MultiLingual

Translation Technology

April/May 2014

CORE FOCUS

- Technology-assisted interpreting
- Language technology drives quality translation
- Post-editing MT: Is it worth a discount?
- Cloud security for SaaS translation providers
- Evolution of cloud-based translation memory

Technology-assisted interpreting

Hernani Costa, Gloria Corpas Pastor & Isabel Durán Muñoz

Unlike translators, for whom a myriad of computer-assisted tools are available, interpreters have not benefited from the same level of automation or innovation. Their work relies by and large on traditional or manual methods. The solutions tailored to the interpreters' needs are few and still far behind.

Fortunately, there is a growing interest in developing tools addressed at interpreters as end users, although the number of these technology tools is still very low and they are not intended to cover all interpreters' needs.

Interpreting modes and opportunities for technology

The main categories of interpreting are simultaneous and consecutive interpreting, which refers to the mode of delivering the original message. In simultaneous interpreting, the target message is given at roughly the same time that the source message is produced, whereas in consecutive interpreting the interpreter waits until the speaker has finished before beginning the interpretation and takes notes in the meantime. Apart from these two main categories, we can also include a third one: liaison interpreting, which can be either simultaneous or consecutive. Liaison interpreters work in both directions for two parties, thus the languages being used become passive and active at the same time.

Other common modes practiced are whispering interpreting, sight interpreting and sign language interpreting. Interpreting modes can be further classified according to the technical equipment used, the settings, the fields of expertise and topics.

However, there is not yet a single, accepted classification. Relevant authors and reputable interpreting institutions such as ITI (www.iti.org.uk) or AIIC (www.aiic.net) have their own classifications. The list below comprises the most frequent interpreting modes encountered in industry literature and offered by company services. By no means is it intended to be exhaustive.

- *Whispered interpreting* (also *chuchotage*) is a subcategory of simultaneous interpreting whispered into the listener's ear for which no specialized equipment is required.
- *Conference interpreting* takes place in multilingual conferences and it can be either simultaneous or consecutive interpreting, depending on the capacity of the conference and on the technical equipment available.
- *Business interpreting* is a subcategory of liaison interpreting used for smaller groups or business meetings, visits to a foreign country, one-on-one interviews and so on.
- *Court interpreting* refers to interpreting services provided in a legal setting such as courts of law. It could be either simultaneous or consecutive, depending on the technical equipment and the audience.
- *Teleinterpreting* (also remote interpreting) is done through a remote or offsite interpreter via telephone (over the phone interpreting) or via video (video remote interpreting), especially in services related to community interpreting. It is mostly consecutive, but it can also be simultaneous.
- *Community interpreting* is another subcategory of liaison interpreting; according to Roda Roberts, its main aim is "to

Left to right: Hernani Costa is supported by the People Programme (Marie Curie Actions) of the European Union's Framework Programme (FP7/2007-2013) under REA grant agreement 317471. Gloria Corpas Pastor is a professor in translation and interpreting at the University of Málaga and a visiting professor in translation technology at the University of Wolverhampton, UK. Isabel Durán Muñoz is a research member at the University of Málaga and holds a PhD in translation and interpreting.



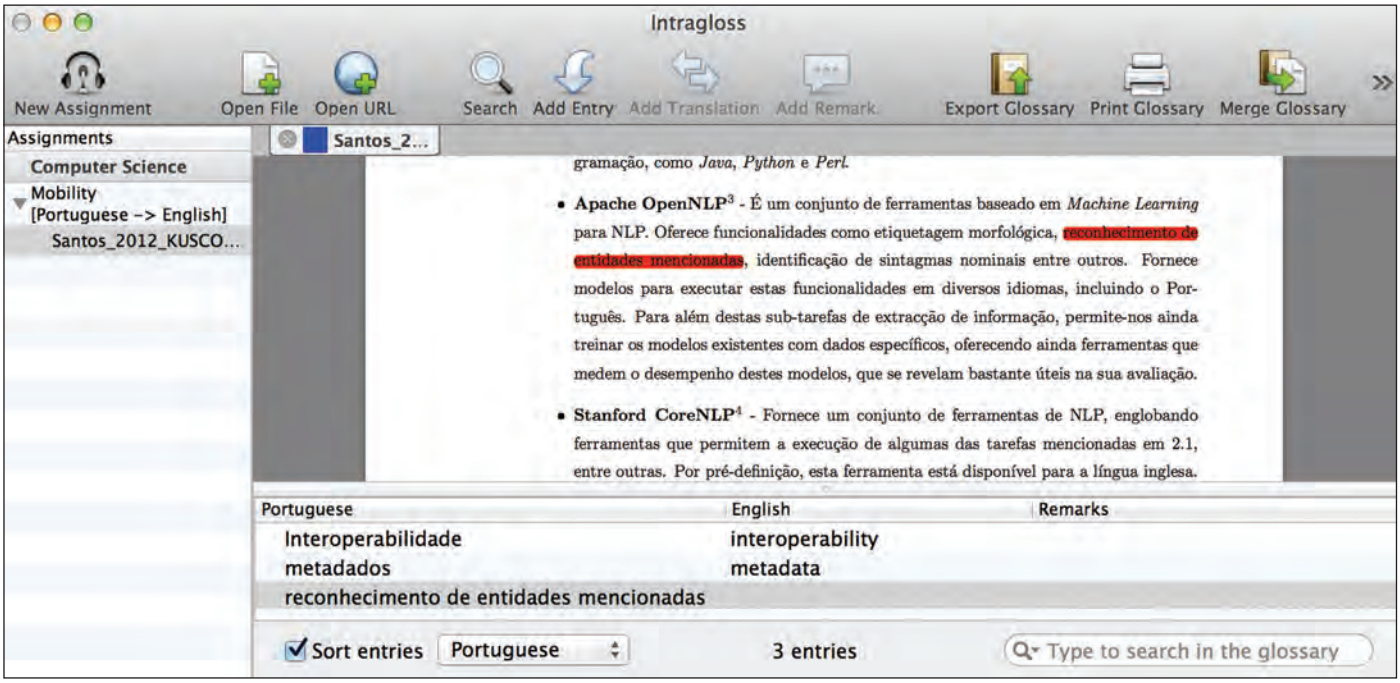


Figure 1: Intragloss screenshot.

enable people who are not fluent speakers of the official language(s) of the country to communicate with the providers of public services so as to facilitate full and equal access to legal, health, education, government, and social services."

There is a manifold of possible interpreting scenarios, and therefore, any technology tools developed for interpreters should necessarily account for this. Most interpreting services (except for teleinterpreting) are on-site, meaning the clients are in the same place that the service takes place. This limits the possibilities of using a suite of tools to assist interpretation. To the best of our knowledge, such a system has not yet been developed. However, thanks to the development of smartphones, notebooks and tablets, interpreters have some useful applications at their disposal.

The chances to develop tools for interpreters increase with regard to the preparation phase prior to any interpreting service, when interpreters need to acquire as much information and specialized knowledge as possible in order to get ready for their work. Once interpreters know the topic, the setting and all the features of the interpreting service, they can start compiling terminological resources such as glossaries, managing documents and so on. The correct management of these tools will usually mean better output. Another scenario prone to technology development is training, where all kinds of software and applications could be used to train interpreters at various stages and in different modes.

Terminology tools for interpreters

Several tools and applications have been implemented to meet the needs of different interpreting contexts and modes. Even though some interpreters still store information and terminology on scraps of paper or excel spreadsheets, there are some specialized computer and

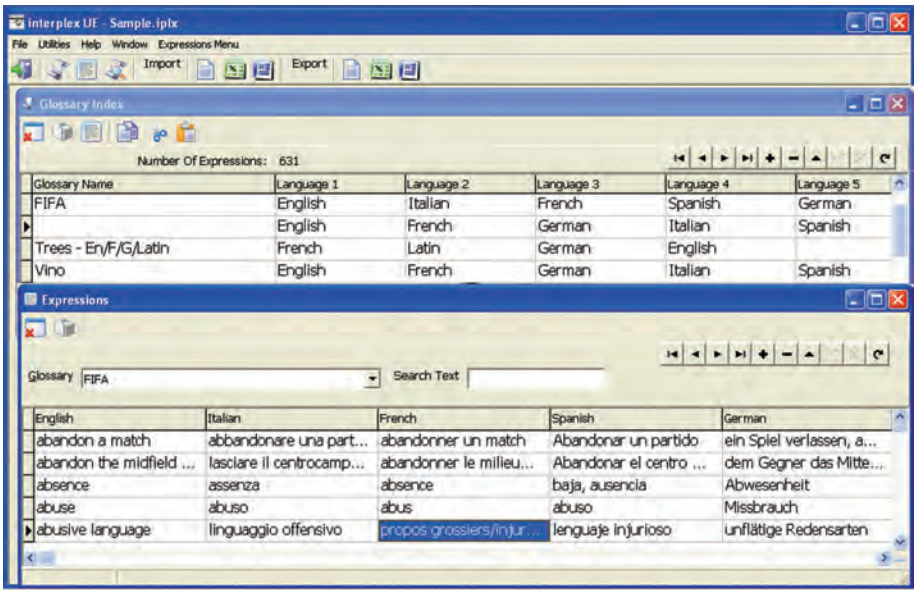


Figure 2: Interplex screenshot.

mobile software that can be used to compile, store, manage and search within glossaries. They can typically be used to prepare an interpretation in consecutive interpreting or in a booth. Those applications are quite similar to the look-up terminology tools currently used by translators. In fact, some of them have been developed to cater to the needs of both translators and interpreters.

Intragloss is a Mac OS X software created specifically to help interpreters when preparing for an event by allowing them to manage glossaries. This application can be simply defined as a glossary and document management tool created to help the interpreter prepare, use and merge different glossaries with preparation documents, in more than 150 different languages. It allows the import and export of glossaries from and to Microsoft Word and Excel formats. Every glossary imported to or created in Intragloss is assigned to a domain glossary, which contains all the glossaries from the subareas of knowledge, named *assignments*. The creation of an assignment glossary can be done in two different ways: either by extracting (automatically or manually) all the terms from the domain glossary that appear in the documents, or by highlighting a term in the document, search for it on search sites (such as online glossaries, terminology databases, dictionaries and general web pages) and adding the new translated term to the assignment glossary. The system allows for adding remarks to the glossary entries (see Figure 1).

In short, Intragloss is an intuitive and easy-to-use tool that facilitates the interpreters' terminology management process by producing glossaries (imported or created *ad hoc*), by searching on several websites simultaneously and by highlighting all the terms in the documents that appear in the domain glossary. However, it is currently platform-dependent and only works on Mac OS X platforms.

