European Perspective on Privacy Issues in ‘Free’ Online Machine Translation Services.

Paweł Kamocki[1, 2, 3]
(1) Institut für Deutsche Sprache, R 5, 6-13, D-68161 Mannheim, Germany
(2) Institut Droit et Santé, Université Paris Descartes, 45 rue des Saints-Pères, 75270 Paris Cedex 6, France
(3) Institut für Informations-, Telekommunikations- und Medienrecht, WWU Münster, Leonardo-Campus 9, 48149 Münster, Germany kamocki@ids-mannheim.de

Résumé.

La perspective européenne sur les questions liées à la protection de la vie privée dans les outils 'gratuits' de traduction automatique en ligne
Suite à la Révolution Numérique, la langue anglaise s’est établie comme la langue internationale; cependant, seuls 28,6% des internautes ont l'anglais pour langue maternelle. La traduction automatique (Machine Traslation, MT) est une technologie puissante qui peut combler ce fossé. En développement depuis le milieu du vingtième siècle, la traduction automatique est devenue accessible à chaque internaute au cours de la dernière décennie, grace aux outils disponibles gratuitement en ligne. Cette étude a pour ambition d’examiner les implications que ces outils peuvent avoir sur la vie privée des utilisateurs dans le contexte de la loi européenne sur la protection des données personnelles. Sont analysés le traitement initial (du point de vue de l’utilisateur et du fournisseur du service de traduction automatique) et le traitement secondaire qui peut potentiellement être entrepris par le fournisseur du service de traduction automatique.

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.

Mots-clés : données personnelles, traduction automatisée, vie privée, Directive 95/46/CE, Google Translate
Keywords: personal data, Machine Translation, privacy, Directive 95/46/EC, Google Translate

1 Introduction

Digital revolution (started with the proliferation of personal computers in the late 1970s and continuing to the present day), just as any revolution worthy of its name, has changed our everyday life in more ways that one may want to admit. The switch from the analog to the digital has multiplied useful technologies that enabled an ordinary person to perform tasks that two generations ago required a considerable amount of time and/or manpower. 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 hills: the language barrier. Even though English has undoubtedly taken advantage of the Digital Revolution to establish itself as a global language (Crystal, 1997); it has recently been estimated that it is used by 55,7% of all websites (German comes second with 6,1%)[1], as a matter of fact only 28,6% of Internet users speak English as a native language (as of December 31, 2013)[2]. 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, as a consequence, can only take advantage of a tiny fraction of the content available on the World Wide Web. Therefore, it cannot be denied that the global communication requires linguistic support systems in order to develop its full potential (Cribb, 2000).

[1] “Usage of content languages for websites,” W3Techs, accessed April 13, 2015, http://w3techs.com/technologies/overview/content_language/all.

[2] “Internet World Users by Language,” Internet World Stats, accessed April 13, 2015, http://www.internetworldstats.com/stats7.htm.

The Digital Revolution has provided a number of tools for linguistic support: it has (arguably) made language acquisition faster and more efficient, it has also helped improve the quality and the availability of human translation[3]. But none of these paths can lead to the erosion of language barriers in digital communication in the way that Machine Translation (MT) can accomplish this task.

1.1 MT in Context

MT (or automated 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.

1.1.1 History

The idea to mechanize the translation process can be traced back to the seventeenth century (Hutchins, 1986); most notably, in 1629 Descartes described a system of codes that would relate words between different languages (Hutchins, 1986), therefore allowing quick translation from one language to another. The first proposals for 'translating machines', however, did not appear until 1933, when Georges Artsrouni (Corbé, 1960) and Petr Troyanskii (Bel'skaya, Korolev Panov, 1959) were issued patents (in France and Russia respectively) for 'automated dictionaries'; both inventions remained nearly unknown and had become outdated before they were brought to the attention of the scientific community in 1950s-1960s.

In March 1949, inspired by the developments in code breaking during the Second World War, Warren Weaver, a researcher at the Rockeffeler Foundation, published a memorandum in which he put forward the idea to use computers for translation (Hutchins, 1999). This document marks the beginning of MT as a scientific discipline.

