Shared Resources, Shared Values? Ethical Implications of Sharing Translation Resources

Jo Drugan
Centre for Translation Studies
University of Leeds
Leeds, LS2 9JT UK
J.Drugan@leeds.ac.uk

Bogdan Babych
Centre for Translation Studies
University of Leeds
Leeds, LS2 9JT UK
B.Babych@leeds.ac.uk

Abstract

The exploitation of large corpora to create and populate shared translation resources has been hampered in two areas: first, practical problems ("locked-in" data, ineffective exchange formats, client reservations); and second, ethical and legal problems. Recent developments, notably on-line collaborative translation environments (Desillets, 2007) and greater industry openness, might have been expected to highlight such issues. Yet the growing use of shared data is being addressed only gingerly.

Good reasons lie behind the failure to broach the ethics of shared resources. The issues are challenging: confidentiality, ownership, copyright, authorial rights, attribution, the law, protectionism, costs, fairness, motivation, trust, quality, reliability.

However, we argue that, though complex, these issues should not be swept under the carpet. The huge demand for translation cannot be met without intelligent sharing of resources (Kelly, 2009). Relevant ethical considerations have already been identified in translation and related domains, in such texts as Codes of Ethics, international conventions and declarations, and Codes of Professional Conduct; these can be useful here.

We outline two case studies from current industry initiatives, highlighting their ethical implications. We identify questions which users and developers should be asking and relate these to existing debates and codes as a practical framework for their consideration.

1 Introduction

Perhaps unsurprisingly, given the diversity of the industry, use of specialist software for translation and localization[1] is decidedly varied. Only a decade after the creation of the first TM tools, an overview of the industry in 2000 listed five "most commonly used" tools (TRADOS Translator's Workbench, STAR Transit, Atril Déjà Vu, SDLX, IBM Translation Manager) and five further proprietary tools (by Alpnet, Lionbridge, Cypresoft, Translation Craft, and LANT) (Esselink, 2000). There were also seven common tools for software localization (Corel Catalyst, AppLocalize, RC-WinTrans, Passolo, Visual Localize, Visual Translate, Venona Translation Toolkit) and six proprietary in-house tools (by Autodesk, Intel, Lotus, Microsoft, Oracle and Symantec) (Esselink, 2000). These tools were all for Windows applications, so further options were needed for Macintosh.

Since 2000, additional tools have become established in the market, including a few Open Source ones: the ATA currently lists 20.[2] This diversity means that TM users have found it difficult to leverage data across platforms. Organisations like the Localization Industry Standards Association (LISA) have developed shared exchange formats such as TMX[3] to address this, but the core design of the tools, different segmentation rules and analysis processes, and inefficient manual database

[1] Localization tools are based on TM technology but offer further functionalities for the translation of software. Esselink (2000: 360) explains that "Translation memory and terminology tools are generally combined in a tool set for the translation of documentation, on-line help, or HTML text. Software localization tools are used to translate and test software user interfaces, i.e. dialog boxes, menus, and messages." In this discussion, TM tools will be used to refer to the broad category of tools used in both translation and localization.

[2] See http://www.ata-divisions.org/LTD/?page_id=89

[3] Translation Memory eXchange.

Ventsislav Zhechev (ed.): *Proceedings of the Second Joint EM+/CNGL Workshop "Bringing MT to the User: Research on Integrating MT in the Translation Industry" (JEC '10)*, pp. 3–9. Denver, CO, 4 November 2010.

maintenance have all combined to make effective sharing of data challenging, even if TM owners were prepared to try. Translation exchange formats are also relatively recent and not universally adopted. Many in the industry are critical of their limitations. Kirti Vashee points out that a specification like TMX is not a standard unless it is widely adopted and used. He concludes that "Translation tools often trap your data in a silo because the vendors WANT to lock you in and make it painful for you to leave."[4]

A final significant obstacle to widespread sharing of translation resources has been client objection. Translation clients have viewed TM content as their property, pace occasional challenges from translators, and have typically safeguarded this content as confidential. Even public bodies have protected "their" taxpayer-funded translation content.

