



# DO YOU SPEAK L10N?

the concise CAT dictionary







---

## What's it all about

I wrote this glossary (initially on my blog) mostly for my own delight, but also to help peers who are trying to cut their way through the translation industry jargon. I encourage you to take everything here with a generous pinch of salt. I did my best to stay tool-agnostic and included the least possible amount of memoQ jargon, but I consciously did not include jargon from any other tools. I simply feel I have neither the knowledge nor the authority to do that, and anyways, thata-way be dragons.

*Gábor Ugray*

memoQ founder



## [activation]

- A short communication between an installed program and the manufacturer's website. The program sends your serial number and a few anonymous details about your computer. The website checks that you own a license or that you are just starting a free trial, and returns a code to authorize the program to run on your computer.

» *See also* [CAL license](#)

## [alignment]

- Often, when you get a document to translate, you receive a set of previously translated documents along with it, or you can find matching pages on websites. Those may contain a

whole lot of translations you could use either as **TM matches** or through **concordancing**. Problem is, the **TM** needs segments, and you have whole documents. Alignment means splitting source and target documents into segments, and algorithmically finding out which target segment corresponds to which source segment. Not a straightforward thing to do! Advanced **CAT** tools have a function to automate the bulk of the work and help you correct the rest.

» See also **LiveAlign**

## [analysis]

Before you accept a job, you need to know how much text there is to translate. But you already have **TMs** with past translations, so you also want to know how much new text there is, and how many fuzzy or exact **matches** you can expect. That's what analysis does it compares your text against your **TMs** and **corpora**, and gives you a neat breakdown expressed in segment, word and character counts. Analysis is sometimes used interchangeably with statistics, which has absolutely no fancy scientific meaning in this context.

## [API; application programming interface]

- A nerdy term to say that a program allows other programs to use its functions, just as if a human was clicking its buttons. If a program has no API, then it's impossible to integrate it with other systems, and humans end up with tendonitis from lots of completely unnecessary clicking. It is particularly important to make sure a cloud-based tool you're considering has an API. If it does not, you may get locked in, with no easy way to retrieve your data if you want to switch.

## [auto save on server]

- When you're working in a memoQ [online project](#), your translations are initially saved only on your computer. You can choose to [synchronize](#) a few times a day, but if you enable auto save, your translations are sent to the server immediately, without holding you up. This way others can see your work \*almost\* as if you were editing a Google doc in real time. What better way to stay consistent?

## [automatic concordance]

If you want to see how an expression has been translated, **concordance** gives you exactly that. But a good **CAT** tool can do more: by looking at the source segment it can find the parts that occur in a lot of other segments, and point you to them. That's effectively saying, "Hey! These phrases seem to be all over the place, it's probably a good idea to concordance them right now!" And if you're very lucky, those phrases also occur as entire source segments, and the **TM** or **corpus** will give you their translation right away.

A small black circle containing a white lowercase letter 'i', used as an information icon.

*There are over 200 artificial languages that have been invented for books, television, and movies.*



## [AutoPick]

- I don't know about you, but I hate to type numbers as I'm translating, and I also hate to lose the flow to select, copy and paste something over from the source. In addition to numbers, source segments also contain other things that can go straight into your translation: **tags**, **non-translatables**, **terms**. If you just press and release Ctrl (in memoQ), AutoPick highlights all the special entities in your source, lets you cycle through them with the arrow keys, and insert the next one with a single keystroke. It also re-formats numbers to match your target language's conventions.

## [auto-propagation]

- Almost every text you translate has **repetitions**: segments that occur multiple times. With some technical texts these may even make up the majority. One responsibility of **TMs** is to pick these up, but CAT tools can do even better. If you enable auto-propagation, then as soon as you confirm a segment, the tool immediately populates all the other occurrences in the document and marks them as **confirmed**.



## [auto-translatables]

Most texts (particularly technical, legal and financial) have recurring entities that follow some pattern. Think of a date: 05/27/1978, to be “translated” as 27.5.1978. Auto-translation rules allow you to create **regular expressions** that recognize and transform such patterns in an incredibly flexible way.



i

*The word Mamihlapinatapai is derived from the Yaghan language of Tierra del Fuego, listed in The Guinness Book of World Records as the “most succinct word”, and is considered one of the hardest words to translate. It allegedly refers to “a look shared by two people, each wishing that the other would initiate something that they both desire but which neither wants to begin.”*



## [BiDi]

- Short for bidirectional text, i.e., the right-to-left scripts used to write languages like Arabic, Hebrew or Farsi. It's bidirectional because numbers and some proper names in Latin letters are written from the left within the overall right-to-left text flow.

## [bilingual Excel]

- » [See multilingual Excel](#)

## [bilingual RTF]

A specially formatted Word document that contains a translation's source and target **segments**, and often also **comments** and other information. This way, a translator can share work in progress with a client or a domain expert who has no **CAT** tool, only a word processor. CAT tools, in turn, can read an edited bilingual RTF with changes and comments, and bring the updates back into the translation environment. Some old formats relied on hidden text and were very easy to ruin with a single misplaced edit. These days it's more common to see a table with three or more columns.