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 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).

At the end of 1950s, Yehoshua Bar-Hillel (the world's first full-time researcher in MT) questioned the possibility of developing a high-quality MT system, basing his argument mostly on semantic ambiguity of certain expressions in natural languages - a phenomenon that machines would never be able to deal with properly (Hutchins, 1998).

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 foreseeable future (Hutchins, 1986). 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).

During the 1980s, Japan found itself in the avant-garde of MT technology, with a particular focus on Japanese-to-English and English-to-Japanese translation (Hutchins, 1986).

In the 1990s, researchers (particularly in Germany) started to work on speech translation; during this period, low-end MT tools started appearing on personal computers. Most notably, in 1995 SYSTRAN launched SYSTRAN Professional for Windows. It was estimated that approximately 1000 MT packages were available for PCs in 1996 (Hutchins (ed.), 2000). BabelFish, the first online translation service (also based on SYSTRAN) was launched in 1997[4]. Google started providing an online translation (initially also using SYSTRAN) service in 2006[5] and Microsoft launched Bing Translator in 2009[6].

When American giants Microsoft and Google developed their proprietary MT systems in the second half of the 2000s (Google switched from SYSTRAN to a proprietary system in 2007; Microsoft Research developed an MT system in 2008 (Microsoft Translator Team, 2008), Europe also wanted to catch up. A German project Verbmobil (http://verbmobil.dfki.de/) was focused on German-English and German-Japanese language pairs. The Quaero project, established initially as a French-German cooperation, aimed at developing a multilingual search engine; MT technology was meant to be an important part of this initiative. After German partners announced their departure from the project, the future of this endeavour remains uncertain.

[3] a form of translation in which human translator uses software to facilitate the process is referred to as Computer-Assisted Translation (CAT) and should be clearly distinguished from Machine Translation (MT).

[4] according to the history of SYSTRAN software posted on www.translationsoftware4you.com.

[5] according to the company's history posted on Google's website.

[6] according to Wikipedia: http://en.wikipedia.org/wiki/Bing_Translator.

1.1.2 Technology and challanges

The technological approach to MT has changed significantly over time. Chronologically the first MT method can be referred to as 'the dictionary method'. In this approach, words in a sentence are translated one-by-one, just like Descartes would have imagined it. This method, not different from a simple re-coding, can rarely produce satisfying results (although it is possible for closely related language pairs such as Dutch and English — Madsen, 2009); the system used in the IBM-Gorgetown experiment described above was in fact not much more complicated than this (the output translation was satisfying mostly because the input sentences were carefully selected — Madsen, 2009).

Everyone who has ever tried to translate a text knows that before it can be translated, it has to be understood first. The rules that humans use to decipher the meaning of linguistic expressions are morphological (i.e. how to build words out of morphemes) syntactic (i.e. how to build sentences out of words) or semantic (i.e. what words mean and what are the relations between them, such as e.g. hyponymy and hypernymy). According to some linguists, computers - in order to be able to provide MT of satisfying quality - have to integrate these rules (Winograd, 1972). This classic approach to MT is known as 'rule-based' or 'knowledge-based' MT.

A rule-based MT system, rather than translating word-by-word, performs a linguistic analysis of the input text (based on information retrieved from language-specific dictionaries, semantic hierarchies etc.) and then on the basis of its output generates a sentence in the target language.

Such a system is in fact not a simple tool, but a conglomerate of tools performing different tasks in this multi-stage process: a morphological analyser (in both source and target language), a syntactic parser, a thesaurus etc. Therefore, it is obvious that the development of such systems is costly and time-consuming, and that they are difficult and expensive to update. Nevertheless, such systems (the best known of which is SYSTRAN, especially in its older versions) were developed and implemented worldwide, starting from the 1970s.

With the development of corpus linguistics and the growth of computational power, another approach, the so-called 'statistical MT', has become more appealing. In this paradigm, revived in 1993 by IBM researchers (Brown *et al.*, 1993), an MT system is based on a set of human-translated bilingual (or multilingual) language corpora. A statistical model derived from the analysis of the bilingual corpus is then used to translate input from source to target language according to the probability distribution.