There has, however, been significant recent progress on many of these points. First, large minable multilingual corpora have been released online. Some parallel multilingual content was already freely available, e.g. English-French Parliament of Canada's Hansard (Brown et al., 1990), UN texts in six languages (Eisele and Chen, 2010), Europarl corpus, currently available for MT developers in 11 European languages (Koehn, 2005). Since the 1990s, these have been aligned on the sentence and word levels and used to build data-driven MT systems. Large-scale statistical MT platforms like Google still rely heavily on these freely available parallel corpora.

The use of such parallel resources (those which are freely available in the public domain) for building MT systems does not involve serious ethical issues: all the data are produced using public funds and from the outset, it is intended they will be available for everyone to consult and use. Importantly, they also do not contain any confidential information. However, the limitations of such resources are well-understood by MT developers: they exist only in a very narrow subject domain and contain very specific language in terms of genres, linguistic constructions, etc.

Apparently, this aspect (the quality of data used for training MT systems) is a very sensitive issue: it has been shown that Statistical MT (SMT) is domain-dependent to a much greater degree than rule-based MT (RBMT) (Koehn, 2010: 61). If an SMT engine is trained on a corpus that doesn't match the domain or genre of the translated text, then the quality of such out-of-domain translation is greatly reduced.

Sharing translation resources from a wide range of text types and subject domains becomes an essential condition for building high-quality SMT systems. However, this, in turn, requires consideration of a much wider range of ethical and legal issues, which include but are not limited to confidentiality of data, trade industrial and state secrets, and intellectual property rights of translators, authors and data owners.

The European Union's move to share substantial TM resources for free from November 2007[5] went a step further than previous decisions to share translation resources, in that the data exist across dozens of languages and in a standard format for leveraging. Second, an industry consortium established the Translation Automation User Society (TAUS) in 2007, with the aim of meeting "ever-increasing commercial and societal demand for translation". They quickly concluded that "a shared industry database of translation memories and terminology was needed as a fundamental building block to support future growth and innovation" and moved to establish such a resource.[6] Third, collaborative approaches to translation, e.g. shared on-line Translation Environment Tools (TenTs) and crowdsourcing, have drawn attention to the radical potential of shared resources to achieve translation more quickly and/or more cheaply.

Despite these developments, discussion of ethical concerns around shared data has been conspicuous by its absence in the industry. Where such concerns have been discussed, it has so far been around the extent to which translators' work is leveraged without their consent or knowledge, and two potential "threats": shared resources "providing the material to improve the quality of MT" (Bell, 2010), or community translation taking work from professionals (Kelly, 2009).

Many of the ethical issues here are familiar from the history of trailblazers like Wikipedia (e.g. ownership, attribution). They involve legal grey

[4] See http://kv-emptypages.blogspot.com/2010_05_01_archive.html.
Vashee is VP Enterprise Translation Sales at Asia On-line.

[5] See http://langtech.jrc.it/DGT-TM.html#Download

[6] See http://www.tausdata.org/blog/about-taus-data

areas, particularly as translation is affected and bound by different national, regional and international laws. Some of the issues discussed in this paper, such as translation ownership and rights, do indeed also have a legal or contractual dimension. It is perhaps understandable that the industry has shied away from confronting such complexity. But this means that users either adopt a naïve "Don't Ask, Don't Tell" stance, leaving them at risk of prosecution (if metadata were used to identify ownership of TM segments leveraged without authorisation, for instance); or they are excessively cautious, passing up potentially valuable and available data.

In contrast, engaging directly with ethical questions means that users can share data confidently, arguing from clearly stated values and precedents in related fields or debates. If TAUS is correct that "The Information Age has led to insatiable demand for translation services that cannot be met with existing proprietary business models and the capacity of around 300 000 professional translators worldwide",[7] such informed sharing of resources is critical.

