All Your Data Are Belong to us . European Perspectives on Privacy Issues in 'Free' Online Machine Translation Services

Pawel Kamocki, Jim O'regan, Marc Stauch

«All your data are belong to us »*. European Perspectives on Privacy Issues in 'Free' Online Machine Translation Services.

Paweł Kamocki[1 2 3], Jim O'Regan[4], and Marc Stauch[5]

[1] L'Université Paris Descartes, 75006 Paris, France
[2] Westfälische Wilhelms-Universität Münster, 48149 Münster, Germany
[3] Institut für Deutsche Sprache, 68161 Mannheim, Germany
[4] Centre for Language and Communication Studies, Trinity College Dublin, Ireland.
[5] Leibniz Universität Hannover, 30167 Hannover, Germany

Abstract. The English language has taken advantage of the Digital Revolution to establish itself as the global language; however, only 28.6% of Internet users speak English as their native language. Machine Translation (MT) is a powerful technology that can bridge this gap. In development since the mid-20th century, MT has become available to every Internet user in the last decade, due to free online MT services. This paper aims to discuss the implications that these tools may have for the privacy of their users and how they are addressed by EU data protection law. It examines the data-flows in respect of the initial processing (both from the perspective of the user and the MT service provider) and potential further processing that may be undertaken by the MT service provider.

Keywords: personal data, Machine Translation, privacy, Directive 95/46/EC, Google Translate

Introduction

The digital revolution, which started with the proliferation of personal computers in the late 1970s and continues to the present day, has (just as any revolution worthy of its name) changed our everyday life in more ways than one may want to admit. Most importantly, new modes of communication developed in this Digital Age allow people to exchange information across the globe within seconds. A live chat with a contractor from another continent or an online search for the most obscure items sold in the four quarters of the Earth is now as easy as pie. Or is it. . . ?

Not yet, and for a reason as old as the hills: the language barrier. Even though English has undoubtedly taken advantage of the Digital Revolution to establish

* This is a paraphrase of '*All your base are belong to us*', a phrase from the 1991 video game *Zero Wing*, poorly translated from the original Japanese. The phrase has become an Internet phenomenon which had its peak in 2004 (see: Know Your Meme, `http://knowyourmeme.com/memes/all-your-base-are-belong-to-us`).

itself as the global language[1]; it has recently been estimated that it is used only by 54.3% of all websites (as of November 24, 2015). Moreover only 26% of Internet users speak English as a native language (as of June 30, 2015) [Internet World Stats]. While it is true that a certain percentage of the remaining Internet users speak (some) English as their second or third language, it remains a fact that a substantial part of the global Internet community does not speak it at all and, thus, can only take advantage of a fraction of the content available on the World Wide Web. In response the Digital Revolution has provided a number of tools for linguistic support. Foremost among these is the technology of Machine Translation (MT).

In this paper, while accepting the importance of MT as a prima facie beneficial technology for enhancing global communication, we aim nonetheless to consider a problematic aspect to it that has so far received little attention: this is the potential incompatibility of the technology – or at least the typical way it is made available to users – with key tenets of data protection law (our focus is EU data protection rules). We shall next describe further the development of the technology, and the way it operates, so as to provide some context for the subsequent legal analysis.

1 MT in Context

MT (or automatic translation) can be defined as a process in which software is used to translate text (or speech) from one natural language to another. This section will briefly present the history of MT and various technologies used in the process.

1.1 History

The idea to mechanize the translation process can be traced back to the seventeenth century [Hutchins, 1986, chap. 1]; however, the field of machine translation is usually considered to have begun shortly after the invention of the digital computer [Koehn, 2010, chap. 1]. Warren Weaver, a researcher at the Rockefeller Foundation, published a memorandum named "Translation" in which he put forward the idea to use computers for translation [Hutchins, 1999], proposing the use of Claude Shannon's work on Information Theory to treat translation as a code-breaking problem[2].

During the Cold War, researchers concentrated their efforts on Russian-to-English (in the US) and English-to-Russian MT (in the USSR). In January 1954 the first public demonstration of an MT system (used to translate more than sixty sentences from Russian to English) took place in the headquarters of IBM (the so-called Georgetown-IBM experiment) [Hutchins, 2004]. In the following

[1] which was pointed out as early as 1997: Crystal [2012, p. 22]

[2] Weaver also cited work by McCullough and Pitts on neural networks, which, due to recent advances in "deep" neural networks, have been applied to Statistical Machine Translation.

years, imperfect MT systems were developed by American universities under the auspices of such players as the U.S. Air Force, Euratom or the U.S. Atomic Energy Commission [Hutchins, 1986, chap. 4].