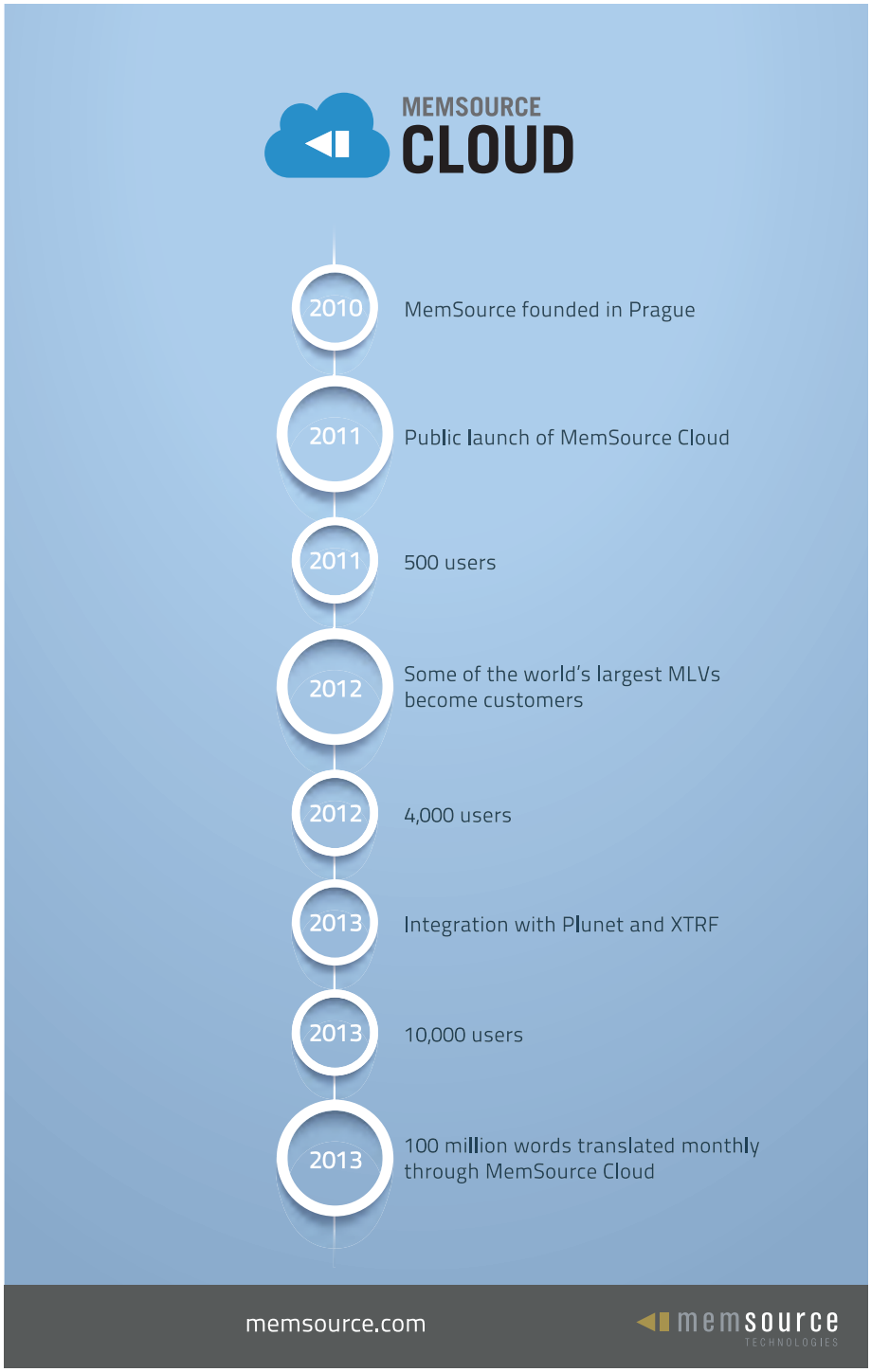
InterpretBank is a simple terminology and knowledge management software tool designed both for interpreters and translators using Windows and Android. It helps manage, learn and look up glossaries and term-related information. Due to its modular architecture, it can be used to guide the interpreter during the entire workflow process, starting from the creation and management of multilingual glossaries (TermMode), passing through the study of these glossaries (MemoryMode), and finally allowing the interpreter to look up terms while in a booth (ConferenceMode).

InterpretBank also has an Android version called InterpretBank Lite. This application is specifically designed to access bilingual or trilingual glossaries previously created with the desktop version. It is useful when working as a consecutive, community or liaison interpreter, when a quick look at the terminology list is necessary.

InterpretBank has a user-friendly, intuitive and easy-to-use interface. It allows us to import and export glossaries in different formats (Microsoft Word, Microsoft Excel, simple text files, Android and TMEX) and automatically proposes translations to terms by taking advantage of online translation portal

services. However, it is also platform-dependent (it only works on Windows), does not handle documents, only glossaries, and requires a commercial license.

Another user-friendly multilingual glossary management program that can be used easily and quickly in a booth while the interpreter is working is Interplex UE. Instead of keeping isolated word lists, it allows users to group all terms relating to a particular subject or field into multilingual glossaries that can be searched in an instant. This program enables us to have several glossaries open at the same time, which is a very useful feature if the working domain is covered by more than one glossary. Similar to the previously analyzed programs, Interplex UE also allows us to import and export glossaries from and to Microsoft Word, Excel, and simple text files (Figure 2).



Interplex UE runs on Windows; nevertheless, it has a simpler version for iOS devices, one named Interplex Lite, for iPhone and iPod Touch, and another named Interplex HD, for iPad. Both glossaries and multiglossary searchers offer the functionality of viewing expressions in each of the defined languages.

In general, Interplex UE has a user-friendly interface and it is regularly updated. It allows us to import and export glossaries from and to Microsoft Word and Excel formats. However, it, too, is platform dependent (Windows and iOS only), does not handle documents, only glossaries, and requires a commercial license.

The next two applications are particularly relevant for conference interpreting (simultaneous mode). LookUp is a commercial multilingual glossary management tool developed for Windows, aiming to be used during simultaneous interpreting and while translating. It offers support for multilingual glossaries (English, German, Spanish, Italian and French), and its main purpose is to consult terminology rapidly while interpreting in a booth. The Interpreter's Wizard is a free iPad application capable of managing bilingual glossaries in a booth. It is a simple, fast and easy-to-use application that helps the interpreter to search and visualize terminology in seconds.

Unit converters could also prove beneficial to interpreters familiarizing themselves with new terminology measures such as temperature, distance, currency, speed and so on. ConvertUnits and OnlineConversion are two illustrative samples. Both seem to be quite comprehensive, providing online conversion calculators for all types of measurement units. Apart from this, interpreters can also find conversion tables for the International System of Units, as well as calculators and converters. For Windows, there's Convert, and for Mac OS X, there's Converto. These are two free and easy-to-use unit conversion programs that convert the most popular units. There are also several mobile applications that can be used, such as Convert Units for Free and Units for iOS devices, or Unit Converter and ConvertPad for Android devices.

Finally, corpora and corpus management tools have proven most beneficial for interpreters as a device to speed up the preparation phase and to improve the quality of the input. A corpus can provide vast amounts of domain expert knowledge and accurate terminological and phraseological information in an efficient, effortless and inexpensive way.

Note-taking applications

Consecutive interpreters use a specific system of taking notes to retrieve part of their source speech understanding from memory while minimizing their processing effort. This supporting technique is usually performed manually (pen and paper) and will continue in this manner for many years to come. However, as more and more interpreters are turning to mobile devices to take notes, it is only natural that those devices become the favorite note-taking and ubiquitous capture tool on the go.

Evernote is a very dynamic and useful tool to keep more effective notes. It allows us to create an agenda note for each event, including any file, snapshot of a handwritten note, audio message, webpage, PDF or Microsoft document. Evernote can also be used to work in a team, to keep event agendas in a shared business notebook so everyone can access the details of upcoming events, and to review action items that result from these events. With Evernote everything is shareable and accessible across all platforms. Inkeness is also a useful tool to write down ideas, take notes and make sketches. Penultimate is similar, but, in addition, it allows the organization of notes in notebooks. Inkeness and Penultimate are only available for iPad devices, and both enable sharing through Evernote and by e-mail. LectureNotes and PenSupremacy are two similar applications for Android. My BIC Notes is an application specially designed for Android and iOS tablets. This application provides a set of tools for holding notes, drawing quick ideas or even doodles. In addition, it offers the functionality of adding sticky notes with personalized text, pictures and geometric shapes to the notes then printing them or sharing them with others via e-mail.

Along the same line, there is a computer-assisted tool for semiautomation of note-taking in consecutive interpreting that Aneta Rafajlovska discusses in her paper *Natural Language Processing Approach for Macedonian-French and Macedonian-English Interpreting based on Oral Sociopolitical Corpora*. This application provides a keyword with the most frequent symbols used by consecutive interpreters, which are linked to two *ad hoc* parallel dictionaries (Macedonian/English and Macedonian/French). By using the keyword, consecutive interpreters can take the same notes as they could on paper, but then they can also convert those notes into a readable message and save it for future reference.

Finally, digital pens appear to be the answer to the demand for dynamic

technology capable of synchronizing writing with ambient sound. Today these pens use real ink and write on real paper. Sky Wifi Smartpen, Echo Smartpen and Livescribe 3 commercialized by Livescribe, Inc., and the Equil JOT are just some examples of smart digital pens. These four pens are capable of linking the written notes with ambient sound and uploading them to a computer over Bluetooth, wireless or USB. Additionally, the provided software can be used to fully exploit the OCR capabilities of the pen and, for example, build glossaries. Another advantage of digital pens is the freedom to focus on listening and participating instead of worrying about catching every word during an event.

Voice recording and interpreter training

There are currently a number of applications that allow voice recording for training practice. Useful applications for managing text and audio files are GoodReader and Documents. Both tools allow the organization, annotation and synchronization of files of text, images, sound or video. They are available for iOS devices. Applications with a dual function are Audacity, Adobe Audition, AudioNote, Notability, QuickVoice, Voice Dictation and Voice Pro, among others. Besides voice recording, they allow the conversion into several audio formats, editing and quality improvement. Some of these tools provide interesting functionalities. For example, AudioNote, developed for multiple platforms (Windows, Mac OS X, Android and iOS) and Notability, for iOS, are interesting note-taking applications. Both are simple but powerful tools that combine the functionality of a notepad with voice recorder – a perfect choice for interpreters requiring a tool to synchronize text, drawings, photos or handwritten notes with audio.

Simpler but equally useful, Voice Dictation for iOS and Voice Pro for Android are two examples of easy-to-use voice recognition applications. Instead of typing, both applications use the microphone to convert audio notes to text automatically.

Text-to-speech apps for iPad can also be successfully applied to teaching and improving language skills. For example, Speak it!, Web Reader HD, Voice Dream Reader, Voxdox and Talk allow users to listen to words, texts and e-mails in several languages and formats. They are also available for Mac OS X, Windows, iOS and Android.

Finally, there is a very limited set of integrated tools that assist interpreters during their services or when training. Black Box is a computer-assisted interpreter training tool designed to help interpreters work with a range of different materials (texts, audio, video and different types of exercises) and store their results for later review. It can be used to practice in different ways: either by interpreting some audio or video clips or by doing some practical interpreting exercises, such as shadowing, cloze exercises or sight translation. It also allows teachers to edit and break down video and audio recordings to create different exercises and adapt authentic conference materials to the students' levels of expertise. Black Box can be considered a suitable training workbench for trainee interpreters.

