

Supporting and developing
language professionals
worldwide

CIOL

Exploring translation technology

A guide for translators

María del Mar Strange
MCIL Chartered Linguist

About the author

María del Mar Strange

I am a Chartered Linguist with the Chartered Institute of Linguists. I have a degree in Applied Linguistics, a PG Dip in Business Administration and an MA in Online-Education. I trained as a Translator at EADS-CASA in Madrid. Since I moved to England in the 1990s I have worked in publishing, finance and as a translator and translator manager. In 2001, I established myself as a freelancer and since then I have done work for translation companies, in-company language training and adult language teaching for colleges. Currently I am researching Machine Translation from a linguistic perspective as part of my PhD studies.



**"IT'S NOT A FAITH IN TECHNOLOGY. IT'S FAITH IN PEOPLE."
"TOOLS ARE JUST TOOLS."
STEVE JOBS**

Exploring Translation Technology


There are three key factors which will influence the use of technology in translation: time, volume and cost. Whichever combination of these three elements is prioritised, the output should always warrant a translation of "acceptable" quality. The difficult issue is to define what we understand as "acceptable". The majority of professional linguists and translators will agree that an "acceptable" translation is understood to be a translation where the target text provides its readers with a comparable reading experience to the one of the source text's readers.

Dealing with large-scale translation projects in multiple languages can expose the limitations of human capability, especially when "ubiquity" and "instantaneity" are part of the requirements. In those cases, human translators can guide Machine Translation so it can reach an acceptable quality with the speed and volume required. However, when cost and quality are a preference rather than speed and time, translation technology can aid human translators to achieve the best results at fair rates. Finally, if speed, volume and cost are not a concern, human translators can dedicate their best creativity skills and expertise to ensure they deliver a masterpiece. Although in commercial translation contexts, the need for this type of translation is usually uncommon.

Nevertheless, translation technology, including machine translation, should be regarded as a set of tools to be used by professional translators whatever the situation. It could have very detrimental results, in any sort of written foreign-language communication, if translation technology was to be used by people with a lack of proficiency in translation. A good quality translation is the result of a comprehensive process of expertise. Starting with the initial research, going through the proper style and terminology search to end with the final revision and proofreading, not to mention the business skills also needed. The following translation tools are intended to be a brief sample of the translation technology currently in the market.

FILE CONVERTER SOFTWARE

Sometimes a client might require a translation in a specific format and other times a translation technology might only work with certain formats. Whichever your requirement (audio, video, spreadsheet, text, etc.), file converter software not only allows you to convert from one format to another, but also to convert multiple documents of the same format at the same time. Usually the input format is identified by their extension e.g. DOC, TEXT, RTF, XLS, PPT, PDF, JPEG, GIF, XML, HTML, HTM, ZIP, etc. Some examples of file converter tools are: Zamzar, Opentext, Online Converter, Convertio, Nchsoftware, Smallpdf, Conversion-tool.com, VideoProc, Handbrake, Cloudconvert, etc.



If you are scanning documents, you can convert them into text through Optical Character Recognition (OCR) software which recognises words digitally from a scanned copy, although they do not reach 100% accuracy. An OCR solution is sometimes integrated within an online platform. Nowadays, advanced OCR systems can extract text not only from scanned copies but also from documents and forms. This is of special importance in language applications which translate directly from a scanned copy or image of a text.

SPEECH TO TEXT

If your typing speed is not great or you need to use a device with poor typing access you can use Speech-to-Text technology to convert your voice into text as you dictate your translation. One of the most popular tools in the market is probably Dragon software. However, there are countless options equally effective like Braina, Apple Dictation, e-Speaking, Windows Speech Recognition, Google Docs Voice Typing, Speechnotes, etc. Some systems even offer automatic transcription features.


Equally, when your need is from text to audio, Text-to-Speech software will convert your text into audio for you. It might simply be that you want to learn how to pronounce a text in a foreign language, or you

want to proofread the translation by hearing the text repeated to you aloud. Some of the software available in the market is Balabolka, Natural Reader, WordTalk, Ispeech, Readerspeaker.com etc.