In 1964 the U.S. government, concerned about the lack of progress in the field of MT despite significant expenditure, commissioned a report from the Automatic Language Processing Advisory Committee (ALPAC). The report (the so-called ALPAC report), published in 1966, concluded that MT had no prospects of achieving the quality of human translation in the foreseeable future [Hutchins, 1986, chap. 8.9]. As a result, MT research was nearly abandoned for over a decade in the U.S.; despite these difficulties, the SYSTRAN company was established successfully in 1968: their MT system was adopted by the U.S. Air Force in 1970 and by the Commission of the European Communities in 1976 [Hutchins, 1986, chap. 12.1].

Research and commercial development in Machine Translation continued in the "rule-based" paradigm, in which a dictionary, a set of grammatical rules, and varying degrees of linguistic annotation are used to produce a translation, until the early 1990s, when a group of IBM researchers developed the first "Statistical Machine Translation" system, Candide [Berger et al., 1994]. Building on earlier successes in Automatic Speech Recognition, which applied Shannon's Information Theory, the group applied similar techniques to the task of French-English translation. In place of dictionaries and rules, statistical MT uses word alignments learned from a corpus [Brown et al., 1993]: given a set of sentences that are translations of each other, translations of words are learned based on their co-occurrence (the *translation model*); of the possible translations, the most likely is chosen, based on context (the *language model*).

"Phrase-based MT" [Koehn et al., 2003] is an extension of statistical MT that extends word-based translation to "phrases"[3], which better capture differences between languages. Although attempts have been made to include linguistic information [e.g. Koehn and Hoang, 2007], phrase-based MT is still the dominant paradigm in Machine Translation. Google started providing an online translation service in 2006[4], initially using SYSTRAN's rule-based system, but switched to a proprietary phrase-based system in 2007 [Tyson, 2012].

1.2 Technology and challenges

Machine Translation is used for two primary purposes: *assimilation* (to get the gist of text in a foreign language), and *dissemination* (as an input to publication, typically post-edited by translators). Free online services, such as Google Translate, are usually intended for assimilation; the translation services in use at the EU, for dissemination. Consequently, systems for assimilation may trade accuracy for broader coverage, and vice versa.

Rule-based systems can be classified into three main categories: *direct translation*, where no transformation of the source text is performed; *transfer-based*,

[3] Contiguous chunks of collocated words, rather than a "phrase" in the linguistic sense.
[4] according to the company's history posted on Google's website.

where the input is transformed into a language-dependent intermediate representation[5]; and *interlingua*, where the input is transformed into a language independent representation. Rule-based systems tend to be costly and time-consuming to build, as they require lexicographers and linguists to create dictionaries and rules. Transfer-based and interlingua, as they operate on abstract representations, generalize well: if a word is in the dictionary, it can be handled in all forms; but they tend to handle exceptions, such as idioms, poorly: specific rules must be written, and multiple rules may conflict with each other.

The prerequisite for building statistical MT systems is the existence of human-translated bilingual (or multilingual) corpora – and the bigger the better. An obvious source of professionally translated multilingual corpora are international organizations such as the United Nations or the European Union, generating a substantial amount of freely available, high-quality multilingual documents (in 24 languages for the EU[6] and in 6 languages for the UN[7]).

Compared to rule-based MT systems, statistical MT systems are cheaper (at least for widely-spoken languages) and more flexible (a statistical system is not designed specifically for one language pair, but can accommodate to any language pair for which a corpus is available). Also, because statistical MT systems are based on human-translated texts, the output of statistical MT is (or at least can be) more natural, and it naturally adapts well to exceptions (if the corpus contains the phrase, it is effectively not an exception).

Zipf's law[8] states that in a given corpus, the frequency of a word is inversely proportional to its frequency rank: the most frequent word will occur (approximately) twice as often as the second, three times as often as the third, and so on. Conversely, the majority of words (40-60%) are *hapax legomena* (words which only occur once). As statistical MT is corpus-based, it therefore suffers from the problem of *data sparsity* due to the high proportion of hapax legomena: longer phrase matches are absent from the translation model; contextual information is absent from the language model, affecting the quality ("fluency") of the output.

Data sparsity is the biggest problem in statistical MT. Although there have been attempts to solve it by using linguistic information, dating back to Candide, the most common approach is to simply add more data[9]. A large amount of websites are available in multiple languages, so crawling the web for parallel text is a common method of collecting corpora [Smith et al., 2013], particularly for the providers of free online MT, such as Google and Microsoft, who also operate search engines and therefore already have access to such data. The use of such

[5] Typically, *lemma* (the citation form of the word), *morphological analysis* (details such as case, number, person, etc.), and possibly *semantic analysis* (subject of a verb, etc.).

[6] Art. 1 of the Regulation No. 1 of 15 April 1958 determining the languages to be used by the European Economic Community.

[7] Rule 51 of the Rules of Procedure of the General Assembly of the United Nations; rule 41 of the Provisional Rules of Procedure of the United Nations Security Council.

