as translators or editors. When constructing a corpus, it might be advisable to indicate what constitutes the main text of a document, and what constitutes paratext, so the DTD will need to specify a

syntax fit to represent this distinction. However, are footnotes essentially different from endnotes, and should this distinction be encoded? Are they like introductions, meaning that a 'paratext' element can cover both? What about marginalia, written, like a footnote, by a supplemental contributor and apart from the main chain of information, but lacking a footnote's expected markers, such as the use of ordinal numbering?

The latter issue may be partly responsible for the relatively small number of diachronic studies in CTS (Malamatidou 2018: 51), although sophisticated examples are available (e.g. Gabrielatos *et al*. 2012). Languages change, but so do the documents and discursive conventions through which they do so. This makes the construction of diachronic corpora particularly challenging. How much information should be provided to adequately characterise the context of an expression, and how can there be consistency when the potential correspondences between different historical situations are limited? This problem is naturally present in CTS, as it mediates between different cultures, yet in studies covering a broad time span the issue recurs on multiple fronts. Can the same set of metadata be applied to a vellum scroll and to a born-digital blogpost? If so, which characteristics should take precedence? A document type is a conventional constraint in place to categorize relevant information that would otherwise be lost in the structural limitations imposed by adaptation to a KWIC concordance interface. When using a DTD for this purpose, there can be many specifications, but little nuance.

Multimodal corpora (e.g. Baldry & Thibault 2008; Jiménez Hurtado & Soler Gallego 2013) operate partly in response to the radical recontextualization presented by textual concordances, and they can include features such as the layout of manuscripts or the situational environment of spoken utterances. Multimodal corpora can be seen as extensions of the textual paradigm, but they also remind us that a corpus does not have to consist of text, as other communicative modalities are available. And even if a corpus consists of text, textual representation may not be the most efficient or productive way to study its contents. Indeed, a concordance interface is not a mandatory mode of access to a corpus. Most browsers come with facilities such as frequency list tools which can provide information about a corpus without ever having to produce a KWIC view. Such basic mathematical

operations are widely used, as are more complex statistical procedures that can provide multifaceted information about the constitution of a given set of translated texts (e.g. Oakes & Meng 2012). Often, statistical manipulation produces information that can be visualised in tables, graphs or charts, without any need for a concordance.

This does not mean, however, that the influence of the concordance constraint is no longer present. A great number of statistical operations applied to corpora today are variants on collocational measures such as z-score, t-score, and mutual information (Cantos, Pascual & Sánchez 2001: 202). A collocate, or co-occurring lexical element, is statistically significant when it accompanies another lexical element more often than can be expected given a degree of randomness assumed in linguistic exchange. Typically, collocation is calculated "within a specified linear distance or span" (Cantos, Pascual & Sánchez 2001: 202; Sinclair 2004). As discussed in the previous subsection, the pre-machinic Firthian notion of collocation was not necessarily restricted to a specified linear distance, nor to the lexical item. These constraints were imposed by the view presented in a KWIC concordance, and thus even when this does not form part of one's research methodology, the medium continues to exert its influence.

The landscape of the discipline is rapidly changing, and the medium around which its research practices have for a long time converged is no longer a given. Precisely at this point it is necessary, as has been attempted in this section, to indicate how an inherited medium has partly shaped practices and principles in CTS. In the second half of this paper, we will discuss mediality in corpus-based translation studies with specific reference to the construction and analysis of a series of corpora built as part of the Genealogies of Knowledge (GoK) project (genealogiesofknowledge.net/about). We will illustrate the negotiation of the issues addressed in Section 2 in conjunction with the development and use of a set of visual tools that accompany a dedicated concordance interface. We begin in Section 3 with a brief overview of this project's aims and resources, provided to set the discussion that follows in Section 4 in its proper context.

3. Genealogies of Knowledge: Aims and Resources

Genealogies of Knowledge was an interdisciplinary research project led by the Centre for Translation and Intercultural Studies at the University of Manchester and funded by the UK Arts and Humanities Research Council from April 2016 to the end of March 2020. Going forward, the Genealogies of Knowledge team continues to develop and expand its activities through a dedicated Research Network (genealogiesofknowledge.net/research-network/). The core objective of this endeavour has been to explore the role of translation and other forms of mediation in negotiating the meanings of key political and scientific concepts as they have travelled across time and space (Baker & Jones, *forthcoming*). The team is interested, for example, in how translators, commentators and other cultural mediators – including editors, historians, philosophers, citizen journalists and bloggers – have contributed to the ongoing evolution and contestation of concepts such as democracy, citizenship, truth, proof and fact when interpreting and adapting their sources for new audiences (Baker 2020; Jones 2019, 2020; Karimullah 2020). To this end, five non-parallel but closely interconnected corpora have been built, of which the largest is the Modern English corpus. This contains over 350 translations, commentaries and original writings by authors as diverse as Aristotle, Cicero, Rousseau, Marx, Wittgenstein, Foucault and Balibar, and totals in excess of 21 million tokens. The other corpora include an ancient Greek corpus (3.3 million tokens), a Latin corpus (1.5 million tokens), a medieval Arabic corpus (3.3 million tokens) and an Internet English corpus (5.6 million tokens). These all comprise similarly diverse collections of texts, written and/or translated at other moments in time over the past 2,500 years, under very different social, cultural, political and ideological conditions, and with the aim of fulfilling a diversity of philosophical, scientific and political purposes.

