Translation Resources and Translator Disempowerment

Joss Moorkens[1], David Lewis[2], Wessel Reijers[1], Eva Vanmassenhove[1], Andy Way[1]

[1]ADAPT Centre, School of Computing, Dublin City University, Ireland,

[2]ADAPT Centre, Trinity College Dublin, Ireland

E-mail: joss.moorkens@dcu.ie, dave.lewis@adaptcentre.ie, wreijers@adaptcentre.ie,

eva.vanmassenhove2@mail.dcu.ie, away@computing.dcu.ie

Abstract

Language resources used for machine translation are created by human translators. These translators have legal rights with regard to copyright ownership of translated texts and databases of parallel bilingual texts, but may not be in a position to assert these rights due to employment practices widespread in the translation industry. This paper examines these employment practices in detail, and looks at the legal situation for ownership of translation resources. It also considers the situation from the standpoint of current owners of resources.

Keywords: Language resources; copyright; ethics

1. Introduction

Statistical Machine Translation (SMT: Koehn et al., 2003, 2007), the most prevalent paradigm currently for automatic translation, requires large amounts of bilingual parallel language resources. These resources are originally created by human translators whose rights with regard to their creation are not always respected, and who are disempowered by the vendor model widespread within the language services industry. While the proportion of freelance or contingent workers in developed countries has increased, reaching 40.4% in the US in 2010 (U.S. Government Accountability Office, 2015), surveys of translators have found the proportion of freelance workers to be in the region of 80% (84% in Kelly DePalma, & Hegde, 2012; 77% in Ehrensberger-Dow et al., 2015). This prevalence of freelance translation has the effect of disempowering translators who otherwise might be in a position to assert their copyright for work created or derived work, as well as for collective bargaining for pay rates and conditions.

The issue of falling pay rates and of unclear ownership of translation databases has grown in prominence as the use of Translation Memories (TMs: Heyn, 1998) as repositories for previously translated work has become widespread, in particular as 'fuzzy match' scoring (Sikes, 2007) against client-provided TMs has become a common discounting mechanism in pricing translation projects. TMs are also the engines. The use of SMT is already widespread in translation projects, so the opportunity for effective leverage of data from a specific TM in translating content from different clients or domains has widened.

In this paper we look in more detail at translators' employment conditions and their association with practices prevalent in the language industry with regard to data ownership. We then examine how copyright might apply to translated texts and TM databases, and finally offer some recommendations from the perspective of translators and for regulation of copyright ownership.

2. Translators' Agency

Translators' have found their profession increasingly limited in several ways: conditions of employment have moved to a freelance model with an associated loss of security and benefits, technologization of the translation industry has reduced translator autonomy, and the related move to the digital domain has made the situation unclear with regard to the ownership and reuse of translated material. In the following sections, we examine each of these issues in turn.

2.1 The Vendor Model

From a high point of "de-commodification" of labour and gains in worker power in the boom years post-WW2 (Munck et al., 2011), many industries have moved to a freelance model, where workers have become self-employed contractors, who have to "buy their own tools and equipment, and bear all the risks of accident, sickness, or lack of work" (Castles, 2011). The translation industry has, to a great extent, moved in this direction. Reliance on freelance translation work has become widespread among language-service providers, as the freelance model is "flexible, scalable, or cost-effective enough to respond to market demands" (Kelly et al., 2012). This may allow translators a degree of autonomy, but for most translators outside of those working for larger public institutions, there wish to continue to translate rather than moving into management within a company. A survey by Moorkens and O'Brien (2016) found an association between translators' age ranges and their working conditions, where those over the age of 30 are far more likely to work on a freelance basis. Many freelancers (31% of the total) work directly for one agency, a situation referred to as *bogus self-employment* in a study of precarious work for the European Commission. When a freelancer's relationship is "with a single source rather than with a range of clients", this represents "economically dependent work" (McKay et al., 2012).

