Chapter 28. Copyright and the Reuse of Translation as Data

Joss Moorkens & Dave Lewis

Dublin City University & Trinity College Dublin/ADAPT Centre

Abstract

Translation copyright was first codified as a derivative work in the Berne Convention of 1886, subject to the rights of the creator of the original work. While most countries are now contracting parties to the Berne Convention, differing interpretations and additional laws and directives mean that rights to ownership of a translation are not consistent in all jurisdictions. The original intention of the Berne Convention was to protect authors' rights and to prevent piracy, and the authors could not have foreseen the large scale reuse of translations, initially via translation memory tools, then as training data for machine translation (MT) systems. Parallel data is repurposed in ever-increasing amounts, but broken down to word and subword levels. At present, rights to ownership are rarely passed to the translator, meaning that, while an initial translation may be costly, secondary uses are very inexpensive. This chapter explores the history of translation copyright and leveraging, and introduces concerns relating to machine learning more generally and applies them to translation.

Keywords: translation copyright, machine learning, machine translation, language resources, translation memory, data dispossession

Introduction

Since the 1980s, translation data in the form of aligned source-and-target sentence pairs have been harvested for reuse, initially in-house within companies and institutions, then among internal and external teams with the first commercially-available translation memory (TM) tools. The increasing prevalence of TM tools and the availability of server-based memories led some scholars to question who should own these translation repositories and to question the ethics of data sharing (Topping 2000, Smith 2008).

As data-driven machine translation (MT) became the dominant paradigm, so grew the necessity to retask translation data for MT testing and training. The purposes to which MT could be applied broadened as output quality steadily increased (Way 2018). The move from statistical to neural MT (NMT; Wu *et al.* 2016, Forcada 2017) has increased data requirements further, and the associated improvements in MT fluency have concomitantly boosted the use of MT not only for assimilation (i. e. gisting) but also for dissemination where MT is what Forcada (2010: 217) calls 'an intermediate step' in the production of publishable-quality translation.

The application of neural networks to translation means that MT is now considered an application of machine learning. Kelleher *et al.* (2017: 3) define machine learning as 'an automated process that extracts patterns from data'. These patterns are extracted progressively by a series of induction algorithms rather than by following an explicit operator instruction (Kohavi and Provost 1998). The application of machine learning to translation has led to a leap forward in MT quality and claims that MT is approaching human parity (Hassan *et al.* 2018). As a result, automation has become a concern for many in the translator community (Vieira 2018), and the assignation of rights to use translation data for MT training has become a burning issue. The commonly-used metaphor of data as oil or gold suggests that training data for machine-learning systems are naturally-occurring.[i] This is, of course, untrue. Notwithstanding the possibility of webcrawling for parallel texts as a method of gathering MT training data (Toral *et al.* 2017), the most valuable and highest-quality data for training MT systems, as with other applications of machine learning based on human data, is aligned human translations, in which source and target texts are computationally linked, usually at the sentence level.

The many uses to which MT output may be applied are unlikely to have been envisaged by the translator who carried out the original translation, and were definitely not foreseen by the authors of the Berne Convention in 1886, which still forms the legal basis of copyright for translation. As a result of changing practices in digital authorship and data exchange along with the lack of harmony between regional interpretations of applicable agreements and conventions, there are increasing calls for an overhaul of international copyright law (Margoni 2016, Giblin 2018).

In this chapter we begin by providing a brief history of international translation copyright, followed by a review of translation leveraging and reuse. We discuss critical issues that are emerging at the time of writing along with some recent suggestions for change in copyright law, and include some further reading on this topic.

Literature Review and Historical Trajectory

A history of translation copyright protection

The International Association of Authors and Artists (ALAI), with members from countries in Europe and North America, successfully lobbied for an international agreement to protect copyright (by giving exclusive rights to copy, sell, and distribute) for artistic works in the late 1800s. This culminated in the signing of the Berne Convention by representatives from eight countries in 1886, a number of signatories that has grown to 176 at the time of writing (WIPO 2018). Prior to this, some

national governments (such as the UK and the US) had domestic copyright laws and could negotiate bilateral agreements with other nations to prevent piracy. Article 5(2) of the Convention decrees that works will automatically be protected in signatory countries without any further formalities once the criteria for protection are fulfilled. The general view of many countries in continental Europe was that copyright should remain with the original author of a work (the 'naturalist' approach), whereas the UK and USA were more concerned with the owner of copyright, 'whether that be author, publisher, broadcaster, individual, or corporation' (the 'instrumentalist' approach; Burger 1988: 7). The Convention brought about a degree of harmonization, but interpretation remained divergent based on these pre-existing views.