Of note here is the fact that the corpora, given the lengthy timespan covered, contain material originally drawn from a variety of media. In principle this variety is greater than in practice. The text of ancient manuscripts, for instance, was not collected for the corpus in its first documented form, but mostly through copies digitized from relatively recent prints, comparable in most respects to the monographs and edited volumes in our Modern

English corpus. The Internet English corpus texts, on the other hand, have been extracted from the highly dynamic, hyperlinked habitat where they, for the most part, first appeared. This difference in textual transmission history has implications for the number of transformations the material underwent before its inclusion in the corpus, and also influences the process of corpus compilation, as principles and practices of access may differ highly between online and offline publishing cultures. The internet provides a media environment increasingly dominated by a rejection of existing copyright laws through models such as creative commons and copyleft licensing, which attempt to assert the "fundamental human right to access our shared knowledge" (Nesson 2012: ix). Such evolutions bear witness not just to a cybercultural ideal of solidarity, but also to a logistic reality requiring a new property logic: a communicative environment constructed around hyperlinks and subject to a clipboard with instant copying capacity cannot incorporate effective controls against copyright infringement. The memetic internet economy is one of sharing and repurposing content. Access issues thus proved much less problematic when designing and building the Internet English corpus: while the research team did request permission from the site administrators of the 36 online media outlets currently represented, this was in most cases freely granted, in marked contrast with the response of the majority of copyright holders contacted during the construction of our print-based corpora.

The GoK resources are made available to the wider research community by means of a suite of open-source corpus analysis tools, which can be downloaded either from the project website (genealogiesofknowledge.net/software/) or via SourceForge (https://sourceforge.net/projects/modnlp/). This software package includes familiar interfaces such as a KWIC concordancer alongside a collection of more experimental data visualisation 'plugins', some of which have been designed specially with the aims and interests of the Genealogies of Knowledge project in mind. The features of these tools and their material implications for research in this field will be discussed in the following section.

4. Genealogies of Knowledge: Medial Environment

In this section, the medial environment constructed by the Genealogies of Knowledge software is illustrated with reference to two English versions of the *Communist Manifesto*. As Marx and Engels wanted the Manifesto's call to arms to be disseminated rapidly, and on a global scale, the *Communist Manifesto* is known for its 'obsession with its own translations', which are continually called for in successive prefaces (Puchner 2006: 3). The text has consequently been translated and retranslated many times, and in the Modern English corpus we have included both the first English translation (by Helen MacFarlane in 1850) and the most widely distributed one (by Samuel Moore in 1888), both of which are now freely available in the public domain. While MacFarlane's translation was first published in *The Red Republican*, it is the reprint produced in *Woodhull and Claflin's Weekly* in 1871 from which the text in the GoK corpus derives. Samuel Moore's 1888 version was approved by Friedrich Engels himself and remains the canonical version to this day. The copy of this text in the corpus derives from a collected volume of Marx's writings, published by Hackett (Simon 1994).

In Moore's version, the first sentence of the Manifesto reads: "A spectre is haunting Europe – the spectre of Communism" (1888/1994: 158). The phrase is immediately recognizable, and remains creatively productive – apart from the phrase being repeated as is, the word *Communism* has often been replaced in blog posts, newspaper features, academic articles and internet memes with a range of topical phenomena: from "the spectre of authoritarian capitalism" (Macfarlane 2020) to the "the spectre of the Unionised Jazz Musician" (Weidler 2013). Passed on through Derrida's *Specters of Marx* (1993), the phrase also helped lay the foundations for the interdisciplinary study of 'hauntology', a term applied in turn to numerous artistic efforts, particularly in the domain of music (Sexton 2012). Despite this lasting cultural prominence, the spectre never recurs in the Manifesto after the first page. Searching for the term *spectre* in both MacFarlane's and Moore's versions using the Genealogies of Knowledge concordance browser therefore returns just four lines (Figure 1).