This situation leaves translators in a difficult position with regard to collective bargaining, negotiation of rates, and

assertion of copyright. Even though the language industry has continued to show year-on-year growth of over 5% through the recent recession (DePalma et al., 2013), freelance translators have complained of their powerlessness in the face of shrinking per-word rates that are often dictated by their agencies (Kelly, DePalma, & Hegde, 2012).

2.2 Translators and Technology

Translator disempowerment has been exacerbated by the technologization of the translation profession since the introduction of TM technology in the early 1990s. While some translators have been early adopters of new technologies, many resent that new technologies are imposed on them (Penkale & Way, 2013; Way, 2013): first TM with its associated fuzzy match discounts, and more recently MT post-editing, which requires them to accept further discounted rates to fix "fundamental linguistic errors that a trained human translator would rarely generate" (O'Brien, 2012). It is rarely made explicit by companies and research groups that specialize in MT that human translation is its necessary basis, with the focus instead on new and better ways to process this trove of pre-existing 'big data' (Kenny, 2011). The gradual limitation of the translator's role has undermined their ability to conform to the ethical code of their profession (Chesterman, 2001) by reducing the translation process to a series of "language-replacement exercises" (Pym, 2003). Furthermore, as the profession has moved from analogue to digital, translators' powerlessness is reflected in continued data dispossession, common for many knowledge workers, and largely unaffected by legal constraints (Huws, 2014). This is a wider problem within the digital domain, where national laws are of little relevance, and assignation of rights is often buried within data-use policies (Reijers et al., 2016).

3. Ownership of Language Resources

Translators typically create a TM file as a by-product of a translation effort. Currently, handing over TM files to an agency after a translation job has become the norm in the translation industry, whether or not ownership has been specified in translation project contracts. In the absence of a contractual agreement regarding ownership of what Smith (2008) has called the "translation family jewels", the actual legal status of a translation or translation artefacts is subject to a variety of often ill-defined national and international laws and is thus unclear (Lewis et al., 2016). For example, authorship of a source text, including the right to decide whether work is translated, may belong to either the employer or employee depending on the country in which the author is contracted, and contractual assignment of authorship is only valid in some jurisdictions (Troussel & Debussche, 2014).

Unless specified in a contract, a translator may be considered the owner of a translated text as a derivative or adapted work, depending on the perceived originality of the translation and subject to the "rights of the author of the original work" (Troussel & Debussche, 2014). In the US, the claimant of copyright must demonstrate a "minimum degree of creativity" (Cabanellas, 2014). This situation becomes more complex when applied to user-generated content or crowd-sourced translation, for which no specific legal framework exists. The copyright for a database, such as a TM file, is considered to belong to the database creator in both France and Germany, depending on the originality involved with its creation, in this case regarding "segmenting and aligning the data" (Troussel & Debussche, 2014). There may be the option of asserting further *sui generis* rights to the creation of a database, if the creator has demonstrated a substantial investment in obtaining, verifying, or presenting that database (Troussel & Debussche, 2014).

The situation with regard to copyright issues internationally appears fluid. Copyright laws have changed over time in many jurisdictions, and within the EU are further complicated by a number of EU-level directives that are intended as a step to harmonize copyright , and to address new issues raised by unexpected technological advances, permitting mass digitization of books, for example. Periodic public consultations have taken place, most recently in 2013, which look to address issues with text and data mining, and user-generated content, and have been followed up with the establishment of European Commission working groups.[1]

The somewhat fluid state of copyright law has not appeared to effect the reality for ownership of translation data, which (to our knowledge) has never been legally tested. Freelance translators continue to deliver TMs to their client or agency without question, as the failure to do so may affect the "translator's standing with that service provider" and "payment problems could ensue" (Smith, 2008). This situation is critical especially for the large proportion of translators who work directly with a single agency.