Other web-based environments have recently been created along similar lines. InterpretaWeb and Linkterpreting provide interpreters and students with a wide range of exercises, and complete speeches to practice simultaneous and consecutive interpreting, along with information resources and news related to interpreting. These websites are of great use to students and for novice interpreters who are willing to practice and improve their interpreting skills.

Conclusion

Technology tools open up a new world of possibilities for interpreters. This paper has presented an overview of tools and applications available for interpreting practice and training. Although the number of these technologies is growing fast due to an increasing interest toward interpreters' needs, they are still insufficient and unable to fulfil all the necessary requirements. There is an urgent need to develop technologies that automate the process, increase the productivity and ease the labor-intensive activities of an interpreter (either in the preparation stage, before their interpreting service or during it). A next step in the right direction could be to gather detailed information to better ascertain interpreters' technology awareness and real needs in order to design new tools and improve existing ones. **M**

Bibliography

Rafajlovska, Aneta. *Natural Language Processing Approach for Macedonian-French and Macedonian-English Interpreting based on Oral Sociopolitical Corpora.* Master Thesis. Université de Franche-Comté, France and Universidade do Algarve, Portugal.

Roberts, Roda. "Community Interpreting Today and Tomorrow." In P. Krawutschke (ed.). *Vistas: Proceedings of the 35th Annual Conference of the American Translators Association,* 1994: 127-138.

Language technology drives quality translation

Aljoscha Burchardt, Arle Lommel, Georg Rehm, Felix Sasaki, Josef van Genabith & Hans Uszkoreit

Language technology is becoming increasingly important as organizations try to deal with the explosion of digital content and increasing demands for localized versions of this content. Fifteen years ago organizations that published content in more than ten languages were considered to be unusual, and those that dealt with more than 50 could probably have been counted on one hand. Today, however, it is not uncommon for organizations to produce content in dozens of languages, and increasing numbers are now dealing with in excess of 200 languages to one extent or another.

This large-scale change has driven interest in language technology because the human-oriented approach that worked well when FIGS (French, Italian, German and Spanish) was considered sufficient for international business have difficulty scaling to deal with 50 or 100 languages. Additional factors driving the shift include the rising importance of user-generated content and social media; the need for multilingual business intelligence; and the exponential increase in the volume of content that has been enabled through digital technologies.

The translation and localization industry has long used technology in the form of translation memory (TM) and terminology management systems, but for a variety of reasons it has not embraced other forms as readily. Most language technologies today have been deployed as monolingual applications without the multilingual support required by translators.

Machine translation (MT) is currently the best-known example to the public at large, driven largely by the success of free services pioneered by AltaVista's Babelfish and then made truly mainstream by Google Translate. The translation and localization community's acceptance of MT for production purposes has been considerably more reluctant and cautious, but even here it is making significant inroads. This increasing acceptance is leading to more interest in other types of language tech, such as grammar checking, personal assistants (such as Siri or Google Now) or opinion mining.

Considering just MT for the moment, it is no secret that it has not always lived up to the claims of proponents, some of whom have been predicting for at least the last 50 years that near-human translation quality is always just ten to fifteen years away. However, in the last decade, more and more organizations have embraced MT as a pragmatic way to help meet their translation requirements, often in combination with human translation and post-editing.

MT is currently at a crux: existing technologies have delivered great rewards, but their rate of progress has slowed as they have matured. This is not to say that technologies such as statistical machine translation (SMT) have played out, but rather that the "easy" gains in productivity and quality have been made and further improvements will require increasing amounts of effort. Many of the advances in recent years have come from combining various approaches and tools to take advantage of

Aljoscha Burchardt works as a senior researcher in the Language Technology Lab (LT Lab) of the German Research Center for Artificial Intelligence (DFKI) in Berlin, Germany. He is the project manager of QTLaunchPad and QTLeap. Arle Lommel is a senior consultant in the LT Lab at DFKI. He works on issues related to translation quality and standardization in QTLaunchPad. Georg Rehm works as a senior consultant in the LT Lab at DFKI and is the network and project manager of META-NET. Felix Sasaki is a senior researcher in the LT Lab at DFKI. He participates in LIDER and coordinated the project MultilingualWeb-LT. Josef van Genabith works as a professor of machine translation at Saarland University and as a scientific director at DFKI. He is founding director of the Centre for Next Generation Localisation (CNGL) in Dublin, Ireland. Hans Uszkoreit works as a professor of computational linguistics at Saarland University and as a scientific director at DFKI. He is the coordinator of META-NET and QTLaunchPad.

the strengths each has to offer. Examples of such combinations include hybrid MT systems (that combine the deep linguistic knowledge of rule-based systems with statistical systems' ability to learn from existing translations) and systems that integrate TM, terminology management and MT into translation cockpits run by human translators.

Where MT has continued to have trouble, however, is in matching the quality expectations set by human translators. Although many human translators use MT output as a reference in their translations, MT is alternately treated as a source of jokes and as an existential threat by many translators. At a recent international translation conference, developers of MT in particular were publicly compared to the "makers of the atomic bomb" for trying to take away translators' livelihoods. But setting such hyperbolic comments aside, it is clear that large amounts of content presently go untranslated because it is not cost- or time-effective to use human translators, especially when it is impossible to predict in advance what content will be needed when and by whom.

Other language technology applications have the potential to contribute significantly to the quality and success of MT, but so far many of them have been implemented only in standalone applications, often developed and maintained as research tools. This lack of interoperability has proved to be a barrier to integrating these tools, as production users seldom have the expertise needed to take often poorly documented open-source projects and combine them. However, the recent development of the Internationalization Tag Set (ITS) 2.0 specification provides one way for tools to interact in a standards-based approach that does not require users to hard code workflows by adapting software to every use case.

These issues are especially critical in the European Union (EU), with 24 official languages, 38 recognized minority languages and substantial communities speaking other non-European immigrant languages. Because EU law requires that EU citizens be able to communicate with their government in the official language(s) of their countries and that they be able to access the law in those languages, the EU invests huge sums of money in translation.

As META-NET's recent study "Europe's Languages in the Digital Age" pointed out, despite this investment, at least 21 European languages face the real possibility of "digital extinction." With little or no language technology development, speakers of many of these languages find themselves unable to communicate using their own languages and are forced to use foreign languages (usually English) or to stay silent. In addition, the discussion about issues of pan-European concern remains largely separated into language communities. As a result, there is lively debate among French speakers about the role of nuclear power in the post-Fukushima world, and German speakers have similar discussions, but they are not speaking with each other unless they engage with each other in English. But what are monolingual speakers of Basque or Maltese to do? If they are to take their place as equals in global society, they will require the assistance of integrated language technology that goes beyond just MT.

High-quality MT

In the last decades, translation quality has emerged as a major business issue for international businesses. Because so much of an organization's public image depends on the quality of the text they produce, they all claim to want the highest quality. Unfortunately, "quality" itself has been an elastic concept that often amounts to subjective and highly variable impressions from individuals. A joint survey conducted by the EC-funded QTLaunchPad project and the Globalization and Localization Association (GALA) revealed that over 60% of language service provider (LSP) respondents either used an internal quality model or had no (formal) quality model at all. These results show that we, as an industry, still lack any systematic approach to the quality question.

Laying aside the problem of a universal quality definition for a moment, Europe is the region where the lack of fast, affordable quality translation hurts most. Although MT systems keep improving, as can be observed in the performance increase of popular, freely available online translation services, their output is generally unusable for almost all outbound translation demands, and often now even as a source for cost-effective post-editing.

The problem is that most current translation services follow a one-size-fits-all-approach. Commercial LSPs and large institutional users of MT instead need to be able to tune MT effectively and at low cost to their production requirements. They need automatic tools to recognize MT output in at least three categories:

- MT output that can be used as is. Such output may not be perfect,

but instead meets requirements and specifications.

- MT output that can easily be fixed to meet specifications (such as post-editable content).
- MT output that should be discarded. In many cases it is faster and more efficient to translate from scratch rather than to post-edit bad MT output.

When tools are able to identify these quality groupings for a variety of scenarios, rather than providing a generic score for a large batch of translations; delivering entire texts of unknown quality to readers; or expecting humans to post-edit all content, then the tools can work with humans to leverage their strengths. Such strengths include resolving the meaning of difficult passages; translating expressions not previously seen by statistical implementations; or translating texts where artistry or careful attention to nuance is needed.

By focusing on how humans interact with MT output, the QTLaunchPad consortium is advocating for a paradigm shift. Instead of simply adjusting existing MT systems to produce marginally better results, it calls for a novel, human-centric approach to MT. This approach systematically addresses the goal of producing quality translations and takes into account the needs and priorities of LSPs, translators and requesters of translations. To pave the way toward a human-centric high-quality MT paradigm, the QTLaunchPad consortium has cooperated with stakeholder organizations such as GALA, Fédération Internationale des Traducteurs (FIT) and LSPs to develop tools and technologies that support this vision:

- The QuEst system provides fully automatic assessment of translation quality. Such assessment can be conceived of as a rough equivalent to the match rates that TM systems use to indicate how close a match is to what has already been translated and can guide translators in decisions about whether to accept, post-edit or reject sentences.
- Improvements to the open-source translate5 tool for editing and reviewing translations.
- An extension of the META-SHARE language technology exchange repository for MT research and development.
- The Multidimensional Quality Metrics (MQM) system for assessing quality with accompanying tools (such as a