As an initial contribution to this engagement, we present below two case studies of current industry initiatives involving shared translation resources. We describe each in turn, focusing on the key ethical questions it raises. We selected these questions with particular reference to relevant established codes of ethics and conventions (Anderson et al., 1992; ITI, 2010; SFT, 2010; UNESCO, 1976). We conclude by outlining the general benefits of such an approach.

2 Case Studies

2.1 Google Translation Toolkit

In 2005, Google launched a successful on-line Statistical Machine Translation platform (available at `http://translate.google.com/`), which works (directly or via a pivot language) for 58 language pairs as of 2010; new languages are regularly added. Like many on-line MT systems of this type, Google MT is primarily used in translation for assimilation scenarios: when MT does not produce high-quality translation for publication (dissemination) purposes, but rather generates imperfect but still comprehensible output aimed at understanding (assimilation) of the translated text. Users are typically not professional translators who do not know the source language, but wish to access news, software manuals, medical forums, etc. in another language.

Professional translators typically remain unconvinced by claims that such systems could be useful for translation workflow (especially on-line systems that lack proper domain customisation, dictionary management and integration with TMs) (Prior, 2010). Even though there are indications that for some language pairs (e.g. translation between closely-related languages) or in certain narrow subject domains (e.g. software manuals, development documentation), post-editing MT output requires less effort than translating the original text from scratch (Babych et al., 2007), it is hard to demonstrate that the MT post-editing scenario is as efficient for more general cases. For example, Guerberof (2009) shows a gain in productivity of up to 25% if experienced translators post-edit English-Spanish MT for technical texts. However, this figure cannot be generalised to MT between more linguistically distant languages, other subject domains or genres.

In the general case it is not possible to extrapolate expected productivity increases from automated MT evaluation scores (such as BLEU), since calibration parameters of these scores (the slope and the intercept of the regression line between human and automated evaluation figures) are individual for each combination of translation direction and the text type (Babych et al., 2005)

It is plausible that for more distant languages or broader domains, any benefits to professional translators of using MT are greatly diminished. In addition, if translators do not have prior MT post-editing experience, productivity can be much lower than if they simply translate from scratch. Even though greater experience can shift the balance in favour of post-editing, many translators simply do not try because they do not have time to learn the skills (something which would not currently guarantee an increase in productivity).

To address some of these concerns and bring MT closer to professional translators' workflow, Google embedded their MT engine in an experimental on-line collaborative translation envi-

[7] *Ibid*.

ronment, Google Translation Toolkit.[8] The current version of this system integrates MT with TM and user dictionary functionality. It allows users to post-edit MT output, but TM matches and entries in the user glossary have priority over machine-translated segments. Collaboration between translators is supported via a framework similar to Google Docs: multiple translators can link to a translation project and edit the text at the same time. Projects are stored on "cloud" servers, but can also be downloaded in various formats to users' local machines at any time. Useful collaborative functionalities include the possibility to add comments to individual words or phrases. At the time of writing, the Toolkit does not support tracking previous versions (which is possible in Google Docs), but this may be added in future.

By default, translators are presented with the post-editing scenario, but can also choose to work in a more traditional way, using TM and glossary only, without MT in the background.

Technologically, Google Toolkit is an advanced and innovative system. However, a number of factors limit its usefulness for professional translators. Importantly, these limiting factors are not technological: they come from lack of consideration of a number of ethical issues in the first place, which are relevant in many typical scenarios in professional translators' workflow.

The central problem is how the Toolkit and the underlying MT platform ensure confidentiality of the user translation project and resources.

There are some translation projects (e.g. community translation, translation of free software documentation), which are not limited by considerations of confidentiality, but in most cases, client confidentiality is essential. Legal and medical translations usually involve sensitive personal data, for example, while technical translations often contain classified commercial or technical information. Documents intended for consumer use (e.g. operation manuals) can contain information, which must not be released before a given date, to protect product functionality or specifications being copied by competitors before the official release.

These confidentiality requirements are not adequately addressed by the technology behind the Toolkit or by the legal licence framework, to the extent that Google's system is not practical for use in most real-world professional translation projects.