TEXT MINING

Text mining transforms unstructured text data into usable information through Natural Language Processing (NLP) by uncovering patterns in a collection of texts. Similarly, text mining can help the translator to discover and retrieve valuable insights related to the source and target domains, during the initial phases of a translation. The main approaches to text mining are the search for: Word frequency (most recurrent term); Collocation (words which appear together); Concordance (setting a word in context); Disambiguation (finding the right meaning for the right context); Text classification (assigns semantic categories to texts); Term extraction (extracts keywords and key phrases, named entities, product characteristics); Document summarisation (creates a summary with the most relevant information); Sentiment analysis (identifies subjective information such as emotions, intent, beliefs, etc.). Studying associations and trends through statistical analysis is also a key aspect in text mining.

Some of the main tools for text mining are, Google Trends which can find information on most commonly searched terms across the web



so you can understand which terms are more relevant to your target text. DeepDyve and Contentmine let you find millions of scholarly, peer-reviewed journal articles if you need to research your topic before translating. Project Gutenberg provides web access to over 59,000 completely free eBooks. FAQfox is a question scraper tool, so if there is a particular site where your target text is most popular you can ask the tools to only search that site. Carrot2 organises your search results into topics which can speed up finding context for a particular term. When dealing with text mining translators should pay special attention to copyright legislation.

PARSING OR SYNTACTIC ANALYSIS


In linguistics, a sentence is parsed into subject and predicate, phrases, modifiers, etc. in order to analyse the relations among them to extract semantic meaning. It is the process of establishing morphosyntactic categories. In computers, parsing is breaking a text in smaller units following rules so it is understood better by a computer. Parsing is an area of Natural Language Processing and it is used in Translation Machine applications. The parser breaks down the data of text in tokens and it gives each token a tag (this is also known as tokenisation). It can detect any syntax errors according to the underlying grammar. The parser produces a "parse-tree" on the source text and translation in the target text at the same time. There are two types of parsers depending

on how the parser-tree is generated: Top-down Parser and bottom-up Parser. Another name for parser is Syntax Analyser.

CORPORA

Corpus Linguistics is an approach to analysing how language is used in different contexts. Nowadays, this analysis is performed through computer programs, so the corpora must be stored in a format which can be read by the machine. A Corpus (plural: corpora) is a large collection of "naturally occurring texts" which have been stored electronically. Corpus linguistics analyses linguistic patterns in the language on the basis of frequency of occurrence of a key word, term or linguistic unit. Researchers might want to analyse for example the level of adjectives used in the texts of a particular author during a particular time in history, or how often a particular preposition is used in authentic texts from social media, etc.

Usually corpus can be compiled either from a collection of monolingual texts belonging to the same genre, topic or time period (specialised corpora), from bilingual or multilingual texts (comparable corpora) or from translated text aligned in the pair language (parallel corpora). The most widely used application of corpus linguistics for translations is parallel corpus which is used as the basis to create Translation Memories for a CAT tool. A translator would usually want to search concordances



to find a particular word in context rather than just its translated equivalent. Other applications for translation are to compare text previously translated in order to study the specific style of a company or writer. There are two sorts of corpora, unannotated and annotated. Unannotated corpora are in “raw” condition and need linguistic expertise to retrieve information. Annotated corpora have been coded by computational linguists so are machine readable.

Popular tools when doing corpora analysis are AntConc, AMALGAM, ANNIS, BSFU, CLiC, Concordance, Coquery.org, WordSmith, SketchEngine, among others.

You can build your own corpus by extracting the bilingual text from the internet in HTML, PDF, Ms-Word or plain text. There are applications, “web spider”, which will crawl and extract a large number of files from web pages, such as TextSTAT or SpiderLing. Some corpus analysis tools also offer this capability.

Examples of Corpora already available are COCA, the Corpus of Contemporary American English; OPUS corpus is a collection of translated text from the web; United Nations Parallel Corpus contains parliamentary documents of the United Nations which are open to the public; the English-Norwegian Parallel Corpus (ENPC)

includes fiction and non-fiction text translated in these language pairs; the Hansard Corpus contains all the speeches from the British Parliament, etc. Also, the book *The Routledge Handbook of Corpus Linguistics* is an extensive resource to explore this subject further.

