



Human-machine interaction: how to integrate plain language rules in the revision cycles of Neural Machine Translation output

Christopher Gledhill, Maria Zimina-Poirot

► To cite this version:

Christopher Gledhill, Maria Zimina-Poirot. Human-machine interaction: how to integrate plain language rules in the revision cycles of Neural Machine Translation output. De Europa: European and Global Studies Journal, 2023, De Europa European and Global Studies Journal, Special Issue - 2022 (1), pp.149-172. hal-04050523

HAL Id: hal-04050523

<https://u-paris.hal.science/hal-04050523>

Submitted on 2 May 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

Human-machine interaction: how to integrate plain language rules in the revision cycles of neural machine translation output

Christopher Gledhill, Maria Zimina

Abstract: Neural Machine Translation (NMT) is fast becoming the most important tool available in the translation industry. But because state-of-the-art NMT does not explicitly learn to capture stylistic variations from the data, it is still necessary to rely on professional human translators to obtain consistent, high-quality translations. Institutional consumers of translation usually set out their translation quality requirements in Style Guides, with recommendations to use normalised language, including Plain Language (PL). For example, the European Union has made PL a key goal for its online multilingual communication. Here we argue that trainee translators must be aware of such requirements. This position raises a number of questions relating to teaching advanced revision skills on the basis of NMT output. We share our experience of building the principles of PL into the 2nd year Masters course Website Translation Towards English. We then examine how the traditional workflow of post-editing/ revision has been profoundly affected by custom NMT. We suggest that the integration of PL is an opportunity to add value to the translation project as well as promoting the notion of high-quality translation for a variety of text types. To achieve these goals, we propose a solution called “Qualitative Translation/ Revision Workflow” (QTRW).
Keywords: Neural Machine Translation, Plain Language, Style Guide, Website Translation Project, Translator Training

Résumé : La traduction automatique neuronale (TAN) devient une technologie majeure dans le secteur de la traduction. Actuellement, les avancées de la TAN ne permettent pas encore de saisir explicitement la variation stylistique lors de l'apprentissage automatique à partir de données. Par conséquent, il est toujours nécessaire de faire appel à un-e traducteur-trice professionnel-le afin d'obtenir des traductions cohérentes et de qualité. Les institutions font référence à des Codes de rédaction pour exprimer leurs attentes en matière de qualité de la traduction, avec des recommandations explicites pour l'utilisation d'une langue normalisée, telle que la « langue claire » (LC). Ainsi la politique linguistique de l'Union européenne repose sur l'utilisation de la LC pour sa communication multilingue en ligne. Dans ce contexte, nous estimons que tou-te-s les apprenti-e-s traducteur-trice-s doivent être formé-e-s pour faire face à ces exigences. Cette position soulève plusieurs questions relatives à l'enseignement de la traduction et l'acquisition de compétences avancées en révision, notamment sur la base des résultats de la TAN. Dans cette contribution, nous partageons d'abord notre expérience de l'intégration des principes de la LC dans le cours de Master de deuxième année en Traduction de site web vers l'anglais. Nous examinons ensuite comment les moteurs spécialisés de la TAN impactent les flux de travail de post-édition/ révision. Finalement, nous suggérons que l'intégration de la LC apporte une valeur ajoutée aux projets de traduction et permet de promouvoir et défendre la notion de service de qualité dans la traduction d'une grande variété de textes. Pour atteindre ces objectifs, nous proposons un système de processus garantissant une qualité optimale en traduction/ révision que nous appelons « Qualitative Translation/ Revision Workflow » (QTRW).

Mots-clés : traduction automatique neuronale, langue claire, Codes de rédaction institutionnels, projet de traduction de site web, formation aux métiers de la traduction

Introduction

In this paper we explore one major impact of recent advances in Neural Machine Translation (NMT) on a Master's degree-level course in translation at Université de Paris: the need to teach advanced revision skills on the basis of NMT output. We argue that a key component of these skills is the ability to address the quality requirements of end-users relating to normalised language, in particular Plain Language (PL). PL is now required by the communication policies and Style Guides of many organisations. Yet the principles of PL are not always well defined, nor implemented uniformly in source texts. As a consequence, NMT output can be deficient in terms of PL, necessitating many cycles of post-editing and revision. For trainee translators, who are also learning to revise and edit texts in their non-primary language, getting to grips with the principles of PL can be an added difficulty.

In the following discussion, we share our experience of integrating the principles of PL into the teaching workflow of a website translation project (a second year Masters course called "*Traduction de site web vers l'anglais*" delivered in French, hereafter TSA). We start off by highlighting the relevance of PL from our perspective as practitioners and teachers of specialised translation. We then discuss the problems of integrating PL and other competencies into a translation course which uses a NMT platform as its methodological focus. This leads us to examine how the traditional workflow of post-editing/ revision cycles has been profoundly affected by the affordances offered by NMT in terms of customisation and translation memory matching. Finally, we conclude by weighing up the challenges of building PL into both the teaching and NMT workflow. We suggest here that PL is an overlooked feature in current debates on the impact of NMT, and that educators, project managers and trainee translators need to treat the integration of PL as a challenge, but also an opportunity for adding value to the translation project as a whole.

1. Why Plain Language?

Plain Language originated as a rights campaign, whose advocates (such as the Plain Language Foundation¹) have long argued for the principles of PL to be applied to official texts in administration, the law, medicine, etc., either from the point of view of citizens' or consumers' rights (campaigning for equal access to health care, readable contracts, etc.). It is significant that as a result of such lobbying, PL has been adopted as an editorial policy in many jurisdictions in the English-speaking world, for example the 2010 'Obama Law' in the USA, affecting federal and state communication with the public; the adoption of PL in all government communications in New Zealand (Cutts 2013), etc. Other language areas have also adopted the principles of PL, notably at the European level (as we see below) or at the national level (texts produced by the *Direction de l'information légale et administrative*, Benoît-Barnet *et alii* 2002).

We acknowledge that research on simplified language has never conclusively found that plain varieties of language are unequivocally beneficial for the producer of the text or the consumer, even though there are undeniable benefits in terms of document harmonisation (O'Brien 2003). Having said this, and given the widespread influence of PL in administration, the law and other areas, it appears to us that

¹ Plain Language Foundation: <https://www.plainenglishfoundation.com/> (accessed 9 July 2021).

language specialists should at least be aware of the concept of PL, and if possible, should acquire knowledge of how to write clearly as a valuable transferable skill. This has certainly been the argument made by legal rewriters such as Balmford (2002) and the many commercial organisations offering to redraft technical texts on demand, as well as offering certification in PL for technical communicators².

It is perhaps worth adding that Plain Language is not Controlled Language (CL). CLs are designed for a specific form of technical communication (including spoken and written varieties) and are prescribed in explicit linguistic terms (O'Brien 2003; Hartig, Lu 2014), PL on the other hand has no linguistic definition. We have previously attempted to set out some of the features of PL summaries written for non-expert medics³. On the basis of corpus data, we found clear evidence for a distinctive plain style, which contrasts with the style of abstracts written for experts. However, this descriptive approach is not widespread: for the most part, when editorial guidelines discuss PL, their recommendations are usually limited to a few general principles (Jelicic *et alii* 2016). Since PL is not a formally defined variety of language, but rather a discourse practice or a set of communicative goals, we can only propose a functional definition for the term: a conscious stylistic strategy designed to present expert knowledge to non-experts in clear style, that is to say in an accessible, user-oriented and self-contained form.

A lack of formal definition should not be seen as an obstacle: as we demonstrate below, there are good practical reasons why trainee translators and editors cannot and should not be tied to one language variety, largely because they are dealing with many different types of text, and thus require a greater degree of flexibility than that afforded by a strictly defined CL.