The Berne Convention was the first time that international copyright was codified for translations explicitly, considering them to be derivative works that 'shall be protected as original works without prejudice to the copyright in the original work' (Article 2, WIPO 2018). The convention refers specifically to translations of literary or artistic work and to official translations of official texts of a legislative, administrative and legal nature, and grants the author of an original work the exclusive right to authorise a translation. This right was replicated in the Universal Copyright Convention of 1952 and has in turn been added to many national laws. Venuti, among others, argues that, in its treatment of derivative works, copyright law based on the Berne Convention 'contradicts its key principle: that authorship consists of original expression, and hence that legal protection is given only to forms, not ideas' (1995: 4). The concept of 'original work' was not defined in the Berne Convention, and interpretations of originality still differ depending on the level of 'the author's own intellectual creation' in some European Union (EU) countries, or if there is a 'minimum degree of creativity' in the USA, based on Supreme Court precedent (Cabanellas 2015: 20). Margoni (2016) notes the focus on 'author's own' (i.e. not copied) in the UK and Ireland, as opposed to the 'intellectual creation' focus in continental Europe. The evaluation of originality applies to the original and to the translation. Troussel and Debussche (2014) argue that the protection of the translation as an original work will depend on the level of originality of the translated work based on the opinion of the WIPO/Unesco Committee of Governmental Experts (1988). They suggest a creative or complex source text will require a more creative translation, and that the translator may take a creative rather than literal approach to their task, increasing copyright eligibility. From a Translation Studies perspective, some scholars would consider that all translation is inherently creative (O'Sullivan 2013). Troussel and Debussche (2014: 100) admit that originality may be found in creative translation of uncreative work, but assert that 'the probability that such work is eligible for copyright protection is most probably remote'. They further note that 'the work's merit or aesthetic do not matter when considering the question of copyright protection' (*ibid*. 101).

The concept of creativity is difficult to define. Franken (1993: 396) defined it as 'the tendency to generate or recognize ideas, alternatives, or possibilities that may be useful in solving problems, communicating with others, and entertaining ourselves and others'. Many definitions tautologically link creativity to originality and vice versa, such as Sternberg (2011: 479), who defines creativity as 'something original and worthwhile'. The legal bar for creativity tends to be rather low. For example, based on legal precedent in the US, Abrams (1992: 17) writes that to be considered creative, an 'arrangement and presentation of facts need not be either innovative or surprising', and 'need not be very creative after all, although it must be more than independent effort and expense in doing what is obvious'.

There is another set of rights applicable to a translation, in that it may not be 'reproduced or further exploited' without authorization from the translation copyright owner (Cabanellas 2015: 63). This means that the original author may not use the translation as the basis for a further translation[ii] without permission, and that the translation copyright owner retains rights even over translation of a text that is out of copyright.

Article 6 of the Berne Convention also introduces the concept of moral rights (for original works from 1928 and for translations from 1948) that are independent of economic or exploitation rights and are linked to the personality and reputation of an author (Venuti 1995). These rights allow a person to claim authorship and to object to distortion or modification of a text that may harm their honour or reputation, with means of redress to be defined at national level. Troussel and Debussche (2014) note that moral rights are not enforced equally in all EU jurisdictions, and that they are not included in the more recent EU Infosoc Directive. The integrity of a text is protected particularly in France and Belgium, where moral rights may not be reassigned, whereas in the UK, while moral rights are protected, they are little-used in copyright cases (*ibid*. 63). The 1971 Paris Act of the Berne Convention included an Appendix particular to developing countries, allowing them to issue licences for translation of otherwise copyrighted work if a translation has not been authorized for the locale within ten years of the date of publication of the original. Silva (2012: 3) argues that this provision is applied inconsistently and does not 'address the needs of linguistic and cultural minorities in both developed and developing countries'. In the US, the fair use doctrine provides a further limitation on copyright for derivative works based on the purpose of the use, the nature of the copyrighted work, the size of the portion used, and the effect on the audience (Cabanellas 2015: 40). This exemption is ill-defined, but has been used successfully for parody, criticism, commentary, and research. Other countries such as Poland, Israel, and South Korea have adopted similar laws.

The World Copyright Treaty in 1952 brought no further clarification on the nature of originality in a text eligible for copyright beyond that in the Berne Convention, but Troussel and Debussche (2014) identify several subsequent EU Directives that may provide guidance. These include the Software Directive and Database Directive, providing protection for a computer program or database when it is the 'author's own intellectual creation', and the Term Directive, that adds that a photograph to be protected must be the creator's 'intellectual creation reflecting his personality' (*ibid.* 32). Decisions of the European Court of Justice have expanded the application of these directives beyond software, databases, and photographs (Margoni 2016). Troussel and Debussche (2014: 20) suggest that several other relevant directives are due for revision, which 'should be of particular interest for the translation industry'.

The EU published the Directive on Copyright in the Digital Single Market in 2016 in an effort to harmonise EU copyright law and to limit the benefit to internet platforms of shared content with no benefit to the original content creator. The Directive was approved in principle in 2018 (European Commission 2018), although at the time of writing it has not yet been enacted.

Translation leverage and reuse

The idea of reuse or leveraging of previous translations was not universally welcomed when translation memory (TM) tools were introduced in the early 1990s. Initial impetus came from early adopters within the translation community who found productivity gains (García 2006, 2009), although within a few years the use of TM tools (and the acceptance of associated discounts)[iii] was expected of freelance translators when in receipt of a job from a language service provider (LSP). As the technology matured, tools became more stable and could work with large TMs and glossaries, which could be shared between members of a team or for client-specific work, possibly via an interchange format (a common file format that allows import and export from different software tools) (see Chapter X on Standards in this volume). Some translators who found themselves working with TMs that they had not themselves created complained that translation quality suffered as a result (García 2009).