TERMINOLOGY EXTRACTION TOOL (TET)

TET is a computer tool which helps translators and linguists build terminology glossaries by identifying possible terms in a large text or web article. A translator might want to build one glossary from scratch or develop a previously built glossary already in use by a particular company for purposes of consistency. Terminology extraction tools can be used either for monolingual or bilingual extraction and can also extract terminology automatically, Automatic Terminology Extraction (ATE). However, unless they are sophisticated enough these tools can produce a lot of “noise” (invalid terms) or “silence” (not enough terms) so the systems might not be very reliable on their own and might need human intervention. Several ATE tools have been developed for specialised domains or for specific language pairs and generally are more accurate when working with consistent terminology.

Some of the most popular applications are Rainbow, SDL Multiterm Extract, TermSuite, Translated s.r.l., Terminus, TerMine, TermExtractor, TermoStat, QTerm memoQ, Acrolinx, Flashterm, Translate5, etc.

TEXT ALIGNER

Text aligners align parallel text to examine the source text and the target text horizontally. The programme also matches source language sentences with target language sentences so you can use this bilingual text as the basis for a Translation Memory to be used in a CAT tool. Some CAT tools like Wordfast or SDL Trados offer alignment capability. After aligning both files the program allows you to download the file in Excel format to be used off-line.

One thing to keep in mind is that no aligner is 100% accurate so the translator might have to check the output. Once the alignment is done you can export the bilingual text in different formats like HTML or TMX. It is worth highlighting that some advanced aligners can align a large number of bilingual files at the same time.

Popular text aligners are NOVA Text Aligner, LF Aligner, AlignFactory, MateCat Aligner, OnlineTextTools, etc.

CAT TOOLS


Computer-assisted programs are widely known and used by translators. CAT tools work on the basis of segments previously translated. So the translator can either build a database of segments (Translation Memory) from scratch or develop other translation memories already provided.

The way CAT tools work is by offering translation suggestions from already translated sentences in the Translation Memory installed. If the program cannot find any, then it will offer suggestions from the Translation Machine Engine enabled within the CAT tool. As a default the CAT tool will use generic Translation Machine Engines. However, you have the option of installing other fee-pay engines by means of an Application Programming Interface key (API key). If you are working on a translation that requires high confidentiality you might need to disable the Translation Machine Engine in use so no information from your translation is dragged into the web (although it will always be disseminated in segments, and not the whole text).

Some CAT tools also offer the possibility of creating and using your own glossaries, importing existing client glossary or using any already embedded proprietary glossary within the CAT Tool (e.g.: SDL MultiTerm, or IATE from European Union terminology). Keep in mind that many proprietary glossaries would not let you add or delete entries.

MACHINE TRANSLATION

The concept of Machine Translation started around the 1940s. At that time translation was considered a mere coding and decoding of words and sentences from a source language to a target language. Since then, the research both in linguistic studies and computational linguistics has influenced the development of more sophisticated systems. Nowadays,




both fields of research agree that although Machine Translation can achieve high accuracy, it is still very difficult to formulate in mathematical terms the implicit meaning of a text. Therefore, unless a human translator is involved at some stage of the process the idea of a high-quality fully automatic Machine Translation is still a vision.

Nonetheless, Machine Translation as a tool to help translators can achieve results that the human translator will never achieve without the machine. The need for communication in different languages has grown beyond human capabilities.

We can distinguish three types of Machine Translation (MT), Generic, Customisable and Adaptive. Generic MT refers to the basic free MT Engines in platforms such as Bing, Google Translate, DeepL, Appertium, Baidu, etc. Customisable are MT tools that can be trained and therefore can achieve higher accuracy in specific domains. Some examples are Cloud Translation, KantanMT, Lucy LT, OmniscienTechnologies, Systran. Adaptive MT is the latest technology and claim to learn and adapt interacting with translators, examples are Lilt and SDL, although many others are also adopting this technology. It is important to highlight that with the advances on translation technology, it is getting more difficult to place MT Engines under one type of Machine Translation or another.

Since the Second World War to present day, there have been four main approaches in the evolution of Machine Translation: 1. Rule-based (developed during the post-war period, it focused on bilingual dictionaries and grammar rules), 2. Example-based (it was introduced during the 80s and takes into account the use of bilingual corpus and equivalences in the target language), 3. Statistical MT (this approach appeared in the late 90s and is based on statistical model, on frequency and probability of the occurrences in the text); 4. Segment-based (an approach from the 2000s centred in the alignments of segments).