2. Style Guides and the requirement to adopt Plain Language

We start our discussion by examining the concept of the 'Institutional Style Guide', looking specifically at how these texts attempt to normalise language output, especially in terms of Plain Language. Of the 35 competencies that the European Master's in Translation (EMT) network defines as necessary for translation students following courses associated with the EMT network, several are relevant to the TSA course:

- 4 [...] mastering [...] presentation standards, terminology and phraseology [...]
- 10 [...] using the appropriate metalanguage and applying appropriate theoretical approaches
- 11 [...] review and/or revise their own work and that of others
- 18 Master the basics of MT and its impact on the translation process [...]
- 19 Assess the relevance of MT systems in a translation workflow [...]
- 23 Work in a team, [...] using current communication technologies [...]
- 29 Clarify the requirements, objectives and purposes of the client [...]

In this paper, we are interested in skills such as 18 and 19 above, but especially skill 5:

² In particular, Balmford was the founder of <cleardocs.com>.

³ That is to say the linguistic characteristics of Plain Language Summaries of expert medical texts (Systematic Reviews) written for non-expert decision-makers (Gledhill *et alii* 2019).

5 Implement the instructions, style guides, or conventions relevant to a particular translation

In other words, all professional translators should at least be aware of the concept of a Style Guide, and project managers should consider building the concept into their design of the project workflow. We also suggest to our students that if their future clients do not have their own Style Guide (some of whom work for small organisations, and are thus not aware of issues such as document or language normalisation), then it is the responsibility of the translator to point this out and propose a default guide of their own.

Assuming we need to integrate a Style Guide into our website translation course, the next question is: which one? There exist a number of freely-available guides which can serve as reference documents (e.g. the 80-page UNESCO Style Manual). However, they are often highly specific to the organisations which publish them. For the website translation project, we find it useful to refer to two sets of documents, which set out recommendations given to EU translators working in English:

- 1 The English Style Guide (ESG), produced by the Directorate-General for Translation, Brussels⁴;
- 2 The Europa Web Guide (EWG), produced by the European Commission⁵.

The English Style Guide (ESG) is relatively detailed, although only the first 60 pages (of 128 pages) are relevant to website translation, including topics such as Punctuation, Spelling, Capitalisation, Names, Titles, Numbers, Abbreviations, Units, Lists, Inclusive Language, etc. The Europa Web Guide (EWG) covers similar ground, but is shorter (25 subsections, each only one or two paragraphs long). Because our students are primarily involved in translating the websites of academic institutions and charitable associations, we give priority to the EWG over the ESG when there are differences in the recommendations (details such as list format, use of dashes, etc.).

In all of the Institutional Style Guides that we have consulted, clear style is one of the first topics they deal with. The European Union has particularly embraced the concept of PL, producing a number of public relations documents about clear writing in various languages⁶. PL is consequently promoted as the preferred writing style in both the ESG and EWG. The EWG presents PL in highly positive terms, offering professional and ethical reasons for adopting PL, and reassuring the writer/translator that PL is not going to make their text sound less elegant. One section of the EWG raises the issue of users with reading difficulties (a question that is addressed by the concept of Easy Language (Hansen-Schirra, Maaß 2020)). Another section “Jargon and plain language alternatives” presents a list of approved and disapproved terms. As far as linguistic structures are concerned, both ESG and EWG

⁴ English Style Guide: A handbook for authors and translators in the European Commission, https://ec.europa.eu/info/sites/default/files/styleguide_english_dgt_en.pdf (accessed 9 July 2021).

⁵ Europa Web Guide: The official rulebook for the European Commission's web presence, <https://wikis.ec.europa.eu/display/WEBGUIDE/02.+Style+guide> (accessed 9 July 2021).

⁶ For example, “Claire’s Clear Writing Tips”:

https://ec.europa.eu/info/sites/default/files/clear_writing_tips_en.pdf, “*Rédiger clairement*”
https://ec.europa.eu/info/sites/default/files/clear_writing_tips_fr.pdf (accessed 9 July 2021).

avoid detailed discussion (referring the reader to style manuals such as Cutts 2013), although EWG presents some practical examples of reformulation. As can be seen in figure 1, this involves features that are often found in other types of Simplified English⁷, such as:

- using active clauses instead of passives (*initiatives aimed at* > *initiatives to*),
- referring to animate participants instead of abstract ones (*initiatives, services* > *you, we, people*),
- using full verbs instead of light verbs (*make representation to* > *contact the authority*), etc.

As we can see in figure 1, the formatting of the example implies that it is possible to improve the target text by boiling it down to its essential components. The informal metalanguage is also worth noting (*Lexical Density* > *Wordiness*). For these and other reasons, our students appear to find the EWG a very useful resource.



The screenshot shows the 'Europa Web Guide' interface with a header for the European Commission. The main content area is titled 'Wordiness' and contains a table comparing two versions of text.

Don't	Do
EU initiatives aimed at protecting human health from the risks which may be caused by dangerous chemical substances	EU initiatives to protect people from dangerous chemicals [50% shorter than the original text]
Where the Commission's services make representations to the authorities of the Member State against which the complaint has been made, they will abide by the choice you have made regarding disclosure of your identity. Where you have not indicated your choice, the Commission's services will presume that you have opted for confidential treatment	If we have to contact the authorities you complained about, we will keep your identity confidential unless you state otherwise [64% shorter than the original text]

Figure 1: Recommendation on 'Wordiness' in the Europa Web Guide

3. Integrating PL into the website translation project

We argued above that there is a clear need to raise awareness of PL among our trainee translators. However, this position raises a number of theoretical and practical questions which we would like to explore in the remaining sections of this paper.

Firstly, how can we integrate PL into our website translation course? As we see in the following sections, this is a course which requires students to acquire a broad range of specialised skills, including professional, technological and linguistic competencies. Obviously, it is not enough to simply present the recommendations on plain writing (as on the EWG webpage) and then move on to some other topic.

⁷ For example, see the rules of ASD Simplified Technical English (ASD-STE100): <http://www.asd-ste100.org> (accessed 9 July 2021).

Instead, we argue that it is necessary to actually build the principles of PL into the core structure of the teaching and assessment workflow.

However, any attempt to integrate PL into the teaching workflow raises a second question: how can we build PL guidelines into a workflow in which NMT is the central pillar? As we see below, there are many linguistic and textual features that are mentioned in PL guidelines that are difficult to be built into a NMT system, given the current state of the technology and the current methods by which NMT is implemented.

Finally, a related question concerns timing: at what point in our teaching workflow should we apply PL guidelines? During the whole project? Before the project begins? Or perhaps – an experiment we are currently trying out – during a final cycle of revision at the very end of the project, thus presenting PL guidelines as a final exercise of ‘normalisation’ or ‘harmonisation’.

3.1 The webpage as text type: a challenge for our students and NMT

All text types vary in style, and few translators or linguists will be surprised to learn that a webpage mobilises very different discourse strategies and patterns of language from other types of text. We can illustrate this by taking an example from the website of the French Federation of Diabetics (FFD), one of our translation projects in 2020-2021.