Examining this concordance, here sorted by the R2 collocate, reveals that all cases of *spectre* derive from one version of the work, as can be understood

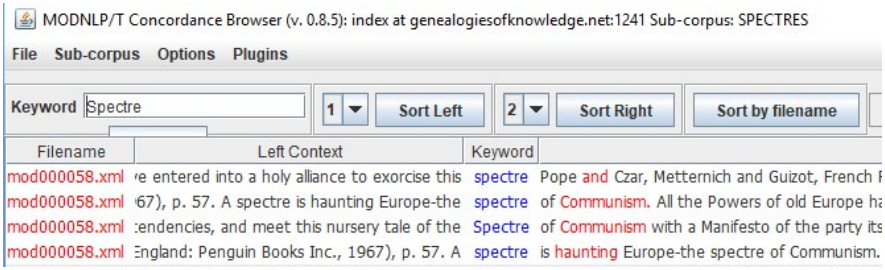

Figure 1: Concordance of **spectre** in MacFarlane's and Moore's translations of the *Communist Manifesto*, sorted by the R2 collocate.

from the identical file numbers on the left of Figure 1. Clicking on the interface's Metadata button reveals that these lines are all drawn from Moore's translation, and that the word *spectre* does not occur in MacFarlane's text. MacFarlane's version commences the treatise as follows: "A frightful hobgoblin stalks throughout Europe. We are haunted by a ghost, the ghost of Communism." A hobgoblin is a folkloristic, mischievous spirit. Moore's spectre simply haunts, but Macfarlane's goblin, or 'bugbear', as she later calls it, stalks. The impersonal Europe is replaced with a personal 'we' as the recipient of the ghost's attention, stressing the ambiguity of the spectral presence in the Manifesto: particularly in MacFarlane's text, it seems to haunt both the communists and their enemies. The possibility of this double reading stems from the fact that both parties intend to expel the ghost, but while their opponents seek to exorcise it, the communists seek to incarnate the spectre.

Such inferences are more difficult to establish in a concordance browser than by simply reading the texts themselves. The rendering of the German *Gespenst* by the English triumvirate of *hobgoblin*, *ghost*, and *bugbear*, or as a single *spectre*, immediately catches the eye in a linear reading, as do the activities pursued by these figures, but the nature of a concordance makes the sequence challenging to process. This is not just because a concordance breaks the narrative of the text, but also because the combination of a set co-textual span (in this case, 130 characters) and a keyword-centred view may produce multiple representations of single tokens, thus skewing the relation between text and data. In the concordance above, for instance, one can see that because the distance between separate occurrences of *spectre* is

shorter than the span, there is a duplication of spectres in the KWIC display. Only four are in the text, while one can spot six in the concordance. Through fragmentation and reordering, the browser conjures up supplementary spectres and amplifies the lexical patterns present in the text (Buts 2019: 93-108). Consequently, as discussed in section 2, the KWIC concordance is designed to spot repetition but disorients when repetition is ubiquitous throughout a corpus or locally concentrated in a specific section of a single text.

Other threads involving more dispersed and numerically significant patterns of repetition can be investigated in a manner more suited to the corpus software. For instance, the frequency list for both translations combined shows that the terms *bourgeois* and *bourgeoisie* are very common in this corpus. At 153 and 141 occurrences in a corpus of only 25,000 words, they take up the nineteenth and twenty-first positions in the list, and both items together occur more than frequent stop words such as *this* and *that*. However, the GoK Metafacet visualisation tool indicates that Moore (100 hits) uses *bourgeois* more than twice as much as MacFarlane (41 hits) (Figure 2).

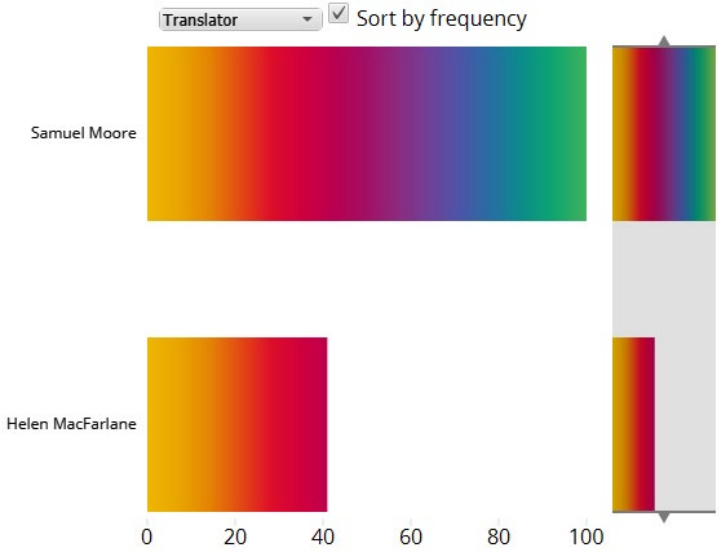

Figure 2: Metafacet visualisation comparing the frequency of **bourgeois** in MacFarlane's and Moore's translations