[8] See, e.g., `https://en.wikipedia.org/wiki/Zipf\%27s_law`

[9] On the other hand, it has been claimed that translation quality can be increased by simply discarding infrequent phrases: Johnson and Martin [2007].

data, however, has its own problems, as such documents are often not just translated, but *localized*: different units of measurement, currency, and even country names [Quince, 2008], because of their collocation, become "translations".

Crowdsourcing, where online communities are solicited for content, is also in increasing use. Amazon's Mechanical Turk, an online market place for work, provides an easy way to pay people small amounts of money for small units of work, which has been used in MT [Zaidan and Callison-Burch, 2011]. Google has made use of crowdsourcing since the earliest days of Google Translate, by providing users with a means of improving translation suggestions [Chin, 2007], by providing a translation memory system [Galvez and Bhansali, 2009], and more directly, with Translate Community [Kelman, 2014].

Finally, the quality of MT output depends on the quality of the input. Even the most banal imperfections such as misspellings or grammar mistakes – not uncommon in electronic communications – even if they are barely noticeable to a human translator, can compromise the most elaborate MT systems [Porsiel, 2012].

1.3 'Free' Online MT Tools

A number of 'free' online MT services are available today. This section will present the most popular of them and try to very briefly evaluate their quality.

1.3.1 Examples

The most popular 'free' online MT service is Google Translate, which is reported to be used by 200 million people every day (in 2013) [Shankland, 2013] and to translate enough text to fill 1 million books every day (in 2012) [Och, 2012]. Launched in 2006, it can now support an impressive number of 80 languages, from Afrikaans to Zulu, including artificial (Esperanto) and extinct (Latin) languages. Google's proprietary MT system is based on the statistical approach. Google Translate is also available as an application for Android and iOS; it is integrated in Google Chrome and can be added as a plug-in to Mozilla Firefox.

Bing Translator (`http://www.bing.com/translator/`) has been provided by Microsoft since 2009. It currently supports 44 languages and is integrated in Internet Explorer, Microsoft Office and Facebook.

1.3.2 Are they 'good enough'?

Erik Ketzan argued in 2007 that the fact that MT had not attracted much attention from legal scholars was a consequence of the low quality of the output [Ketzan, 2007]. He predicted that *'if MT ever evolv[ed] to "good enough," it [would] create massive copyright infringement on an unprecedented global scale'*. While this article is not about copyright issues in MT, the question *'is MT good enough?'* remains relevant.

The user's expectations related to such services have to be reasonable, but we believe that they can be satisfied to a large extent. 'Free' online MT tools can definitely help users understand e-mails or websites in foreign languages.

Moreover, it is apparent that the quality of MT tools is improving. For example, in his article Ketzan quoted *'My house is its house'* as machine translation of the Spanish proverb *'Mi casa es su casa'*[10].

If we take into account Moore's law (according to which computers' speed and capacity double every 18 months [Moore, 1965]), as well as the exponentially growing number of digital language data that may be used to increase the accuracy of statistical MT systems, the future of MT technology looks promising. The number of users of 'free' online MT services will probably keep growing – it is therefore important to discuss the impact that such tools may have on user privacy.

2 Data protection issues in respect of data processed in 'free' online MT services

2.1 Background

'Free' online MT services allow users to translate texts of different length: from single words and phrases to multiple paragraphs (while Bing Translator is limited to 5000 characters, Google Translate can handle several times more). These texts can be of various types, including private and professional correspondence, blog entries, social media content, newspaper articles, etc. It is therefore not surprising that these texts may contain information that is sensitive from the point of view of privacy, and more specifically, constitute personal data. If we take into account the fact that MT is an integral part of such privacy-sensitive services as Gmail or Facebook, this becomes even more obvious.

The concept of personal data is defined in art. 2(a) of the Directive 95/46/EC on the protection of individuals with regard to the processing of personal data and on the free movement of such data (hereinafter: the Directive). According to this, personal data shall mean *'any information relating to an identified or identifiable natural person'*. This definition has been further analysed by the Article 29 Data Protection Working Party (hereinafter: WP29) in its Opinion 4/2007, which advocates a broad understanding of the concept[11]. In particular, according to WP29's analysis it covers not only 'objective' information (i.e. facts), but also 'subjective' information (i.e. opinions and assessments)[12]. The information 'relates to a person' not only if it is 'about' a person (the 'content' element), but also if it is used to evaluate or influence the status or behaviour of the person (the 'purpose' element), or if it has an impact on the person's interests or rights (the 'result' element)[13].

[10] This result was obtained on August 6, 2006 using Google Translate (then based on SYSTRAN); using the current version of Google Translate (October 23, 2014) the proverb is translated correctly as *'My house is your house'*.

[11] Article 29 Data Protection Working Party, *Opinion 4/2007 on the concept of personal data adopted on 20th June 2007*, 01248/07/EN, WP 136

[12] *Idem*, 6.