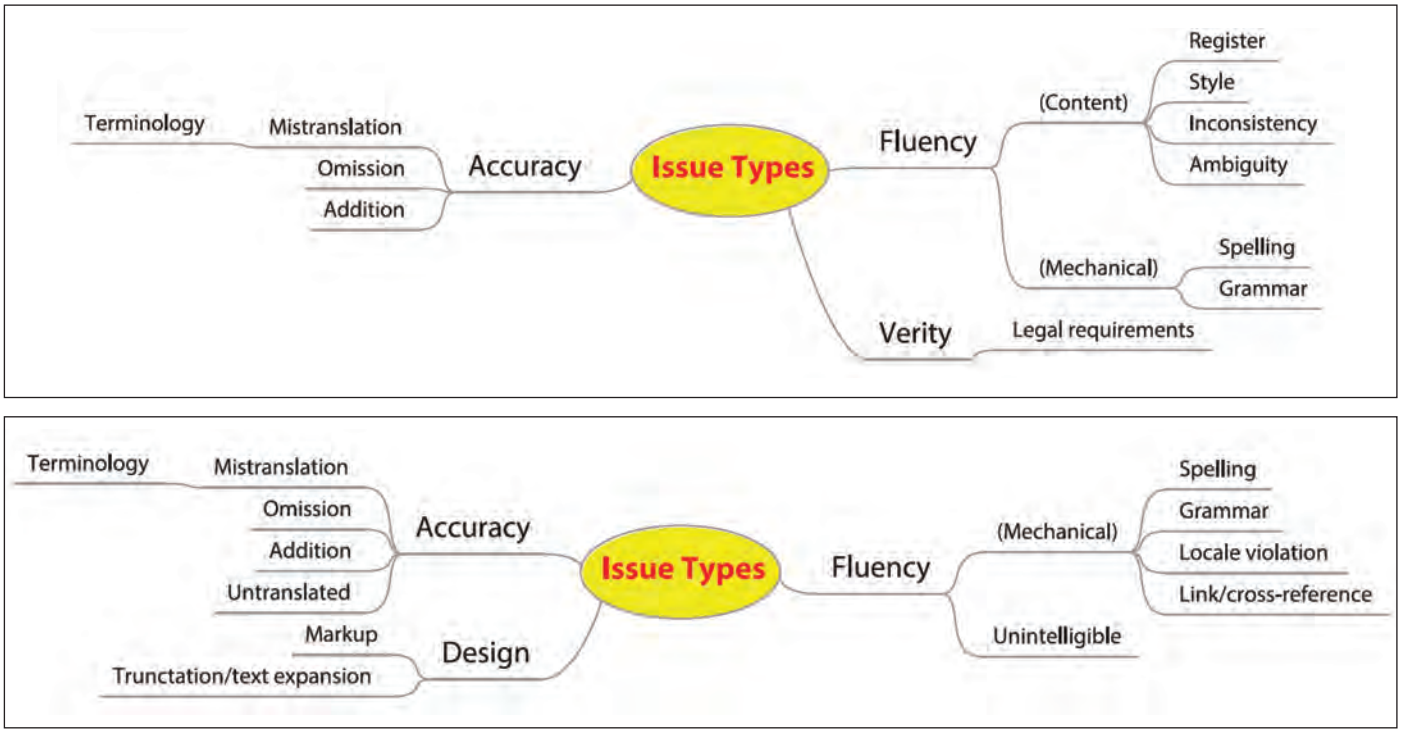


Figure 1: Two possible metrics defined by MQM. The top one, for legal translation, addresses style, ambiguity and legal requirements, among other features. The bottom one, for automated translation of an online help system, checks whether markup is processed correctly, whether text is truncated and whether cross-references are correct.

translation quality score card and an implementation within translate5).

The last item addresses the long-running disconnect between various one-size-fits-all systems for assessing translation quality. Rather than imposing yet another list of errors that all translations must avoid, MQM works within a framework defined by the ISO/TS-11669 standard to link quality assessment to translation requirements defined at the earliest stages of the translation process. It builds on existing specifications, such as SAE J2450 and the LISA quality assurance model, to create a flexible model for defining quality metrics, which vary to meet specific needs (see Figure 1). By using a shared framework, users can compare their results and tune them to their needs rather than being forced to use a generic model that may or may not apply. It also addresses issues in both source and target texts to allow the causes of problems to be identified and fixed.

MQM was developed to address both human and MT and bring them both under one set of quality metrics. While MQM will not replace MT metrics such as BLEU and METEOR, it is currently being used to drive research in the quality barriers that impact MT to identify those factors that help differentiate quality translations from those that are not usable. For example, it has helped identify certain grammatical features that are of particular concern and that correlate most strongly to human assessments of quality. The focus of this research has been on high-quality and "almost good" translations. Here it is enabling research rather different from traditional MT research, which has tended to emphasize quality improvements at the low end of the quality spectrum. The QTLaunchPad project is currently in the process of releasing MQM to the GALA CRISP program, where it will be maintained and developed by the industry as a free and open specification.

A second project, QTLeap, is focusing on improving MT by providing better integration between SMT methods and deep linguistic knowledge. SMT systems seem to have reached a point where it is difficult to achieve further quality improvements in a purely data-driven way. Despite widespread recognition of the advantages that linguistic knowledge can add to statistical methods, there has been a relative deficit in principled research in this direction. This lack of research is partially due to the fact that SMT systems that focus on the textual surface with little linguistic knowledge have done comparably well to this point. But theoretically, systems that focus on structure and meaning should be able to deliver better results and be less sensitive to the particularities of individual languages.

In order to pave the way for higher-quality MT, the goal of QTLeap is to deliver an articulated methodology that explores deeper, more semantic language engineering approaches such as using the sophisticated formal grammars that have become available in recent years. This new approach is further supported by progress in lexical processing that has been made possible by enhanced techniques for referential and conceptual ambiguity resolution, and supported by new types of datasets recently developed as linked open data. In order to ensure that the MT developments include performance improvements in a realistic scenario, the project consortium includes a company in the process of making its PC help desk services multilingual. The goal is to have a monolingual helpdesk database and to machine translate user requests and answers from the database.

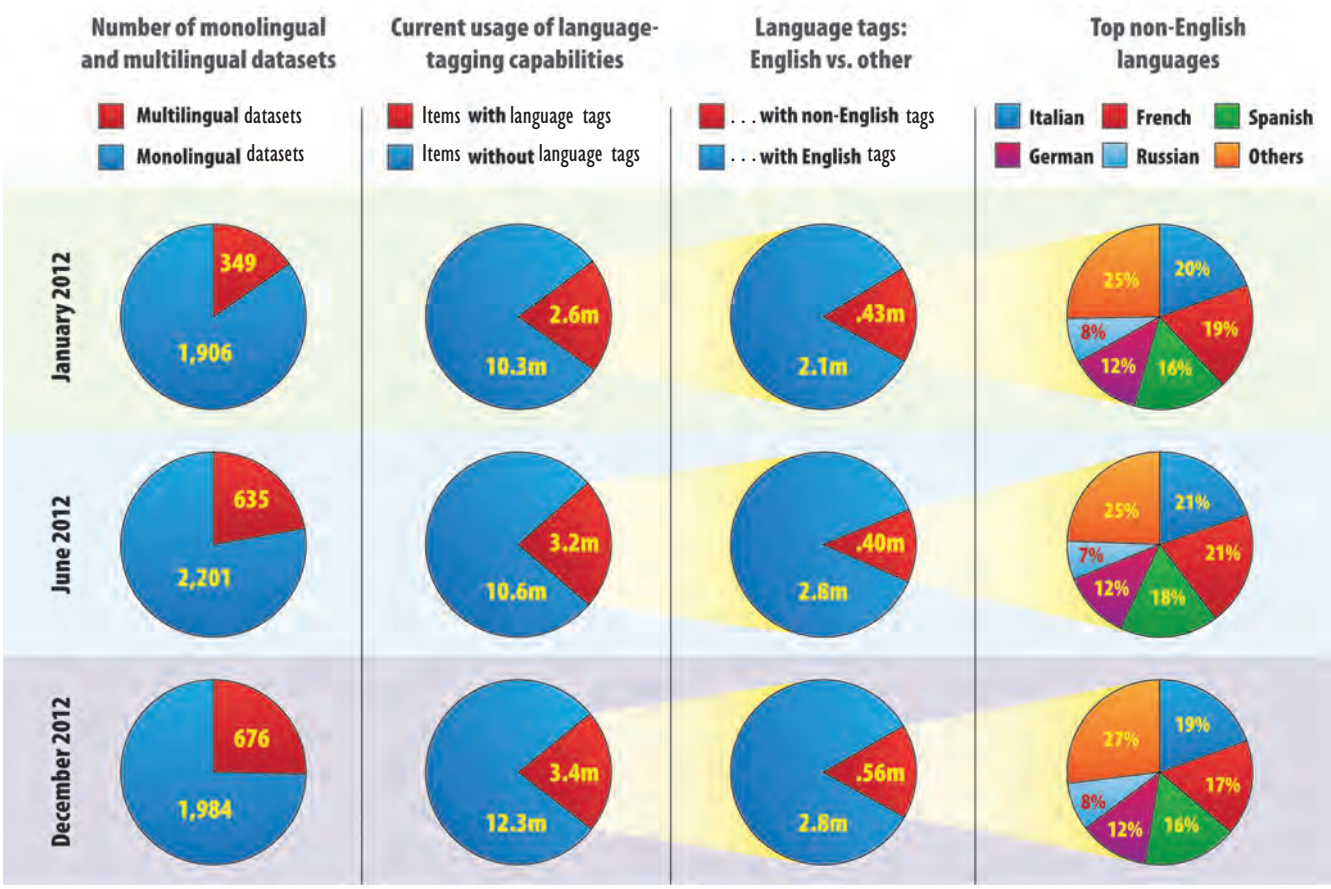


Figure 2: A small but increasing percentage of online linked data is identified specifically by language and multilingual sets.

Taken together, QTLaunchpad and QTLeap are pointing toward a new future in which MT takes the best of current developments and combines them into a human-centric approach.

Content analytics

Although translation is the most visible language technology application – because we immediately realize it when we cannot access a web page or enjoy a YouTube video in a foreign language – it is not the only application that impacts our industry. MT can be used only for content that is already known to be relevant, but cannot directly assist us in cases where we do not know that certain content is relevant. We are used to browsing through the first few hits on Google and other search engines in order to find answers to questions, but we are seldom aware of or care about the content we don't find because we do not see it, especially if it is "hidden" in multimedia content, database applications (the so-called "deep web") that cannot be found through simple keyword searches, especially when multilingualism is a factor.

Thus we may miss the most relevant content simply because it is in another language that does not exist in a textual format. For example, someone in the Basque Country may be looking for particular information on nuclear energy policy in Europe but not find the information because it is in a German-language YouTube video. Because such content will not be found, it will also generally not be translated. Similarly, if we have technical problems, the answers may exist in user forums, but finding the right answers from hundreds of incorrect, outdated or simply irrelevant search results is already a significant problem, even before different languages are factored in.

The answers to some of these difficulties can be found in recent developments in content analytics, a set of technologies for making sense of data. This definition is as broad as the diverse set of application scenarios that content analytics applies: sentiment analysis, business intelligence, opinion mining, intelligent web search and many others.

There are some basic technologies that are common to all content analytics applications. Named entity recognition helps to identify unique concepts and allows for disambiguation ("Paris" the city vs. "Paris" the mythological figure). Relation extraction identifies relations between entities (the city "Paris" is located in the country "France").

The actual implementation needed for content analytics technology is often quite language specific. It is no surprise that technology support for English is predominant. One current challenge in content analytics is to make such implementations available for a wide range of languages.