Metafacet allows for a comparison between the texts included in a corpus selection on the basis of any characteristic recorded in the metadata, such as year of publication, translator or outlet. Metafacet interacts directly with the concordance, and upon request removes lines associated with, for instance, one translator's output from view. It thus allows the user to alternate rapidly and dynamically between concordances generated from the two different translations. Although the Metafacet tool stands in direct connection with the concordance, it is a purely numerical, frequency-based application providing a count of concordance lines without relying upon or indicating what is presented within these lines. This distinguishes the tool from the Mosaic, a different visualisation tool meant to represent lexical items and their co-text occurring within the concordance. The Mosaic display is informed by principles of visualisation theory, as well as by the didactic work of Sinclair, who used similar diagrams in his example analyses (Luz & Sheehan 2014, 2020; Sinclair 2003). The user can request a Mosaic based on frequency, or on several variants of popular collocation measures such as mutual information and z-score. Restricting the search to MacFarlane's translation using the Metafacet tool, and then switching to the Mosaic visualisation's Column Frequency view reveals that *bourgeois* commonly precedes the words *regime*, *society*, *freedom*, *property*, and *socialism* in MacFarlane's translation (Figure 3).

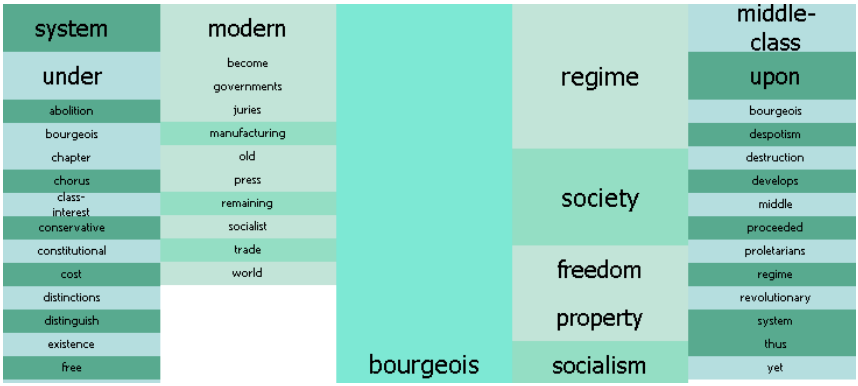

Figure 3: Column Frequency (No stopwords) view of the Mosaic tool displaying the most frequent L2-R2 collocates of **bourgeois** in MacFarlane's translation

Mosaic provides an alternative means of engaging with a keyword in context by grouping the lexical items that occur in each word position to the left and right of the search term and allocating them a differently shaded rectangle in columns placed immediately adjacent to this keyword. The bigger the word tile in the Column Frequency view, the more frequent the collocation. Figure 3 shows that many collocates listed in the R1 column are nouns. The concordance lines confirm that this pattern derives from MacFarlane using the term *bourgeois* almost exclusively as an adjective. By contrast, Moore, while generally conforming to this pattern, also uses the word nominally, as in 'the individual bourgeois'. One could perhaps further infer from Figure 3 that the term *bourgeois* is intimately associated with *socialism* and *freedom* for the writers of the Manifesto. However, references to 'bourgeois socialism' in the Manifesto are imbued with a strongly negative aura of meaning, given that in Marx's view this form of socialism seeks to redress social grievances only "in order to secure the continued existence of bourgeois society" (Marx & Engels 1994: 181).

The text as a whole presents an antagonistic dichotomy between proletariat and bourgeoisie, and vocabulary is part of the battlefield. Both parties may use terms such as *socialism*, but supposedly one class produces false representations, while the other holds onto true aspirations. Closer scrutiny of the co-text is necessary here to adequately interpret the visual representations, meaning that, at the very least, use of the Mosaic tool must be integrated within a workflow that additionally exploits the affordances of the concordance display. Nevertheless, the Mosaic counts, orders and highlights with remarkable efficiency, and its ability to shift between several measures of collocational importance can strongly impact the user's interaction with the data. It is important to note here that some functionalities of the GoK software package such as Frequency List are based upon the whole corpus under investigation, while tools such as Metafacet and Mosaic only take into account data present in the concordances retrieved for a specific keyword. Furthermore, whereas Metafacet gives information about the contextual provenance of the lines returned in a search, the Mosaic presents an alternative visualisation of their textual constitution. In short, statistics and their visualisations can function independently, but often depend on the generation of KWIC concordances, or at least on the selection of a keyword.

It should further be noted that the *keyword* in *keyword in context* is itself a flexible designation. In the GoK browser environment, as in many corpus tools, a variety of orthographic sequences corresponding to a search pattern can be retrieved by means of a sequencing grammar, as well as more complex regular expressions. When corpora are queried through such metalanguages, one makes use of a means of interpretation that is heavily dependent upon the affordances of the medium. In what way the metalanguages facilitating digital research correspond to the data being processed and queried, and ultimately to the text studied, is a matter requiring continued attention. Similarly, in what way the output of visualisation tools designed for corpus analysis, such as KWIC, Metafacet, and Mosaic discussed above, can be considered mutually equivalent representations, and to what extent they can function independently, is a question CTS must dare to ask.