However, the most recent developments in Machine Translation are the Neural Machine Translation which has evolved as a result of the new research on Natural Language Processing and Deep Learning. Neural Machine Translation uses trained neural network models combining with statistical methods in an attempt to simulate the human brain when acquiring a language. Neural Machine Translation systems make use of the concept of Back-Translation to create what is known as "synthetic data". When there is not enough amount of translation data in a particular language, the automatic translation from that language into another language is used as "synthetic data" to train translation models. Back translation was used initially in Statistical MT but recently has been proven to be more efficient in Neural Machine Translation.



If you want to extend your knowledge about this subject the book *Machine Translation* by Thierry Poibeau, from MIT Press Essential Knowledge series, will give you a good and concise overview. You can also check the [Translation Automation User Society](#) (TAUS) website to be up to date with the latest information and news about machine translation and language data. Equally, the website of the [European Association for Machine Translation](#) (EAMT) offers information about this sector and has a mailing list open to the public for all aspects of translation technology, as well as organising workshops and conferences through the year.

MACHINE TRANSLATION PRE-EDITING AND POST-EDITING

Pre-editing and post-editing are two processes usually included in Machine Translation systems as the aid of human translator is crucial to guarantee an acceptable quality translation. While pre-editing focuses on altering the source text, post-editing is used for reviewing and correcting the translation-output of the machine.


The main purpose of pre-editing is to make changes in the source text in order to eliminate ambiguities and facilitate the translation by the machine. The tasks involved can range from simple spelling and grammar checking to a more complex method of Controlled

Language. In general any word or structure which is considered unclear or challenging to translate by the machine will be replaced for clearer language without changing the meaning. If Controlled Language is being applied, it implies the use of computer software which has been programmed to detect and filter rules and terms which can cause ambiguity. Usually pre-editing is applied to a text which is going to be translated into several languages.

During the process of post-editing, a translator will check the grammar, spelling, style, meaning etc. The post-editing process can be done in two different levels depending on the purpose of the translation. A light post-editing looks only at major and obvious issues to make the translation accurate, correct and intelligible. Full post-editing is a more in-depth revision where style, terminology, tone, context and syntax etc. are also examined. After full post-editing, the final translation should read as if it had been done by a human translator. The general rule for both levels is to keep as much Machine Translation output as possible. However, training is often given to post-editors so they can be fully aware of this process and the guidelines of the client.

MACHINE TRANSLATION EVALUATION

The core elements which are usually evaluated in a Machine Translation are adequacy and fluency. Although evaluation has been done with



human translators in the past, it has proven more efficient and faster to use automatic measures known as “metrics”. However, they are not completely satisfactory and they need to be reviewed from time to time.

There are three main evaluations: BLEU score (Bilingual Evaluation Understudy), NIST score (The National Institute of Standards and Technology), METEOR score (metric for evaluation of translation with explicit ordering). The results of the evaluation are usually displayed in Quality Dashboards such as TAUS Quality Dashboards, where quality and analysis of translation performance is displayed.

LOCALISATION AUTOMATION

Localisation tools share some features with CAT tools. However, in order to maintain some source codes intact (e.g. codes related to default language, currency, dates, time, specific cultural or regional features, etc.), the localisation tools apply a “text string” technology so these codes are not changed. Filtering these variables manually would be a draining task for translators. The tools adapt the software to the geographical users helping translators to carry on with the translation of documents or content without having to worry about unwanted changes. In short, localisation tools allow you to distinguish between a code and a localisable term, evaluate and count localised terms, keep track of changes, apply visual editing, read different formats and localise

large volume of sentences. Specific patterns and rules are often outlined for consistency in multilingual jobs.


Once automated localisation has been accomplished, companies will also apply a testing process, known as localisation testing, to the localised content in order to check the accuracy of the content and assess the quality of the regional localisation.

The full localisation process might sometimes be integrated as part of a Content Management System (CMS) or a Translation Management System (TMS). Some examples of Localisation tools include Localazy, Multilizer, Transifex, Weblate, Mojito, Ranorex, Phrase, Weblate, etc.

If you want to learn more about localisation automation visit the [Globalisation and Localisation Association \(GALA\)](#) website. It is also worth reading the LocWorld News and Blog in the Locworld site.