At this juncture, ownership and sharing of translation data became a point of discussion among translators and translation buyers. For example, Topping (2000) discussed whether TMs should belong to the end client, the LSP, or the translator (as owner of the TM tool); whether data sharing was ethical, and whether bespoke TMs created for one purpose would be useful for another. Topping made it clear that she is negatively disposed to TM sharing outside of a standard agreed framework or process and reported a small amount of online TM sharing via an exchange server.

Gow (2007) addresses the competing layers of copyright applicable to TMs as databases of segments for which they can at best claim a portion of copyright, considering reuse only within a TM workflow. She suggests that there is little to gain by asserting copyright over the compiled contents of a TM database. Smith (2008: 23) notes that, despite unclear copyright ownership, TM files are usually delivered to clients, and were they to be withheld by a translator their 'standing with that service provider may well suffer and payment problems could ensue'. His advice is to have a prior written agreement between translator and client. This may be part of an employment contract or a project-specific agreement to transfer or retain rights. An example of this is in the standard terms and conditions of the Netherlands Association of Interpreters and Translators for translation work, which states that unless it is 'expressly stated otherwise in writing, the translator reserves the copyright on translations and other texts produced by the translator' (NGTV 2017).

The international Agreement on Trade-Related Aspects of Intellectual Property Rights (TRIPS) between members of the World Trade Organization, effective from 1995, stipulated the software and compilations of data should be protected as literary works in the Berne Convention if they can be classed at intellectual creations independent of the data contained therein (World Trade Organization 1994). Gow (2007) is of the opinion that a TM file would not qualify as a sufficiently original creation in the US or Canada on this basis. The European Database Directive from 1996 allows copyright and/or *sui generis* rights to apply to an individually- or jointly-created database, interpreted more broadly than in the TRIPS agreement and thus likely to encompass TM files (Troussel and Debussche 2014: 118). In some EU countries copyright over a database created as part of employment will automatically transfer to the employer.

According to the Database Directive, *Sui Generis* rights may be asserted where there has been 'a substantial investment in either the obtaining, verification or presentation of the contents' (European Parliament 1996: Article 7), continuing for 15 years from the final edit of the database. Troussel and Debussche (2014) advise that this substantial investment must be independent of the resources used to create the translations themselves. *Sui Generis* protection gives a holder the exclusive right to reutilise 'by the distribution of copies, by renting, by on-line or other forms of transmission' or to extract the whole or a substantial part of a database. There are exceptions for private, teaching, public service, or research use (*ibid*.). Troussel and Debussche (2014: 129) concur with Gow (2007) and others in being unable to give clear direction on TM ownership based on contractual agreements and the competing rights for source text, target text, subsequent translation, and database rights. They suggest that these may only be unpicked on a case-by-case basis. As mentioned, these competing rights are not usually considered and TM data is passed from

the translator to the client along with the translated files for review while the copyright which is potentially attributable to the translator is hardly ever asserted.

Copyright and Machine Learning

Data-driven MT overtook rule-based systems in the late 1990s as detailed by XX in Chapter 7 of this volume, with a requirement for aligned bilingual texts for system training. Metadata that might identify the translator who created each target segment and the date of translation are generally removed prior to MT training. Moorkens *et al.* (2016) recommended the use of TM metadata to explicitly assign or limit usage rights for translators who are concerned with their work being used to train MT systems, but this has not become commonplace. As data privacy has become more of a concern, especially after the implementation of the General Data Protection Regulation (GDPR) in the EU in 2018, the removal of metadata has the added benefit of anonymising what may have been considered personal data, attributable to one or many translators.

TM tools save translations as segments of text (see Chapter X on TM in this volume) . These segments can be a sentence, a headline, or a list element. SMT systems are trained on these segments, and phrase-based SMT systems create a table of possible output 'phrases', usually of two to three words. The wide scale and rapid adoption of NMT from 2016 means that previous translations are reused not only at the segment level (as happens in a TM system) or phrase-level (as in SMT), but at the word-, subword-, or even character-level in MT output. NMT systems encode and output words, one by one, followed by sentence-ending punctuation. However, in an attempt to increase the vocabulary of the system, it is common to encode parts of words or subwords in order to better produce compound or morphologically complex words that do not appear in the training data (Sennrich *et al.* 2016). In addition, requirements for training data are larger than for SMT. Wu *et al.* (2016), for example, use 5 million segment pairs for EN-DE training and 36 million for EN-FR. This means that tracing the intellectual contribution of a single translator, whose rights are unclear prior to training, is incredibly difficult or practically impossible. This is also problematic for crowdsourced work, for which there is 'no specific legal framework' (Troussel and Debussche 2014: 104). It would be very difficult to argue that a word or part thereof should be considered an intellectual creation reflecting the creator's personality. Furthermore, an attempt to unpick copyright ownership and intellectual contribution would hinder research progress for MT contrary to the intention of copyright law, according to the US Constitution is to 'promote the Progress of Science and useful Arts' (U.S. Const. art. I, sect. 8).