The prerequisite for building statistical MT systems is therefore the existence of human-translated bilingual (or multilingual) corpora - and the bigger the better. Such corpora of satisfying quality may not exist for every language pair. An obvious source of human-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[7] and in 6 languages for the UN[8]).

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 providing that the appropriate human-translated bilingual corpus can be 'loaded' into the system). Also, because statistical MT systems are based on human-translated texts, the output of statistical MT is (or at least can be) more natural. Finally, it can be assumed that, with the growth of available multilingual data, the quality of statistical MT will quickly improve.

In the last decade, there was a tendency for online MT providers to switch from the rule-based approach to the statistical approach (marked especially by Google Translate's shift from SYSTRAN to its proprietary, largely statistical MT system in 2007). Nowadays, the two approaches are often combined and it can be assumed that hybrid MT systems are the future of MT.

MT still has to face some important challenges, the most serious of which has always been disambiguation, i.e. choosing the 'right' meaning of an ambiguous input sentence. The two approaches described above can help the system to make the right choice, but they are still imperfect. It has also been put forward that MT cannot properly handle non-standard language, such as e.g. poetic expressions (but see: Genzel, Uszkoreit, Och, 2010); the expectations towards MT should therefore be mitigated. Some researchers claim even today that MT will never achieve the quality of human translation (Madsen, 2009).

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 system.

1.2 '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.

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

[8] 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.

1.2.1 Examples

The most popular ‘free’ online MT service is Google Translate. 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.

Google Translate is the most popular ‘free’ online MT service; it’s been reported to be used by 200 million people every day (in 2013[9]) and to to translate enough text to fill 1 million books every day (in 2012[10]).

SYSTRAN-based Yahoo! Babel Fish (formerly www.babelfish.yahoo.com) was chronologically the first ‘free’ online MT service. Opened by Systran and AltaVista in 1997, the service could support up to 38 languages. Alas, it no longer exists - in May 2012 it was replaced by Bing Translator. A company funded in 1995 named BabelFish still provides a ‘free’ online MT service at www.babelfish.com, supporting a humble number of 14 languages.

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.

An Israeli public company Babylon Ltd. and a French company Reverso-Softissimo also provides ‘free’ online MT tools. Babylon (http://translation.babylon.com) supports 30, and Reverso (www.reverso.net) 13 languages.

As a final example, a free/open-source MT platform Apertium is available at www.apertium.org. The engine, the tools and the data are licensed under CC BY-SA 3.0 or GNU GPL 3.0 and can be freely shared and re-used. Unfortunately Apertium offers mostly translation for closely-related language pairs.

1.2.2 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.

First of all, given that organizations such as U.S. Air Force or the European Commission use high-end MT systems we assume that the technology is (and was already in the 1970s) far from being useless altogether. Whether low-end, ‘freely’ available online MT tools are ‘good enough’ is a slightly different issue.

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, we remark 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*’. He obtained this result on August 6, 2006 using Google Translate (which was at that time based on SYSTRAN). Today (April 15, 2015) the proverb is translated correctly as ‘*My house is your house*’.

If we take into account Moore’s law (according to which computers’ speed and capacity double every 18 months — More, 1965)[11], 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 their privacy.

2 Nature of the Data Processed in ‘Free’ Online MT Tools

‘Free’ online MT services can be used to process all sorts of texts - private and professional correspondence, commercial offers, blogs, Internet fora, social media content. The input may be of different length -- from several paragraphs (e.g. Bing Translator is limited to 5000 characters, but Google Translate can handle several times more) to single words.

In the light of the above it is not surprising that information entered by users in ‘free’ online MT tools can be sensitive from the point of view of privacy. In fact probably much more often than the users would like to acknowledge it, the input may be qualified as personal data.