More and more multimedia content is being created on the web, which leads to another challenge: current content analysis technologies and implementations are largely limited to working with textual data. These tools will need to be enhanced to cover non-textual content such as audio. For example, consider the query, "find me all movies with kiss scenes at the end." So far no video on demand portal has such information

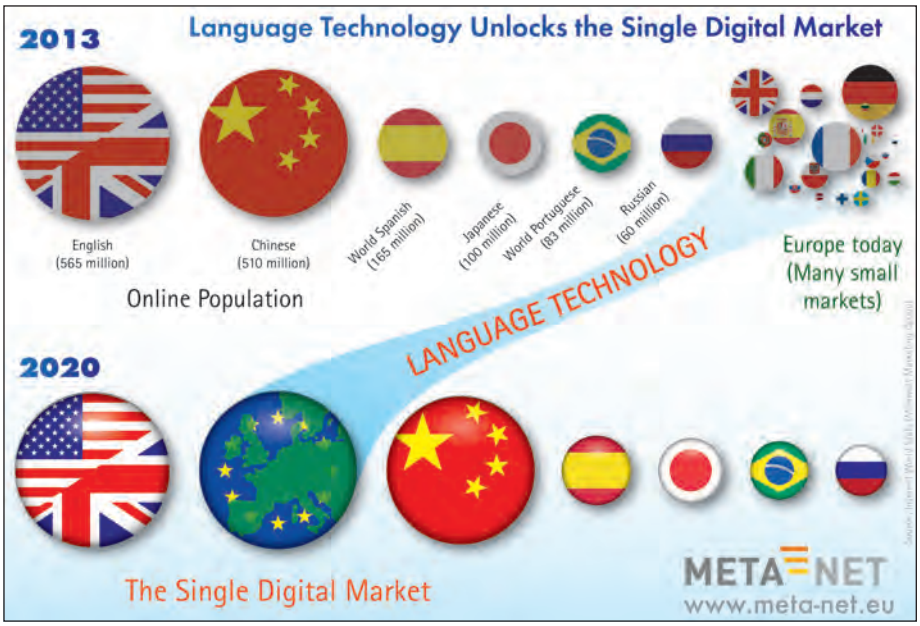


Figure 3: Language technology has the potential to unlock the European Single Digital Market by lowering language barriers, creating the second largest online market, with benefit for business both within and outside of Europe.

available, but content analytics can help create it. Since the quality of audio or video analysis is still rather limited, multimedia content analytics also rely on textual information available, such as subtitles and closed captions.

It now becomes clear that for improving multilingual and multimedia content analytics, one needs information about the content. More and more structured data sources are currently being created, without necessarily having content analytics in mind. A prominent example is Wikipedia's infoboxes. These provide structured information that can serve as seed information to improve content analytics. Wikipedia is also useful since it provides links between languages. In this way, it can build the path to truly multilingual content analytics.

Other structured and partially also multilingual knowledge resources that are relevant to content analytics include Freebase, the Wikipedia-based Wikidata effort and BabelNet. In terms of linked data, a cloud of more and more data sets is being created, and a growing (although still rather small) portion of this linked data is multilingual.

The main challenge for realizing the opportunities for content analytics and linked data is that, so far, the relevant communities are not aware of each other. Language experts have multilingual data available in various forms, such as lexicons, term bases and TMs. Linked data specialists create structured data out of resources such as Wikipedia, resulting in DBpedia, the linked-data counterpart to the resources created by language specialists. Finally, providers or consumers of multilingual and multimedia content may have ideas about requirements for processing multimedia items, but are generally not aware of the possibilities that content analytics may give them.

In all these cases, different groups are facing the same issues in different ways. Language specialists don't know how to convert data into multilingual linked data since established approaches to achieve this conversion do not exist yet. Linked data specialists, on the other hand, are generally unaware of the requirements imposed by multilingual data and often design systems that do not work properly with it. In any event, there are not yet many content analytics applications that make use of linked data in the way described here.

Linked data and multilingual/multimedia content analytics is the core topic of the LIDER project (http://lider-project.eu/). LIDER will build the path toward a linked-data cloud of linguistic information to support content analytics tasks in unstructured multilingual cross-media content. The LIDER consortium consists of key research groups in the realm of both language technologies and linked data. With input from various industries, LIDER is creating a roadmap around industry-focused content analytics use cases with a view toward defining needed research steps. To ensure that it gathers input from a range of communities, LIDER's outreach and dissemination efforts are taking place via the World Wide Web

Consortium (W3C) and its Multilingual-Web initiative. This initiative has proven highly successful in bringing together diverse groups of people interested in multilingual issues.

Developing language technologies

In an effort to combat Europe's linguistic fragmentation and to support the goals of the European Commission toward a single digital market, the EU funded the development of META-NET, comprised of 60 research centers in 34 European countries dedicated to the technological foundations of a multilingual, inclusive and innovative European society. META-NET created the Multilingual Europe Technology Alliance (META), with more than 750 organizations.

META-NET worked to support monolingual, crosslingual and multilingual technology support for all European languages (Figure 3). The future paths laid out in its Strategic Research Agenda (SRA) for Multilingual Europe 2020 are connected to application scenarios that will provide European research and development with the ability to compete with other markets and subsequently achieve benefits for European society and citizens as well as opportunities for the European economy. Two themes focus upon core technologies and resources for Europe's languages and a European service platform for language technologies.

The goal of many of these projects and currently planned actions is to turn META-NET's joint vision into reality and enable large-scale opportunities for the whole continent.

An important aspect of META-NET's suggestions centers around the idea of providing high-quality translingual technologies instead of focusing on tools for gist translation. Projects are already working actively on the topic by systematically identifying barriers for quality translation and pushing their boundaries. In addition, META-NET has worked to lower the barriers to access for current language technology applications and resources through META-SHARE, an online portal that provides access to these resources.

After years of development in disconnected projects, language technolgy is finally being adopted by users around the world to meet their requirements for access to content and to interact around the world. While there is still a long way to go, a variety of developments, many of them centered in Europe, are starting to break through the barriers. As new projects appear, the shift is toward a user-centric perspective and toward adoption and integration. **M**

Acknowledgements

This article has received partial funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement numbers 296347 (QTLaunchPad), 610516 (QTLeap) and 610782 (LIDER).

Post-editing MT: Is it worth a discount?

Alessandro Cattelan

Machine translation (MT) is undeniably an amazing, albeit controversial, technology. At times, its poor performance and errors in dealing with ambiguous texts make us chuckle; other times, MT output is so unintelligible that it leaves us puzzled.

Nevertheless, internet users love MT and are using it to a great extent. Interestingly enough, however, while over 200 million web users use MT from Google Translate alone every month, only a few translators and language service providers admit to using it for their work. Professional translators often point to how it actually reduces translation quality and productivity. Such reasoning, however, is often based on anecdotal evidence and on using the wrong approach when it comes to integrating MT in the professional translation workflow.

We have been using MT extensively for many years, and are currently using it for light post-editing (around 10% of Translated.net's turnover) and for over 50% of our standard translation projects where it is used as an extra suggestion along with translation memories (TMs). While constantly exploring new ways to further integrate MT in all of our processes, we are developing software to make it the standard suggestion source in computer-aided translation (CAT) tools.

Software alone, however, is not enough to guarantee that MT will turn out to be successful in the translation workflow.

Alessandro Cattelan is the Director of Localization Operations at Translated.net. He is also Product Manager of the MateCat translation tool and is responsible for the overall coordination of its development.

There are other factors to bear in mind, among which is whether translators are willing to use MT and how such technology affects their productivity and income. We tend to assume that using MT results in savings for customers and lower rates for professional translators. However, this is not always true and it seems that we still need to understand whether it makes sense to apply discounts for MT post-editing both for customers and translators and, if so, to what extent.

MT and translation providers

The translation industry has not proven too keen on adopting MT on a wide scale and many professionals in the industry still dismiss it as a laughable, mostly useless technology and refuse to adopt it for their work – or do they?

In fact, translators are probably not as disapproving and opposed to MT as it would appear by reading the comments that are so often published in online public forums. Observing the success that the MyMemory SDL Trados plug-in had combining MT and collaborative TMs, it seems that translators are actually quite willing to use MT – at least, when such technology is well integrated and not imposed on them, and when they can reap the benefits of using it.

Professional translators, language service providers (LSPs) and clients alike clearly understand that MT can prove an effective means to improve productivity and therefore reduce turnaround times and translation costs. Sure, one could debate whether MT should indeed be used in any translation project or whether it should be restricted to specific projects. Yet the decision on whether to adopt MT usually boils down to one simple question: assuming that the desired quality level is guaranteed and that the processes allow for the use of such technology, will MT improve translators' productivity?

MT quality is not always predictable. It depends on a number of factors related to linguistic and technological issues: some language pairs are inherently more difficult than others to translate

via MT, while for some other languages or domains there are not enough training data to build an effective MT engine. This results in an inconsistent effort required by translators in order to produce a translation starting from MT output. As a consequence, there is no one-size-fits-all approach to setting translators' and clients' rates for projects where MT is used.

Translated.net carried out a first attempt to solve the problem through an analysis of the purchase order acceptance rate when offering an MT post-editing option. During this experiment, we sent out purchase orders offering two options: translators could either be paid their full per word rate for a 1,000 word translation starting from scratch or the same rate for a lower number of weighted words, such as 700 (1,000 raw words - 300 words discount due to MT matches = 700 words), for post-editing MT output. We started offering 500 weighted words and slightly increased the number of words paid for the post-editing job until over 75% of translators were opting for the post-editing job over the standard translation job.

The number that prompted translators to switch to post-editing varied depending on the language pairs: for English to French and English to Italian, it was around 730 words – which meant that MT matches allowed the translator to achieve a 27% discount. The opposite happened with English to German, where the number had to be increased up to 1,100 words. If we wanted our translators to accept post-editing jobs in this language pair, we actually would have had to pay them more than for a standard translation.

This approach to defining the appropriate rate for post-editing jobs is quite fair as it gives translators full control over the best way to increase their productivity. However, it was only effective as a means to empirically understand how much MT was helping translators, but proved unfeasible for broader implementation. A more practical and usable way to measure how much MT improves productivity was needed.

Defining productivity

Defining whether MT would be useful in a specific project and to what extent it would reduce the turnaround times and costs is indeed a delicate task. Some research has been carried out on the subject, and a number of metrics have been proposed with the goal of predicting the quality of machine translation for a given project and hence the usefulness of MT. However, the key relevant element for all stakeholders involved is productivity as related to the actual effort to produce the desired output, be it a ready to publish translation or a "good enough" post-edited text.

Productivity can be expressed in terms of two performance indicators:

- Time to edit: the average number of words processed by the translator in a given timespan.
- Post-editing effort: the average percentage of word changes applied by the translator on the matches provided.

The first indicator directly expresses the time labor required by the translators, and hence improvements on this figure are directly related to cost savings. The second indicator measures the quality of the matches provided by the TM and MT. This corresponds to computing a distance score between matches provided by the system and the post-edited version submitted by the user. The indicator is indeed an estimate of the percentage of edit operations performed in the whole set of translated segments.

The ability to understand to what extent MT can increase productivity allows the identification of when such technology can be integrated into standard TM tools. Measuring productivity using the abovementioned performance indicators requires the collection of a large amount of data from real translation projects that have leveraged MT. In order to reliably measure productivity gains and collect post-editing data, specific technologies are required to record the translators' editing patterns and interactions with the software during translation, and the time needed to perform a given post-editing job. Together with the research organization Fondazione Bruno Kessler, the University of Le Mans and the University of Edinburgh, Translated.net is working on MateCat, a European Union funded project that has among its goals the development of an enhanced web-based CAT tool integrating new MT functionalities.

The MateCat tool is an enterprise level CAT tool that can be used in real translation projects to collect information on the editing patterns and time to edit of each segment post-edited or translated by professional translators. It is able to collect:

- Matches provided by the TM server (if any) with their relative quality match.
- Matches provided by the MT engine (if any) with their relative quality match.
- Target segments edited by the translator.
- Time taken to edit each segment (measured by adding the time used to perform multiple edits on the same segment).
- Post-editing effort measured by the word edit distance between the first match provided and the final translation.