NMT is usually considered to be an example of machine learning or of weak artificial intelligence (for a specific rather than general purpose) due to a process that uses the result of an equation in a

series to decide on the next transformation to take place, rather than following an explicit user instruction. In common with other applications of machine learning, human-created data is required for training in vast quantities. Webcrawling for parallel text is a cost-effective option for data-gathering and may be useful when high-quality data is scarce (Toral *et al.* 2017). However, neural networks ideally require large amounts of high-quality training data and this is particularly true of NMT since the popularization of the system architecture known as the transformer model, which has exhibited efficiency gains (Vaswani *et al.* 2017).[iv] As a result, human translation data has become a valuable resource. This may be illustrated by high valuations for language data brokers, who sell language data for machine learning system training, and media stories of a 'boom in language data' (Diño 2018). Some publicly available translation data, such as that from the Directorate-General of Translation in the EU, is used as a basis for MT training, although a greater amount and breadth of data from other sources is required for an MT system to have a competitive advantage.

The lack of explicit human involvement in machine learning presents a problem with regard to copyright ownership of NMT output. The UK Government Copyright, Designs and Patent Act (1988) states that 'in the case of a literary, dramatic, musical or artistic work which is computer-generated, the author shall be taken to be the person by whom the arrangements necessary for the creation of the work are undertaken'. The final report of the US National Commission on New Technological Uses of Copyrighted Works (CONTU; 1979) stated that the 'author is the one who employs the computer'. At the time of writing, only humans may be granted copyright. The CONTU report further stated that the 'development of this capacity for "artificial intelligence" has not yet come to pass', and therefore considerations of non-human copyright were 'too speculative to consider at this time' (*ibid*.). Troussel and Debussche (2014: 103) believe that MT output 'would not be protected under copyright given that it leaves no room for human creativity and would therefore be deprived of originality'. Authors such as Dormehl (2016) suggest that computers *can* come up with solutions that may be considered counterintuitive, uninhibited by human bias, and possibly creative. Bridy (2012: 2) argues that 'all creativity is inherently algorithmic' in that human creations are based on a combination of influences, and that 'works produced autonomously by computers are […] less heterogeneous to both their human counterparts and existing copyright doctrine than appearances may at first suggest'. Copyright laws have not yet adopted this position, with the Irish legal stance that work 'generated by computer in circumstances where the author of the work is not an individual' should belong to the person (or company) 'by whom the arrangements necessary for the creation of the work are undertaken' typical of international law (*ibid*. 27).

This debate, which Wu (1997: 154) called 'one of the most puzzling problems in copyright law', continues. A European Parliament resolution on Civil Law Rules on Robotics (2017) called for

guidance in this regard, although the resolution is more concerned with liability than copyright. The risks for self-driving cars are obvious, but publication of raw MT output carries reputational, legal, and communicational risk, potentially causing harms to the public (such as unusable products), injury, or even death (Canfora and Ottman 2018). At present an automaton may not have work protected as only humans are legally considered to be creative, meaning that copyright will revert to the programmer if they have exercized choice or intellectual labour in the preparation for translation. Cabanellas' (2015) interpretation is that otherwise uncreative raw MT output will not be subject to copyright. This mirrors the judgement in a landmark 1991 case in the US, which set a precedent whereby information alone cannot be protected by copyright without a minimum of original creativity. The judge in the Feist Publications vs. Rural Telephone Service (1991) case ruled that there is a 'narrow category of works in which the creative spark is utterly lacking or so trivial as to be virtually nonexistent', and that such works are 'incapable of sustaining a valid copyright.'

Emerging Issues

Liability for machine learning applications is an ongoing issue for which no solution has yet been adopted. Other emerging issues relate to the difficulty for legislators in keeping up-to-date with technological developments in copyright, personal data, and new models of ownership. We consider some of these in this section.

Many authors and academics have described the current copyright regime as outdated and unworkable in the digital era. Barlow (1994) wrote that intellectual property (IP) laws are based on physical objects rather than ideas, Lunney (2001: 814) believed that IP laws increasingly 'serve the private interest of those who create and disseminate' work, and more recently Giblin (2018: 370) opined that copyright is based on outdated assumptions, such as the high cost of copying and distribution, and that these are embedded via treaties 'into domestic laws worldwide'. Giblin and Weatherall (2017) suggest moving away from the dichotomy of the instrumentalist and naturalist approaches to copyright, as these are fundamentally meant for social or economic benefits (respectively) but are both unrealistic. Many updates and overhauls have been proposed. Reese (2017), for example, believes that in an age where almost every person creates content via social media, not every act of creation should be subject to copyright, and that a threshold based on the level of independent creativity required and the size of the creation should be considered. Based on his suggestions, much non-creative translation work would become non-copyright, thus democratising access to data. Giblin (2018) suggests that creators should have to actively register for copyright protection, and that any rights assigned via contract should revert to the creator after 25 years. If material to be translated was not registered, this would also improve access to data. The

upcoming EU Directive on Copyright in the Digital Single Market does not include any of these suggestions, however, focusing instead on fair payment for content creators, limits to news aggregation or sharing websites that do not add value or produce content, and requires providers to prevent, remove, and make unavailable copyrighted and unlicensed material.