## [CAL license]

- CAL is short for “concurrent access license.” While individual licenses allow a single person to use a program, an organization can purchase CAL licenses instead, which can be handed out to any end user on an on-demand basis. The limitation is how many people can use the tool at the same time; it doesn’t matter who they are or where they work from.

## [CAT; Computer Aided Translation]

- Software that helps translators and reviewers work more efficiently and in good quality, even if the work involves many people working simultaneously on the same large text. Sometimes the name is reduced to “**TM** tool” because TMs were the first function that CAT tools focused on. Jost Zetzsche

prefers translation environment tools, or TEnT; I tend to agree. **Translation Management System (TMS)** is often used synonymously because the border, really, is quite blurry.

## [CCJK]

Stands for the three East Asian languages, Chinese, Japanese and Korean. There are two Cs because Chinese can be written either with simplified characters (PRC and Singapore) or traditional ones (Hong Kong and Taiwan).

## [check out]

When a translator or reviewer checks out an **online project** in memoQ, the tool downloads the assigned documents and sets up a correctly configured working environment. This eliminates saving email attachments and going through an error-prone series of steps, saving time and ensuring all project participants work with the right resources and settings.

## [CMS; content management system]

- Software used to edit, organize and publish large amounts of content. A CMS typically tracks who is responsible for what content, whether it is approved or obsolete, what the content applies to, and much more. Often, CMSes break down content into smaller chunks, which are reused in several related documents. CMSes are important because an incredible amount of text that gets translated comes from them, typically as small chunks of XML and often in the DITA format.

## [comment]

- In a CAT tool you can mark entire documents, source or target segments, or just a short part within a segment. You can add a remark, or use the function to simply highlight something. This way you can communicate with other translators, reviewers or even clients, or just bookmark something for yourself to return to later. You can keep comments private, or choose to export them as part of the finished translation.

## [concordance]

A function of **translation memories** and LiveDocs **corpora** — that allows you to search for a word or expression, retrieving all translated segments where it occurs. This is nothing short of a small wonder, allowing you to “Google” existing translations. memoQ also highlights the expression’s most probable translation within the target segments, just like Linguee, but from your own private data.

## [confirmed]

» See **segment status** —

## [context ID]

Usually a short machine-readable text that identifies a **string** — that belongs to a specific place in an app or a game. It’s crucial to distinguish between, say, “Open” on a label (translated into German as “Offen”) or on a button (translated as “Öffnen”). The **TM** stores the ID and returns a **context match** if the same text occurs with the same ID later.

# CONCORDANCE

ANNO 1230

Where does "misericordia" occur in the Bible?



It's in passages

~~~~~ W W  
on pages...



ANNO 2017

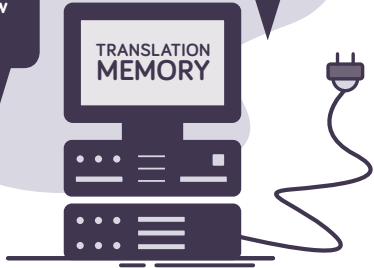
Where does "liner regression" occur in my translation memory? & How was it translated?



I found these segments:

~~~~~ W → ○○○  
~~~~~ W → ○○○  
~~~~~ W → ○○○

TRANSLATION  
MEMORY





## CONCORDANCE GIVES YOU



All sentences where an expression occurred

+Metadata: when they were translated, by whom, for which client,...

## YOU CAN

Search for words or expressions

Use wildcards

bat\* finds bat and batman

\*bat finds bat and acrobat

\*bat\* finds bat, acrobat, batman and sabbatical

Search source text or translations



## CONCORDANCE IS AWESOME

Because it lets you “Google” past translations. You don't need to research the same expression again

## [context match]

- » See [TM match types](#)

## [CSV; comma-separated values]

- A seemingly simple text-based format that stores several values in each line, separated by commas. It's still widely used to exchange [glossaries](#), and sometimes even for translatable content. In spite of its apparent simplicity it's very easy to mess up; the most common problem is using the wrong code page instead of [Unicode](#).

## [custom fields]

- » See [metadata](#)

*The language with the largest alphabet in the world belongs to the Cambodian language Khmer and is 74 characters long. The shortest alphabet is 12 characters long, and belongs to Rotokas.*





## [deliver]

The translator or reviewer's action to signal that they are finished with their task. In memoQ, delivery is not a symbolic step, as it usually triggers a series of actions like automated QA checks or emailing the finished translation back to the end client.

## [dictation]

Technology that allows you dictate text, instead of typing it on a keyboard. Dictation is preferred by a minority of translators; they, however, report a productivity boost of 50% or more over typists.