5. Conclusion

Translation studies is accustomed to the importance of presentation, style, and rhetoric, and does not question the idea that different translatorial strategies lead to different translation solutions, and thus to different interpretations. In this respect, corpus-based studies of translation have facilitated the meticulous investigation of small shifts that aggregate into paradigm changes. Nevertheless, despite this sensitivity to matters of transmission and transformation, CTS has not brought into focus its own medial qualities. The transformation of text into data allows for many competing approaches and representations, none of which should individually be accepted uncritically as equal and accurate renderings of the object of study. The representation of text, for instance, is not a mandatory constraint for the interpretation of text. Yet at present, given the human familiarity with this form of expression and the influence KWIC design has exerted on the development of alternatives, the triangulation of multiple software tools, often integrated with the KWIC concordance, seems to be a sensible approach for CTS to pursue. Yet, as tools become more abundant and easier to work with, one risks losing sight of the particular choices that govern their inception and implementation. Efficient research tools tend to draw attention away from themselves, and from the choices they impose on the form that an object of analysis takes.

One may be reminded here that Sinclair, speaking about the lexicon, argued that "there is no distinction between form and meaning" (Sinclair 1991: 7). The statement illustrates a functional view of language, one in which use is the determining factor: some words 'fit' in certain situations. Corpus research, a mode of enquiry that finds its origins in the declaration of equivalence between meaning and form, cannot leave the research environment out of its purview. From a similarly reflective perspective, translation studies, like translation, cannot merely repeat, copy, or reproduce information. When interpretation takes place in a customized medial environment, the affordances of this environment should be interrogated. This article has attempted to promote wider awareness of this issue in CTS, and to illustrate the influence of the medium on corpus analysis with reference to the representation of the *Communist Manifesto* in the Genealogies of Knowledge Modern English corpus. We have drawn particular attention to the interdependence between textual analysis, theoretical development, and corpus construction, for instance with reference to the convention that a corpus should contain ample textual variation, rather than consist of repeated utterances. The short case study analysis illustrated that one of the causes for the avoidance of duplication in corpus construction may be that the concordance view is well suited to call attention to dispersed repetition, but less so to represent concentrated repetition. Such examples go beyond the established critique that corpus analysis tends to disregard the integrity of the text as a communicative unit. The fruitful alliance between corpus and discourse studies has paid ample attention to the balance sought between closer and more distant forms of reading, and has indeed recently turned a critical eye to methodological choices and their relation to the tools used for research (Taylor & Marchi 2018). This article has argued that CTS should be at the forefront of this ongoing critical engagement, as a translational perspective may aid in explicating the various transformations that turn text into data, and that make data suitable for interpretation.

Examining this dusty corner of the discipline is all the more important as the gap between analysis and presentation widens. When concordance evidence was the incontestable 'stock-in-trade' of the corpus analyst, research articles could reproduce a large part of the investigative flow, thus ensuring transparency and replicability. Today, the use of very variable corpora,

statistical operations and visualisation tools is rapidly multiplying, and it is often a struggle to represent methodological pathways in the classic format of research papers. In effect, this procedure often requires back-translation: textual information processed to facilitate corpus research is once again adapted to representation on the page, be it printed or digital. Finally, then, we suggest that further research into the question of how different media shape our interaction with textual data must also begin to consider the demands of publishing cultures and the extent to which existing models for publication may need to evolve in order to meet the requirements of the expanding digital humanities.

References

ANTHONY, Laurence. (2018) "Visualisation in Corpus-based Discourse Studies". In: Taylor, Charlotte & Anna Marchi (eds.) *Corpus Approaches to Discourse: A critical review*, London: Routledge, pp. 1-15.

ARMSTRONG, Guyda. (2020) "Media and Mediality." In: Baker, Mona & Gabriella Saldanha (eds.) 2020. *Routledge Encyclopedia of Translation Studies*. London: Routledge, pp. 310-315.

BAKER, Mona. (1993) "Corpus Linguistics and Translation Studies: Implications and applications." In: Baker, Mona, Gill Francis & Elena Tognini-Bonelli (eds.) 1993. *Text and Technology: In honour of John Sinclair*. Philadelphia & Amsterdam: John Benjamins, pp. 233-250.

BAKER, Mona. (2020) "Rehumanizing the Migrant: The translated past as a resource for refashioning the contemporary discourse of the (radical) left." *Palgrave Communications* 6:1.

BAKER, Mona & Henry Jones. (*forthcoming*) "Genealogies of Knowledge: Theoretical and methodological issues." *Palgrave Communications*.

BAKER, Paul. (2006) *Using Corpora in Discourse Analysis*. London: Continuum.

BAKER, Paul, Costas Gabrielatos, Majid KhosraviNik, Michal Krzyzanowski, Tony McEnery & Ruth Wodak. (2008) "A Useful Methodological Synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK Press." *Discourse & Society* 19:3, pp. 273-306.