As machine learning becomes increasingly able to profitably leverage interlinks between data sets and their annotation, the partial protections offered by copyright for participants in increasingly varied and complex data value chains may serve to undermine the attractiveness and economic benefits of those chains. César *et al. (*2017) argue that reliance on copyright and other partial forms of protection such as the EU Database Directive and GDPR will continue to add complexity and uncertainty to data sharing contracts that could be avoided by a simple universal data ownership right. They propose a non-exclusive data ownership right that a party can assert merely by recording their contribution to a data set. This offers an interesting opportunity for translators, as well as other language workers such as terminologists and quality reviewers, to secure future benefit from their labour without having to anticipate the value arising from future machine learning from that data in order to secure contract terms. Such data can therefore flow more freely before the future value of the data contribution is fully known, using low cost means for logging contributions such as standard provenance metadata (Filip *et al.* 2013) or distributed ledgers (Nolan and Adair 2016).

In their book *The End of Ownership*, Perzanowski and Schultz (2016) discuss the ramifications of the trend from purchasing to licensing of digital media and consider that this trend tends to strengthen the long-term power of the rightsholder at the expense of the consumer. They also believe that copyright law has not been updated for the digital era, which 'could harm both the public and creators in the long run' (*ibid*. 36). Mulligan (2017) believes that limitations to IP rights are required to untangle the competing interests of digital media, such as in the case of a rightsholder non-exclusively licensing the creation of a derivative work (such as a translation) and maintaining whole or partial ownership of both the original and the derivative work. She suggests that the interest (and potential veto) of the original rightsholder should be exhausted once the derivative work has been published, as currently such works are potentially subject to claims from multiple owners and 'are particularly vulnerable to anticommons problems' (*ibid*. 268).

The anticommons refers to a situation wherein there are so many competing claims on a resource that it becomes impossible to use or exploit it, despite foreseeable benefits (see Heller 2008). Hess and Ostrom have also identified problems with digital information that is 'enclosed, commodified, and overpatented' (2007a:4). Their suggested solution is a digital commons that is managed by an online community based on 'collective action and self-governing behaviours; trust and reciprocity'

(2007b: 43). Returning to the theme of data as a natural resource, a commons is a shared resource available to all, such as air or water. Hess and Ostrom identify different options for access to this digital commons that range from the right to view only to the right to sell, lease, or access. We applied this possibility to translation data and used Hess and Ostrom's institutional analysis and development (IAD) framework to analyse potential outcomes (Moorkens and Lewis 2019). We suggest that, while somewhat utopian, this model, administered via professional translator organizations, would improve the sustainability of the translation industry by improving redistributional equity within translation production networks, circumventing any future legal challenge to the current ambiguous copyright status quo, providing ownership rights and ongoing royalties to the translation community, and offering a resource-anchored hub for improved mutual professional support activities (*ibid*.). Defining the boundaries of the translation community is difficult as many translators work part-time, without relationships with national organizations or networks, some with translation training and others without (Koskinen and Dam 2016). Nonetheless, there are sporadic efforts being made by translator societies and others to define the boundaries of the profession with an eye to royalty payment. Were these translators to form a guild or to sign up to a collective agreement, rights to reuse their non-sensitive translation data could be managed or leased by representatives of the community for translation leverage or machine-learning training on a non-profit basis, with royalties accruing to translators. Royalties for translation work are not usually paid to non-literary translators, although Kuo (2015) reports that a minority of respondents to a survey about working conditions for subtitlers reported receiving regular royalties for secondary use of their work, often operating as part of joint collective societies representing audiovisual translators in Nordic countries.

The survey respondents in Kuo (2015) more commonly find that they have little power for negotiation about rates or royalties when working individually, a finding reported for translators of other text types (Moorkens 2017). Abdallah and Koskinen (2007) note that freelance translators working via LSPs often have a single intermediary, with little access to other points in the production network and few opportunities to build goodwill, loyalty, and a relationship of trust with their client. Trust breakdown is thus a threat to translation production networks, and a trend of increased workplace monitoring is unhelpful in this regard. Workplace monitoring, using swipe or 'clocking in' cards have been common in some workplaces, but new forms of monitoring store more intimate or social data for aggregation and comparison (Moore 2017). For translators, this is operationalized as a tendency towards 'instrumentation' of translation tools, recording logs of translation activity and times where the translator is inactive, particularly within web-based editing interfaces (Moran and Lewis 2011). More broadly, as MT providers and language data brokers amass large troves of data

for training MT engines, data privacy must be a concern, if TM data with attribution to individuals were to be considered personal data, a breach within the EU would risk serious financial penalties under the terms of the GDPR. LSPs that collect translation activity data via monitoring interfaces incur further risk as such data is difficult to anonymise. See Part V of this volume for examples of translation user activity gathering for research purposes. Typing patterns captured or streamed to an external service provider may be used as a passive form of identification, and thus data that is ostensibly anonymized may be considered personal data. Typing patterns and mouse movement patterns have also been used to infer sensitive attributes about an end user when aggregated with other forms of gathered data. These attributes may not necessarily considered as personal data (and thus may not be covered by data privacy laws), and yet may allow parties with access to the aggregated data and analytical capabilities to make sensitive inferences about the user regarding attributes such as sexual orientation, political opinions, or mental health (Wachter and Mittelstadt 2019).