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). For the purpose of the Directive, personal data shall mean ‘*any information relating to an identified or identifiable natural person*’. According to Article 29 Data Protection Working Party (hereinafter: WP29), the definition should be interpreted quite broadly. For example, it covers not only the ‘objective’ information, but also ‘subjective’ information such as opinions or assessments[12]. Moreover, 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 behavior of the person (the ‘purpose’

[9] according to http://www.cnet.com/news/google-translate-now-serves-200-million-people-daily/.

[10] according to http://googleblog.blogspot.co.uk/2012/04/breaking-down-language-barriersix-years.html.

[11] this growth rate, however, which continued steadily for more than half a century, is expected to stop very soon.

element), or if it has an impact on the person's interests or rights (the 'result' element). The person that the data relate to (i.e. the data subject) can be not only identified, but also identifiable by any means likely reasonably to be used by the data controller or any other person[13]. It has to be taken into account here that the providers of 'free' online MT tools often provide a wide range of internet services (email, social media platforms, applications...); they may therefore be in possession of vast datasets which, cross-referenced with the data entered into an MT system, may help identify the data subject[14]. Finally, the data processed in 'free' online MT tools may concern not only natural, but also legal persons (e.g. information on the financial condition of an employer can be found in the employee's private e-mails), which would exceed the definition of personal data, but may still be concerned by privacy/confidentiality considerations.

Users might be inclined to believe that what MT tools do is nothing more than simple automatic re-coding; in fact, as it has been shown above, MT tools perform an analysis of the input text. This analysis can certainly be qualified as 'processing' in the sense of art. 2(b) of the Directive[15].

In our view, the processing of data in 'free' online MT tools can be divided into two stages: first, the information is entered in order to be translated (this stage can be called 'initial processing'); then, the MT provider may want to further process the input data for different purposes (these purposes may range from scientific research on the development of the system, the evaluation of the system, statistics to direct marketing). The following sections will examine these two stages in the context of the Directive.

3 Initial Processing

The two actors at this stage of processing are: the user (who enters the data into the system) and the MT provider. It is not clear whether both of them can be regarded as data controllers in the sense of the Directive[16], or whether only the user is the controller, and the provider - a processor[17]. In our view - given that the provider plays a crucial role in determining the functioning of his MT system - both actors can be regarded as processors[18].

3.1 The User's Perspective

By entering data into the MT system the user may not only process information concerning him, but also information concerning other natural (or even legal) persons. It seems, however, that merely entering the data into a freely available MT tool for the purpose of obtaining an imperfect translation (to be able to better understand the text) may in most cases be qualified as a 'purely personal or household activity' and as such exempted from the Directive under art. 3(2).

It is not clear how to understand 'purely personal or household activities'. In our view, the use of freely available MT tools may fall in the first category even if the tool is used for professional or academic purposes (in fact, the Directive mentions 'private activity', and not 'private purposes'), unless the user is a professional translator. It is difficult to argue that human-translating a text without publishing the translation - just to be able to understand the original - may amount to unfair personal data processing, even if the text contains someone else's personal information. This should also be the case if MT is used.

3.2 The Provider's Perspective

By definition, machine translation is carried out in an automated way. Of course, the Directive still applies to such processing (art. 3(1)).

International MT providers such as Google or Microsoft could argue that the Directive does not apply to them because they are not established on the territory of an EU Member State 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 and data protection authorities)[19], however, will soon become invalid as the proposed text of the new General Data Protection

[12] 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, 6.

[13] Recital 26 of the Directive 95/46/EC.

[14] see: the concept of 'linkability' in: Article 29 Data Protection Working Party, Opinion 05/2014 on anonymisation techniques adopted on 10 April 2014, 0829/14/EN, WP 216, 11.

[15] 'processing' is defined in art. 2(b) of the Directive 95/46/EC as *'any operation or set of operations which is performed upon personal data, whether or not by automatic means, such as collection, recording, organization, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, blocking, erasure or destruction'.*

[16] Art. 2(d) of the Directive 95/46/EC defines the controller as *'the person who determines (alone or jointly with others) the purposes and means of the processing of personal data'.*

[17] Art. 2(e) of the Directive 95/46/EC defines the processor as *'the person which processes personal data on behalf of the data controller'.*

[18] cf.: CJEU judgement in Case C-131/12 Google Spain SL, Google Inc. v Agencia Espan#ola de Protección de Datos, Mario Costeja Gonzàlez, 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'.*

Regulation extends its applicability to the processing activities related to the offering of services (such as MT services) to data subject in the EU (Tene, Wolf, 2013).