[13] *Idem*, 10–11

As regards the processing of personal data, the effect of the EU data protection regime is that processing this requires a legal basis, either in the form of the consent of the person in question (the 'data subject'), or another legitimate ground specified under the Directive.[14] In addition, the processing must respect certain principles of fairness, set out in art. 6 of the Directive. For its part, 'processing' is another broad concept defined in the Directive: in fact, every operation performed on data (be it manual or automatic) is 'processing' in the sense of art. 2(b).

As we have seen in the previous sections, MT services perform a series of automatic operations on input data which certainly qualifies as 'processing' of data. Though, as noted, in principle MT software installed on the user's computer could perform the translation locally (with data never leaving the computer), in practice most MT services – including the main 'free' applications offered by providers such as Google – operate by transferring the data to a remote location for processing. As described in Part 1, this is partly to take advantage of the far higher processing capacity available (allowing a superior and faster service), but is also an innate part of the 'data-capturing' business model of those providers. From the perspective of data protection law, it is the latter model, involving as it does a transfer or disclosure of the data by the MT service user to the service provider, that undoubtedly presents the key challenges.

For the purposes of this paper, we shall accordingly focus on remote MT services. The processing involved in such services can in fact be divided into two discrete stages. The first concerns the activity of the user of the MT service when he enters data into the MT tool in order to have it translated. The second stage relates to the processing that is then performed by the MT service provider, who then performs a series of operation on the input data and sends the translated output back to the user. The processing operations at this stage may reflect different purposes; most obviously, there will be a need to perform the translation as requested by the user. However, in addition there will generally be further processing (referred to below as secondary processing), where the MT service provider processes the input data for purposes, other than the return of the translation. This may be part of the evaluation and development of the service, including through the use of statistical techniques (see Part 1); another possibility is of user profiling with a view to direct marketing. The following sections will analyse the data protection implications of these various processing operations separately, as they present substantially different legal considerations.

2.2 Processing by the MT service user

For each stage of processing, it is essential to identify the data controller, i.e. 'the person who determines (alone or jointly with others) the purposes and means of the processing of personal data'[15]. It may seem that as far as initial input

[14] Art. 7 of the Directive; as discussed below, there are also 'special categories' of personal data, which are subject to more stringent conditions for processing under art. 8 of the Directive.

[15] Art. 2(d) of the Directive.

of text data into the MT tool is concerned, the user should be regarded as the controller, whereas the MT service provider is merely a processor (i.e. a person who processes data on behalf of the controller[16]). However, given that the MT provider plays a crucial role in determining the functioning of an MT service (including the required format of input data, etc), he can also be regarded as a controller[17]. Indeed, the Directive, in art. 2(d), expressly allows for there to be more than one controller in respect of a single processing operation. Moreover, it seems possible that – to the extent that the MT provider makes use of third party software in the process – even more data controllers could be identified (albeit this hypothesis will not be considered in detail in this study).[18]

From the MT service user's perspective, two main categories of personal data can be processed at this stage: data concerning the user himself and data relating to a third person. For example, a Polish person wishing to book a hotel room in Italy, may feed the sentence, "Please could I have a single room for the second half of August?" into the MT service, and communicate the result to the hotel (first-person processing). Quite often, though, there may also be an element of third-person processing, insofar as the user introduces another person within the communication, as in, "Please could I have a double room for myself and my wife for the second half of August?" A further instance of third-person processing would be when the guest writes to the Italian hotel in Polish, and the hotel (as MT service user) utilises the translation service to understand the message.

As regards the case of pure first-person processing this does not seem to raise any particular concerns as far as the lawfulness of the user's own activity is concerned. In fact, it seems that a person can always process his own data; in this situation the rights and duties of the data controller, the data processor and the data subject are all merged in one person and the idea of informational self-determination can be realised to its fullest extent. Equally, from the privacy perspective, and recalling that an aim of the Directive is to concretise the right to private and family life under art. 8 of the European Convention on Human Rights,[19] it can be argued that the legal rules are not directed at the use made by individuals of data about themselves.

Turning to cases where the MT service user processes data of another person, matters are more complex. As noted, such a scenario is presented where our hotel seeker refers to his wife as well as himself in his booking request. Here a preliminary question is whether the user might invoke the so-called 'household exemption'. According to this, processing of third person data may be exempted from the Directive if this is done in the course of a purely personal or household

[16] Art. 2(e) of the Directive.

[17] See the CJEU judgement in Case C-131/12 Google Spain SL, Google Inc. v Agencia Española de Protección de Datos, Mario Costeja González, par. 23: 'The operator of a search engine is the 'controller' in respect of the data processing carried out by it since it is the operator that determines the purposes and means of that processing'.

[18] See the Article 29 Data Protection Working Party, *Opinion 02/2013 on apps on smart devices adopted on 27 February 2013*, 00461/13/EN, WP 202, 9-13.