The information is displayed in real time on a web interface and is also available in a CSV file, which allows for in-depth analysis of the results of each field test. Such data can then be analyzed to draw up statistics on the performance, and hence predict the usefulness of machine translation in specific language pairs and domains.

We believe that MT will become the predominant technology for production and that it will be integrated with current TM technology, so as to be used in the broadest range of projects. To this day, however, there are still no industry standards or common practices on a fair payment scheme for post-editing jobs, as there are for translation jobs where TM is used. Current CAT tools do not integrate the time-to-edit or post-editing effort measurements to allow for a fair and effective MT quality evaluation.

Translated.net is approaching the problem by developing technologies to measure the average time-to-edit in post-editing projects so as to understand what is to be expected in terms of productivity improvements from adopting MT. This will eventually provide a solid basis of statistical data to draw up accurate payment and cost schemes. Our initial results show that post-editing data rich and morphologically simple languages, such as English, French, Italian and Spanish, require an effort comparable to fixing a 75-99% TM fuzzy match (and by consequence would be paid about 60% of the full rate for new translations). Morphologically rich languages such as German and Czech do not appear to allow room for any discounts.

As of today, however, the available quality metrics and tools do not help much in predicting whether, and to what extent, MT is useful for translation providers and buyers alike. An open discussion with translators and customers still seems to be the only viable solution for LSPs. **M**

Cloud security for SaaS translation providers

Shannon Zimmerman

We don't have to go far to find someone affected by, and justly concerned with, the ongoing news blitz surrounding Edward Snowden's security leaks. From those of us worried about our stored customer account information on consumer websites to multibillion dollar enterprise organizations worried about exposed sensitive data, the notion of cloud-based data security is on a lot of minds lately. At the same time, the lure of cloud computing – stemming from ease of management and scalability – has resulted in more than 90% of all organizations at least discussing cloud use in 2013, up from 75% one year prior, according to a survey by Symantec.

Of course, in the language services industry we're handling extremely sensitive and confidential client data every day. We are responsible for countless gigabytes of it in various forms: translation memory (TM) files, terminology bases and mountains of source content including proprietary information. This places the language services vendor that offers cloud-based software as a service (SaaS) front and center in the discussion around cloud security.

One of the most important areas we address with clients is how secure our data storage and systems infrastructure are, both cloud-based and physical storage. Enterprise organizations expect the same level of sophistication that their own operations run on. At the same time, we have to acknowledge today's increasingly common attitude of circumspection around cloud-hosted data.

Even though cloud security seems to have made its way into the common consciousness, companies that are seeking translation management system technologies don't always think to address the issue when comparing vendors. For that reason, it's valuable to point out the steps that the more tech-forward language service providers are taking to ensure reliable cloud-based translation technology.

Security concerns can lead to lost clients

Whether or not a company can trust a vendor to protect its sensitive data from prying eyes can make or break a business relationship. We recently had a situation in which an enterprise-level organization came to us, reeling from a previous vendor's lack of system security. The company learned belatedly that the language service provider (LSP) had been storing clients' TMs on a public file transfer protocol site. This led to all of the LSP's clients having access to one another's TMs. As just about anyone would agree, sharing intellectual property doesn't exactly lend itself to gaining a competitive edge.

Naturally, our early talks with this company included how to make sure that this kind of unintentional data sharing never happens again. This frustrating and alarming experience led to the decision to pack up and move on. Not all translation service providers with cloud-accessible software follow the same standards, but many do abide by common best practices. From a client perspective, it's critical to find out as much as possible about a potential vendor's security system. After all, no one wants to be in the position of realizing too late that his or her data has been compromised.

Thus, it's always a good idea to ask as many questions as possible when evaluating a vendor's translation technology, especially regarding how it's hosted. Many people are turned off initially by the term "cloud-based." Because the phrase appears

Shannon Zimmerman is cofounder and chief executive officer of Sajan, a global language services and technology provider.

in countless news articles and gets tossed around with abandon, the actual meaning and distinctions within it can sometimes become lost or hazy. Some might assume files are just floating out in the internet ether, unprotected and exposed. This isn't necessarily true, and vendors are taking some security measures to guard against the data sharing liability I mentioned before.

A translation vendor with cloud-based software doesn't do itself any favors by not offering industry standard 128-bit encryption for data transfers between itself and the client. However, it's pretty rare for companies not to take this commonplace precaution. It's a way to prevent unauthorized interception of data during the file transfer process. While this may seem like common knowledge, and even a given that a translation company has this in place, not every translation buyer knows to ask about it.

Saying a system is accessible from just about anywhere sounds very appealing. But the initial feeling of intrigue can turn into wariness if a potential client views that from the perspective of vulnerability ("does that mean anyone can tap into it from anywhere in the world?"). This is where controls and credentials come into play.

Accessibility isn't worth very much on its own without the ability to control who accesses what. Role-based accounts provide for greater security because it serves as a gateway for everyone who might touch the translation process, from linguists to project managers. Each system user is set up with a profile that lays out what he or she can see within a translation management system, for instance.

In the coming months, we will likely see heightened sophistication with how much these role-based accounts can be fine-tuned. Some providers of cloud-based workflow technologies are working on getting more detailed with who can access what information once logged in to the system – such as translation project requestors in a given department only having access to certain types of projects.

One change we may see in particular is authenticator integration with other systems. In effect, it allows a user of another system to log in to a translation management system using his or her credentials for, say, the user's organization's intranet authenticator. The main benefit of this kind of login compatibility is that users don't have to remember another password, in addition to the sheer convenience of it.

While not every translation buyer may request it, another important way for a vendor to demonstrate data security is by offering up its cloud-based system for hacking. It's considered a best practice for companies in our industry to put this on the table. Either the translation buyer or a third party can attempt to physically hack the system, the results of which can quickly determine whether it's up to the organization's standards. One of our clients asked to do this when they were first getting familiar with our solution and found that our security infrastructure even exceeded their own.

It's also a wise practice for a SaaS vendor to have a third party perform penetration testing. We do this every year as a matter of course. The external company tries to figuratively scale the walls of our tools and break into our internal systems from the outside. During the process, they check for any vulnerabilities that require attention. For a potential client, it can be valuable to have access to these reports, which spell out how exactly the third party company conducted the tests and how it arrived at its results.

Frequent data backups also lend an extra measure of security – and reassurance – for any companies that might be leery about cloud computing. While this is also standard across service providers, some clients may not be aware of the frequency of data backups and plans in place in the event that any security breaches or power losses occur. Daily information backups, both onsite and offsite, in addition to having another cloud-based server to push data onto, help ensure that client information won't be subject to loss or theft. These are things we often educate buyers about if they have reservations about how the data is stored and protected online and offline.

Bringing cloud security down to earth

Does it take a veritable fortress of impermeability to ensure that client data won't be compromised? Absolutely not. While cloud accessibility may seem inherently risky, we in the language services industry do have capabilities to lock down data, however it's accessed and stored. I believe we will begin seeing even more sophisticated measures to strengthen the virtual gatekeeper for cloud-based systems, especially as investment in IT and software engineering increases. **M**

Dreams of better terminology tools

Tatiana Gornostay

Terminology is at the very heart of our linguistic landscape. In everything we do – fixing our cars, preparing meals, taking medication, even enjoying our hobbies – we come into contact with specialized language units. In language science, these units are called terms. Terms are not just important for scientists or professional language workers; they play a significant role for all of us. By being more aware of terminology and its evolution, we can take better care of the treasures of our language.

As language workers, we see that correct, consistent terminology is becoming more important and complex than ever, thanks to the multilingual environment we live in. For instance, we have 24 official languages in the European Union. In many spheres of the linguistic landscape, texts must be translated in each of the official languages. Terminology is the key to making translation clear, consistent and precise.

With the rise of web technologies and the boom in online data, we are also seeing a huge increase in the number of texts that need to be translated. This is putting pressure on professional translators, who form the backbone of the linguistic landscape. Multilingualism is an important heritage feature that we are all struggling to preserve; our task as language professionals is to support these efforts on a professional level. Of course, each translator has his or her own unique knowledge, for instance, of a specific subject field. A translator cannot be an expert in everything. Therefore, the way we organize our professional activities, as well as acquire and manage our knowledge, is supreme.

This thriving multilingualism is originally what led me to become interested in language science and to devote my life to terminology. During my childhood, growing up in a multicultural family in the Ukraine, several languages always surrounded me: Ukrainian, Polish, Belarusian, Russian and, a bit later, Latvian. Sometimes I got these languages mixed up, and I am still not sure which was my first – it was probably a mixture of languages. For precisely this reason, we always had a lot of dictionaries at home. I loved to compare words in them, leafing through the entries, and clearly recall how the thickly bound volumes sat in a row on a bookshelf in my family's home.

In today's multilingual world, I've come to the realization (as have others in the language field) that we must take a new look at trends in terminology. We must think beyond the conventional praxis and static models that no longer fit user requirements. Changes are required, and innovation is being brought into focus to introduce novel patterns of language work. We need new tools to reflect, and to integrate, these profound changes into our terminology work. This raises a few questions, of course. What would our "dream" terminology tool or workstation look like? How would it work?

First and foremost, a dream tool should be friendly to its user. A language worker uses various language tools. Text editors, spelling and grammar checkers, electronic dictionaries and databases, computer-assisted translation tools, machine translation systems, voice recognition devices – these have become indispensable tools in our professional life. It is important that we enjoy the tools we use and the way we communicate with them. We want the tool to be friendly, even exciting. The less time we spend on routine operations (for example, term extraction and lookup), the more we have for our core tasks.

Tatiana Gornostay is a terminology service manager at Tilde. She also works as a translator trainer and an English–Russian freelance translator. Tatiana received a PhD in computational linguistics in 2010 in St. Petersburg.

Language workers can spend up to one-third of their time on terminology work. In some cases, terminology can consume an even greater share of their working time. A terminologist studies a concept and creates a term or identifies it in a text. A writer utilizes the term in the text he or she is creating, and must use terminology consistently to prevent contradictions. A translator communicates the concept by means of a translation equivalent in a target language. Even two languages can pose a problem if your team is working with 24 official European languages. Without a doubt, a terminology workstation should guarantee a collaborative work process, ensuring that a language worker is no longer alone in his or her task.

A dream terminology workstation would also save us time and money. Diligent terminology work is time-consuming and therefore expensive. The more professionals make use of existing terminology, and the more they are involved in its elaboration, the higher the return on investment is. Conventional media for terminology work, such as desktop- and server-based tools, are not sufficient for engaging language workers of different profiles. Cloud-computing technology is one of the relatively recent revolutionary advances in information and communication technologies that allow for constructing flexible services. This is now becoming a novel pattern in language work.

Though a number of tools currently exist to support terminology work, there is no single solution that could cover all the major steps within a term life cycle, from identification to translation and further exploitation in other language applications. Existing or available tools are not adjusted to new trends in terminology work. For example, few tools integrate facilities for corpus work, most tools have limited language coverage, few tools have sharing facilities and are adherent to ISO standards, and no tool is based on cloud computing.