## [dicto]

- A neologism derived from typo. It means an error by the dictation software. Unlike typos, which are nonsensical on an elementary level, dictos are insidious because they are valid phrases that sound like what was intended, but mean something completely different, like an immature middle school joke.

## [DITA]

- Short for Document Information Typing Architecture, DITA is exactly as unsexy as it sounds, but tremendously useful. It is an open standard that defines how to structure and reuse content in CMS systems. The format is based on XML, and if your CAT tool supports it, you can deal with a huge share of the content coming from several different CMSes.

## [DTP; desktop publishing]

DTP tools include the likes of FrameMaker and InDesign, used to produce professionally typeset printed documents. In the industry DTP typically means an activity after translation and review. Translated text looks really bad in the original format unless you adjust the typesetting to accommodate longer paragraphs, different special characters, or even a complete left/right directional swap.



i

*Every two weeks, another language dies. Or, perhaps, a dialect. By the next century nearly half of the roughly 7000 languages spoken on Earth will likely disappear, as communities abandon native tongues in favor of English, Mandarin, or Spanish.*



## [edit distance]

- A number that expresses how different one text is from another, usually derived from the number of insertions, deletions and swaps needed to get from here to there. While similar TM matches are qualified by **fuzzy** match rates, edit distance is sometimes used to measure the extent of a reviewer's changes.

## [ELM license]

- » See **CAL license**

## [exact match]

- » See **TM match types**

## [export]

- » See **file format filter**

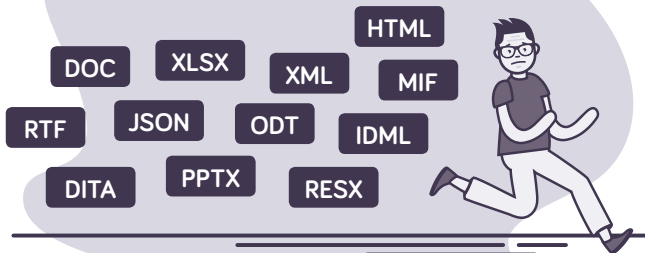


## [file format filter]

One key benefit of **CAT** tools is that you always translate in the same familiar editor, regardless of the file format your text came in. That means CAT tools must somehow extract the text from all the different file formats. The component that does this is called a file format filter: it “filters” text from all the other stuff in the file. Bringing the text into the CAT tool is called importing a file; retrieving the translation in the original format is called exporting it.

Every filter comes with its own options that affect how it works (“Do you want to extract the hidden text from this Word file?”), and for some formats, notably **XML**, these settings make an enormous difference.

# IMPORT • EXPORT FILTERS

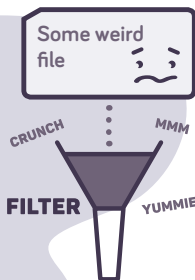


All these file formats OMG!  
How am I supposed to translate them??!



Don't worry!  
I have a FILTER  
for all of them!

TRANSLATE  
ME!





When you **IMPORT** a file, you use a **FILTER** in your translation tool to extract the **SOURCE TEXT**.

Ok-kay... But I have an Excel sheet and I'm supposed to ignore pink cells!



Easy. Filters know all about file formats and you can tell them what to do through **IMPORT OPTIONS**.

## WHEN YOU'RE DONE TRANSLATING

**EXPORT**

**YOUR  
TRANSLATION**



Some weird  
file\*



\*looks just like the original, but has your translations in it



Files also have a lot of important stuff that's **NOT** text.

formatting, images, references, ...

These produce **<TAGS>** when you import a file

## [find/replace listing]

- In memoQ, the Find function has an option that puts all occurrences on a separate list, instead of walking through them one by one from the pop-up window. The outcome is the find/replace listing, where you can review each **segment** comfortably and decide where to replace and where to leave as is.

## [font substitution]

- Many file formats, particularly from **DTP** tools, tend to use fonts that look really good, but cannot draw a lot of special characters. If your target language happens to have a lot of these, the translated file will look ugly, or skip letters outright. Font substitution is a function of **file format filters** that tweaks the file, replacing the original font with one that has the right glyphs for your target language.

## [fragment assembly]

If there is no exact or fuzzy match for a segment in your **TMs** — or **corpora**, a lot of the segment's parts may still have a match from a **term base**, **non-translatables** or **auto-translatables**. Fragment assembly takes all of these and just replaces them with their target equivalents, giving you a patchwork segment that might still take a lot less work to brush up than translating it from scratch.

## [fuzzy]

First impressions are correct here: this is one of the fuzziest words in the entire industry jargon. Initially a fuzzy match was used in contrast to an exact match from a **TM**: you get a translation that is fully legit, except it's the translation of something more or less different from your current source **segment**. Just how different is expressed by the fuzzy match rate. Eventually fuzzy matching was also extended to **terminology**, where it can be pretty useful if your language is in the habit of changing letters in the middle of words.