SEO TRANSLATION

Search Engine Optimisation (SEO) uses keyword research to improve visibility on the internet. Any website or social media platform wants its content to reach as large an audience as possible and increase traffic to their website. Through the use of keywords companies can see what topics people are searching and in which regions they are most popular.



The majority of online translations are produced for global exposure but at the same time they are intended to have an influence locally. However, when a company applies a localisation strategy they also need to consider how to localise the SEO techniques. This is the main reason why many companies are incorporating SEO strategies in their translation process.

SEO techniques usually need to be added before the translation of the website or online content is done. Therefore, it is important the translator is involved in producing the domain-keywords in the target language, which can also be adapted to the SEO approach.

As a result, multilingual SEO is becoming a growing concept when a company is trying to reach international markets. Frequently, this can be facilitated by applying a unified SEO approach to multiple languages and using multilingual keyword terms.


Some tools to help you with translations for SEO and keyword search are Google Trends, Keyword SurferSite, Keyword Explorer, Keyworddit, Rank Math, Keyword Hero, etc.

EDITING AND PROOFREADING TOOLS

Currently there is an extensive range of tools in the market to help translators check, edit and proofread the target language. Though, you might find that some tools are limited to certain languages.

One of the most popular and well known technologies is Grammarly, which can be used for many languages. It can be applied to any digital text, from a document to an email or a twitter message and detect errors as you go along. Grammarly has a plug-in which can be installed in many applications like messenger, google docs, twitter, facebook, gmail and many more. Grammarly not only corrects spelling mistakes but it can also check the style and tone.

A great number of editing and proofreading tools are platforms in iCloud. Their various features not only help you with grammar issues or spelling errors but they can also suggest a selection of styles to choose from, highlight issues like repetitive words, ambiguity, bad grammar structures, over-use of adverbs as well as integrating synonymous and thesaurus tools. The underlined concept of “intelligent tools” means that many of these technologies learn from the data you input.



Among the features offered you may find encryption, a notion which is welcome if you are dealing with documents of confidential nature. Other features include tracking of changes, the use of markers to add comments and annotations for other proof-readers to see, the possibility of side by side checking of the texts, as well as the use of dashboard for synchronised projects.

Some of the tools available on the market are ProWritingAid, Grammar, PageProof, Language Tool, Easywrite, Wordrake, Polishmywriting, WhiteSmoke, Hemingway, Ginger, Slick Writer, Writefull, Stilus, etc. You will need to check if the tool can be used for your target language.

CROWDSOURCING TRANSLATIONS

Crowdsourcing Translations applies the idea of crowdsourcing where the talent of “the many” is seen as a great potential to capitalise on mass-collaboration. One of the most well-known examples is Wikipedia. Crowdsourcing can be applied at a professional level to gain experience and recognition or as an amateur teamwork where people are willing to work to help their community.

Through the internet, a company or social media platform may prompt collaboration from translators world-wide to help with large projects

enabled by Crowdsourcing Translations process. One major issue is how to assess final quality, although some platforms have mechanisms in place to evaluate the competence of translators. Other criticisms raised are lack of consistency, confidentiality and copyright issues.

Some of the most popular crowdsourcing translation platforms are Amara, Global Voices, TED Talks, Facebook, Duolingo, etc.

TRANSLATION MANAGEMENT SYSTEM (TMS)

TMS is a platform where the whole translation process is automatised. It is mostly used when complex projects in many different languages running parallel need centralisation to improve efficiency. TMS helps translation companies with high volume of translations manage jobs, vendors and teams from different countries, avoid the depletion of tasks, manage workflows, and even check progress through the different stages of the translation project and its revision. TMS also provides statistics on productivity and performance.

Initially, these systems helped to maintain databases of translators and clients. With the advance of technology, more productive business features have been included such as invoicing, quotes generation, complaint procedures, analytics, etc. As they are cloud-based systems they are easily accessed anywhere. Frequently, numerous translation

technologies are integrated within the system such as CAT tools, Terminology, Corpora, and Machine Translation Engines and so on.

Some examples of TMS are Lingotek, XTRF, Ontram, Consoltec, POEditor, Translate5, Zelenka, QTRM, Rulingo, Space TMS, Alisa TMS, memoQ, SDL, Localize, TextMaster, etc.

