Data Privacy and MT Engines

I know some of you might not be enthusiastic about me writing again about data privacy when using generic machine translation (MT) engines like Google, Microsoft, and DeepL. This is partly because I've done so a number of times already.¹ Also, I think many might be using the data privacy issue as a kind of marketing ploy that's just too good to let go—even though it's not exactly truthful (more on that below).

Now, I'm under no illusion that whatever I write here or elsewhere holds more weight than whatever someone else might write. But I want to make really sure I understand the admittedly very important data privacy issues, so I'm just taking you (once again) on that journey with me.

The question is this: Is my clients' data privacy assured when I, as their translator, use services like Google Translate, Microsoft Translator (or whatever it might be called at this particular point in time), or DeepL?

Let's start with times when using these engines is not safe or ethically defensible. (Note that I'm not going to talk about the use of MT in general, just about whether it's safe to trust Google, Microsoft, or DeepL to use the data you transmit to them only for the purpose of suggesting an MT-generated translation to you and nothing else.)

First, it's not ethically defensible if your client expressly prohibits it. That's it as far as that point is concerned. It might be that the client is ill-informed about why they prohibit this, but that's clearly not your concern. If they say don't do it, you don't do it.

Second, it's not safe to use any of those services if you use their web interface at **translate.google.com**, **bing.com/translator**, **deepl.com/translator**, or through apps of any of those companies that offer MT for free (exception: Microsoft Office products—see below). These companies expressly say that they very well might use your data to improve their services.

- Here's what Google says: "We also collect the content you create, upload, or receive from others when using our services (...) And we use your information to make improvements to our services. For example, understanding which search terms are most frequently misspelled helps us improve spell-check features used across our services."² While this doesn't specifically pinpoint translation services, it's my understanding that they are included (as well as Gmail and myriad other Google services). If you've been using the web interface for Google Translate while logged into Google, you can select the History icon at the bottom of the page to see what Google has actually stored in the last three or so months.

- Here's what Microsoft says: "Microsoft Translator processes the text, image, and voice data you submit, as well as device and usage data. We use this data to provide Microsoft Translator, personalize your experiences, and improve our products and services."³

- And here's what DeepL says: "When using our translation service, please only enter texts that you wish to transfer to our servers. This is necessary

Is my clients' data privacy assured when I, as their translator, use services like Google Translate, Microsoft Translator, or DeepL?

in order for us to produce the translation and offer you our service. The transfer of these texts is necessary for us to carry out the translation and offer you our service. We process your texts and the translation for a limited period of time to train and improve our neural networks and translation algorithms. If you make corrections to our proposed translations, these corrections are also forwarded to our servers to verify the accuracy of the corrections and, if necessary, to update the translated text to reflect your changes. We also store your corrections for a limited period of time to train and improve our translation algorithm."[4]

So far so good. Good? Yes, I think this is good for us because it differentiates the casual user of MT from those of us who use MT as one of our resources during professional translation. Because what we (should!) do is access MT from those sources via their application programming interface (API—how different programs exchange information). And if we access it within a translation environment (e.g., Trados, memoQ, Memsource, etc.), that's exactly what we're doing.

Here's what the different systems say about that:

- **Google:** "Google does not use any of your content for any purpose except to provide you with the Cloud Translation API service."[5]

- **Microsoft**: "Azure Cognitive Services Translator is a cloud-based machine translation service and is part of the Azure Cognitive Services family of cognitive APIs for building intelligent apps. Customer data submitted for translation to Azure Cognitive Services Translator (both standard and custom models), Speech service, the Microsoft Translator Speech API, and the text translation features in Microsoft Office products are not written to persistent storage. There will be no record of the submitted text or voice, or any portion thereof, in any Microsoft data center. The audio and text will not be used for training purposes either."[6]

- **DeepL:** "When using DeepL Pro, the texts or documents you submit will not be permanently stored and will only be kept temporarily, to the extent necessary for the production and transmission of the translation. Once you have received the translation, all submitted texts or documents and their translations will be deleted. When using DeepL Pro, your texts will not be used to improve the quality of our services."[7]