» See also **TM match types**

# {G}

## [global find & replace]

- In the olden days, the find function only worked if you first opened a document. Global find & replace searches through all documents in your project: a massive difference if, for instance, your job entails hundreds of tiny XML files from a CMS.

## [glossary]

- » See [term base](#)



i

*Java, one of the most popular language for programming, was originally called Oak and designed for interactive TV.*



[highlight]

» See [comment](#) —

[homogeneity]

A garden-variety [analysis](#) tells you how much of your text has fuzzy or exact [matches](#) from your existing [TMs](#) and [corpora](#). But even if you start with an empty TM, as you progress in a document, you will start getting matches from your own new translations! The homogeneity function quantifies these “internal” matches as part of the analysis, going beyond the mere detection of [repetitions](#). —

## [horizontal layout]

- A two-column grid layout where you see source on the left and target on the right has engulfed **CAT** tools like a flash flood washing away a hapless creekside camper's stock of ABC soup. But many translators still prefer to see their target text below the source. The horizontal layout option reshuffles the active segment's dominos, so source and target show up one below the other.

i

*Hungarian is a very productive language: one standard noun could, in theory, generate tens of thousands of forms via various forms of compounding. Hungarians therefore like to brag about their impossibly long words like megszentelteleníthetelenségétekért - that ultimately have absolutely no use in an everyday conversation.*





[import]

» See [file format filter](#) —

[internationalization; #i18n]

**Localizing** a product entails more than just translation: it includes things like showing dates in the right format, displaying temperatures in Celsius vs. Fahrenheit, writing first name last or vice versa, and the like. It requires extra effort to enable a product to do all this; that effort is called internationalization. —

[interoperability]

The ability of CAT **tools** to understand each other's formats and **APIs**, and to support standard formats well, so that people using software from different manufacturers can work together without drama, tears and major tragedies. —



## [join segments]

— » *See segments*



## [KWIC; keyword in context]

— A layout for **concordance** results where the search term is in the middle, with preceding and following text on both sides, row after row.





## [leverage]

To “leverage” past translations is fancy talk for: the tool gives me what I already translated, I don’t need to do it again. “Leverage” as a noun is fancy talk for the extent that happens: if a tool promises to enhance leverage, you should expect to type fewer new characters while translating the same text.

## [light resources]

This is memoQ lingo for things like **non-translatables**, **segmentation rules**, and a lot more. As opposed to heavy resources, which mean **TMs**, LiveDocs **corpora** and **Muses**, light resource have much less data. But while in many other tools they are “settings,” in memoQ they are resources: they have a name; they can be exported and imported; you can re-use them in different projects; and they can be shared online through memoQ server.

## [linguist]

- This term is probably the single biggest crime of the translation industry against proper English usage. For every educated person, a linguist means someone like Noam Chomsky, William Labov, Daniel Everett, or Arrival's Amy Adams: a scientist studying language in the mind, or language in society. In the translation industry, "linguist" is sloppy shorthand for translator or reviewer.

## [LiveAlign]

- memoQ's approach to **alignment**, where you simply throw a bag of source and target documents at the tool, and start translating. The tool aligns first the documents, then their segments, and indexes them in the background so they immediately give you lookup results in the editor. There will inevitably be errors, but you only spend time fixing those that actually give you matches.

## [LiveDocs corpus]

memoQ's alternative to **TMs**. While a TM holds a homogeneous mass of translated segments in no particular order, a LiveDocs corpus preserves entire translated documents, but gives you the same kinds of **matches**. If you want to check the context of a past translation, you can jump directly to the full document from the translation editor. TMs have one big advantage: they only store every translation once. If your content has a lot of repetitions, LiveDocs can become cumbersome.

## [localization; #l10n]

Sometimes used as a synonym for translation, localization entails a bit more: it includes showing dates in the right format, money in the right currency, and the like. In order to localize a product, it must first enable doing all this, which is called **internationalization**.

## [localization engineer]

- A person who knows the ins and outs of **CAT** tools, nasty **file formats**, **regular expressions** and other arcana. Many are not shy to code either. They make sure that before a complex project is launched, all the content is imported correctly, the **segmentation** is right, untouchable segments are **locked**, and a lot more. Without localization engineers, complex projects would never be finished on time and budget, and translators would tear out their hair and move to a farm to raise pigs.

## [lock segments in different languages]

- One thing that no **CAT** tool copes with well is mixed languages in a source document. In memoQ there is a well-hidden feature in the mundane function to **lock** segments. This inconspicuous option will algorithmically detect each segment's language, and if it's different from your document's source language, lock it.

## [locked]

In practically every CAT tool **segments** have a status like **new**, **pre-translated** or **confirmed**. Independently from this, segments can also be **locked**, which makes them read-only. If a **localization engineer** has populated some segments with translations approved (and mandated) by the client, then locking makes sure these do not get changed accidentally. Locked segments are also easy to exclude from the word count during **analysis**.



i

*IKEA kitchen accessories are named after fishes, mushrooms and descriptive words, while children's products after Mammals, birds and descriptive words.*



## [LQA; linguistic quality assurance]


- In addition to merely reviewing and correcting translations, human reviewers can also mark every error they find, indicating the error's type from a pre-defined list; the error's severity; and possibly other details. This information can later be evaluated to assess quality objectively. LQA is the function that facilitates this in CAT tools.