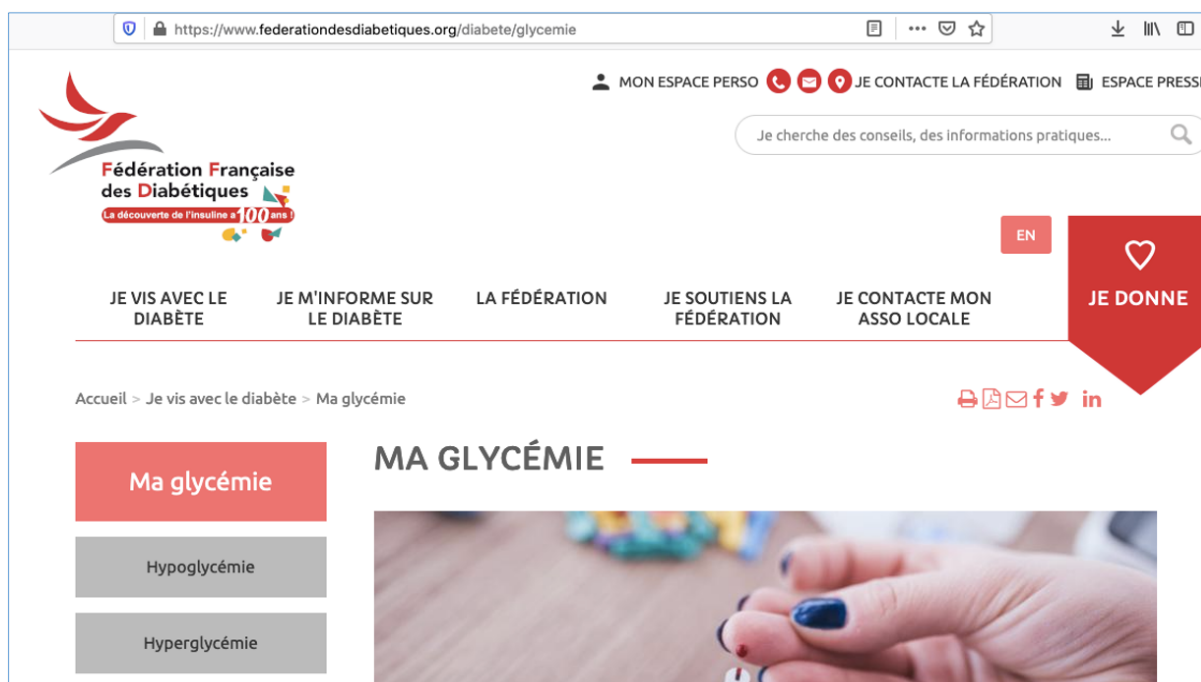


Figure 2: Source webpage in French (www.federationdesdiabetiques.org)

In figure 2 we can see that each lexical and grammatical choice in French is adapted to the interpersonal context of the website (calling for help, addressing the reader as a participant) as well as the textual context (fitting short phrases into small menu boxes). It is also notable that most of the information on this page has an intertextual function (helping the reader navigate somewhere else).

Thus, website translation presents many challenges including many varieties of discourse, and the average website project might involve translating the following:

- Menu items/main text items
- Links to interconnected websites
- Search forms, links to administrative documents
- Job adverts, lists of personnel, personal weblogs (blogs), biographies, obituaries, etc.
- Meta-texts, including guidelines, legal notices, contact information, etc.

It comes as a surprise to some of our students (especially those who have followed traditional translation courses), that they have to deal with all aspects of a webpage, including the *péritexte* (via text-navigation features). However, it soon becomes clear that there are just as many interesting translation problems to be solved, especially when it comes to post-editing and revising: the results of automatic translation of short text segments, such as menu items, demonstrate that the use of extended context is crucial to improve NMT technology (Tiedemann, Scherrer 2017). In figure 3 you can see the FFD webpage (cf. Fig. 2) automatically translated from French into English using a generic neural MT engine: SYSTRAN Pure Neural® Server (SPNS)⁸. SPNS is a translation platform which seamlessly integrates with existing website architecture and helps our trainee translators handle website content translation/ revision without extensive knowledge of HTML/ XML formatting rules. It is deployed on a university intranet available to our Masters students. It is important to point out that this page was initially translated using a generic model (FR-EN) which had not been trained for a specific domain or topic. Thus a brief look at figure 3 reveals several examples of low-quality machine translation into English. For example, the informal abbreviation in French “ASSO” is inappropriately translated as “ASSO” in English. Another typical error concerns the menu items expressed as first-person clauses in French (“JE VIS”, “JE M’INFORME”, “JE CONTACTE”, etc.). In particular, one of the main menu items on the French page of the association “[heart symbol] JE DONNE” (cf. Fig. 2) is clumsily translated by “[heart symbol] I GIVE” (cf. Fig. 3). But rather than point out such errors immediately, our question to the students is as follows: “Are these the best translations for these segments?”. If not, we will clearly need to reconsider the other translation choices on the same page.

⁸ See the official product website: <https://www.systransoft.com/translation-products/systran-pure-neural-server> (accessed 9 July 2021).

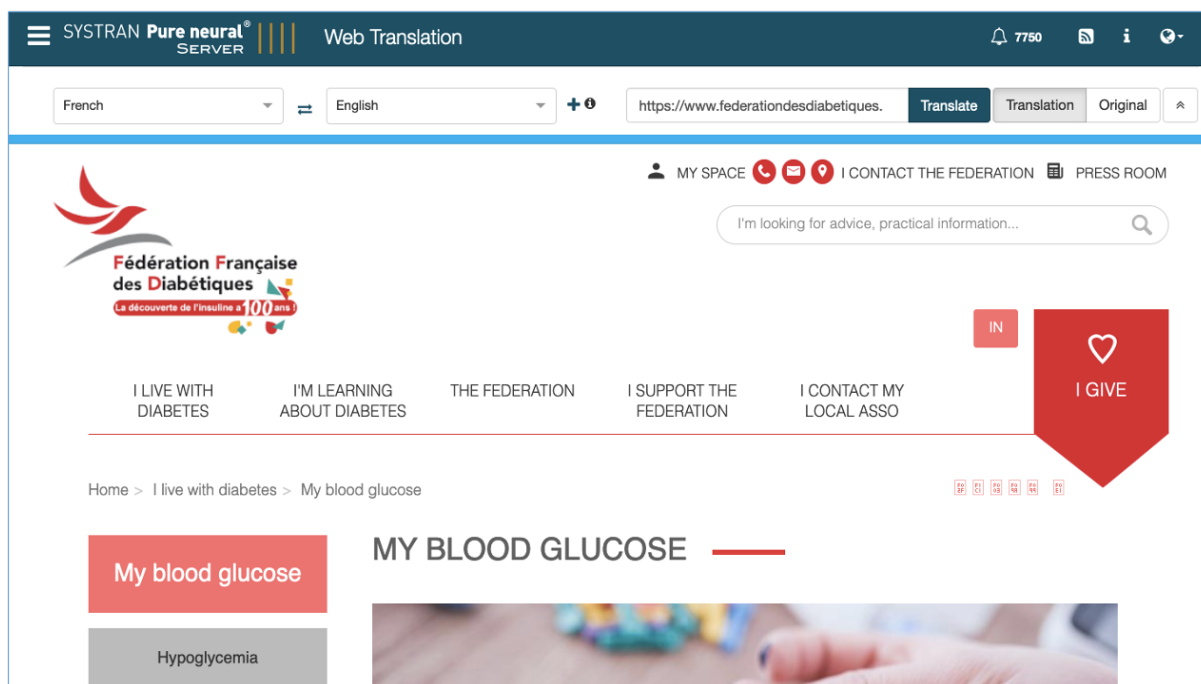


Figure 3: Initial NMT output in English (www.federationdesdiabetiques.org)

In the specific case of “[heart symbol] I GIVE”, the students found that this is a typical discourse pattern on French charitable web pages, but not in English. Figure 4 shows how the page looks after our students researched alternative possibilities, including “DONATE” (an unfamiliar form of the verb) and “MAKE A DONATION” (which would be too long for the menu item, etc.) before they arrived at a participle verb (“DONATING”). This solution led to an interesting debate in the group about the tendency for English to prefer imperatives such as “SUPPORT” and “CONTACT” for links to procedural pages, as opposed to participles such as “LIVING WITH” for links to descriptive pages. Finally, it is notable that there is still a persistent error on the published page “I CONTACT THE FEDERATION” which requires further revision (a point which we return to below, cf. Fig. 10).