Even during initial encounters with Google machine translation, professional translators expressed concern that it was not clear whether and how confidentiality of any source texts uploaded to Google servers could be guaranteed. It is technically possible to pass data via encrypted channels and store encrypted user spaces on the servers, but this is implemented neither in Google MT, nor in the Toolkit. Appropriate user agreements and licences should also support this functionality, but are not as yet in place.

A less obvious but more serious problem is with the current default settings of the Toolkit and the existing licences. By default, all new translations go into the "Global TM" to be shared with all other users. It is possible to build a non-shared TM, which is available only to collaborators on a particular project, but the user licence states that Google can use all translations—even those in private TMs—to improve their MT engine. This means that human developers will not look at the translated texts. These texts will be automatically broken down and aligned at the phrase level and reused by the automatic system. Still, for many translation projects this type of sharing is a step too far, especially if their translated text contains proper names—people, locations, products, company names.

Google Toolkit fails to acknowledge other enduring ethical issues where the limiting factors are not technological.

Some of these ethical issues concern recognition and perhaps compensation of translators' work. Where a translator is first to translate a new technical term (e.g. for a software feature), the challenging nature of this work has often been recognised. Making it available immediately through MT could share translators' work without their consent, or appropriate credit or remuneration. Similarly, translators' new segments in a TM are typically marked with their username or initials. This is important because a translator knows whether future matches are his own or those of known, trusted colleagues. Equally, if the translator is compelled to use a poor translation which has previously been approved in a client's TM, it is clear that the fault or inferior quality translation is not hers. This can be essential where a flawed translation has legal consequences, for instance.

[8] See `http://translate.google.com/toolkit`

Similar collaborative translation tools do already support such features (e.g. MyMemory: see http://mymemory.translated.net/doc/).
Finally, contracts with clients or agencies normally stipulate ownership of TMs and permitted future use. In Google Toolkit, segments saved in the "Global TM" for future access, to be shared by all users, raise ethical questions regarding attribution. Is it clear whose work is being used? Has permission been given?

None of these issues is new or unique to Google Toolkit. Translation agencies, professional associations like the ATA, tools developers and translators themselves have considered these questions and found workable technological and legal solutions in other contexts.

To conclude, Google Translation Toolkit does not have an adequate legal and licensing framework to address the needs of professional translation projects, so in practice cannot be used for most real-world translations. Our case study demonstrates a more general issue when potentially useful and innovative technology does not take into account practical user-based scenarios, in part because it is not supported by an appropriate ethical framework. These issues are not technological in nature and have previously been addressed by translation specialists.

2.2 TAUS Language Search Engine (LSE)

As part of a strategy to share TMs, TAUS Data Association (TDA) developed their Language Search Engine, an on-line tool for searching uploaded TMX data.[9] It generates parallel "concordances" (matches for key-words found inside TM segments).

The LSE works as an "intelligent dictionary" rather than a translation engine. It can search and display translation equivalents for a particular word or phrase; searches can be limited by industry, domain, product, etc. This allows users to address important issues such as translation of terminology or near-terminological phrases in specific subject domains. Where a term does not appear in existing terminological dictionaries, but was previously translated in any TMX-format file shared in TDA repositories, the LSE can identify and highlight translation equivalents for the relevant terminological unit, using parallel concordance and word alignment techniques.

An important issue for TDA is managing user expectations: according to LSE developers, logged searches include full sentences (which naturally do not produce results). The LSE can search for discontinuous phrases, but it is not a Google-type MT or search engine; like a dictionary, it produces meaningful output only for shorter phrases or individual words.

The ethical framework for the LSE lies in the agreements with participants, who all agree to share TM resources. The agreement makes explicit that data owners consent to other users accessing or re-using their TM segments. Responsibility for ensuring that shared TMs contain no confidential data (or that such data are anonymized) rests with the data owner.