Conclusion: A summary and implications

Translation copyright was first codified as a derivative work in the Berne Convention of 1886, subject to the rights of the creator of the original work. While most countries are now contracting parties to the Berne Convention, differing interpretations and additional laws and directives mean that rights to ownership of a translation are not consistent in all jurisdictions. The original intention of the Berne Convention was to prevent piracy in the production of physical artefacts, and the authors could not have foreseen the large scale reuse human data in current applications of machine learning. Source texts and translations in the form of parallel texts have been recycled, initially via translation memory tools, then as training data for MT systems, with data requirements (and data value) growing exponentially in the case of NMT. End users of MT are unlikely to assume that this translation is based on amassed human translation data.

Copyright ownership for translation is based to a great extent on the Berne Convention, although there are possible conflicting claims from translators as creators or derivative work and from database maintainers, particularly if a degree of creativity or intellectual effort has been expended in this work. At present, benefits are rarely passed to the translator, meaning that, while an initial translation may be costly, secondary uses are very inexpensive. Several commentators and academics have suggested an overhaul to the current copyright regime in order to rebalance benefits away from rightholders and to simplify claims on derivative works. The upcoming EU Directive on Copyright in the Digital Single Market attempts to address this perceived imbalance, but

is unlikely to have a great effect on translation copyright. At present, the attribution of translation copyright (and the reuse of translation as data) is subject to a number of conflicting and inconsistently-interpreted laws and conventions and thus remains somewhat unclear.

List of related topics in this volume (based on the chapter list provided)

7. Machine translation: From RBMT and SMT to NMT

9. Multinational language service provider as user

10. Freelance translator as user

14. Technology and technical translation and localization

29. Translation technology evaluation research

30. Translation workplace-based research

31. Translation technology research with eye-tracking

32. Translation technology research and HCI

33. Sociological approach to technology

34. Information theoretic approach to translation and technology

References

Abdallah, K. and K. Koskinen (2007) 'Managing Trust: Translating and the Network Economy', *Meta* 52(4), 673–687.

Abrams, H. B. (1992) 'Originality and creativity in copyright law', *Law and Contemporary Problems* 55 (2): 3-44.

Barlow, J. P. (1994) The Economy of Ideas. Wired 3/1/1994. Available online: https://www.wired.com/1994/03/economy-ideas/ [last access 10 October 2018]

Bridy, A. (2012) 'Coding Creativity: Copyright and the Artificially Intelligent Author', *Stanford Technology Law Review*, 5: 1-28.

Burger, P. (1988) 'The Berne Convention: Its History and Its Key Role in the Future', *Journal of Law & Technology* 3: 1–69.

Cabanellas, G. (2015) *The Legal Environment of Translation.* London: Routledge.

Canfora, C. and A. Ottmann (2018). 'Of ostriches, pyramids, and Swiss cheese: Risks in safety-critical translations.' *Translation Spaces* 7(2), 167–201.

Cerda Silva, A. (2012) 'Beyond the Unrealistic Solution for Development Provided by the Appendix of the Berne Convention on Copyright', *PIJIP Research Paper no. 2012-08* American University Washington College of Law, Washington, D.C.

César, J., J. Debussche, and B. Van Asbroeck (2017) 'White Paper - Data ownership in the context of the European data economy: proposal for a new right', Bird & Bird. Available online: https://www.twobirds.com/en/news/articles/2017/global/data-ownership-in-the-context-of-the-european-data-economy [last access 14 December 2018].

Diño, G. (2018). 'Korean Voice Assistant Highlights Tech's Insatiable Hunger for Language Data.' *Slator*, Oct. 8, 2018. https://slator.com/technology/korean-voice-assistant-highlights-techs-insatiable-hunger-for-language-data/ [last access 9 October 2018].

Dormehl, L. (2016) *Thinking Machines: The inside story of Artificial Intelligence and our race to build the future.* New York: Random House.

European Commission (2018) Joint statement by Vice-President Ansip and Commissioner Gabriel on the European Parliament's vote to start negotiations on modern copyright rules. *Press Release 12 September 2018.* Available from http://europa.eu/rapid/press-release_STATEMENT-18-5761_en.htm

European Parliament (1996) Directive 96/9/Ec of the European Parliament and of the Council. Available online: https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31996L0009:EN:HTML [last access 9 October 2018].

European Parliament (2017) European Parliament resolution of 16 February 2017 with recommendations to the Commission on Civil Law Rules on Robotics. Available online: http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//TEXT+TA+P8-TA-2017-0051+0+DOC+XML+V0//EN#BKMD-12 [last access 9 October 2018].

Feist Publications, Inc. v. Rural Telephone Service Co., 499 U.S. 340 (1991). Available online: https://www.law.cornell.edu/copyright/cases/499_US_340.htm [last access 14 December 2018].