Baldry, Anthony & Paul J. Thibault. (2008) "Applications of Multimodal Concordances." *Hermes – Journal of Language and Communication Studies* 41, pp. 11-41.

Berry, David. (2011) "The Computational Turn: Thinking about the digital humanities." *Culture Machine* 12, pp. 1-22.

Buts, Jan. (2019) *Political Concepts and Prefiguration: A corpus-assisted enquiry into democracy, politics and community*. Manchester: University of Manchester. Unpublished PhD Thesis.

Cantos, Pascual & Aquilino Sánchez. (2001) "Lexical Constellations: What collocates fail to tell." *International Journal of Corpus Linguistics* 6:2, pp. 199-228.

Cronin, Michael. (2013) *Translation in the Digital Age*. London & New York: Routledge.

Derrida, Jacques. (1994) *Specters of Marx: The state of the debt, the work of mourning and the new international*. Peggy Kamuf (trans.). New York & London: Routledge.

Dixon, Dan. (2012) "Analysis Tool or Research Methodology: Is there an epistemology for patterns?" In: Berry, M. David (ed.) 2012. *Understanding Digital Humanities*. Basingstoke: Palgrave Macmillan, pp. 191-209.

Fenlon, John F. (1908) "Concordances of the Bible." In: Herbermann, Charles G., Edward A. Pace, Condé B. Pallen, Thomas J. Shahan & John J. Wynne (eds.) 1908. *The Catholic Encyclopedia, Volume 4*. New York: Robert Appleton Company, pp. 195-196.

Firth, John R. (1968) "Linguistic Analysis as a Study of Meaning." In: Palmer, Frank R. (ed.) 1968. *Selected Papers of J. R. Firth*. London & Harlow: Longman, 12-26.

Gabrielatos, Costas, Tony McEnery, Peter J. Diggle & Paul Baker. (2012) "The Peaks and Troughs of Corpus-based Contextual Analysis." *International Journal of Corpus Linguistics* 17:2, pp. 151-175.

Hermans, Theo. (2007) *The Conference of the Tongues*. Manchester: St. Jerome Publishing.

Hoey, Michael. (2007) "Lexical Priming and Literary Creativity." In: Hoey, Michael, Michaela Mahlberg, Michael Stubbs & Wolfgang Teubert (eds.) 2007. *Text, Discourse and Corpora: Theory and analysis*. London & New York: Continuum, pp. 31-56.

HOEY, Michael. (2011) "Lexical Priming and Translation." In: Kruger, Alet, Kim Wallmach & Jeremy Munday (eds.) 2011. *Corpus-Based Translation Studies: Research and Applications*. London & New York: Bloomsbury, pp. 153-168.

HUNSTON, Susan. (2007) "Semantic Prosody Revisited." *International Journal of Corpus Linguistics* 12:2, pp. 249-286.

JIMÉNEZ Hurtado, Catalina & Silvia Soler Gallego. (2013) "Multimodality, Translation and Accessibility: A corpus-based study of audio description." *Perspectives* 21:4, pp. 577-594.

JONES, Henry. (2018) "Mediality and Audiovisual Translation." In: Pérez-González, Luis (ed.) 2018. *Routledge Handbook of Audiovisual Translation*. London & New York: Routledge, pp. 177-191.

JONES, Henry. (2019) "Searching for Statesmanship: A corpus-based analysis of a translated political discourse." *Polis: The Journal for Ancient Greek and Roman Political Thought* 36, pp. 216-241.

JONES, Henry. (2020) "Retranslating Thucydides as a Scientific Historian." *Target* 32:1, pp. 59-82.

KARIMULLAH, Kamran. (2020) "Editions, Translations, Transformations: Refashioning the Arabic Aristotle in Egypt and metropolitan Europe, 1940–1980." *Palgrave Communications* 6:3.

KENNY, Dorothy. (2011) "Translation Units and Corpora." In: Kruger, Alet, Kim Wallmach and Jeremy Munday (eds.) 2011. *Corpus-Based Translation Studies: Research and Applications*. London & New York: Bloomsbury, pp. 76-102.

KLEIN, Hans K. & Daniel Lee Kleinman. (2002) "The Social Construction of Technology: Structural considerations." *Science, Technology & Human Values* 27:1, pp. 28-52.

KRESS, Gunther. (2003) *Literacy in the New Media Age*. London & New York: Routledge.

LITTAU, Karin. (2006) *Theories of Reading: Books, bodies, and bibliomania*. Cambridge: Polity Press.

LITTAU, Karin. (2011) "First Steps Towards a Media History of Translation." *Translation Studies* 4:3, pp. 261-281.

LITTAU, Karin. (2016) "Translation and the Materialities of Communication." *Translation Studies* 9:1, pp. 82-96.