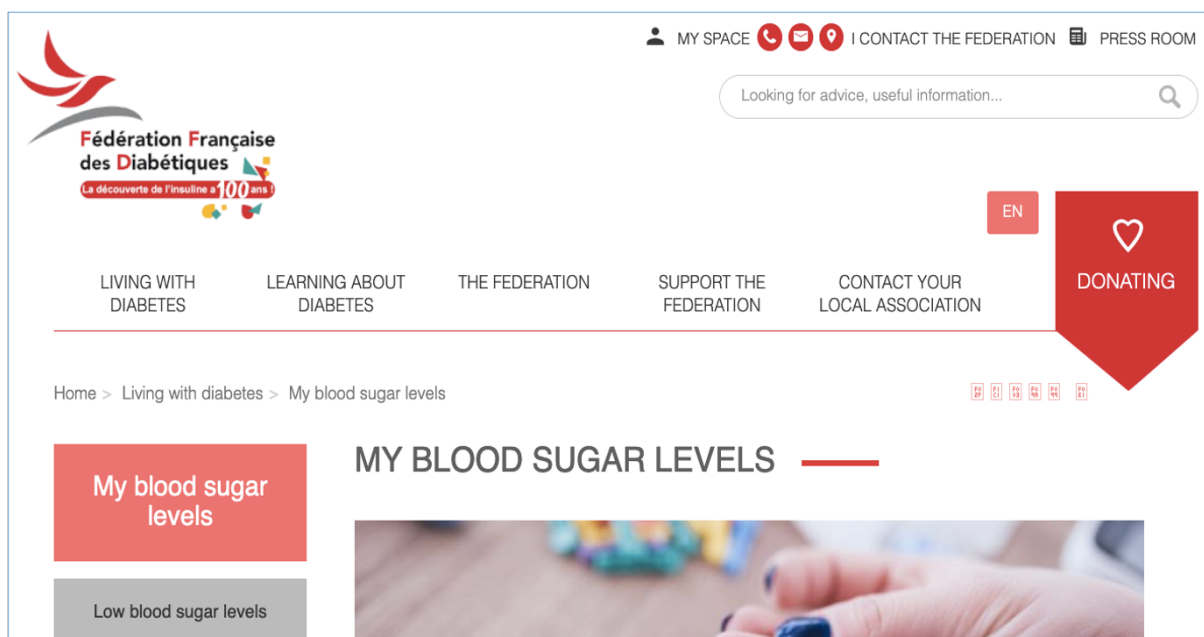


Figure 4: First revision of the webpage translation (www.federationdesdiabetiques.org)

Overall then, webpage translation is complex and requires several cycles of revision. Because every segment is context-sensitive, it follows that the quality and accuracy of NMT can vary significantly from one text segment to another, especially if the model used is untrained for this type of text.

It should come as no surprise therefore that, given such problems, as well as the sheer volumes of text involved, not all website owners are able to maintain high quality standards of translation, especially if they have to update their pages regularly, or deal with multiple languages⁹. In this case, an exclusion or limitation clause, or disclosure strategies are used to indicate that the MT output may contain errors of content or form. It is also considered good practice to clearly indicate that this translation process involves no human post-editing or revision. This type of statement can even be found on the European Union website regarding the use of eTranslation:

Machine translation can give you a basic idea of the content in a language you understand. It is fully automated and involves no human intervention. The quality and accuracy of machine translation can vary significantly from one text to another and between different language pairs. The European Commission does not guarantee the *[sic]* accuracy and accepts no liability for possible errors.¹⁰

However, this is not the paradigm we attempt to follow with our students. We inform them that it is our policy to adhere to the following quality statement (as set out by the European Commission Directorate-General for Translation):

⁹ We would like to thank an anonymous reviewer who mentioned the relevance of single-sourcing software (Hysell 2001) for website translation projects. Although this resource would be clearly relevant to this kind of project, not all institutional websites rely upon this type of software design.

¹⁰ Use of machine translation on Europa. Exclusion of liability: https://ec.europa.eu/info/use-machine-translation-europa-exclusion-liability_en (accessed 9 July 2021).

DGT's Quality Management Framework calls for our translations to be fit for their intended communicative purpose to satisfy the expressed or implied needs and expectations of our direct customers, our partners in the other EU institutions, the end-users, and any other stakeholders.

Since our objective is to achieve these quality requirements for all of the different texts we translate, we propose a solution that we call “Qualitative Translation/Revision Workflow” (QTRW). We set out the details of this process in the following two sections.

3.2 Getting the best out of NMT: from generic output to customised NMT

The first stage in implementing QTRW involves a discussion of the different tools used in the project, most notably the professional translation platform SYSTRAN Pure Neural[®] Server. SPNS offers a broad range of features relevant to the website translation project, including¹¹:

- Managing collaborative translation
- Identifying website updates
- Maintaining consistency
- Instantly showing how translations will look online
- Enabling Neural MT specialisation with custom resources
- Using Translation Memory and User Dictionary
- Producing output translation in HTML/XML format
- Helping with localisation

We now explore some of these features to explain to what extent they can be used to influence writing style and to improve translation quality. First, it is important to stress that it is not enough to use generic NMT models to obtain quality output: rather it is necessary to create custom NMT workflow (Senellart *et alii* 2003). However, this requires effort and skills, in particular a thorough understanding of the available data and linguistic resources (such as training corpora, translation memories, terminology, etc.). Thus, our students need to be involved at every stage of the process of NMT implementation, as we demonstrate below.

Figure 5 shows a custom NMT profile that delivers a reasonable translation quality improvement using a selection of available linguistic resources (normalisation dictionaries, bilingual terminology, translation memory combined with a translation model (FR-EN) selected in the SPNS ‘Profiles’ menu.

¹¹ Financed by a “Projet Pédagogique Université Paris-Diderot” (five-year contract: 2019-2023). Funding extension granted within the framework of the call for projects “Scientific Platforms and Equipment”, Université de Paris, 2021: PAPTAN/DL4MT@UP (Deep Learning for Machine Translation at Université de Paris).