Finally, MT providers may try to invoke a liability limitation, e.g. under art. 14 of the Directive 2000/31/EC on e-commerce[20], claiming that all they do is to provide a service that consists of processing data provided by the users. This argument also has little chance of success in court -- the CJEU held recently[21] that search engine providers are responsible for the processing of personal data which appear on web pages published by third parties.

Therefore, MT providers are bound by the provisions of the Directive, 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)).

3.2.1 Consent

Art. 2(h) of the Directive defines 'consent' as '*any freely given specific and informed indication of his wishes by which the data subject signifies his agreement to personal data relating to him being processed*'. According to WP29, it can be '*any signal, sufficiently clear to be capable of indicating the data subject' s wishes and to be understandable by the data controller'*[22]. It seems therefore that consent may be concluded from the data subject's behavior. The use of an MT service can, in our view, be interpreted as consent for processing, but only to the extent necessary for the translation (and not for any further processing).

3.2.2 Performance of a contract

Another ground for lawfulness that can be imagined in the context of 'free' online MT services is performance of a contract to which the data subject is party.

The MT provider offers an MT service to the users; by entering data in the service in order to obtain a translation the user accepts the offer. Without entering into details of contract law theory, we believe that these circumstances 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 -- see below).

The processing of data is therefore necessary for the performance of such a contract - which in itself may constitute a valid ground for lawfulness.

4 Further processing

Contrary to what some users may imagine, the data entered in a 'free' online MT system do not 'disappear' once the MT task is accomplished. In fact, some MT providers expressly state in their Terms of Service that by entering data in their services the users grant them a copyright license[23], which indicates the provider's intention to re-use the data.

In our view, the purposes for which MT providers may further process the data may be divided into two categories. Firstly: non-commercial purposes (including statistics, evaluation and improvement of the system) and secondly: commercial purposes (including e.g. direct marketing).

4.1 Non-commercial Purposes

This section examines possible grounds for further processing of data entered into online MT services for non-commercial purposes. In our view, these grounds include: processing for purposes compatible with the initial consent and legitimate interest of the data controller. In addition, such processing may also be carried out on the basis of a statutory exception.

4.1.1 Purpose Limitation

According to art. 6(1)(b) of the Directive, personal data must be 'collected for specified, explicit and legitimate purposes and not further processed in a way incompatible with those purposes'. This principle, referred to as 'purpose

[19] cf.: CJEU judgement in Case C-131/12; Déliberation de la Commission Nationale Informatique et Libertés no. 2013-420 prononçant une sanction pécuniaire à l'encontre de la société Google Inc.

[20] 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.

[21] in case C-131/12 (cf. *supra*).

[22] Article 29 Data Protection Working Party, Opinion 15/2011 on the definition of consent adopted on 13 July 2011, 01197/11/EN, WP187, 11.

[23] cf. e.g. Google's Terms of Service, last modified April 14, 2014, http://www.google.com/intl/en/policies/terms/.

limitation', allows further processing for purposes compatible with the initial purpose (provided that this purpose is specified, explicit and legitimate).

As discussed above, MT can be analyzed as a specified (i.e. sufficiently defined), explicit (i.e. unambiguous) and legitimate (i.e., among others, carried out on the basis of one of the grounds listed in art. 7) purpose[24]; the possible grounds for legitimacy are consent of the data subject and performance of a contract to which the data subject is party (see above).

The next step is to assess whether the purposes such as evaluation and development of the tool or scientific research can be regarded as compatible with the initial purpose. According to WP29[25], the key factors to be considered during the compatibility assessment include:

– the relationship between the initial purpose and further purposes:

In our view, evaluation and improvement of the MT tool (further purposes) are very closely related to the translation (initial purpose) - indeed, it can be said that such further processing may be regarded as a logical next step of the initial processing.