Filip, D., McCance, S., Lewis, D., Lieske, C., Lommel, A., Kosek, J., Sasaki, F. and Y. Savourel (2013) 'Internationalization Tag Set (ITS) Version 2.0 W3C Recommendation 29 October 2013', World Wide Web Consortium (W3C). Available online: http://www.w3.org/TR/its20/#provenance [last access 14 December 2018].

Forcada, M. L. (2017) 'Making Sense of Neural Machine Translation', *Translation Spaces* 6(2): 291–309.

Franken, R. E. (2006) *Human Motivation* (6th ed.), Boston: Cengage Learning.

García, I. (2009) 'Beyond Translation Memory: Computers and the Professional Translator', *Journal of Specialised Translation* 12: 199-214.

García, I. (2006) 'Translators on translation memories: A blessing or a curse? Pym, A., A. Perekrestenko, B. Starink (eds.) *Translation Technology and its Teaching*, Tarragona: Servei de Publicacions, 97-106.

Giblin, R., K. Weatherall (2017) If we redesigned copyright from scratch, what might it look like? Giblin, R., K. Weatherall (eds) (2017) *What if we could reimagine copyright?* Acton: ANU Press, 1-24.

Giblin, R. (2018) Reclaiming Lost Culture and Getting Authors Paid. *Columbus Journal of Law and Arts* 41, 369-410.

Gottlieb, H. (1994) 'Subtitling: Diagonal translation', *Perspectives* 2(1), 101–121.

Gow, F. (2007) ‘You Must Remember This: The Copyright Conundrum of ‘‘Translation Memory’’ Databases’, *Canadian Journal of Law and Technology* 6 (3): 175–192.

Heller, M. (2008). *The Gridlock Economy: How Too Much Ownership Wrecks Markets, Stops Innovation, and Costs Lives*, New York: Basic Books.

Hess, C. and E. Ostrom (2007a). ‘Introduction: An Overview of the Knowledge Commons.’ Charlotte Hess, Elinor Ostrom (eds). *Understanding Knowledge as a Commons: From Theory to Practice*. Boston: MIT Press, 3–26.

Hess, C. and E. Ostrom (2007b). ‘A Framework for Analyzing the Knowledge Commons.’ Charlotte Hess, Elinor Ostrom (eds). *Understanding Knowledge as a Commons: From Theory to Practice*. Boston: MIT Press, 41–82.

Kelleher, J., B. Mac Namee and A. D'Arcy (2015) *Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies*. Boston, MIT Press.

Kohavi, R. and F. Provost (1998) ‘Glossary of terms’, *Machine Learning* 30: 271–274

Koskinen, K. and H. V. Dam (2016) ‘Academic boundary work and the translation profession: Insiders, outsiders and (assumed) boundaries’, *Journal of Specialised Translation* 25: 254-267.

Kuo, A. S-Y. (2015) ‘Professional Realities of the Subtitling Industry: The Subtitlers’ Perspective’. Díaz Cintas, J., R. Baños Piñero (Eds) *Audiovisual Translation in a Global Context: Mapping an Ever-changing Landscape.* London: Palgrave, 163-191.

Lunney Jr, G. S. (2001) ‘The Death of Copyright: Digital Technology, Private Copying, and the Digital Millennium Copyright Act’, *Virginia Law Review* 87(5), 813-920

Margoni, T. (2016) ‘The Harmonisation of EU Copyright Law: The Originality Standard’. Mark Perry (Ed) *Global Governance of Intellectual Property in the 21st Century: Reflecting Policy Through Change*. Berlin: Springer, 85–105.

Moore, P. V. (2017) *The Quantified Self in Precarity: Work, Technology and What Counts,* London: Routledge.

Moorkens, J., D. Lewis, W. Reijers, E. Vanmassenhove and A. Way (2016) ‘Translation Resources and Translator Disempowerment.’ In *Proceedings of ETHI-CA² 2016: ETHics In Corpus collection, Annotation and Application*, 49-53.

Moorkens, J. and D. Lewis (2019) Research Questions and a Proposal for the Future Governance of Translation Data. *Journal of Specialised Translation*, 32.

Moran, J. and D. Lewis (2011) ‘Unobtrusive methods for low-cost manual evaluation of machine translation’, In *Proceedings of Tralogy 2011*, Paris, 1-9.

Mulligan, C. (2017) ‘A Numerus Clausus Principle for Intellectual Property’, *Tennessee Law Review* 80, 235-290.

NGTV (2017) General Terms and Conditions of The Netherlands Association of Interpreters and Translators for Translation Work. Available online: https://ngtv.nl/application/files/1215/2846/6121/NGTV_logo_algemene_voorwaarden_Engels_20170915.pdf

Nolan, P. and M. Adair (2016) 'Blockchain Technology: Emerging from the Shadows', Thomson Reuters Practical Law UK Articles. Available online: https://uk.practicallaw.thomsonreuters.com/4-634-8506?__lrTS=20170328154357505&transitionType=Default&contextData=(sc.Default)&firstPage=true&bhcp=1 [last access 14 December 2018].

O'Sullivan, C. (2013) 'Creativity', Gambier, Y., L. van Doorslaer (Eds) *Handbook of Translation Studies (Volume 4).* Amsterdam: John Benjamins, 42-46.