## [LSC]

- » See [automatic concordance](#)

## [LSP; language service provider]

- A business that sells translation services to its clients.



*The size of the language technology industry in 2016 is estimated at €29 billion.*





## [MT; machine translation]

A computer system that transforms a sequence of characters into a different sequence of characters that is recognizable, to a human, as text in another language, with relevant clues about the information in the original text. MT systems come in three main flavors. In rules-based (RBMT) systems, humans hand-craft grammatical rules. In statistical (SMT), a statistical system is trained on large amounts of human-translated text. Neural (NMT) systems are also trained on human-translated data, but they need a lot more computation, and have been reported to produce superior results.

Tools like Google MT are generic. In the translation industry it is more common to train specialized systems that do really well on a single type of content. This needs MT experts and a body of relevant, high-quality human translations for the training data. »

One way in which MT is used is human post-editing the MT system produces a rough translation that is often incorrect and ungrammatical, but is cheaper/faster to fix by humans than to translate from scratch. Another way is interactive MT, where the human translator receives suggestions from the MT system while she is translating text in a **CAT** tool.

## [machine translation post-editing]

- One way to utilize machine **translation**, where the MT system produces a rough “translation” that is often incorrect and ungrammatical, and is then post-edited by a human to remove major errors and improve fluency.

## [master TM]

- » See **working, master and reference TM**

## [match rate]

- » See **TM match types**



## [MatchPatch]

A memoQ function that improves **fuzzy** matches from a **TM** or a **corpus** by replacing the phrases that are different, relying on **term base** matches, **auto-translatables** and **non-translatables**. —

## [metadata; metainformation]

Additional details about a piece of stored information, like —  
“who translated this, when, and for what client” in the case of a **translation unit**, or “what source did this come from and did the client approve it” for a **term base** entry. **CAT** tools usually support a set of standard fields like the ones above, but also allow users to define their own custom fields and categories for more detail.

## [mobile or floating license]

» See **CAL license** —

## [monolingual review]

- A function where you **export** your translation into its original format, make changes outside the **CAT** tool, and can then bring those changes back into the translation environment from the edited target-language file. It is particularly useful when you need to send your work for client review but even a Word-based bilingual **file** is “too complicated.”

Why do you want to bring such changes back into the CAT tool? To make sure your **TM** contains only final, approved translations. Otherwise you may end up with **trash in, trash out**.

## [MQXLIFF]

- An **XLIFF** file than contains additional, non-standard information specific to memoQ, such as **segment statuses**, **QA warnings**, **LQA errors**, **comments** etc.

## [multilingual Excel]

An Excel file with source text, translations, comments and other information. Sometimes it's a small, innocuous file with two columns for source and target text, but we have reliable eyewitness reports of files out there with 50,000 rows and 25 columns for different languages. Such monstrous files often come from computer games.

## [MultiTrans XML]

The XML-based format used by SDL MultiTrans to export and import terminology. Although not an official standard, it is widely used for terminology exchange even between completely different systems.

## [Muse]

One of the resources powering predictive typing in memoQ. A Muse is built by analyzing existing TMs and corpora, with the aim of extracting words and phrases that correspond to each other in the two languages. When you translate a new source segment, the Muse looks at the phrases in it and gives you a list of suggestions that might be the translation of a phrase in the source text.



## [NMT; neural machine translation]

— » *See* [machine translation](#)

## [non-breaking space]

— A special character that looks like a normal space but acts differently because it doesn't allow a line break to intervene between the word on its left and right. A non-breaking space is a must before a colon in French (you don't want “:” to start a line), and between a number and a unit of measurement (you don't want “cm” to start a line either). In most word processors you can type it by pressing Ctrl+Space.

## [non-printing characters]

Spaces, non-breaking spaces, tabs, and newlines. Also, a few other invisible characters used in **bidirectional** text. The point is, they are all blanks and you normally don't see them. Just like Word, **CAT** tools have an option to show them, so that you don't accidentally type two spaces, or a normal space where a non-breaking one is warranted.

## [non-translatables]

Somewhat similar to **terms**, except that they are identical in all languages. Most often they are brand names that are to be left alone.





## [OCR; optical character recognition]

- Software whose original function is to turn an image (e.g., a scanned page) into editable text, usually a Word document. In translation OCR is used to turn documents in the one-way PDF format into a Word document that you can edit or import into a CAT tool.

## [online project]

- A memoQ project that stores documents in a server, allowing multiple people to simultaneously translate and review them, working together in real time. Online projects also make it really simple to assign work because they eliminate sending files around in email, and they prevent trivial errors because they make sure everyone in the project uses the right settings and resources.