It seems relatively clear to me, but a) I'm not a lawyer, and b) all too often fellow translators or other technology providers like to throw shade on those provisions by pointing to other sections in the

legal thickets of those companies that might read like loopholes to those conditions. If the skepticism arises out of real doubt about whether that data might be treated differently than outlined in the legal statements above, it's not only justified but laudable. But in other cases, I seem to notice a stubbornness borne either of wanting to sell a product or service that in some way competes with those generic MT offerings (a sales pitch masquerading as moral high ground), or just a general rejection of MT in all its forms (or any combination of the two). I think we have to be careful about taking stands that might be hard to defend, especially when it comes to the core of our business as translators or translation technology providers.

Plus, it has always seemed kind of preposterous to assume that professional translators have so much to add to the ongoing collection of data that it would even make a dent in the billions of times non-API users access the data and enter text. (Remember, we're only talking about source data here, unless you would be using a tool's interface to make corrections to the translation data.) Would these companies really embarrass themselves by not keeping what clearly seems to be a contractual promise?

Either way, I thought it would be helpful to actually reach out to some people from these organizations to see what they actually know about their company's

plan for data submitted through their APIs. I did contact someone at Google who essentially confirmed the contractual agreement, though he was very eager not to go on record with anything that could get him into hot water with Google's legal team. (I remember when interviewing the former head of Google's MT years ago, two members of the legal team sat right next to him and weighed every word that came out of his mouth). But I was very grateful to Microsoft's Chris Wendt—or rather former Microsoft employee Chris Wendt, who happened to retire just days after I asked him (Happy Retirement!). Here's what he said:

"When using the Translator API, free or paid, or a commercial application like Office, no customer content will be stored by Microsoft. When using a Microsoft consumer app, the Microsoft Translator app for the phone or bing. com/translator, Microsoft may save the customer content and use it for quality improvement. We recently changed the phone app to specifically ask for permission before storing customer content.

There is a difference between customer personal data and customer content. Customer content is the payload of the translation request. Customer personal data identifies the customer, like the subscription ID, email address, physical address, the internet provider the request came from, and similar

information. The services, including Microsoft, do maintain personal data in order to send the bill, ensure fairness, and throttle the service. That's why the explanation of what happens with personal data is somewhat lengthy. What I say above is about customer content (payload). Not about the metadata associated with the use of the service."

And, just for clarification, I asked again: "Is it correct that when using the paid API services to obtain translation from Microsoft (with or without Custom Translator), there's no case where the source data will be used by Microsoft?" And Chris' answer: "That's correct. Not the translation either."

And all of the above is by no means me arguing that you or anyone should use MT. I have no dog in that fight (it's really not a fight in the first place), but I think it's really important to be clear about the legal ramifications. Most of the articles written about MT are about customized MT systems. It's possible to use customized systems—either provided by clients or through systems like the ones above that we ourselves can train. Although the fact is that most translators don't have access to customized systems (either because the clients don't provide them or because translators work in too many different fields and sub-fields to spend time training engines), so it's these kinds of systems that many are using. And it's good to know exactly what that means. **ATA**

NOTES

1 Zetzsche, Jost. "Data and the Fine Print, or How to Create a Sh*tstorm," *The ATA Chronicle* (March 2015), *http://bit.ly/MT-engines*.

2 Google Privacy Statement, *https://policies.google.com/privacy*.

3 Microsoft Privacy Statement, *https://privacy.microsoft.com/en-us/privacystatement*.

4 DeepL Privacy Statement, *www.deepl.com/en/privacy*.

5 Google Cloud Data Usage FAQ, *http://bit.ly/google-data-usage*.

6 Microsoft Confidentiality Statement, *https://bit.ly/Microsoft-confidentiality*.

7 DeepL Privacy Statement, *www.deepl.com/en/privacy*.

Jost Zetzsche is chair of ATA's Translation and Interpreting Resources Committee. He is the author of *Characters with Character: 50 Ways to Rekindle Your Love Affair with Language*. *jzetzsche@internationalwriters.com*

This column has two goals: to inform the community about technological advances and encourage the use and appreciation of technology among translation professionals.