[19] Recitals 10 and 11 of the Directive.

activity[20]. It is not clear how to interpret this category. Textbook examples of such activities include private correspondence and keeping of address books; in its 2009 Opinion on online social networking, the Article 29 Working party also implied that sharing of information among a limited circle of 'friends' could be covered.[21] This would suggest that the use of MT tools in order to translate information about one's private activities into another language may be exempted from the Directive, as long as neither the input nor the output data are made public. The scope of the 'household exemption', however, has been recently interpreted narrowly by the CJEU in the *Rynes case*[22] concerning a camera system installed by an individual on his family home, which also monitored a public space. In its ruling the Court emphasised that the exclusion concerns not all the personal and household activities, but only those that are of purely personal and household nature. The camera system in question was *'directed outwards from the private setting'*[23] and as such it was not covered by the exemption. It is possible to draw an analogy between such a camera system and the use of an MT system, which is also 'directed outwards', as its functioning involves (which may however not be obvious to an ordinary user) transmitting data over a network. It is possible, therefore, that the private user of an MT service will have to comply with the Directive. In other cases, where a person uses the service in the course of a business (as in the example of the Italian hotel translating a message received in Polish), there will be no basis to argue the exemption in the first place.

Assuming the Directive applies, the user will need to comply with the requirements relating both to the lawfulness and the fairness of data processing. As far as the grounds for lawfulness of processing are concerned, the default legal basis for processing should be the data subject's consent. Consent is defined in art. 2(h) of the Directive as *'any freely given specific and informed indication of [the data subject's] wishes by which [he] signifies his agreement to personal data relating to him being processed'*. This definition does not require that consent be given e.g. in writing. Rather, as the WP29 suggested in its 2011 Opinion on consent, it allows for consent implied from the data subject's behaviour[24]. Therefore, it might be argued for example that if the user receives an e-mail in a language that he does not understand, he can imply the sender's consent to enter it into an MT system, especially if the sender knew that the addressee was unlikely to understand the language in which the message was written. After all, the purpose of sending an e-mail is to communicate, i.e. to be understood.

In our view, however, there are several difficulties with implying consent in this way. First of all, it will likely miss the 'informed' element, as it may be doubted that the sender of the email (assuming he has the knowledge of an

[20] art. 3(2) of the Directive.
[21] WP29 Opinion 5/2009 (WP 163), at p. 6.
[22] C-212/13, 11 December 2014
[23] idem, par. 33.
[24] Article 29 Data Protection Working Party, *Opinion 15/2011 on the definition of consent adopted on 13 July 2011*. 01197/11/EN, WP187, 11.

average computer user) fully understands the implications of having his data entered into an online MT service[25]. As such he arguably cannot validly consent to the processing unless this information is given to him up front. Secondly, the Directive sets forth other conditions for consent, namely that it is *unambiguous* (i.e. leaving no doubt as to the data subject's intention[26]; art. 7(a)). Moreover, when it comes to processing of special categories of 'sensitive' data (relating to the subject's health, sex life, political or religious beliefs, etc.), the Directive requires consent to be *explicit* (art. 8.2(a)), which excludes implied consent. Indeed this points up an underlying problem in the context of MT, namely that the recipient of a message in an unfamiliar language is not in a position to assess the contents of the data he will process. This problem would arguably persist even in a case where the sender of the message expressly invites the recipient to have it translated, for this consent can only concern processing of data relating to the data subject, and not a third person. And yet how is the recipient to tell the difference? For example, if a hotel owner uses a 'free' online MT system to translate a message "Ground floor please, my wife has an artificial hip", he may suddenly find he is unlawfully processing data concerning the sender's wife's health.

It is true that the Directive also allows alternative legal bases for processing (other than the data subject's consent), in particular when processing is necessary for the the performance of a contract to which the data subject is party or in order to take steps at the request of the data subject prior to entering into a contract (art. 7(b)); or for purposes of the legitimate interests of the data subject, the data controller or a third party (art. 7(f)). In our view, though, the same fundamental difficulty remains of ruling out (in the event sensitive data are present) that the more stringent processing conditions under art. 8 will apply.

2.3 Processing by the MT service provider

We now turn to the processing of the data by the service provider. What are the potential rights and obligations of the latter under the Directive? Here some of the major players on the 'free' online MT market (such as Google and Microsoft) could attempt to argue that the Directive does not apply to them because they are not established on the territory of an EU Member State, and nor do they use equipment situated on the territory of such a state (art. 4 of the Directive). This argument (which has already been rejected by European courts[27] and data protection authorities[28]), however, will soon be precluded as the proposed text of the new General Data Protection Regulation extends its applicability to the

[25] For a description of the discrepancy between the user expectations and reality regarding online privacy on the example of Facebook see, e.g., Liu et al. [2011]

[26] idem, 21.

[27] cf.: CJEU judgement in Case C-131/12 Google Spain SL, Google Inc. v Agencia Española de Protección de Datos, Mario Costeja González.