3.1 Consequences for Reuse

Although these potentially conflicting claims of copyright for written or translated material are currently ignored, they may create difficulties for enterprises offering MT and, to a lesser extent, collectives sharing MT. For translators, the re-tasking of TM as parallel text for training MT engines is a particular concern (Moorkens & O'Brien, 2016).

The leverage of TMs from previous translations is well understood by translators. They understand the role it plays in avoiding unnecessarily retranslation of similar segments and the resulting role played by matching scores between available TMs and the source of incoming translation projects in price discounting. The practice of individual translators retaining TMs from previous projects independently of vendors is widespread, as modern desktop

[1] See Text and Data Mining Working Group website at: https://ec.europa.eu/licences-for-europe-dialogue/en/content/text-and-data-mining-working-group-wg4.

translation tools allow them to use these as reminders of previous translations and for term concordancing. These are useful features for individual translators even if the level of useful TM matching leverage with a personal TM is low. These practices seem to indicate a tacit approval by translators of the use of TM leverage. There seems to be an appreciation that they benefit from the prior work of other translators captured in a TM in the same way that other translators will benefit from their work in future. We can assume there is a degree of collegiality at play here, since even if translators producing and consuming translation via TM may not know each other's identities directly, the poor level of TM leverage across domains or client content types means benefitting translators can be assumed to be working in the same broad domain as those who produced the content.

The use of TMs for MT training erodes this traditional acceptance of TM leverage, since translators perceive that the resulting MT system can be used by vendors and clients for translation in very different domains. In particular MT is seen to be useful in classes of translation tasks where little or no translator input is required (cf. Way, 2013), contributing to the misconceived perception that the spread of MT endangers the livelihood of translators.

Although TM data interoperability standards, such as Translation Memory eXchange (TMX) [2] and XML Localization Interchange File Format (XLIFF) [3] enable translator provenance to be recorded, such metadata is typically stripped from TMs before being returned to clients or used between projects by vendors. The traditional acceptance of TM leverage means that, outside of a specific translation project, the tracking of the provenance of individual translation to specific translators is not practised, and is not strongly demanded by translators. However, the loss of this provenance data means that there is no way for individual translator contributions to large aggregated TMs to be differentiated, and hence translators are denied the opportunity to specify any preferences on the rights they wish to declare over the use of TMs they return to vendors and clients.

The situation in Public Service Institutions, with regard to the collection and sharing of resources, may be somewhat simpler, depending on where they were created. The EU has a harmonized directive for re-use of information that was enacted in 2003 (directive 2003/98/EC)[4] and updated in 2013, which stipulates that written texts, databases, audio files and film fragments held within public repositories (with some exceptions) may be reusable for commercial and non-commercial purposes. These purposes need not relate to the initial intended purpose of the data. The only difficulty remaining in this instance is whether the data was created by an external party, in which case they may not have been made aware of the Public Service Information directive, nor have supplied materials such as parallel data that were created during the process of completion of their task. In this case, there may be a requirement to negotiate the release of data ownership retrospectively.

The situation at present in which laws of copyright are effectively bypassed in content collection, curation, and exploitation, permits resource holders to retain data at a cost to disempowered human writers and translators, and also at a cost to end-users of translated content. The disconnect between the MT services and the human translated corpora might further alienate translators from their work, and add to existing mistrust in MT and in data sharing.

4. Recommendations

Working largely independently within the vendor model with increasing imposition of translation technology, there are nonetheless possibilities for freelance translators to maximize their agency through collective bargaining. This could be via a national or international translators' organization such as FIT (The International Federation of Translators)[5] or online groups such as proz.com.