A terminology as a service (TaaS) project presents a brand new solution that brings sophistication and advanced approaches to terminology work. It proposes an automated approach to terminology identification applying linguistic intelligence. One of the main advantages of the new terminology service model over other existing terminology extraction tools is its capability for processing languages with rich morphology. Other functionalities include translation lookup using major terminology resources (for example, EuroTermBank and IATE) and web data; facilities for collaborative terminology refinement and approval; export in popular formats used by the community, such as TBX (TermBase eXchange), CSV (comma-separated value) and TSV (tab-separated value); refinement of raw monolingual and bilingual terminological data; and sharing and using the resulting terminology.

We foresee the necessity for an interoperable working environment supporting the evolution of the internet and emerging Web 3.0 technologies. It is therefore compulsory to implement standards that can be used to exchange terminological data between different applications and systems – for example, updated XML-based standards that allow for interoperability with the Linked Open Data community. Thus terminological data will be an important part of the semantic web and will be accessible not only by typical terminological applications.

Enabling smaller languages in emerging markets

This new service model could be particularly beneficial for language professionals who work in emerging markets. Many of our emerging markets have smaller languages. These areas have to rely on even more translation tasks and volumes to make their voices resonate across the world. For example, here in Latvia, where I work, our language is spoken by just over one million people. Therefore, translation is the only way we can make our language heard across Europe. Likewise, we are constantly inundated with texts from the major languages – such as English, Russian and German – that need to be swiftly translated into Latvian.

In these emerging markets, translation is often the way in which new terminology enters the languages for the first time. Translators are thus endowed with a great responsibility: to introduce terminology into their countries. The new TaaS terminology service is an effective solution for ensuring that the introduction of terminology is sound, consistent and logical, and that the same terms are chosen by a large number of translators.

These developments for terminology, and indeed for language as a whole, are something I could have only dreamed of as a language-loving child growing up back in the 1980s, leafing through those dusty dictionaries at my family home. **M**

Evolution of cloud-based translation memory

José Gambín

Cloud-based sharing of translation memories (TMs) has occurred at a much slower pace than we first expected when we started to learn about this technology, partially due to lackluster adoption by freelancers. A previously unpublished survey answered by 1,302 participants was conducted in February through Proz.com, one of the major internet portals for professional translators, to document the topic from the linguist's perspective. Language service providers (LSPs) will need to make an effort to address their concerns if we want to keep working with the best translators on a cloud-based setup.

In the mid-1980s, we first found software whose main capacity was the creation of a database (or TM), fed with the work of human translators. Sharing of TMs has been happening since the very conception of computer-aided translation (CAT) tools and all company-level editions of this kind of software incorporated the ability of sharing databases over a local area network.

The idea of sharing TMs over the internet was the next logical step in the development of CAT tools, and from the beginning of this century, the first solutions connecting linguists through the internet entered the market. ForeignDesk from Lionbridge was one of the first solutions approaching this type of collaborative work. It was not based on a centralized TM, but instead it was a repository of projects on each linguist's computer connecting to the rest of the team. Other pioneers were Telelingua with T-Remote, an add-on that was able to connect, for example, a Trados TM over the internet, and Logoport, a web service based on the software as a service (SaaS) model that connected team members to a central TM hosted at Logoport servers.

Today, ForeignDesk is an open source solution, thanks to the generosity of Lionbridge. T-Remote never had a real impact on the market and the company stopped its development in 2005. Logoport was acquired in 2005 by the developer of ForeignDesk.

Idiom WorldServer had a modern and singular approach to online TM sharing. It incorporated two ways of sharing a TM. Firstly, connecting through a desktop application to a central TM; the singularity being that the central TM was not fed in real time. Translators received 100% and fuzzy segments on a local TM and were able to do a concordance search in the central TM. Team members could update the TM from time to time. This approach was designed to overcome the most important factor affecting the adoption of online TM sharing: infrastructure. Real-time TM reading and writing over the internet needs stable and powerful internet connections, and Idiom's approach minimized the impact of this, making its approach practical when working with low-quality connections. For real-time sharing of TMs, WorldServer included a web-based interface, which we would now call a cloud-based solution.

While ten years ago we only had four or five solutions using online TMs, today we can find more than 40. We now classify CAT tools in two different categories: desktop-based or cloud-based. For higher flexibility, we find both approaches being addressed by some developers such as WordFast or Kilgray (memoQ).

One of the most important changes that we can appreciate in these ten years is the proliferation of this type of solution, with a clear trend toward cloud-based solutions. Proliferation means competition and competition means lower and

José Gambín, sales and marketing manager at AbroadLink, has worked as a freelance translator, in-house translator, desktop publisher and project manager.

flexible pricing. Today a group of translators can offer this technology to its clients while in the past only big corporations had the infrastructure and money to do so. It is still true that the level of development is very different and that this will be reflected in the price, but high-end solutions are still more affordable today than they used to be.

OmegaT, probably the open source solution with the greatest impact on the CAT tool industry, now offers the possibility of sharing an online TM. Even omnipresent internet giant Google has entered the game with Google Translator Toolkit, with the "fee" consisting of them having the right to use your translations to power their statistical machine translation solution. Many of the big LSPs also have their own proprietary systems.

Despite all of this, we can state that this technology has not yet fulfilled its whole potential and the barriers are still the same – internet infrastructure and the inherent hurdles involved in team work with a freelance base.

In a recent report commissioned by the Internet Corporation for Assigned Names and Numbers, The Boston Consulting Group analyzes the main factors that hinder the full realization of e-business. They call them e-frictions.

Infrastructure accounts for 50% of e-frictions, broken down into access, speed, price, traffic and architecture. The infrastructure e-friction will be the factor we need to consider with regard to the implementation of workflows involving the use of online TMs. This has been one of the main reasons why this technology is not more widespread in our industry, and the reason why it has grown in importance in recent years, with a lot of players appearing on the market. In our particular case, it was what prevented us from investing in this technology in 2005, after having the experience of working with an online TM in a translation project for another LSP and realizing that we had to spend a considerable amount of unpaid time waiting for the responses from the TM.

Nowadays, this is still the same if we want to work with a team located in countries such as Morocco, Pakistan or Nigeria. The abovementioned e-friction index model establishes a classification for 65 countries according to their infrastructure. The top country in this classification is Sweden, with an e-friction score of 14, and the last is Nigeria, with 85 points. This type of classification can help project managers to establish areas of collaboration where connection speed and stability are not a source of risk for their projects.

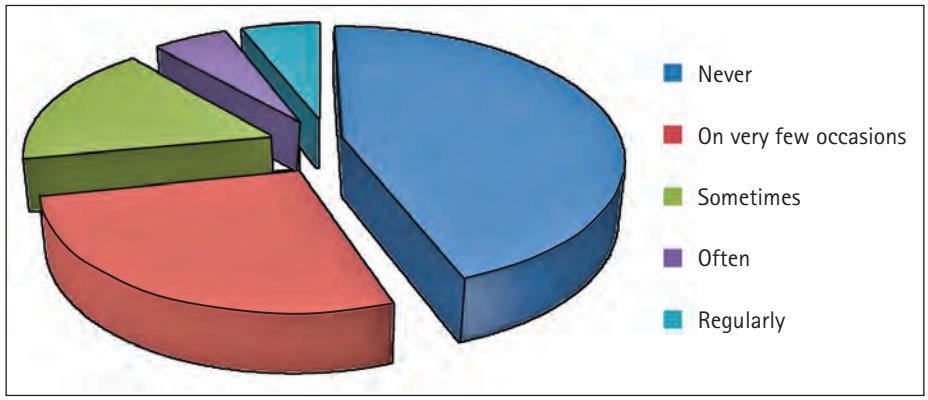


Figure 1: Answers to the question about how often freelancers work on projects that involve sharing an online TM.

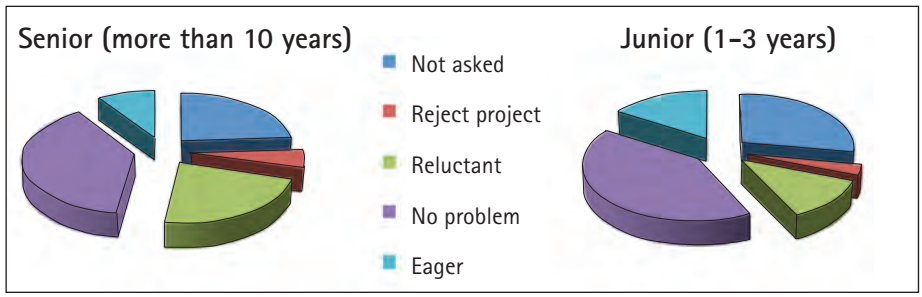


Figure 2: Breakdown of attitudes toward cloud-based translation tools by senior (left) and junior (right) fulltime translators.

Survey results and analysis

Many highly qualified translators are reluctant to work on a model that shifts the power balance to the LSPs. LSPs need to understand that if they want to work with the best qualified translators they will need to address all the freelance translators' concerns. These concerns, from a linguist's point of view, are still the same as those named by Garry Levitt over a decade ago in a 2003 *MultiLingual* article.

Proz.com's survey on the subject reached the same conclusions. Out of a total of 1,302 freelance translators who responded, 854 (65.6%) were full-time translators, 270 (20.8%) part-time translators and the rest were people taking translation work as a parallel activity. A full presentation of the collected data is available at www.abroadlink.com/onlineTMsurvey.pdf.

From the collected data, we have an indicator of the penetration rate of online TM sharing in our industry. According to the survey results, less than 6% of the translators regularly work on projects involving TM sharing (Figure 1).

If we delve deeper into the analysis of the data, we can see that senior translators do not participate in this type of project as often as junior translators. When filtering the responses by fulltime freelancers with more than ten years in the market, we observe that 48% answered that they were willing or eager to work with this technology and 11% actually work often or regularly with it. In comparison, 57.63% of full-time translators with one to three years of experience are willing to do so, and 16.87% already work often or regularly on such projects. See Figure 2.