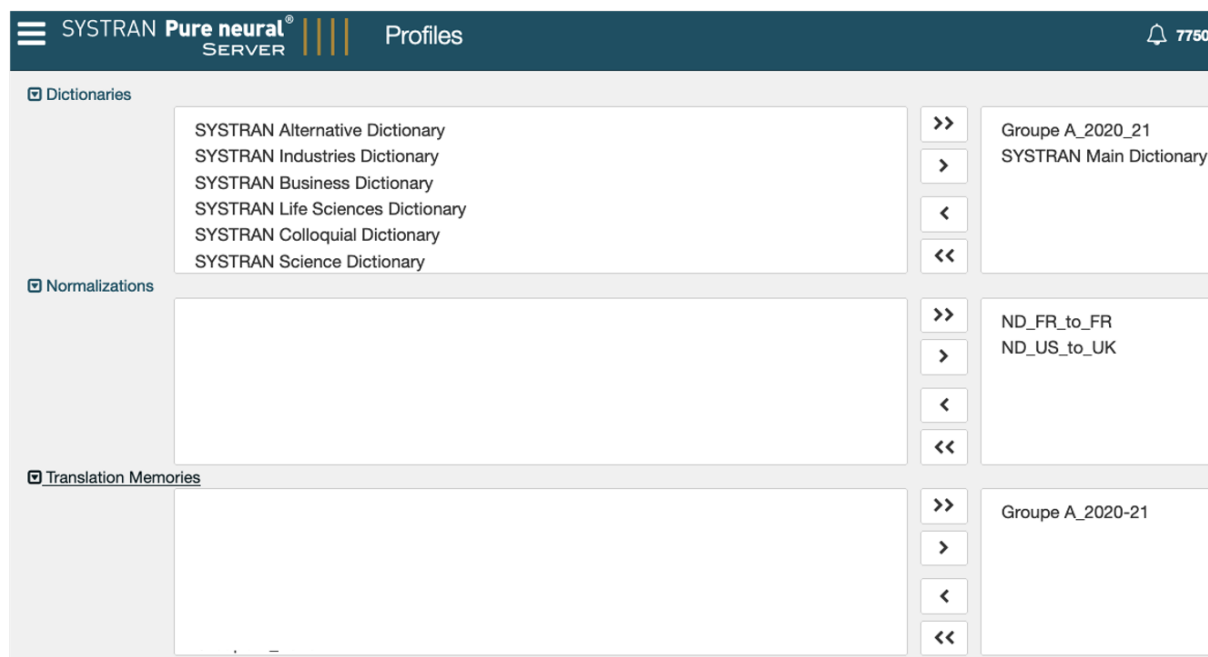


Figure 5: Neural MT specialisation: custom translation profile with selected linguistic resources

Along with rapid NMT customisation, a ‘specialised’ translation model can be developed to update an existing NMT engine, which will then deliver higher quality translation output than generic machine translation (Servan *et alii* 2016). For example, figure 6 shows the output of generic NMT compared to that of ‘specialised’ NMT for medical discourse. It can be noted that the target segment “fasting and 2 hours after meals” becomes “on an empty stomach and 2 hours after meals” in the output of the domain-specific model.

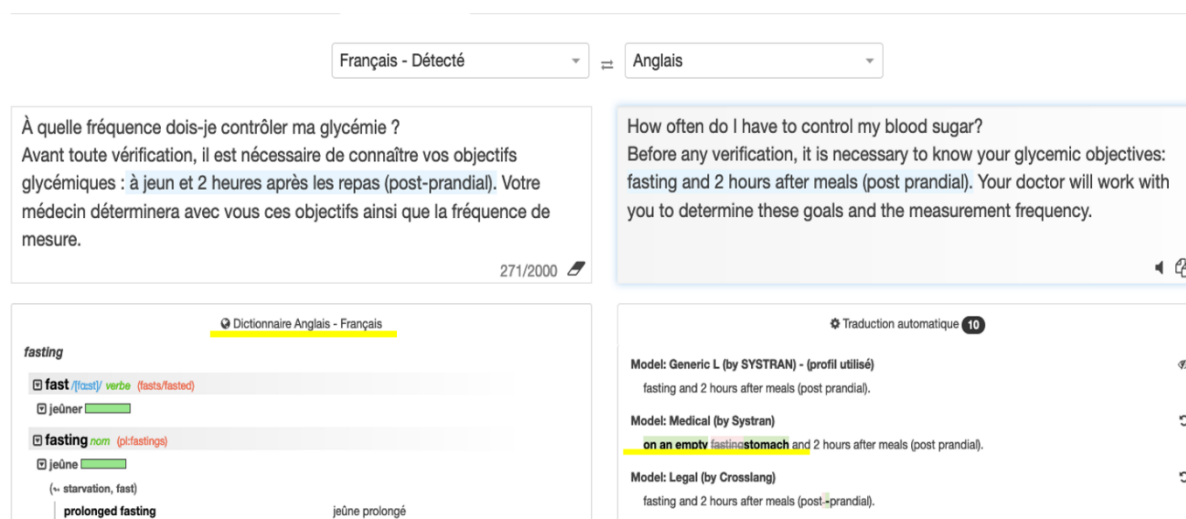


Figure 6: Generic NMT compared to that of ‘specialised’ NMT for medical discourse

As mentioned, customising NMT requires large amounts of domain-specific training data and resources. It is said that approximately 100 000 aligned text segments are

needed to create a translation model from scratch¹². Thus, every customisation is different, and requires a considerable amount of investment in terms of teaching time as well as learning curve.

As further demonstration of NMT specialisation, figure 7 shows a paragraph from the FFD website translated into English using three different translation models. The first one is a generic model. The second was customised for medical English. The third is a domain-specialised model maximising access to factual information relating to the coronavirus. This example shows the impact of taking into account writing style. Figure 7 also shows that if the writing style does not match purpose, then the content must be revised (as you can see on the left-hand side in “Student translation: work in progress”).

À quelle fréquence dois-je contrôler ma glycémie ?
Avant toute vérification, il est nécessaire de connaître vos objectifs glycémiques : à jeun et 2 heures après les repas (post-prandial). Votre médecin déterminera avec vous ces objectifs ainsi que la fréquence de mesure.

273/2000

Student translation: work in progress
How often do I need to control my blood sugar levels?
Before monitoring your blood glucose levels, it is necessary to know your blood sugar level targets: when fasting and 2 hours after meals (postprandial). Your doctor will agree with you on the target level that you need to aim for and the frequency of measurement.

How often do I have to control my blood sugar?
Before any verification, it is necessary to know your glycemic objectives: fasting and 2 hours after meals (post prandial). Your doctor will work with you to determine these goals and the measurement frequency.

Model: Generic L (by SYSTRAN) - (profil utilisé)

How often do I need to control my blood sugar?
Before any verification, it is necessary to know your blood sugar objectives: on an empty stomach and 2 hours after meals (post prandial). Your doctor will work with you to determine these objectives and the measurement frequency.

Model: Medical (by Systran)

How often should I monitor my blood sugar?
Before any tests, it is necessary to know your glycemic goals: fasting and 2 hours after meals (post-prandial). Your doctor will determine with you these goals and the frequency of measurement.

Model: CoronaCrisisCorpus (by TAUS) - (profil utilisé)

Figure 7: A paragraph in medical French translated using three different translation models

An additional point is that ‘specialised’ NMT models can be combined with normalisation dictionaries to correct spelling, units of measurement, formatting issues, etc. Such custom resources do not implement general linguistic patterns or rules. Normalisation allows for one-for-one replacements of the matched character strings in the source text (pre-editing step) and in the target text output (post-editing step). In this respect, state-of-the-art NMT is quite different from rule-based MT.

For example, figure 8 shows how UK spelling (a feature of normalisation that is imposed by current EU Style Guides) is implemented on a case-by-case basis.

¹² Guidelines for successful corpus collection prior to Specialization. SYSTRAN Pure Neural® Server for End-Users Expert Users and Administrators. Systran, 24 February 2020, Université de Paris.

Nom du fichier :

ND_US_to_UK

Langue

Anglais

Ajouter

Supprimer

		Source	POS	Cible	
<input type="checkbox"/>		bicolor	Adjectif	bicolour	4
<input type="checkbox"/>		bisulfate	Nom	bisulphate	4
<input type="checkbox"/>		bisulfite	Nom	bisulphite	4

Figure 8: Custom NMT: SPNS Normalisation Dictionary for UK-EN spelling

Custom NMT models can also incorporate a User Dictionary (UD) to include project terminology and a bilingual glossary. The SPNS dictionary module implements one-for-one replacements with linguistic coding rules (singular, plural, verb conjugation, “do-not-translate”). Examples of this can be seen in figure 9, with items of project-specific terminology outlined in blue: the French-English UD entries are automatically propagated from the bilingual glossary database to the translated text. The use of a fine-tuned customised model adhering to terminology constraints considerably reduces post-editing effort and improves translation consistency (Michon *et alii* 2020), especially in collaborative translation. However, although they are particularly suitable for controlled languages, UD are limited when it comes to polysemous terms with multiple translations.