## [online TM]

A **translation memory** shared through a server or in the cloud. —  
They allow organizations to store their translations centrally (and always find them when they are needed). They also make sure that translators working together in real time on different parts of a project get to see each other's translations instantly, ensuring their work will be consistent.

## [on-the-fly filter]

A function present in all advanced **CAT** tools (though usually called differently) that allows you to filter the segments of the document you're working in. It is "find" on steroids: you can quickly skim segments that contain a particular word or expression, and make changes if you changed your mind about a translation. It's also useful to eliminate, say, segments that are already **confirmed** so you can just focus on what needs work. —



## [PDF]

- Portable document format by its maiden name, it is meant to make sure a document looks exactly the same no matter where you view or print it. The price of that consistency is that it's extremely hard (nigh-impossible) to change the text inside it. In other words, it's a one-way format, which makes it one of the biggest nuisances for the translation industry. Apart from a few innovative solutions like TransPDF, your best bet is to convert a PDF into a Word file with an **OCR** tool, then translate that. Or, if you have the chance, to get the source (InDesign, FrameMaker or similar **DTP** file) from your client and work on that.



## [PEMT; post-edited machine translation]

» See [machine translation post-editing](#) —

## [penalty]

Some translations are to be trusted less than others. They —  
may be too old, coming from the wrong translator, or applicable to a different client or domain. A penalty means reducing the translation's natural [match rate](#) so it gets ranked lower than others.

## [perpetual license]

A license that allows you to use a piece of software you have —  
purchased forever. Perpetual licenses typically belong to a specific version of the software: to make sure the developer stays in business, it needs to finance its work by charging an [upgrade fee](#) for new versions.

## [plugin]

- A small module in a larger piece of software that performs a specific task. The point, usually, is that anyone can develop a plugin without the need to involve the main software's developer. A typical example in CAT tools is **machine translation** plugins: the providers of the machine translation make their service available through plugins for the CAT tools used by translators, **LSPs** or companies. Whatever the purpose, you can create plugins if the CAT tool's developer opened up a part of their application by creating and publishing an **SDK**.

## [populate number-only segments]

- Quite often a source document has lots of **segments** that all contain only numbers – think financial reports. Numbers are funny. They don't need translation in the traditional sense, but they also cannot go straight into the translated text: the target language's conventions for decimal separators and the like are different. This function processes all the segments with numbers in one go, adjusting their format along the way.

## [predictive typing]

One of CAT tools' most sexy functions with the least sexy name. Predictive typing makes you both faster (by eliminating keystrokes) and more consistent (by offering the right things to type). It looks at the characters that you have typed so far and offers a list of continuations from the current segment's **term base** matches, **auto-translatables**, **non-translatables**, **Muse** hints and other such sources. —

## [pre-translated]

» See **segment status** and **pre-translation** —

## [pre-translation]

A function that processes every **segment** in a document and automatically inserts the best translation from the project's **TMs** and **corpora**. —

# PRE-TRANSLATION



Can you translate this file by tomorrow?



Say yes! Most of it is the same as last week's file! I remember those segments.

TRANSLATION  
MEMORY

# LATER THAT DAY...



"Bla bla bla"

Got that!  
Translation: [...]



"Bla bla bla"

Got that!  
Translation: [...]



"Bla bla bla"

Got that!  
Translation: [...]

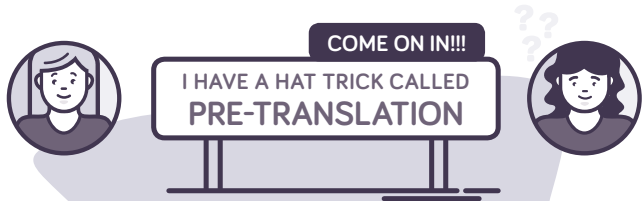


"Bla bla bla"

Got that!  
Translation: [...]



this is taking way too long if we go one segment at a time!!!



- 1) Put documents in hat
- 2) Click pre-translate
- 3) Check documents

A single click?  
On a hat?



If a segment has a fuzzy or exact match, it is now inserted. Everywhere!

**You can do this...even for hundreds of documents**

for 10 segments  
or 10 thousand

IT'S FAST

**YOUR TEXT AFTER  
PRE-TRANSLATION**

| SOURCE   | TARGET | STATUS |
|----------|--------|--------|
| Yodda    |        | 101%   |
| Bla Bla  |        | 100%   |
| Badaboom |        | --     |
| Dada 12  |        | 78%    |
| Hoopla   |        | 100%   |

⇒ Each segment's status shows if it was pre-translated

⇒ You can see the match rate

**DON'T FORGET TO  
FIX FUZZY MATCHES!**

## [preview]

- A screen area in your CAT tool that shows the document in its original format, or a close approximation. A premise of CAT tools is to rip text to segments and let you translate these in the same efficient environment, regardless of the original file format. But one thing is lost through this: visual context. The solution to this self-inflicted pain is the preview, which magicks visual context right back into your environment.