[28] see e.g. Déliberation de la Commission Nationale Informatique et Libertés no. 2013-420 prononçant une sanction pécuniaire à l'encontre de la société Google Inc.

processing activities related to the offering of services (such as MT services) to data subject in the EU [see Tene and Wolf, 2013, p. 2].

A second possibility, if MT providers were seeking to avoid the effect of the Directive, might be to invoke the liability limitation of art. 14 of the Directive 2000/31/EC on e-commerce (hereafter the E-commerce Directive)[29], claiming that all they do is to provide a service that consists of processing data provided by the users, and lack control over the content that is being processed in these services. However, this argument also has little chance of success in court – the CJEU held recently that a search engine provider is responsible for the processing of personal data which appear on web pages, even if they were published by third parties[30].

Therefore, MT providers are bound by the provisions of the Directive (which cannot be altered or waived by a contractual provision, e.g. in Terms of Service[31]), such as those according to which processing may only be carried out on the basis of one of the possible grounds listed in its art. 7. In our view, the only two grounds that can be taken into consideration here are: the data subject's consent (art. 7(a)) and performance of a contract to which the data subject is party (art. 7(b)). It is, however, not clear if such a consent in case of MT services would be regarded as sufficiently informed, given the fact that in practice no information about the functioning of the service is given to the user[32]. Thus, in the case of Google Translate, neither the Terms of Service nor the Privacy Policy can be regarded as sufficiently informative, especially given that, since 2012, they attempt to cover all the services provided by Google (Gmail, Google Docs, Google Maps. . .) in a composite way.

Another legal ground that can be thought of in the context of 'free' online MT services is performance of a contract to which the data subject is party. In fact, the MT provider offers an MT service to the user who, by entering data in the service accepts the offer[33]. Without entering into details of contract law theory, we believe that these circumstances (offer and acceptance) may be sufficient for a contract to be formed, at least in jurisdictions that do not require consideration (i.e. something of value promised to another party) as a necessary element of a contract (however, the data itself may be regarded as consideration for the translation service).

The processing of data is therefore necessary for the performance of such a contract – which in itself may constitute a valid legal basis for processing.

In reality, however, these legal bases are only valid for the processing of data relating to the user. Once again, by processing data relating to a third party,

[29] this provision has received rather extensive interpretation from the CJEU, especially in case Google France SARL and Google Inc. v Louis Vuitton Malletier SA (C-236/08) concerning the AdWords service.

[30] Case C-131/12.

[31] Article 29 Data Protection Working Party, *Opinion 02/2013 on apps on smart devices adopted on 27 February 2013*, 00461/13/EN, WP 202

[32] Article 29 Data Protection Working Party, *Opinion 02/2013 on apps on smart devices adopted on 27 February 2013*, 00461/13/EN, WP 202, 15.

[33] cf. idem, 16.

the MT service provider is potentially in breach of the Directive. Just as in the case of processing by the user, processing by the MT provider fits with difficulty within the framework of the Directive.

2.3.1 Data processing obligations of the MT service provider

What then are the processing obligations of the MT service provider in respect of the data fed into the translator by the user? In the first place, the provider will undertake processing of the text to provide the specific translation for the user. Such primary processing is exactly what the user expects, and consents to when he uses the service and is not further problematic. Some users may imagine that the data entered in a 'free' online MT service 'disappear' once the MT process is accomplished. In fact, MT service providers are interested in keeping the data and re-using them in the future. For example, Google's Terms of Use expressly state that by entering content in one of Google services, the user grants Google *'a worldwide [IP rights] license to use, host, store, reproduce, modify, create derivative works (...), communicate, publish, publicly perform, publicly display and distribute such content'*. The text further specifies that *'the rights [the user] grant[s] in this license are for the limited purpose of operating, promoting, and improving our Services, and to develop new ones'*.

In this regard, the business model behind 'free' online MT services is likely not very different from the one on which SNS (Social Networking Services) are based[34]: they allow the service provider to harvest data from users which can then be re-used (either directly by the MT provider or by a third party) for direct or indirect marketing or advertising purposes (in Google's case: advertising is actually one of the services that the quoted passage refers to). Naturally, the data can also be used to improve the tool (by enriching the corpus on which the translation model can be based). In this model, the data (together with additional input from the user provided e.g. by accomplishing crowdsourced tasks, such as resolving Captcha challenges or proposing an alternative translation, or choosing the most appropriate one) are in fact a form of payment for the service (hence, the services are not really 'free'). Such services can be called 'Siren Services', by analogy with J. Lanier's 'Siren Servers' [Lanier, 2014]. These services (or servers) lure users (just like sirens lured sailors in The Odyssey) into giving away valuable information in exchange for services. In this scheme the users receive no payment (or any other form of acknowledgment) for the value they add to the service.