		Source	POS	Target
<input type="checkbox"/>		sommaire	Noun	contents
<input type="checkbox"/>		stylo	Noun	insulin pen
<input type="checkbox"/>		traitement oral/injectable	Expression	oral or injectable medication
<input type="checkbox"/>		Équilibre	Proper Noun	Équilibre

Pour mesurer votre glycémie, il vous faut :

- un lecteur de glycémie : l'appareil n'est remboursé par la Sécurité sociale tous les 4 ans qu'en cas de diabète traité par insuline ou de rétinopathie diabétique.
- des bandelettes ou électrodes (stockées à sec).
- un stylo autopiqueur à usage strictement personnel.
- des lancettes (fines aiguilles) : à usage unique, à adapter à l'autopiqueur.

392/10000

To measure your **blood sugar levels**, you need:

- a **glucometer**: the device is only reimbursed by the Social Security every 4 years in case of diabetes treated by **insulin** or **diabetic retinopathy**.
- strips or electrodes (stored dry).
- **lancing device insulin pen** for strictly personal use.
- lancets (thin needles): for single use, adapted to the **lancing device**.

Figure 9: Project-specific terminology in NMT output

Finally, along with translation glossaries, it is possible to upload Translation Memories (TM) for a rapid customisation of a translation profile using database files (*.TMX) consisting of the original source texts and their respective translated

segments. The machine translation system automatically reuses any segments from the database that are the same or similar (called ‘fuzzy’ matches) on new projects. TM resources are thus especially meaningful for human translators when it comes to NMT customisation. For group translation projects, sharing translation memory also guarantees total consistency of terminology, as it ensures that all translators are translating keywords and phrases uniformly¹³. However, before using a TM in a group project, an editor needs to review all included segments to ensure consistency. In the future, new generation TM search engines may be able to identify similarity not only of lexical items but of longer stretches of discourse, such a routine lexicogrammatical patterns (Zimina 2019). For example, figure 10 presents TM matches for the source segment identified in the revised version of the webpage that we analysed earlier (“*Je contacte la Fédération*” > “I contact the Federation”, cf. Fig. 3).

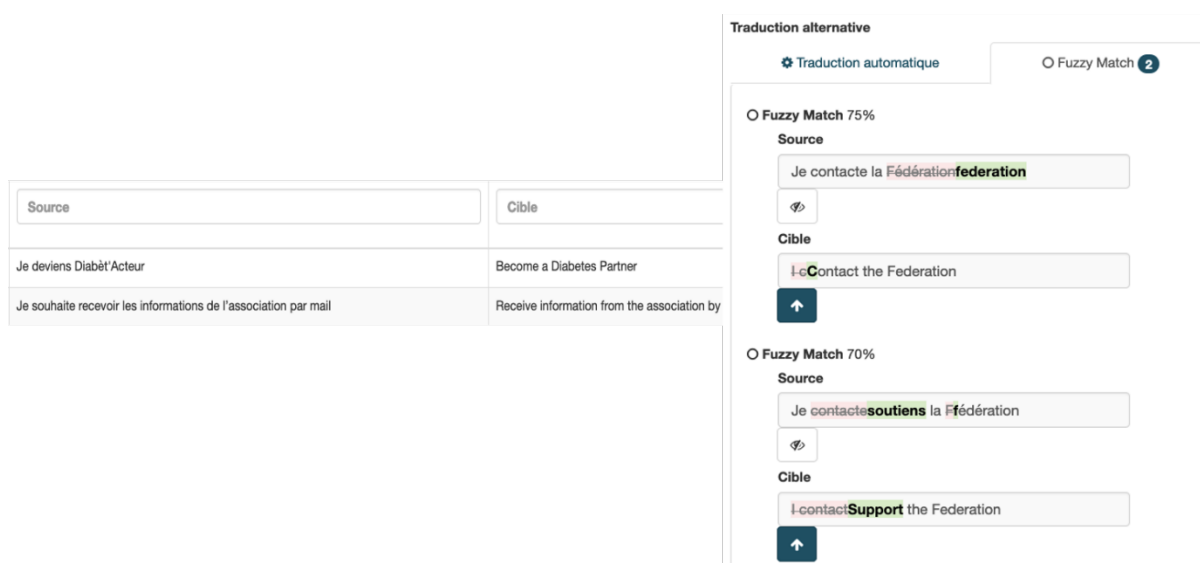


Figure 10: Finding a match for “*Je contacte la Fédération*” (“Contact the Federation”) with a TM management system

For segments to be reused, the TM management system relies on lexical similarities, in this case fuzzy matches of previously translated similar phrases, such as “*Je soutiens la fédération*”. If the system encounters a novel segment, such as “*Je deviens Diabète'Acteur*” (literally, “I become a Diabetes Actor”) or “*Je souhaite recevoir les informations*” (literally, “I want to receive information”) which have both been translated using an imperative (“Become”, “Receive”, etc.), it should ideally be able to spot not just the formal similarity (no lexical matches in this case), but also the similar rhetorical function of these constructions.

The goal of the NMT search engine is to reduce human translator efforts by leveraging high-potential sentences through pattern matching. In other words, the translator should have the final say, but only when presented with a judicious selection of potential translation solutions. This point brings us back to the question of how to achieve efficient human-machine interaction within a Qualitative Translation/Revision Workflow.

¹³ On TM features in computer-aided translation:

<https://www.memsource.com/features/translation-memory> (accessed 9 July 2021).

3.3 Optimising human skills in the website translation project

It would not be accurate to divide a translation project like TSA into ‘custom machine translation’ on one hand and ‘human intervention’ on the other. As we have seen, our students are heavily involved in the project on both sides of the equation, and in two different but interdependent capacities. First, they have to be text analysts and data managers, in that they need to learn how to customise the NMT itself. Yet even after the custom NMT models have produced detailed and accurate output, there are still many tasks which are not ‘solved’ by the existence of NMT output, including revision, task management, delivery, and so on. Thus, our students have to be both language experts and project managers; they need to learn how to validate the quality of the NMT output, at the same time as complying with the specific quality requirements of the clients.

It is sometimes difficult for our students to appreciate both of these aspects simultaneously. One response to this problem has been to develop specific roles for each participant. Therefore, at the beginning of the project, we assign different job titles to each of the students, enabling the overall project to be broken down into manageable microtasks. For example, the auditing team determines the overall architecture of the website and decides which pages on the website can be translated; the terminologists manage not only technical terms but also the many repeated segments that need to be translated consistently across the website and implemented in the User Dictionary (such as the menu items we saw in paragraph 3.2); the task managers record which pages have been assigned to which translators and revisers, etc. In future iterations of the TSA course, we plan on asking the students to change roles, so that they will benefit from different vantage points on the project at different times.

Space precludes a full discussion of the many other aspects which need to be considered in terms of human resource management (as reported in Gledhill, Zimina 2019). So, we focus here on the one aspect of the project which requires considerable linguistic confidence on the part of the translators, and demonstrates very clearly the limits of machine translation, even intensively trained custom NMT: the need to re-write the target text in plain style.