One of the main objectives of conducting this survey was to give the freelance translator community a voice with regard to this technology. Question 7 presented the major identified issues from the translators' perspective and asked them to classify them in order of importance. The indicated issues were the following:

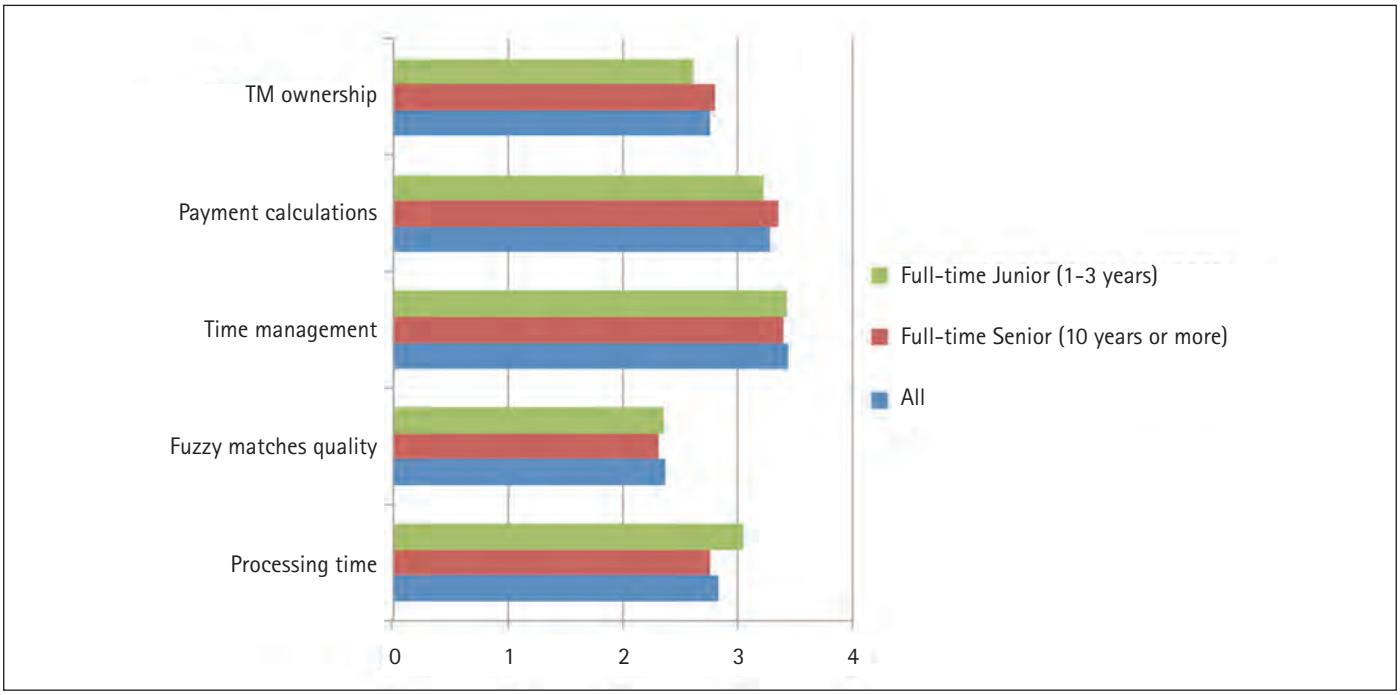


Figure 3: Answers to the question: "Please evaluate the following drawbacks of working with online TM solutions from 1 (most important) to 5 (less important)."

- Getting low-quality fuzzy matches from other translators working on the project that I will need to fix or that will appear as done by me
- Payment calculation
- Higher processing time that lowers my translation productivity
- Uncertainty of how long the project will take
- Not being able to keep a TM of my own translations

The order of the questions was set up randomly to avoid the conditioning of answers. Respondents were forced to choose a different value for each question. Translators were asked to evaluate these drawbacks of working with online TM solutions from 1 (most important) to 5 (less important). See Figure 3.

LSP and software development companies should take action to respond to these concerns. Regarding ownership, we have already solved this in most of the desktop-based applications where translators can keep their own TM locally (for example, in the case of SDL Studio or memoQ). This issue mostly affects cloud-based-only interfaces. In any case, this is a concern that can easily be solved technically if there is an agreement on that.

But the most important issue according to freelance translators is still the quality of fuzzy matches received from other linguists working on the project. As a matter of fact, this is an issue even when freelancers accept discounts for fuzzies in a TM sent to them to work locally. Working with online TM sharing makes this problem bigger as the TM is fed in real time. If translators are being paid according to fuzzies and new words calculated as they translate, it is important that all segments introduced in the TM by linguists are final, so that when the

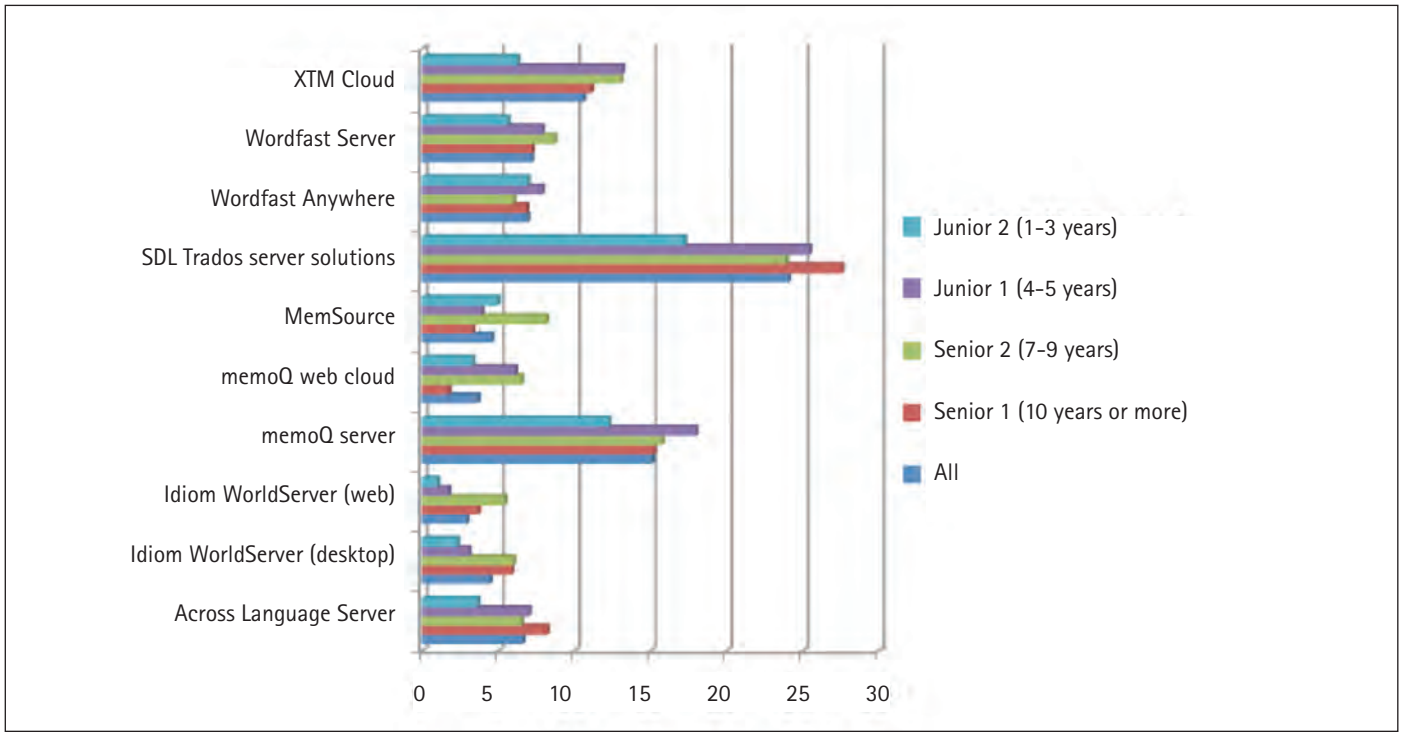


Figure 4. Answers to the question "Which of the following online TM solutions have you worked with in a team project over the Internet?"

other colleagues working on the same project find a fuzzy they can work with it. On the other hand, if translators translate a segment and then "sleep on it" before confirming the segment and introducing it into the TM, they may be preventing others from benefiting from their work, losing possible fuzzies and creating inconsistencies.

The rest of the aforementioned issues need the attention of LSPs, as their good handling of these matters will ensure successful projects and will guarantee talent retention. For example, LSPs should ensure good management of company IT resources and be aware of freelancers' internet connections to avoid time-consuming delays that annoy the end users of the system, lowering their productivity in a way that can affect the final delivery date of the project.

In regard to the software companies offering this solution, the survey ratifies SDL Trados server solutions as the most used, with 24.06% of respondents reporting having participated in a project using this technology. MemoQ Server is the second most used solution with 15.09%, and XTM Cloud the third with 10.60%. Figure 4 shows the most popular commercial software.

Like other technical solutions providing faster deliveries with a higher guarantee of quality, online TM sharing is here to stay. Competition on the software development arena and the SaaS model will improve these types of solutions and make them more affordable, enabling smaller players to compete for high volume projects. LSPs and freelance translators will keep improving their capacities, adapting to the new challenges for the sake of satisfying the needs of their clients. **M**

An invitation to subscribe to

MultiLingual

Ever-growing, easy international access to information and goods underscores the importance of language and cultural awareness. What issues are involved in reaching an international audience? Are there technologies to help? Who provides services in this area? Where do I start?

Savvy people in today's world use *MultiLingual* to answer these questions and to help them discover what other questions they should be asking.

MultiLingual's eight issues a year are filled with news, technical developments and language information for people who are interested in the role of language, technology and translation in our twenty-first century world. A ninth issue, the annual *Resource Directory and Index*, provides valuable resources – companies in the language industry that can help you go global. There is also an index to the previous year's magazine editorial content.

Two issues each year include a *Core Focus* such as this one, which are primers for moving into new territories both geographically and professionally.

The magazine itself covers a multitude of topics including these shown below:

Translation

Translators are vital to the development of international and localized software. Those who specialize in technical documents, such as manuals for computer hardware and software, industrial equipment and medical products, use sophisticated tools along with professional expertise to translate complex text clearly and precisely. Translators and people who use translation services track new developments through articles and news items in *MultiLingual*.

Localization

How can you make your product look and feel as if it were built in another culture for local users? Will the pictures and colors you select for a user interface in France be suitable for users in Brazil? How do you choose what markets to enter? What sort of sales effort is appropriate for those markets? How do you choose a localization service vendor? How do you manage a localization project? Managers, developers and localizers offer their ideas and relate their experiences with practical advice that will save you time and money in your localization projects.

Internationalization

Making content ready for the international market requires more than just a good idea. How does an international developer prepare a product to be easily adaptable for multiple locales? You'll find sound ideas and practical help in every issue.

Language technology

From systems that recognize your handwriting or your speech in any language to automated translation on your phone – language technology is changing day by day. And this technology is also changing the way in which people communicate on a personal level – affecting the requirements for international products and changing how business is done all over the world.

MultiLingual is your source for the best information and insight into these developments and how they affect you and your business.

Global web

Every website is a global website because it can be accessed from anywhere in the world. Experienced web professionals explain how to create a site that works for users everywhere, how to attract those users to your site and how to keep the site current. Whether you use the internet for purchasing services, for promoting your business or for conducting fully international e-commerce, you'll benefit from the information and ideas in each issue of *MultiLingual*.

Managing content

How do you track all the words and the changes that occur in your documents? How do you know who's modifying your online content and in what language? How do you respond to customers and vendors in a prompt manner and in their own languages? The growing and changing field of content management, customer relations management and other management disciplines is increasingly important as systems become more complex. Leaders in the development of these systems explain how they work and how they interface to control and streamline content management.

And there's much more

Authors with in-depth knowledge summarize changes in the language industry and explain its financial side, describe the challenges of communicating in various languages and cultures, detail case histories that are instructional and applicable to your situation, and evaluate technology products and new books. Other articles focus on particular countries or regions; specific languages; translation and localization training programs; the uses of language technology in specific industries – a wide array of current topics from the world of multilingual language, technology and business.

If you are interested in reaching an international audience in the best way possible, you need to subscribe to *MultiLingual*.

Subscribe to *MultiLingual* at www.multilingual.com/subscribe