The growing number of precarious workers in all industries – especially for well-publicized technology companies using a crowdsourcing model such as Uber and Amazon's Mechanical Turk – has made precarious work a topical issue. 30% of paid jobs in the EU between 1987 and 2007 were temporary work, and the percentage of flexible employment contracts issued in Greece rose from 21% in 2009 to 41% in 2011 (McKay et al., 2012). In the US, the number of contingent employees more than doubled between 1969 and 1993 (Cummings & Kreiss, 2008). Concern over this issue has led to sporadic moves to allow contingent workers the right to organize, with legislation for limited collective bargaining on behalf of freelance workers progressing towards being enacted in law in Ireland in 2016 (Houses of the Oireachtas, 2016), and collective bargaining agreements are already in place for several categories of contingent workers in Washington State since 2013. One of these categories of workers is Language Access Providers, defined as 'any independent contractor who provides spoken language interpreter services for Department of Social and Health Services appointments or Medicaid enrollee appointments' (Washington Federation of State Employees, 2013). This bargaining agreement defines rates of pay, payment deadlines, and a grievance procedure. If these agreements are considered successful, there may be grounds for expanding to other categories and professions.

A second recommendation for translators is to inform themselves about their legal rights for translation. This could be encouraged via a conversation in the wider language service industry, and volunteer translation organizations such as Translators Without Borders[6] and The Rosetta Foundation[7] could also raise awareness by

[2] https://www.gala-global.org/tmx-14b
[3] http://docs.oasis-open.org/xliff/xliff-core/v2.0/xliff-core-v2.0.html
[4] http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32003L0098
[5] http://www.fit-ift.org/
[6] http://translatorswithoutborders.org/
[7] http://www.therosettafoundation.org/

explicitly using an open or standard data ownership policy and allowing volunteers to control the ways in which the content that they translate is leveraged.

A third recommendation is that translators use TM metadata more effectively to both identify the translations and translation alignments in which they had a creative input and to explicitly assign usage rights to those assets. While such metadata can be captured in existing TM data standards (TMX and XLIFF), population and maintenance of this metadata needs to be integrated into translation workflows. In addition, better shared models for differentiating used of assets is required. For example, Lewis et al. (2016) suggest an extension to the existing metadata vocabulary for expressing usage rights to allow differentiated usage rights between traditional TM leverage and TM use in MT training to be declared. A clear and legally defensible definition to allow this differentiation to be unambiguously established in any given case is still required, however.

Recent efforts to harmonize copyright laws in the EU are welcome and any agreed ethical code for collection and reuse of human translations will need to be universally agreed. The potential financial implications of this in an industry valued at US$34.778 billion in 2013 (DePalma et al. 2013) are likely to make agreement difficult to achieve.

5. Conclusion

The prevalence of the vendor model in the translation industry shows no sign of abating. As noted by Linder (1999), once cost-cutting employment practices become commonplace in an industry, other players are pushed into following those same practices in order to remain competitive. This does not necessarily mean that the outlook for translators is poor. The industry continues to grow, and governments and society are beginning to realize that they need to legislate for the protection of contingent workers and to allow collective bargaining.

Steps towards harmonization of copyright laws are being made, but legislation is particularly uneven in the digital domain, where working groups and consultations are taking place in an effort to keep up with technological changes. These developments are likely to have significant ethical implications for people working in the translation industry.

For translators, it is in their best interests to act collectively where possible, to maximize bargaining power and to share information, particularly with regard to making best use of the metadata possibilities of current interchange formats. Ideally, any agreement for collection, ownership, and reuse of translation data will come about via consensus, but more empowered translators may become emboldened to pursue copyright claims as described in Section 2, as a legal challenge on behalf of a translator could have massive repercussions in an industry where the norm is usually unchallenged.

Acknowledgements

This work has been supported by the ADAPT Centre for Digital Content Technology which is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund, by the European Commission as part of the FALCON project (contract number 610879), and by the Dublin City University Faculty of Engineering & Computing under the Daniel O'Hare Research Scholarship scheme.

6. References

Cabanellas, G. (2014). *The Legal Environment of Translation*. Abingdon: Routledge.

Castles, S. (2011). Migration, Crisis, and the Global Labour Market, *Globalizations*, 8(3), pp. 311—324.