## [Project home]

- The screen in memoQ where you can add or remove documents to translate, pick TMs, term bases, Muses and other resources, and fiddle with your working environment in countless other ways, whenever you have an urge to procrastinate.

## [project management system]

- Software that keeps track of jobs, prices, customers, vendors, deadlines, invoices, and a host of other things that you need to run a translation business or department.

## [project templates]

If you get recurring or at least similar translation jobs (and you do), you are forced to do the same things over and over again: pick the right **TMs**, **term bases**, **light resources**, settings, people etc., and also perform the same actions like **analysis**, **pre-translation** and the like. Project templates define rules for all of these and a lot more, so you don't make embarrassing mistakes, don't get tendonitis from incessant clicking, and have a fighting chance to stay sane in the midst of it all. Also, project templates allow you to reuse the work of experts like **localization engineers** and make it a lot simpler for new hires to get up to speed.

## [pseudo-translation]

Translation is the fun part, but if you're dealing with complex file formats from esoteric systems, you need to make sure your work will also make it back to the original system at the end and not crash your client's multimillion-dollar flagship app right before the deadline. Pseudo-translation allows you to test the whole process without actually translating anything. It replaces source text with funny characters, words spelled backwards, and made-up stuff to inflate **strings**.

## [PTA; post-translation analysis]

- PTA for short, it is similar to **analysis**, but is performed once the translation is finished, not up front. Every large text will yield a lot of “internal” **TMmatches**, both **fuzzy** and exact. When two or more translators work together, there’s no way to say in advance who will translate a segment from scratch, and who will get a match because someone else was there first. memoQ keeps track of this in its **online projects**, and gives a precise and fair breakdown when all the work is done. Incidentally, the numbers in the pre-translation analysis match very closely those from **homogeneity**. The difference is that PTA’s breakdown shows who got how many of the internal matches that homogeneity predicted at the start.



i

*Nine languages don't have words for colour - they only differentiate between black and white.*





## [QA; quality assurance; automatic QA]

Machines cannot even come close to humans in crafting a message that resonates, but humans are really bad at getting numbers and other boring and repetitive things right. Quality assurance checks come to the rescue: they verify that you got your numbers right, that you didn't type two spaces, that you used MegaCorp Ltd's official **terminology**, that you translated the same thing the same way throughout the text, that you didn't forget a mission-critical **tag**, and a lot more. And if the dumb machine didn't get it right, you always have the option to ignore (suppress) individual QA warnings.



## [reference TM]

— » See [working, master and reference TM](#)

## [regular expressions; regex]

— Even texts written by humans are full of patterns that have well-defined moving parts. Think dates: `Number(1-12)/Number(1-31)/Number(four digits)` is a date in the US. For German, you have the same numbers in there, you just need to rearrange them and use dots instead. Regular expressions are a super counter-intuitive but super-useful way to describe exactly these kinds of patterns. No wonder **CAT** tools support them across the board, from defining [file format filters](#) through [auto-translatables](#) to [find/replace](#). You can get half a [localization engineer](#)'s career just out of knowing your regex.

## [rejected]

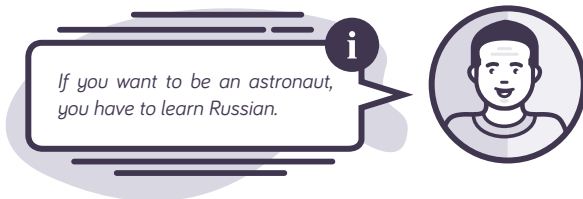
» See [segment status](#) —

## [repetition]

Any [segment](#) that occurs at least twice in your source text — is a repetition. They are a delight because usually you need to translate the same thing only once. That's how repetitions gave rise to [auto-propagation](#) and [exact matches](#). And for the cases where the same thing must be translated differently, you have [context IDs](#) to differentiate.

## [RTL; right-to-left]

» See [BiDi](#) —





## SDK; software development kit]

- A set of tools and documentation that allows developers to build their own module to work together with a different piece of software. SDKs are what allows third parties to develop **plugins** for **CAT** tools, for instance.

## [segment]

- When we translate text, we almost always proceed sentence by sentence. If you try to get to the bottom of it, however, nobody really knows what a sentence precisely is. Also, when you translate a single word in a bullet-point list, is that a sentence? **CAT** tools decided to sidestep this can of worms altogether, so we speak about segments instead.

Generally (though not always), a segment is the essential unit of translation: you proceed segment by segment in the editor, and you store the translation of segments in the **TM**. Your **TM** and **corpus** matches also refer to the segment you are translating at the moment.

Segments are born with the active cooperation of **regular expressions**, in a special incarnation called segmentation rules. As all regex, they look gibberish to the uninitiated eye, but they basically elaborate a single theme: “If you find sentence-final punctuation like a period followed by one or more spaces followed by a capital letter, start a new segment right there. Except if the last word before the period is a known abbreviation.” Segmentation normally happens quietly, behind the scenes, when your CAT tool's **file format filter** imports a source document.