– the context in which the data have been collected and the reasonable expectations of the data subjects as to their further use:

We assume that an average reasonable person imagines that the data entered in a 'free' online MT system 'disappear' once the task is accomplished (Porsiel, 2012) and does not expect them to be further processed.

A link to a document containing detailed information about the service's privacy policy does not seem to influence the assessment of this factor: it is no secret that an average user does not read such documents.

It has to be noted, however, that the awareness and reasonable expectations of users may vary over time. While nowadays an average user does not seem to realize that his personal data are in fact a currency that he can use to pay for 'free' online services (European Data Protection Supervisor, 2014), it may become obvious a decade from now. This may be the case especially if, being aware of the existence of payable alternatives, the user chooses deliberately to use a 'free' service[26].

– the nature of the data and the impact of the further processing on the data subjects:

This factor is difficult to assess a priori, as it depends on the circumstances of each case. In fact, circumstances in which a 'free' online MT service is used to translate texts containing sensitive information (e.g. about health, sex life, religious or philosophical beliefs etc.)[27] cannot be excluded. Whether the processing of such data for non-commercial purposes mentioned above can have adverse consequences for the data subject is another issue.

According to WP29, in assessing the possible impact on the processing on the data subject, several elements different from the nature of the data should also be taken into account. These include: whether the data are processed by a different controller, publicly disclosed to a large number of persons or combined with other data. In our view, further processing strictly limited to the 'non-commercial' purposes mentioned above may pass this test; this will definitely not be the case of processing for commercial purposes.

– the safeguards applied by the controller to ensure fair processing and to prevent any undue impact on the data subject:

Once again, the assessment on this factor will be different for different MT systems. While some providers of paid MT services respect high security standards (such as ISO/IEC 27001:2013), this may not be the case of all 'free' MT services.

The experience shows that even companies believed to conform to high security standards may suffer from 'privacy incidents'; their approach to 'privacy by design' is not flawless (Rubinstein, Good, 2013), and their policies and practice in this field have been criticized for lack of compatibility with EU law[28].

An important element to take into consideration here is pseudonymization or anonymization of the data. The extent to which MT service providers apply these safeguards remains uncertain (Toubiana, Nissenbaum, 2011).

As far as processing for research purposes is concerned, art. 6(1)b of the Directive contains a specific provision on further processing for 'historical, statistical or scientific purposes'. It enables the Member States to authorize such processing, as long as appropriate safeguards are provided. In practice, it seems that Member States set the threshold for

[24] Article 29 Data Protection Working Party, Opinion 03/2013 on purpose limitation adopted on 2 April 2013, 00569/13/EN, WP 203, 12.

[25] *idem*

[26] Article 29 Data Protection Working Party, Opinion 06/2014 on the notion of legitimate interests of the data controller under Article 7 of Directive 95/46/EC adopted on 9 April 2014, 844/14/EN, WP 217, 47.

[27] Art. 8(1) of the Directive 95/46/EC.

[28] cf.: Déliberation de la Commission Nationale Informatique et Libertés no. 2013-420; WP29's letter to Larry Page (CEO of Google Inc.) of 23 September 2014, accessed October 23, 2014, http://ec.europa.eu/justice/data-protection/article-29/documentation/other-document/files/2014/20140923_letter_on_google_privacy_policy.pdf.

'appropriate safeguards' rather high[29]; therefore, it seems that providers of 'free' online MT services will rarely be able to meet it.

In the light of the above, it remains uncertain whether further processing of data for non-commercial purposes can be regarded as compatible with the initial purpose. In fact, it may be easier for MT providers to rely on art. 7(f) to legitimize such processing.

4.1.2 Legitimate Interests

Art. 7(f) of the Directive provides an open-ended ground for privacy by allowing 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*'.

WP29 provides guidelines for carrying out the balancing test: in order to be regarded as 'legitimate', the controller's interest has to be lawful, sufficiently concrete and it has to represent real and present interest[30]. In our view, the evaluation and the development of the tool which can be achieved via additional processing of input data can pass this test.