LOUW, Bill. (1993) "Irony in the Text or Insincerity in the Writer: The diagnostic potential of semantic prosody." In: Baker, Mona, Gill Francis & Elena

Tognini-Bonelli (eds.) 1993. *Text and Technology: In honour of John Sinclair*. Philadelphia & Amsterdam: John Benjamins, pp. 157-176.

LUHN, Hans Peter. (1960) "Key Word-in-Context Index for Technical Literature (Kwic Index)." *American Documentation* 11:4, pp. 288-295.

LUZ, Saturnino. (2011) "Web-Based Corpus Software." In: Kruger, Alet, Kim Wallmach & Jeremy Munday (eds.) 2011. *Corpus-Based Translation Studies: Research and Applications*. London & New York: Bloomsbury, pp. 124-149.

LUZ, Saturnino & Shane Sheehan. (2014) "A Graph Based Abstraction of Textual Concordances and Two Renderings for their Interactive Visualisation." In: 2014. *Proceedings of the International Working Conference on Advanced Visual Interfaces*. New York: ACM, pp. 293-296.

LUZ, Saturnino & Shane Sheehan. (2020) "Methods and Visualization Tools for the Analysis of Medical, Political and Scientific Concepts in Genealogies of Knowledge." *Palgrave Communications* 6:49, pp. 1-20.

MACFARLANE, Laurie. (2020) "A Spectre is Haunting the West – The spectre of authoritarian capitalism." *Open Democracy*. <https://www.opendemocracy.net/en/oureconomy/a-spectre-is-haunting-the-west-the-spectre-of-authoritarian-capitalism/>

MALAMATIDOU, Sofia. (2018) *Corpus Triangulation: Combining data and methods in corpus-based translation studies*. London & New York: Routledge.

MANOVICH, Lev. (2001) *The Language of New Media*. Cambridge, MA & London: MIT Press.

MARX, Karl and Friedrich Engels. (1850) "German Communism – Manifesto of The German Communist Party." Translation by Helen MacFarlane. *The Red Republican* 21:1.

MARX, Karl and Friedrich Engels. (1850/1871) "German Communism - Manifesto of The German Communist Party." Translation by Helen MacFarlane. *Woodhull and Claflin's Weekly* 4:7. Online version: <http://iapsop.com/archive/materials/woodhull_and_claflins_weekly/>

MARX, Karl & Friedrich Engels. (1888/1994) "The Communist Manifesto." Translation by Samuel Moore. In: Simon, Lawrence H. (ed.) 1994. *Karl Marx: Selected writings*. Indianapolis: Hackett Publishing Company, pp. 157-186.

MCINTYRE, Dan & Brian Walker. (2019) *Corpus Stylistics: Theory and practice*. Edinburgh: Edinburgh University Press.

MCLUHAN, Marshall. (1964) *Understanding Media: The extensions of man*. New York: McGraw-Hill.

MUNDAY, Jeremy. (2011) "Looming Large: A cross-linguistic analysis of semantic prosodies in comparable reference corpora." In: Kruger, Alet, Kim Wallmach & Jeremy Munday (eds.) 2011. *Corpus-Based Translation Studies: Research and Applications*. London & New York: Bloomsbury, 169-186.

NESSON, Charles R. (2012) "Foreword." In: Dulong de Rosnay, Melanie & Juan Carlos de Martin (eds.) 2012. *The Digital Public Domain: Foundations for an Open Culture*. Cambridge: Open Book Publishers, pp. xi-xiii..

OAKES, Michael P. & Ji Meng. (eds.) (2012) *Quantitative Methods in Corpus-Based Translation Studies: A practical guide to descriptive translation research*. Amsterdam & Philadelphia: John Benjamins.

PARTINGTON, Alan. (1998) *Patterns and Meanings: Using corpora for English language research and teaching*. Amsterdam & Philadelphia: John Benjamins.

PÉREZ-GONZÁLEZ, Luis. (2014) "Multimodality in Translation and Interpreting Studies: Theoretical and methodological perspectives." In: Bermann, Sandra & Catherine Porter (eds.) 2014. *A Companion to Translation Studies*. Chichester: Wiley Blackwell.

PUCHNER, Martin. (2006) *Poetry of the Revolution: Marx, manifestos, and the avant-gardes*. Princeton: Princeton University Press.

SEXTON, Jamie. (2012) "Weird Britain in Exile: Ghost Box, hauntology, and alternative heritage." *Popular Music and Society* 35:4, pp. 561-584.

SHERIDAN, Mary P. (2016) "Recent Trends in Digital Humanities Scholarship." In: Dejica, Daniel, Gyde Hansen, Peter Sandrini & Iulia Para (eds.) 2016. *Language in the Digital Era: Challenges and perspectives*. Warsaw & Berlin: De Gruyter Open, pp. 2-13.

SIMON, Lawrence H. (ed.) (1994) *Karl Marx: Selected writings*. Indianapolis & Cambridge: Hackett Publishing Company.