Perzanowski, A. and J. Schultz (2016) *The End of Ownership: Personal Property in the Digital Economy*. Boston: MIT Press.

Reese, R. A. (2017) What should copyright protect? Giblin, R., K. Weatherall (Eds) *What if we could reimagine copyright?* Acton: ANU Press, 111-146.

Sennrich, R., B. Haddow and A. Birch (2016) 'Neural Machine Translation of Rare Words with Subword Units', In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016),* Berlin, ACL, 1715–1725. Available online: https://aclweb.org/anthology/P16-1162 [last access 14 December 2018].

Smith, R. (2008) Your Own Memory. *The Linguist* 47 (1): 22-23.

Sternberg R. J. (2011) *Cognitive Psychology* (6th ed.). Boston: Cengage Learning.

Topping, S. (2000) 'Sharing translation database information: Considerations for developing an ethical and viable exchange of data', *Multilingual Computing and Technology* 11(5): 59– 61. Available online: https://multilingual.com/all-articles/?art_id=1105 [last access 12 November 2018].

Toral, A., M. Esplá-Gomis, F. Klubička, N. Ljubešić, V. Papavassiliou, P. Prokopidis, R. Rubino and A. Way (2017). 'Crawl and crowd to bring machine translation to under-resourced languages.' *Language Resources and Evaluation* 51(4), 1019–1051.

Troussel, J.-C. and J. G. Debussche (2014) *Translation and intellectual property rights.* Report by Bird & Bird for the European Commission DG Translation. Luxembourg: European Commission. Available from https://publications.europa.eu/en/publication-detail/-/publication/e079e290-e250-482d-9d4f-dae566ba67ff/

UK Government (1988) Copyright, Designs and Patent Act. Available online: https://www.legislation.gov.uk/ukpga/1988/48/contents [last access 14 October 2018].

United States Constitution (Article I, Section 8, Clause 8).

US National Commission on New Technological Uses of Copyrighted Works (CONTU) (1979) Final Report. Available online: http://digital-law-online.info/CONTU/

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin (2017) 'Attention is All you Need', Paper presented at *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, 6000—6010.

Venuti, L. (1995) 'Translation, Authorship, Copyright', *The Translator*, 1(1), 1–24.

Vieira, L. N. (2018) 'Automation anxiety and translators', *Translation Studies* (online first).

Wachter, S. and K. Mittelstadt (2019) A Right to Reasonable Inferences: Re-Thinking Data Protection Law in the Age of Big Data And AI. *Columbus Business Law Review* preprint.

Way, A. (2018) 'Quality Expectations of Machine Translation', in J. Moorkens, S. Castilho, F. Gaspari and S. Doherty (eds) *Translation Quality Assessment*, Cham: Springer, 159–178.

World Trade Organization (WIPO) (1886) Berne Convention.

World Trade Organization (WIPO) (1994) Agreement on Trade-Related Aspects of Intellectual Property Rights (TRIPS). Available online: https://www.wto.org/english/docs_e/legal_e/27-trips.pdf

Wu, A. J. (1997) From Video Games to Artificial Intelligence: Assigning Copyright Ownership To Works Generated By Increasingly Sophisticated Computer Programs. *AIPLA* 25(1), 131-155.

Wu Y, Schuster M, Chen Z, Le QV, Norouzi M, Macherey W, Krikun M, Cao Y, Gao Q, Macherey K, Klingner J, Shah A, Johnson M, Liu X, Kaiser L, Gouws S, Kato Y, Kudo T, Kazawa H, Stevens K, Kurian G, Patil N, Wang W, Young C, Smith J, Riesa J, Rudnick A, Vinyals O, Corrado G, Hughes M and J. Dean (2016) Google's neural machine translation system: bridging the gap between human and machine translation. arXiv:1609.08144

Further reading

Cabanellas, G. (2015) *The Legal Environment of Translation.* London: Routledge.

Troussel, J.-C. and J. G. Debussche (2014) *Translation and intellectual property rights.* Report by Bird & Bird for the European Commission DG Translation. Luxembourg: European Commission. Available from https://publications.europa.eu/en/publication-detail/-/publication/e079e290-e250-482d-9d4f-dae566ba67ff/

These are key texts on translation copyright, both following a similar path through the conventions, laws, directives, and legal precedents that affect copyright ownership. Some authors had previously published articles on this topic, but none in such detail. The Troussel and Debussche report is available to download free of charge, and focuses on member states of the European Union. Cabanellas's book takes a broader international view, adding discussions of confidentiality and contracts related to translation.

---

[i] Dormehl (2016: 156) commented that if 'data is the oil of the digital economy, then we need to place a proper valuation on it'.

[ii] This practice, in which translations are 'produced not from the original, but from an existing translation in another language' is common for minor language subtitling (Gottlieb 1994: 117).

[iii] LSPs commonly expect or impose discounts for 'perfect and near matches' from a TM when paying a translator per word for a translation (García 2006: 97).

[iv] Early NMT systems usually comprized several types of neural network and an 'attention mechanism', which predicted likely collocates for words. The transformer model focuses on this attention mechanism, dispensing with many of the other neural networks. Vaswani *et al.* (2017) found this model to produce superior results, leading to its popularization within the MT research community.