The important difference with Google's Toolkit is that here the agreement to share translation resources is explicit – users do not have to search through small print buried in the user licence; nor must they research and change multiple system default settings. Instead, the incentive to share one's own resources comes from gaining reciprocal access to other participants' shared resources, something which provides a much clearer ethical framework and even a model for other collaborative uses of data.

However, even if the TDA initiative goes further than Google Toolkit in considering ethical issues, both leave key questions unaddressed. Again, these issues are ethical rather than technological in nature.

First, even if client confidentiality is taken into account, broader ethical issues around ownership and consent are less clear-cut. TDA describes itself as a "community of users and providers of translation technologies and services"[10] but members are all large-scale clients, agencies and technology providers rather than individual translators or end-users of translated material. TDA lists three types of member: content owners (e.g. Adobe, DELL), practitioners (e.g. Lionbridge, HiSOFT) and technology (e.g. SDL, SYSTRAN). The size of these companies and organizations makes it highly unlikely that all clients gave their informed consent to

[9] See http://www.tausdata.org/index.php/language-search-engine

[10] See http://www.translationautomation.com/about-taus/mission.html

this use (far less the translators who originally produced the data being shared).

Of course, translators sign standard contracts waiving their future rights: there is no legal problem here. Rather, issues identified as problematic in various industry codes of ethics are relevant, such as "taking credit for others' ideas or work, even in cases where the work has not been explicitly protected by copyright, patent, etc."[11] Does a translator signing a standard contract with an agency realize his work may be shared in this way in future? Just as TDA has a clear agreement with its members regarding use of their data, translators might expect such clarity from agencies and clients. Section III (c) of UNESCO's 1976 Recommendation on translators' status and rights offers a practical suggestion here, that contracts might make provision for a "supplementary payment should the use made of the translation go beyond the limitations specified in the contract" (UNESCO, 1976).

Even where translators' contracts are clear, they effectively have little choice in an industry where default translation and TM ownership resides with the paying client. Translators' organisations have long debated such issues in ethical terms (Blésius, 2003). Given that ultimate responsibility for a translation lies with the translator, who has invested time, money and intellectual effort in learning languages and translation skills, might she not have some ethical claim to ultimate ownership of the data she creates, even if paid to do so?

A further relevant ethical principle in most codes is that of avoiding harm (Anderson et al., 1992; ITI, 2010; SFT, 2010). Even well intentioned actions may cause harm (e.g. in this instance to translators' rates or reputation). Codes of ethics can prove a useful practical resource here, suggesting steps to mitigate this (e.g. the importance of considering impact carefully at the design stage).

Also relevant here are translation quality and subsequent use of shared TM content. For example, when accepting exact or fuzzy matches in sensitive translation domains (e.g. medical devices), it is important to know what QC steps are standard in the data provider's workflow, and arguably, who provided the original translation. If subsequent misuse of TM data caused harm, who would bear responsibility? Possible candidates include the original translator, the original client who commissioned the translation, the original agency, which managed the translation, the new translator, the new client, the new agency or TDA. Addressing issues of attribution and conditions for reuse is thus important on ethical and legal grounds, as well as the purely technical or copyright-related.

To conclude, while the LSE addresses some ethical issues more effectively than Google Toolkit, neither confronts some broader ethical questions directly. Our case study highlights some general points where existing ethical codes can offer valuable suggestions.

3 Conclusion

A final reason to highlight ethics in this context is more positive in nature. A compelling case can be made that these two initiatives are profoundly ethical in their general aims and ambitions when they are considered in the light of ethics codes.

Ethical principles which are relevant here include the importance of professional review and accessing informed critiques from peers to improve quality and raise standards; improving public understanding; contributing to society and human well-being; respecting human diversity; supporting fellow professionals; contributing to the standing of one's profession; and enhancing quality of life.[12] These initiatives share resources with neglected linguistic communities and attempt to satisfy communication needs, which might otherwise remain unmet.

Directly engaging with ethical issues is thus valuable not simply on defensive grounds (e.g. avoiding prosecution or justifying actions against potential criticism): it can allow for a strong positive case to be made for action rather than inaction.