Chesterman, A. (2001). Proposal for a Hieronymic Oath, In Pym, A. (Ed.), *The Translator*, 7(2), pp. 139-154.

DePalma, D. A., Hegde, V., Pielmeier, H., and Stewart, R. G. (2013). *The Language Services Market: 2013 (Report)*. Common Sense Advisory, Boston MA.

Heyn, M. (1998). Translation Memories – Insights & Prospects. In L. Bowker, M. Cronin, D. Kenny and J. Pearson (Eds.) *Unity in Diversity? Current Trends in Translation Studies*, Manchester: St. Jerome, pp. 123—136.

Huws, U. (2014). *Labor in the Global Digital Economy: The Cybertariat Comes of Age*. New York: Monthly Review Press.

Kelly, N., DePalma, D. A., and Hegde, V. (2012). *Voices from the freelance translator community (Report)*. Common Sense Advisory, Boston MA.

Kenny, D. (2011). The ethics of machine translation. In *Proceedings of the New Zealand Society of Translators and Interpreters Annual Conference 2011*. Auckland, New Zealand.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, M., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: open source toolkit for statistical machine translation. *ACL 2007: proceedings of demo and poster sessions,* Prague, Czech Republic, pp.177—180.

Koehn, P., Och, F.J., and Marcu, D. (2003). Statistical phrase-based translation. In *HLT-NAACL 2003: conference combining Human Language Technology conference series and the North American Chapter of the Association for Computational Linguistics conference series*, Edmonton, Canada, pp. 48–54.

Lewis, D., Fatema, K., Maldonado, A., Walshe, B., and Calvo, A. (2016). Open Data Vocabularies for Assigning Usage Rights to Translation Memories. In *Proceedings of the 10th edition of the Language Resources and Evaluation Conference (LREC)*, Portorož, Slovenia.

Linder, M. (1999). Dependent and Independent Contractors in Recent U.S. Labor Law: An Ambiguous Dichotomy Rooted In Simulated Statutory Purposelessness.

Comparative Labor Law & Policy Journal 21, pp. 187—230.

McKay, S., Jefferys, S., Paraksevopoulou, A., Keles, J. (2012). *Study on Precarious work and social rights: Carried out for the European Commission*. Working Lives Research Institute, London Metropolitan University.

Moorkens, J., O'Brien, S. (2016). Assessing User Interface Needs of Post-Editors of Machine Translation. In D. Kenny (Ed.), *Human Issues in Translation Technology: The IATIS Yearbook.* Abingdon: Routledge.

Munck, R., Schierup, C. U., and Wise, R. D. (2011). Migration, Work, and Citizenship in the New World Order, *Globalizations,* 8(3), pp. 249--260.

O'Brien, S. (2012). Translation as human–computer interaction. *Translation Spaces*, 1, pp. 101–122.

Penkale, S., and Way, A. (2013). Tailor-made Quality-controlled Translation. In *Proceedings of Translating and the Computer* 35, London, UK, 7pp.

Pym. A. (2003). Translational Ethics and Electronic Technologies, In *Proceedings of the Profissionalização do Tradutor Conference*. Lisbon, Portugal.

Reijers, W., Vanmassenhove, E., Lewis, D., Moorkens, J. (2016). On the need for a global declaration of ethical principles for experimentation with personal data. In *Proceedings of the ETHI-CA 2016 Workshop,* Portorož, Slovenia.

Sikes, R. (2007). Fuzzy matching in theory and practice. *Multilingual*, 18(6):39 – 43.

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

Troussel, J. C., Debussche, J. (2014). *Translation and Intellectual Property Rights* (Report by Bird & Bird for the European Commission DG Translation). Luxembourg: European Commission. doi:10.2782/72107

Way, A. (2013). Traditional and Emerging Use-Cases for Machine Translation. In *Proceedings of Translating and the Computer* 35, London, UK, 12pp