As outlined above, we require our students to be familiar not only with the most relevant recommendations of two EU Style Guides (ESG and EWG), but also with the general principles of plain writing, which are implicit in these guidelines. We achieve this by asking our students to integrate many of the general principles of PL into the preparation of custom NMT (relevant terminology, spelling conventions, units of measurement, etc.). We also ask the students to act as editors and perform one ‘final’ cycle of revision (called ‘language harmonisation’) in which they are meant to apply the different recommendations of the Style Guide and PL requirements to the latest version of the translated website. Once again, this task is broken down into microtasks, so that one team of students will be tasked with checking for gender-inclusive language (a requirement of the ESG), another team looks at excessive nominalisation (so-called ‘zombie-nouns’, as recommended by the EWG), etc.

There are two objectives to this exercise. Firstly, we want our students to adopt a systematic approach to the different types of linguistic errors and other types of

recommendations that are involved in PL (and in the revision process in general, of course). Secondly, we want our students to be aware of the metalanguage in which these features are defined, and to adopt a systematic approach to classifying them. Although our students are often good at spotting inconsistencies, they are often excessively vague about the status of the different errors they have encountered, simply treating all problems as “not sounding right” (in English), “*bizarre*” (in French), etc. In an attempt to correct this ‘blanket’ approach to error analysis, we therefore ask them to categorise all of the edits they make in terms of four very broad categories, as set out here¹⁴:

- a) Phonology-graphology
- b) Lexico-grammar
- c) Discourse-semantics
- d) Context (of culture/ situation)

In the following subsections, we present examples of how different features of PL (or other items that belong to the requirements of our two style manuals: ESG and EWG) have been identified and classified by our students in previous translation projects (Gledhill, Zimina 2019). The point of this final exposition is to highlight the range of different categories of linguistic features that are simply not covered by NMT, thus requiring the linguistic skills of our students in identifying them as well as sometimes also finding appropriate translation solutions.

a) Revisions at the level of phonology-graphology¹⁵

This level of analysis covers problems of formatting, orthography, punctuation, typography, etc. Most of the recommendations in the EU Style Guides are concerned with formal problems of this type, and most of the errors identified by our students are to be found at this level also. Superficially, it may seem that these features are trivial, but issues such as adopting the appropriate date format, respecting the conventions on capitalisation in English, the formatting of lists/ headings, etc. can collectively have an impact on text harmonisation.

Some issues are straightforward when viewed out of context (such as the requirement to write the date *14 April 2021* instead of *April 14th 2021*, etc.). But as the following example shows (PG1: taken from a student evaluation form), even the question of whether to use an initialism requires concentration and clear-thinking on behalf of the translator/ editor:

PG1

Enriquillo-Plantain-Garden is used twice before the abbreviation EPG appears in brackets. From then on EPG is used. This is inconsistent. Changed English so that Enriquillo-Plantain-Garden fault (EPGF) is used once, followed by its abbreviation. From then on, the abbreviation is employed. (J. W., TSA student project reporting form).

¹⁴ The four ‘strata’ of language and other aspects of our metalanguage are based on Systemic Functional Grammar (Halliday, Matthiessen 2014).

¹⁵ Although we are dealing with written documents, we retain Halliday’s term ‘phonology’ so that we can potentially include audio translation. This also allows us to discuss the relation between punctuation and prosodic features, etc.

In addition, it is important for our trainee translators to be aware of the fact that ‘high value texts’ (such as the policy documents produced by the EU) must obey strict rules of formatting, especially when these texts are to be used as the primary texts for translation into many other languages.

b) Revisions at the level of lexico-grammar

As we go up one level, we arrive at the stratum of the lexico-grammar (the lexical and grammatical “resources for construing meanings as wordings” (Halliday, Matthiessen 2014: 131)). At this level many students are less sure of themselves in English, choosing to report simple morphological errors in the initial NMT output. However, more confident students are happy to engage with sophisticated issues of syntax and phraseology (selecting the appropriate combinations of words according to discourse type). Example LG1 shows a simple example where the student has chosen to apply the PL recommendations on ‘avoiding the passive’, as we can see in the following example (LG1, where FR: original French, EN0: machine translation output, EN1: first revision, etc.):

LG1

FR: *Le sud de l'île d'Haïti est traversée par une faille décrochante majeure : la faille d'Enriquillo-Plantain-Garden (EPGF).*

EN0: The south of the island of Haiti is crossed by a major stalling fault: Enriquillo-Plantain-Garden fault.

EN1: The Enriquillo-Plantain-Garden fault (EPGF), a major lateral fault, cuts across southern Haiti [...]

Example LG2 shows a more sophisticated type of reformulation: ‘de-nominalisation’, the rewording of a process noun by a clause, thus making the participants in the clause explicit. LG2 is complex because it involves an increased degree of nominalisation in the NMT output (EN0 and EN1), followed by interventions by the translators/ editors who progressively reformulate the nouns as clauses (EN2). This example gives a good picture of how this type of sophisticated reformulation is typically introduced over a series of different cycles of revision rather than all at once:

LG2

FR: *Ces résultats amènent donc à repenser la structure et le fonctionnement de ces écosystèmes [...]*

EN0: These results lead to a rethinking of the structure and functioning of these particular ecosystems [...]

EN1: These results thus allow researchers to reconsider the composition and functioning of these specific ecosystems [...]

EN2: These results thus allow researchers to reconsider what these specific ecosystems were made of, and how they worked [...]

c) Revisions at the level of discourse-semantics

As we go up another level, we arrive at the level of discourse-semantics (“the set of strategies for construing, enacting and presenting non-language as meaning” (Halliday, Matthiessen 2014: 189)). This allows students to discuss problems of cohesion/ coherence as well as terminology (conceived here as exophoric reference). This level of analysis also includes reformulations above the level of the clause (seen as the upper limit for problems of lexico-grammar). The following example (DS1, identified by student M. H.) shows a distinct problem at this level of analysis: over-complex phrase structure, in this case involving multiple subordination in FR (which is carried over into EN):

DS1:

I found this sentence in French quite difficult to follow in order to understand the different steps. Therefore I decided to split the sentence so as to make the English version smoother. (M. H., TSA student project reporting form).

FR: *Toutes les lunes formées en dessous de l’orbite synchrone, dont Phobos, chutent vers Mars, mais Deimos s’étant formé juste au-dessus de l’orbite synchrone il est repoussé vers l’extérieur.*

EN0: All the moons formed below the synchronous orbit, including Phobos, fall to Mars, but Deimos having formed just above the synchronous orbit it is pushed back to the outside.

EN1: All moons that were formed below the synchronous orbit, including Phobos, are being pulled towards Mars. // Deimos, however, was formed just above the synchronous orbit and is therefore being pushed outward.

EN2: All the moons that were formed below the synchronous orbit, including Phobos, are being pulled towards Mars. // Deimos, however, was formed just above the synchronous orbit and is therefore being pushed outward.

As we can see in DS1, our students are usually very reactive when it comes to ‘chopping’ texts down to size like this (here signalled by a double slash //). And as discussed above, this type of intervention is entirely in keeping with the recommendations of the EWG on sentence length/complexity.

d) Revisions at the level of context

At the highest stratum of analysis, we have context of culture/ context of situation. These terms refer to errors which are significant because of cultural expectations, or due to a specific, transient feature of the situation¹⁶. Problems at this level are sometimes quite serious, but also consequently not frequent, and students sometimes have difficulty identifying them. With experience however, it is almost always possible to identify contextual problems, right across the website.

Example CS1 represents a very clear problem of cultural context, in that the NMT has assigned a gendered determiner “his” in English to translate the French possessive “sa”:

CS1

FR: *une entreprise, qui confie à un doctorant un travail de recherche objet de sa thèse ;*

¹⁶ Both concepts originate in the contextualist theory of J.R. Firth, as mentioned by Anne Condamines (this volume).