SIGN TRANSLATION SYSTEM: THE CASE OF GOOGLE AND KINDLE

A sign translation system can produce a translation from a text in an image. First, the system detects text in an image, then captures the image by video, transfers the text of the image into readable text through OCR software and finally applies automatic translation into the target language. Three technologies are involved in this system: sign detection, text extraction through OCR, and automatic translation.


One of main applications of this system can be found using the Google Translate app. Once downloaded via a mobile, you have the choice of selecting the language pair to translate (source and target). Then you place the camera over the text-image to obtain a fully automatic translation into the desired target language (although the machine translation engine is usually of a generic type). Google Translate offers 103 languages to translate, a feature for offline translations, translation

produced in real-time and capability to build a phrasebook by saving translated words and phrases, in addition to speech technology to hear the translation aloud.

The eBook platform Kindle uses sign translation for their electronic books. Kindle allows instant translations of words, sentences, paragraphs or whatever you highlight in the text. Through the technology “Word Wise” you can also access definitions and synonyms of words which can be handy if you are reading in English as a foreign language. Similarly, you can access monolingual or bilingual dictionaries in common language pairs. According to Amazon-Kindle, Word Wise is only available in books in English but not in all books.

AUGMENTED TRANSLATION

Augmented translation was a term coined by the research company Common Sense Advisory (CSA) which inferred this term from the “augmented reality” technology. According to the CSA Research, augmented translation uses Artificial Intelligent (AI) to give individuals access to information about their surroundings, which can offer linguists context and guidance for their translation. As the research claims, “this computing power will help language professionals be more consistent, more responsive, and more productive... and it puts linguists and translators in the centre”.



The idea is that all the technology tools interact with each other. They are directed by the translator in an environment where all the translation technology is integrated, and where the machine is learning from the professionals. So the translators are not part of the process anymore, but they control the process.

If you want to know more about how the future of translation industry is going to be shaped by the new augmented technology, you can read an interesting article “How Augmented Translation Will Redefine the Value of Translators” by Norbert Oroszi, CEO of memoQ, article edited by Slator.

ARTIFICIAL INTELLIGENCE AND THE FUTURE OF TRANSLATIONS

Translation will always be linked to language communications needs. There is no doubt therefore that the progress of Artificial Intelligence in Natural Language Processing will influence how we use technology to translate.

Currently the vision for computer experts is to make a robot interact effectively in a human communication. Once this is conquered the next natural step will be for a robot to translate like a human. However, even with the new advances on Machine Learning, where

the machine can see and learn from patterns blinded to the human eye, the current attempts on conversational chat box have shown that something is still missing.

There is no question that Artificial Intelligence can achieve great results in translations by automatising many repetitive processes and it can help the translator to be more efficient and faster. Nevertheless, when it comes to the unexpected and the exception to the rule, we must remember that Artificial Intelligence Algorithms follow mathematical binary systems, it is either 0 or 1. Ambiguity, uncertainty, deviation, and interpretation have shown so far, a particularly challenging task for the machines. Translation conveys the art of “decoding” not only words in the text, but also para-text cues related to culture and social context which the human translator combines with his or her knowledge and expertise to arrive at a translation decision.

Having said all that, we must remember that our mobiles carry a more sophisticated technology than the one used to get to the moon. So, in decades to come the robot might be talking and translating like humans. However, the need to communicate and translate might have also developed in a way that will make Artificial Intelligence fall behind human creativity again.

Supporting and developing
language professionals
worldwide



About CIOL

CIOL (Chartered Institute of Linguists) is the leading professional body for people using foreign language skills at work, setting the standard for linguists worldwide. Membership offers professional recognition, and our range of membership grades leads to Chartership, recognised worldwide as the gold standard for practitioners, whatever their profession, publicly understood as a badge of quality and competence. We offer many benefits and services, such as mentoring and access to networks. Our awarding organisation, CIOL Qualifications, delivers fully regulated professional language qualifications, such as the Diploma in Public Service Interpreting and the Diploma in Translation, recognised by government departments, agencies, business, universities and schools.

For more information on membership and qualifications, visit ciol.org.uk

Contact

CIOL | Chartered Institute of Linguists
7th Floor, 167 Fleet Street
London EC4A 2EA
United Kingdom
+44 (0)20 7940 3100
info@ciol.org.uk