Moreover, the processing has to be necessary to achieve the interest pursued. The word 'necessary' shall not be interpreted as synonymous with 'indispensable'[31]; instead, it should be examined whether there are other less invasive means to achieve the intended purpose[32]. It may seem at first sight that indeed there are other ways of improving or evaluating an MT system than by enlarging the corpus that the system is based on. Also, the data necessary for the enlargement of the corpus may well be obtained from different sources. This part of the test may therefore be difficult to pass.

Establishing balance between the interests of MT providers and the rights of the data subject is not an easy task; many factors, some of them similar to those examined above in the context of the purpose limitation principle, should be taken into account. In our opinion, the fact that the benefits that the whole community may derive from the advancement in free online MT are so important that they outweigh the interests of the data subject - especially in view of the fact that the risks of adverse consequences for the data subject remain relatively low.

The fact that it is definitely not easy - if possible - for data subjects to exercise their right to object to the processing does not work in favor of MT service providers. The recent judgement of the CJEU[33], however, allows us to believe that the right to be forgotten will soon have to become a standard for online services, including MT.

To conclude, it is not clear whether MT providers can further process personal data entered into their MT systems for the purposes of research, improvement and evaluation of these systems. It is our belief, however, that the interest of the whole Internet community in the development of free online MT tools outweighs the possibly negative impact that such processing may have on the data subject. Therefore, in our view, art. 7(f) of the Directive may be a sufficient ground for lawfulness of such processing. Of course, other principles relating to data processing (laid down in art. 6 of the Directive) should still be respected.

4.1.3 Research exception

Unlike the Directive, the new General Data Protection Regulation contains (in its art. 83) a research exception[34]. According to the wording proposed by LIBE in October 2013, such processing can be allowed if it passes the necessity test (i.e. if its purposes cannot be achieved by processing anonymized data - see above), or if the information allowing identification of the data subject is separated from the data and kept separately under the highest technical standards. It is too early to say how this provision can be interpreted in practice (e.g. whether compliance with an ISO standard will be necessary to meet the 'highest technical standards' requirement). In any case, it seems that it will not bring much relief to MT providers, unless they implement robust anonymization (or at least pseudonymization) mechanisms.

4.2 Commercial purposes

MT providers may want to further process the data entered into MT services for purposes such as direct marketing (e.g. sending advertisements of local Italian restaurants for those who had used the free online MT service to translate a recipe from Italian), consumer profiling etc. All these purposes will be jointly referred here as 'commercial'.

[29] cf. e.g. in Germany: art. 13(2)8 BDSG (requiring that the scientific interests must significantly outweigh the interests of the data subjects), in France: chapter IX CNIL (containing a derogatory framework for health research).

[30] Opinion 06/2014 (cf. *supra*), 25.

[31] Judgement of the European Court of Human Rights in Silver & Others v. United Kingdom of 25 March 1983.

[32] Opinion 06/2014 (cf. *supra*), 29.

[33] in case C-131/12 (cf. *supra*).

[34] Proposal for a regulation of the European Parliament and of the Council on the protection of individual with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation). Compromise Amendments on art. 30-91: COMP Art. 83 of 17.10.2013.

According to the rules presented in the preceding section, it is obvious that the data collected from users of MT tools cannot be further processed for such purposes.

In particular, the argument that by using an MT service the user gives his consent for such processing just because the Terms of Use or the Privacy Policy stipulate so must be declared invalid[35]. Such consent could not be regarded as sufficiently informed and would not meet the requirements of the Directive.

Other possible grounds (such as art. 7(f) of the Directive - see above) would also have to be rejected. Finally, commercial purposes cannot be regarded as compatible with the initial purpose of the processing (translation).

5 Conclusions

While 'free' online MT services provide convenient and quick translation of rather satisfying quality, they may pose a threat to privacy of users. In fact, data entered into such a service do not disappear once the translation is accomplished; instead, the MT providers may want to use the data for various purposes including not only the evaluation and improvement of the tool, but also direct marketing.