EN0: a company, which entrusts to a Ph.D. student a research work that will be the subject to his thesis;

EN1-2: a company which entrusts research work to a Ph.D. student on the subject of his/her thesis;

This gender-specific usage is proscribed in official documents in many English-speaking countries, and likewise the recommendations of both the ESG and EWG stipulate that when gender does not need to be specified, then the authors must adopt an alternative strategy (i.e. gender-neutral pronouns/ determiners).

Example CS2 represents a different but also very typical problem, this time relating to the context of situation:

CS2

FR: *Les résultats seront affichés en ligne en janvier 2017.*

EN0: Results will be posted online in January 2017.

EN1-2: Results will be published online as soon as possible.

Here the source text has not been updated, and so the NMT output has just translated the out-of-date information as presented. In the absence of immediate feedback from the website owner, it is often necessary for our students to be proactive, and so in this case it was decided to provide a neutral ‘holding’ statement.

Finally, example CS3 shows a problem of editorial policy. As this relates to the expectations of the end-users and website owners, it is analysed as context of culture:

CS3

FR: *un récit des principales observations du Laboratoire de Physique du Globe effectuées [...]*

EN0: an account of the principal observations of the Physics laboratory of the Earth [...]

EN1-2: a narration of the main observations of the Laboratoire de Physique du Globe [...]

Generally speaking, generic NMT model will attempt to translate the names of institutions (especially less well-known ones). However, in this case it turns out that the organisation wants to maintain its identity and French name, so the translator/ reviser has to intervene. Here the NMT output can be adjusted with a User Dictionary, as explained in paragraph 3.2.

Conclusion

Neural Machine Translation (NMT) is a working tool, and is fast becoming the most important tool available in the translation industry. But the human translator still has a vital role to play. As we have set out in a number of instances in this paper, even when NMT output is customised, there can be significant discrepancies between this output and the principles set out in an Institutional Style Guide, in terms of cohesion, consistency, variations, standards and drafting codes, etc.

Controlled Languages are rule-based and can be implemented using machine translation fairly consistently. But as we have seen, the guidelines for Plain Language are essentially subjective and open to a degree of interpretation. It follows that if

the customer's text is to be written in 'clear style', 'plain writing', etc., then the translators/ revisers actively need to carry out a number of interventions. As demonstrated above, such edits are difficult to implement in NMT systems, and even if there were a means of customising the NMT to produce Plain Language output, this would in many ways defeat the object: custom NMT output is by definition adapted to a specific customer's purpose, while the concept of Plain Language is necessarily flexible, contingent and can vary not only from project to project, but even from one page of a website to another (and even conceivably from one section of the page to another...). For the time being, therefore, we must conclude that it is necessary to raise awareness of Plain Language among trainee translators: they need to learn how to mobilise both their knowledge of the language system and of editorial standards.

Regarding 'Human-Machine interaction', our present position is this: in any professional translation project such as the translation of a whole website, the place of the machine is to produce an initial translation. After this step, the role of revision as a process and as a human skillset is critical to the successful completion of such a project. But because of the increasing complexity of this process, it is just as important for trainee translators to acquire the appropriate skills in designing and implementing a translation/ revision workflow that attempts at every turn to promote quality output, as it is for them to intervene pro-actively in project management and in the several complex cycles of post-editing and revision which are still necessary for this type of project. It is this dual relationship and responsibility which we have attempted to characterise here as a translation project design philosophy, and which we term "Qualitative Translation/ Revision Workflow" (QTRW).

References

Balmford Christopher (2002). "Plain language: beyond a movement". Proceedings of the Plain Language Association International (PLAIN) Fourth Biennial Conference, Toronto, Canada. September 26-29, 2002. Copian Library. Available online: <http://en.copian.ca/library/research/plain2/movement/movement.pdf>¹⁷

Benoît-Barnet Marie-Paule, Collette Karine, Laporte Danielle, Pouëch Françoise, Rui-Souchon Blandine (2002). *Guide pratique de la rédaction administrative*. Paris: Ministère de la fonction publique et de la réforme de l'état.

Cutts Martin (2013). *Oxford Guide to Plain English*. Oxford: Oxford University Press.

Direction de l'Information Légale et Administrative (DGLFLF/OQLF) (2006). *Rédiger... simplement*. Québec: Editions de la Délégation générale à la langue française et aux langues de France/ Office québécois de la langue française.

Gledhill Christopher, Martikainen Hanna, Mestivier Alexandra, Zimina Maria (2019). "Towards a linguistic definition of 'Simplified Medical English': applying textometric

¹⁷ Last accessed 9 July 2021 (all the links in References).

analysis to Cochrane medical abstracts and their plain language versions.” *Lingue Culture Mediazioni/ Languages Cultures Mediation*, 11, 91-114.

Gledhill Christopher, Zimina Maria (2019). “The Impact of Machine Translation on a Masters Course in Web Translation: From Disrupted Practice to a Qualitative Translation/Revision Workflow.” *Proceedings Translating and the Computer 41*, November 2019. Editions Tradulex, Geneva, pp60-73. Available online at : <http://www.tradulex.com/varia/TC41-london2019.pdf>.

Halliday Michael, Matthiessen Christian (2014). *Introduction to Functional Grammar*. 4th revised edition. London: Routledge.

Hansen-Schirra Silvia, Maaß Christiane (eds) (2020). *Easy Language Research: Text and User Perspectives*. Berlin: Frank and Timme.

Hartig Alissa, Xiaofei Lu (2014). “Plain English and legal writing: Comparing expert and novice writers”. *English for Specific Purposes*, 33, 87-96.

Hysell Deborah A. (2001). “Single sourcing for translations”. Proceedings of the 19th annual international conference on Computer documentation (SIGDOC'01). Association for Computing Machinery, New York, NY, USA, 89-94. Available online: DOI:<https://doi.org/10.1145/501516.501535>

Jelicic Kadic Antonia, Fidahic Mahir, Vujcic Milan, Saric Frano, Propadalo Ivana, Marelja Ivana, Dosenovic Svjetlana, Puljak Livia (2016). “Cochrane Plain Language summaries are highly heterogeneous with low adherence to the standards”. *BMC Medical Research Methodology*, 61/16. Available online: <https://bmcmmedresmethodol.biomedcentral.com/articles/10.1186/s12874-016-0162-y>

Michon Elise, Crego Josep, Senellart Jean (2020). “Integrating Domain Terminology into Neural Machine Translation”. Proceedings of the 28th International Conference on Computational Linguistics, 3925-3937. Available online: <https://doi.org/10.18653/v1/2020.coling-main.348>

O’Brien Sharon (2003). “Controlling Controlled English. An Analysis of Several Controlled Language Rule Sets”. Proceedings of EAMT-CLAW-03, Dublin City University: Dublin, 15-17 May 2003, 105-114.

Senellart Jean, Boitet Christian, Romary Laurent (2003). “XML Translation Workflow”. Machine Translation Summit IX, September 2003, New Orleans, United States. Available online: <https://hal.inria.fr/inria-00487747>

Servan Christophe, Crego Josep, Senellart Jean (2016). “Domain specialization: A post-training domain adaptation for Neural Machine Translation”. Submitted to EACL 2017. Available online: <http://arxiv.org/abs/1612.06141>

Tiedemann Jörg, Scherrer Yves (2017). “Neural Machine Translation with Extended

Context”. Proceedings of the Third Workshop on Discourse in Machine Translation, 82-92. Available online: <https://doi.org/10.18653/v1/W17-4811>

Zimina Maria (2019). “Vers une Mémoire de Traduction dynamique et multidimensionnelle”. *Des mots aux actes*, 8, 221-236.