Apart from raising ethical concerns, such behaviour of MT service providers is also of doubtful conformity with the Directive. Firstly, art. 6.1 (e) prohibits data storage for periods *'longer than necessary for the purposes for which the data were collected'*, which in itself may be a barrier to any form of secondary processing of MT data. Secondly, given that an average user may well be unaware of this processing taking place, the requirement of informed consent is arguably missing. Equally, the user will be denied the practical possibility to exercise

[34] Article 29 Data Protection Working Party, *Opinion 5/2009 on Online Social Networking adopted on 12 June 2009*, 01189/09/EN, WP 163, 4-5.

rights that are granted to him by the Directive, such as the right of access (art. 12) or the right to object (art. 14).

From the point of view of the Directive, two scenarios for such re-use of data by the MT-providers should be distinguished: firstly, secondary processing for such purposes as research, evaluation and development of the MT service (translation model); secondly, secondary processing for marketing and advertising purposes. For the sake of simplicity, these two scenarios will be referred to as 'service-oriented' and 'commercial' secondary processing.

2.3.2 Service-oriented secondary processing

In our view, in some cases service-directed secondary processing may be allowed by the Directive even without the additional consent of the data subject. First of all, art. 6.1 (b) interpreted *a contrario* allows for further processing of data for purposes which are compatible with the purposes for which they were initially collected. The same article specifies that processing for historical, statistical and research purposes shall not be considered incompatible with the initial purpose. Therefore, it may seem that the processing for the purposes of statistics and research (including, arguably, the improvement of the translation model) may be allowed by the Directive. However, according WP29's opinion one of the key factors in assessing purpose compatibility should be *'the context in which the data have been collected and the reasonable expectations of the data subjects as to their further use'*[35]. As explained above, any form of secondary processing of MT data does not seem to meet 'reasonable expectations' of MT users, as most of them may simply expect the data to be deleted after the MT is accomplished. Independently of this point, to the extent it is no longer necessary – for the purpose of the secondary processing – any reference to real data subjects, the service provider should anonymise the data, as set out in art. 6(e) of the Directive.

Additionally, MT service provider may also seek to rely on another 'safety valve', provided by art. 7(f) of the Directive, which allows for processing of personal data *'necessary for the purposes of the legitimate interests pursued by the controller or by the third party or parties to whom the data are disclosed'*. The problem here, however, is that this provision of the Directive further specifies that fundamental rights and freedoms of the data subject may override other legitimate interests; therefore, the fact that in case of secondary processing users cannot exercise their rights, and in particular their right to be forgotten, may lead a court to reject art. 7(f) as a valid legal ground for such processing[36]. Finally, it might be argued that service-oriented data use falls within the scope of processing for research purposes, even though the Directive does not provide a specific research exemption (apart from the one in the art. 6.1 (b) quoted above), some Member States, relying on more general principles of the Directive (such as e.g. the art. 7(f)), introduced it in their national laws. This is the case

[35] Article 29 Data Protection Working Party, *Opinion 03/2013 on purpose limitation adopted on 2 April 2013*, 00569/13/EN, WP 203, 24.

[36] cf. CJEU case C-131/12 (Google Spain), para 91.

of UK law as well as the laws of German federal states[37]. In each case, though, additional measures, including anonymisation of the data would again appear mandatory to adequately safeguard the user's interests.

2.3.3 Market-oriented secondary processing

The providers of 'free' online MT data may want to further process the input data for commercial purposes, such as direct and indirect marketing or advertising. It is clear, however, that the Directive does not allow for such form of secondary processing, which falls neither within the scope of art. 6.1 (b), nor art. 7 (f). Such processing would therefore necessitate the data subject's consent, distinct from that given for primary processing, which this time certainly cannot be implied. In particular, in order to validly consent for such secondary processing, the user would need to be thoroughly informed. Even if such detailed information is provided to the user, some forms of commercial secondary processing may, in our view, fail to meet the requirement of fairness, distinct from the one of lawfulness (art. 6.1 (a) of the Directive), and therefore violate the principles of the Directive.

3 Conclusions

MT is a very useful and constantly improving technology which may contribute in a very efficient way to crossing the language barrier in digital communications. Nonetheless, the use of this technology also raises some important privacy risks, of which many users may be insufficiently aware. The current EU data protection framework, if applied and respected by all actors involved should serve to shield the users from many of those privacy risks. However, it some cases it may also place bona fide users or providers of these services in danger of breach of law.

More generally, solving the ethical dilemmas raised by 'siren services' such as 'free' online MT remains a difficult challenge. It is true that these services are - at least in part - built on users' data and other input. Realistically speaking, more information about the functioning of 'free' online MT services should be provided to the community. Private users should consider translating only those bits of texts that do not contain any information relating to third parties (which in practice may limit them to translating text into, and not from, a different language to their own). Businesses in particular may find such a limitation rather constricting and to protect their own data and the data of their clients, may prefer to opt for a payable offline MT tool instead of a 'free' online service.