[11] See "1.6 Give proper credit for intellectual property", *ACM Code of Ethics and Professional Conduct*, op. cit. The SFT Code (SFT, 2010) goes further, stating that translators have authorial rights.

[12] See (Anderson et al., 1992), sections 1.1, 1.4, 2.4, 2.7, 3.2, 3.6; (ITI, 2010), sections 3.2.1, 4.4.4; and (SFT, 2010), sections 3, 4.

References

Anderson, Ronald E. Gerald Engel, Donald Gotterbarn, Grace C. Hertlein, Alex Hoffman, Bruce Jawer, Deborah G. Johnson, Doris K. Lidtke, Joyce Currie Little, Dianne Martin, Donn B. Parker, Judith A. Perrolle, and Richard S. Rosenberg. 1992. *Association for Computing Machinery Code of Ethics and Professional Conduct*, http://www.acm.org/about/code-of-ethics.

Babych, Bogdan, Anthony Hartley and Debbie Elliott. 2005. Estimating the Predictive Power of N-Gram MT Evaluation Metrics Across Language and Text Types. In *The Tenth Machine Translation Summit (MT Summit X),* Phuket, Thailand, pp. 412–418.

Babych, Bogdan, Anthony Hartley and Serge Sharoff. 2007. Translating from Under-Resourced Languages: Comparing Direct Transfer Against Pivot Translation. In *The Eleventh Machine Translation Summit (MT Summit XI)*, Copenhagen, Denmark, pp. 29–35.

Bell, Andrew. 2010. Proactive professionals. *ITI Bulletin. The Journal of the Institute of Translation & Interpreting*. March–April 2010:14–16.

Blésius, Corinne. 2003. Copyright and the Translator: Who Owns your Translations? *ITI Bulletin. The Journal of the Institute of Translation & Interpreting*. November–December 2003:9–12.

Brown, Peter F., John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer and Paul S. Roossin. 1990. A Statistical Approach to Machine Translation, *Computational Linguistics*, 16(2):79-85.

Desillets, Alain. 2007. Translation Wikified: How will Massive Online Collaboration Impact the World of Translation? In *Translating and the Computer 29*. London, UK: Aslib, NP.

Eisele, Andreas and Yu Chen. 2010. MultiUN: A Multilingual Corpus from United Nations Documents. In *LREC 2010: Proceedings of the Seventh International Conference on Language Resources and Evaluation,* Valletta, Malta, pp. 2868–2872.

Esselink, Bert. 2000. *A Practical Guide to Localization*. John Benjamins, Amsterdam and Philadelphia, PA.

Guerberof, Ana. 2009. Productivity and Quality in MT Post-Editing. *The Twelfth Machine Translation Summit (MT Summit XII), Beyond Translation Memories: New Tools for Translators Workshop*, Ottawa, ON, Canada, 8pp.

ITI, 2010. *Institute of Translation and Interpreting Code of Professional Conduct* (Individual Members), http://www.iti.org.uk/pdfs/newpdf/20fhcodeofconductindividual.pdf.

Kelly, Nataly. 2009. Myths about Crowdsourced Translation. *Multilingual*. December 2009:62–63.

Koehn, Philipp. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *The Tenth Machine Translation Summit (MT Summit X)*, Phuket, Thailand, pp. 79–86.

Prior, Marc. 2010. Google Translate (Put Down your Crucifixes). *ITI Bulletin. The Journal of the Institute of Translation & Interpreting*. May–June 2010:8–11.

SFT, 2010. *Code de déontologie des adhérents de la Société française des traducteurs* [French Translators Society Code of Ethics], http://sft.fr/code-de-deontologie-des-traducteurs-et-interpretes.html.

UNESCO, 1976. *Recommendation on the Legal Protection of Translators and Translations and the Practical Means to Improve the Status of Translators*. (22 Nov. 1976) http://portal.unesco.org/en/ev.php-URL_ID=13089&URL_DO=DO_TOPIC&URL_SECTION=201.html