In our view, the existing EU data protection framework in itself is not an obstacle to the functioning of such services - the exemption of 'purely private activities' and the possibility to interpret certain behaviors of data subjects as consent for processing seem to strike balance between the interests of the user and those of the MT provider.

Furthermore, in our view the existing framework provides for sufficient protection against further processing of the data for 'commercial' purposes. Whether these rules are respected in practice, however, is a different issue.

Acknowledgements

The author would like to thank Dr. Marc Stauch, Mr. Jim O'Regan and Dr. Piotr Bański for their expert advice and encouragement.

References

BEL'SKAYA I. K., KOROLEV L. N., PANOV D. YU. (1959). *Переводная машина П. П. Троянского: сборник материалов о переводной машине для перевода с одного языка на другие, предложенной П. П. Троянским в 1933 г.*. Moscow: Изд. Акад. Наук.

BROWN P. F., DELLA PIETRA S. A., DELLA PIETRA V. J., MERCER R. L. (1993). The mathematics of statistical machine translation: parameter estimation, *Computational Linguistics* 19, 263-311.

CORBÉ M. (1960). La machine à traduire française aura bientôt trente ans. *Automatisme* 5, 87-91. CRIBB M. V. (2000). Machine Translation: The Alternative for the 21st Century? *TESOL Quaterly* 34, 560-569. CRYSTAL D. (1997). *English as a global language.* Cambridge: University Press.

EUROPEAN DATA PROTECTION SUPERVISOR (2014). *Privacy and competitiveness in the age of big data: The interplay between data protection, competition law and consumer protection in the Digital Economy*, Brussels : European Data Protection Supervisor.

GENZEL D., USZKOREIT J., OCH F. (2010). « Poetic » Statistical Machine Translation: Rhyme and Meter. Proceedings of the *2010 Conference on Empirical Methods in Natural Language Processing, MIT, Massachusetts, USA, 9-11 October 2010*, 158–166.

HUTCHINS J. W. (1986). *Machine translation: past, present, future*. New York: Halsted Press. HUTCHINS J. W. (1998). Bar-Hillel's survey, 1951. *Language Today* 8, 22-23.

HUTCHINS J. W. (1999). Warren Weaver memorandum: 50th anniversary of machine translation. *MT News International* 22, 5-6.

HUTCHINS J. W. (ed.) (2000). *Compendium of Translation Software, Machine Translation Systems and Computer-aided Translation Support Tools. 1st edition.* Geneva : European Association for Machine Translation.

[35] cf. Article 29 Data Protection Working Party, Opinion on the use of location data with a view to providing value-added services adopted on November 2005, 2130/05/EN, WP115, 5: *'This definition* [in art. 2(h) of the Directive] *explicitly rules out consent being given as part of accepting the general terms and conditions for the electronic communications service offered'* ; see also Déliberation de la Commission Nationale Informatique et Libertés no. 2013-420.

HUTCHINS J. W. (2004). The Georgetown-IBM experiment demonstrated in January 1954. Proceedings of the *6th Conference of the Association for Machine Translation in the Americas, AMTA 2004, Washington, DC*, 102-114.

KETZAN E. (2007). Rebuilding Babel: Copyright and the Future of Machine Translation Online. *Tulane Journal of Technology & Intellectual Property* 9, 205-234.

MADSEN M. W. (2009). *The Limits of Machine Translation*. PhD diss., University of Copenhagen. MOORE G. E. (1965). Cramming More Components onto Integrated Circuits. *Electronics* 38, 114-117.

RUBINSTEIN I., GOOD N. (2013). Privacy by Design: A Counterfactual Analysis of Google and Facebook Privacy Incidents. *Berkeley Technology Law Journal* 28, 1322-1414.

TENE O., WOLF CH. (2013). *Overextended: Jurisdiction and Applicable Law under the EU General Data Protection Regulation*. Washington : Future of Privacy Forum.

TOUBIANA V., NISSENBAUM H. (2011). Analysis of Google Logs Retention Policies. *Journal of Privacy and Confidentiality* 3, 3-26.

WINOGRAD T. (1972). *Understanding Natural Language.* New York: Academic Press.