[37] A research exception was also contained in art. 83 of the General Data Protection Regulation as initially proposed by the Commission; after numerous amendments introduced by the Parliament, its future, however, remains uncertain. If adopted, such an exception would under certain conditions allow for some forms of secondary processing of MT input data.

Bibliography

A.L. Berger, P.F. Brown, S.A. Della Pietra, V.J. Della Pietra, J.R. Gillett, J.D. Lafferty, R.L. Mercer, H. Printz, and L. Ureš. The Candide system for machine translation. In *Proceedings of the workshop on Human Language Technology*, pages 157–162. Association for Computational Linguistics, 1994.

P.F. Brown, V.J. Della Pietra, S.A. Della Pietra, and R.L. Mercer. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2):263–311, June 1993. ISSN 0891-2017.

J. Chin. Suggest a better translation. `http://googleblog.blogspot.com/2007/03/suggest-better-translation.html`, March 2007. Accessed: November 2, 2014.

D. Crystal. *English as a Global Language*. Canto Classics. Cambridge University Press, 2012. ISBN 9781107394605.

M. Galvez and S. Bhansali. Translating the world's information with Google Translator Toolkit. `http://googleblog.blogspot.com/2009/06/translating-worlds-information-with.html`, June 2009. Accessed: November 2, 2014.

W.J. Hutchins. *Machine Translation: Past, Present, Future*. Ellis Horwood Series in Computers & Their Applications. Prentice Hall, 1986. ISBN 9780135435212.

W.J. Hutchins. Warren Weaver memorandum: 50th anniversary of machine translation. In *MT News International issue 22*, pages 5–6. MT News International, 1999.

W.J. Hutchins. The Georgetown-IBM experiment demonstrated in January 1954. In *Machine Translation: From Real Users to Research*, pages 102–114. Springer, 2004.

Internet World Stats. Internet World Users by Language. `http://www.internetworldstats.com/stats7.htm`, 2013. Accessed: October 23, 2014.

J.H. Johnson and J. Martin. Improving translation quality by discarding most of the phrasetable. In *Proceedings of EMNLP-CoNLL'07*, pages 967–975, 2007.

S. Kelman. Translating the world's information with Google Translator Toolkit. `http://googletranslate.blogspot.com/2014/07/translate-community-help-us-improve.html`, July 2014. Accessed: November 2, 2014.

E. Ketzan. Rebuilding babel: Copyright and the future of online machine translation. *Tulane Journal of Technology & Intellectual Property*, 9:205, 2007.

P. Koehn. *Statistical Machine Translation*. Statistical Machine Translation. Cambridge University Press, 2010. ISBN 9780521874151.

P. Koehn and H. Hoang. Factored Translation Models. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLT, pages 868–876, Prague, June 2007. Association for Computational Linguistics.

P. Koehn, F.J. Och, and D. Marcu. Statistical Phrase-Based Translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, volume 1 of *NAACL '03*, pages 48–54, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

J. Lanier. *Who Owns the Future?* Simon & Schuster, 2014. ISBN 9781451654974.

Y. Liu, K. P. Gummadi, B. Krishnamurthy, and A. Mislove. Analyzing Facebook privacy settings: user expectations vs. reality. In *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*, pages 61–70. ACM, 2011.

G.E Moore. Cramming more components onto integrated circuits. *Electronics*, 38(8):114, April 1965.

F. Och. Breaking down the language barrier–six years in. `http://googleblog.blogspot.co.uk/2012/04/breaking-down-language-barriersix-years.html`, April 2012. Accessed: October 23, 2014.

J. Porsiel. Machine Translation and Data Security. `http://www.tcworld.info/e-magazine/content-strategies/article/machine-translation-and-data-security/`, February 2012. Accessed: October 23, 2014.

M. Quince. Why Austria is Ireland. `http://itre.cis.upenn.edu/~myl/languagelog/archives/005492.html`, March 2008. Accessed: November 1, 2014.

S. Shankland. Google Translate now serves 200 million people daily. `http://www.cnet.com/news/google-translate-now-serves-200-million-people-daily/`, May 2013. Accessed: October 23, 2014.

J.R. Smith, H. Saint-Amand, M. Plamada, P. Koehn, C. Callison-Burch, and A. Lopez. Dirt cheap web-scale parallel text from the Common Crawl. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1374–1383, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

O. Tene and C. Wolf. Overextended: Jurisdiction and applicable law under the eu general data protection regulation. Technical report, The Future of Privacy Forum, Washington, DC, January 2013.

M. Tyson. Google Translate tops 200 million active users. `http://hexus.net/tech/news/software/38553-google-translate-tops-200-million-active-users/`, April 2012. Accessed: October 23, 2014.

O.F. Zaidan and C. Callison-Burch. Crowdsourcing translation: Professional quality from non-professionals. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies – Volume 1*, HLT '11, pages 1220–1229, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics. ISBN 978-1-932432-87-9.