SINCLAIR, John. (1991) *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

SINCLAIR, John. (2003) *Reading Concordances: An introduction*. London: Pearson Longman.

SINCLAIR, John. (2004) *Trust the Text: Language, corpus and discourse*. Ronald Carter (ed.) London & New York: Routledge.

SINCLAIR, John. (2008) "The Phrase, the Whole Phrase, and Nothing But the Phrase." In: Granger, Sylviane & Fanny Meunier (eds.) 2008. *Phraseology: An interdisciplinary perspective*. Amsterdam & Philadelphia: John Benjamins, pp. 407-410.

STEWART, Dominic. (2009) "Safeguarding the Lexicogrammatical Environment: Translating semantic prosody." In: Beeby, Allison, Patricia Rodríguez Inés & Pilar Sánchez-Gijón (eds.) 2009. *Corpus Use and Translating*. Amsterdam & Philadelphia: John Benjamins.

STEWART, Dominic. (2010) *Semantic Prosody: A critical evaluation*. New York & London: Routledge.

STUBBS, Michael. (1996) *Text and Corpus Analysis: Computer-assisted Studies of Language and Culture*. Cambridge: Blackwell.

STUBBS, Michael. (2009) "The Search for Units of Meaning: Sinclair on empirical semantics." *Applied Linguistics* 30:1, pp. 115-137.

SVARTVIK, Jan. (1992) "Corpus Linguistics comes of Age." In: Svartvik, Jan (ed.) 1992. *Directions in Corpus Linguistics*: *Proceedings of Nobel Symposium 82 Stockholm, 4-8 August 1991*. Berlin & New York: Mouton de Gruyter, pp. 7-13.

TAYLOR, Charlotte and Anna Marchi (eds.) 2018. *Corpus Approaches to Discourse: A critical review*. London: Routledge.

TOGNINI-BONELLI, Elena. (2001) *Corpus Linguistics at Work*. Amsterdam & Philadelphia: John Benjamins.

TYMOCZKO, Maria. (2010) "Translation, Resistance, Activism: An overview." In: Maria Tymoczko (ed.) 2010. *Translation, Resistance, Activism*. Amherst: University of Massachusetts Press, 1-22.

VENUTI, Lawrence. (1998) *The Scandals of Translation*: *Towards an Ethics of Difference*. London & New York: Routledge.

VINAY, Jean-Paul & Jean Darbelnet. (1995) *Comparative Stylistics of French and English: A methodology for translation*. Amsterdam: John Benjamins.

WHITSITT, Sam. (2005) "A Critique of the Concept of Semantic Prosody." *International Journal of Corpus Linguistics* 10:3, pp. 283-205.

WEIDLER, Arndt. (2013) "A Spectre Is Haunting Europe – The spectre of the unionised jazz musician!" Translation by Paul McCarthy. *Goethe Institute Blog*. <https://www.goethe.de/en/kul/mus/gen/jaz/jah/20454965.html>

ZANETTIN, Federico. (2011) "Hardwiring Corpus-Based Translation Studies: Corpus encoding." In: Kruger, Alet, Kim Wallmach & Jeremy Munday (eds.) 2011. *Corpus-Based Translation Studies: Research and Applications*. London & New York: Bloomsbury, pp. 103-123.

BIONOTE

JAN BUTS is a postdoctoral researcher attached to the QuantiQual project (https://adaptcentre.ie/projects/quantiqual/) at Trinity College Dublin, and a co-coordinator of the Genealogies of Knowledge Research Network (https://genealogiesofknowledge.net/research-network/). He works at the intersection of translation theory, conceptual history, corpus linguistics, and online media.

HENRY JONES is a lecturer in translation and intercultural studies at Aston University, UK. He is a co-coordinator of the Genealogies of Knowledge Research Network (https://genealogiesofknowledge.net/research-network/) and co-editor of the *Routledge Encyclopedia of Citizen Media* (2021). His current research interests include corpus-based translation studies, translation history, media theory and online translating communities.

NOTICES BIOGRAPHIQUES

JAN BUTS est chercheur postdoctoral attaché au projet QuantiQual (https://adaptcentre.ie/projects/quantiqual/) au Trinity College de Dublin, et un des coordinateurs du Genealogies of Knowledge Research Network (https://genealogiesofknowledge.net/research-network/). Il travaille à l'intersection de la théorie de la traduction, de l'histoire conceptuelle, de la linguistique des corpus et des médias en ligne.

HENRY JONES est maître de conférences à Aston University, Royaume-Uni. Il est un des coordinateurs du Genealogies of Knowledge Research Network (https://genealogiesofknowledge.net/research-network/) et un des éditeurs du *Routledge Encyclopedia of Citizen Media* (2021). Ses intérêts de recherche comprennent la traductologie de corpus, l'histoire de la traduction, la théorie des médias et les communautés virtuelles.