No matter how elaborate, segmentation rules will inevitably get it wrong from time to time. To help get around this, CAT tools have a function to join neighboring segments, and to split a single segment into two.

## [segment status]

- From the moment you **import** a document into your **CAT** tool all through the steps of translation, review, client review, proofreading after a good night's sleep, and additional review by your pet, up to the point of exporting it for **delivery** to your client, the text lives in the form of **segments**. In this form, there's a whole lot to know about segments beside the text itself: does the target come from a **TM**, or from you? Have you confirmed it already? Was it rejected by a reviewer? Is it halfway edited but not quite finished yet? That is the kind of information that you can see in the form of colors and icons within the translation environment. For several years after they arrived, my friends from Mars were convinced translators were in the business of turning empty (grey) segments into confirmed (green) ones, and they thought this was a terribly appealing job. By now they know they were wrong, but they still think the job is awesome.

## [segmentation]

- » See **segment**

## [simultaneous translation and review]

A function of online collaborative **CAT** tools that allows several people to edit the same document together in real time. You can think of this as Google Docs on steroids, customized for the two-column, source-and-target world of translation. —

## [SMA; support & maintenance agreement]

While a license agreement entitles you to use a piece of software, the SMA that usually goes along with it grants you access to support from a human and to new versions of the software. Normally, perpetual licenses have a one-off fee; SMA, in contrast, is charged on an annual basis. —

## [SMT; statistical machine translation]

» See **machine translation** —

# SEGMENTATION

Yodda, Yabada? Hoolla, ragga filmwork etc. Dada



A paragraph can be a lot of text.  
You want to focus on shorter chunks:

~~SENTENCES~~

SEGMENTS

## WHY THE WEIRD NAME?

what is a  
"sentence" even?



DUNNO

COMPUTER



PROFESSOR



Let's just split text whenever I see **!?** followed by a  
**CAPITAL** letter That will be a segment



## WELL IT'S A LITTLE MORE COMPLICATED

### ABBREVIATIONS

etc.

ca.

Dr.

Dec.

### NUMBERS

'stop! 2 to go!

### GERMAN

All those Capitalized Nouns...



Don't worry!



- 1 You can add new abbreviations
- 2 You can split & join segments if I mess up
- 3 You can tweak segmentation rules using

## REGULAR EXPRESSIONS



## [split segment]

- » See [segment](#)

## [SRX; Segmentation Rules eXchange]

- An XML-based standard that allows different [CAT](#) tools to read each other's [segmentation](#) rules.

## [subsegment leverage]

- This is a strong contender for the industry's most fuzzy word, right there after [fuzzy](#) itself. When a [CAT](#) tool vendor uses it, they basically want to say, "We're doing something extremely advanced and useful here." In prosaic terms it means lookup results and suggestions (aka [leverage](#)) that refer to a shorter bit of the source [segment](#). In all earnestness, often the machinery that generates such matches really is pretty advanced, extrapolating knowledge from past translations in ways that are far from obvious.

## [statistics]

- » See [analysis](#)

## [string]

In developer-talk, a string is a sequence of characters. When you translate the user interface of a software application or a game, all the chunks of text that appear in different places are called “strings.” Typically, a string shows up as a single **segment**, and it has an associated **context ID** to disambiguate it. —

## [structural tags]

» See **tags** —

## [synchronize]

When you work in a memoQ **online project**, you have the option not to **save** every translated segment in the server immediately, but instead gather a lot of changes locally, and exchange news with the server in one go. That action is called synchronizing the project. —



## [table RTF]

— » See [bilingual RTF](#)

## [tag error]

— Some inline [tags](#) are optional: maybe that bold formatting in the source text is not needed in your translation at all. Others, however, are mission-critical: if they represent N in the sentence “You have N enemies left”, then if you omit the tag, the translated game will crash and the outrage of gamers will put your client out of business. To avoid such an outcome, the [QA](#) module of [CAT](#) tools gives you a tag error right in the editor, and won't let you [deliver](#) your translation until you fix it.

## [tag insertion mode]

Tags can be a real nuisance as you translate: you need to think about where they must go, you need special shortcuts to insert them, and generally, they throw you out of the flow. So in memoQ you can just focus on translating a segment's text first, then activate tag insertion mode and sprinkle your target segment with tags in the right places.

## [tag soup]

An unfortunate but all too frequent situation when a document that you have just imported is chock full of tags that are unexpected, pointless, or both. This most often happens with Word documents that an OCR tool produced from a PDF because it wanted to make sure everything is shown exactly in the right place, down to a hundredth of a millimeter. You can make things better by tweaking the OCR tool's settings, running a cleanup macro like Dave Turner's CodeZapper, or pestering your CAT tool's developers to do something about it. Only the first two have been conclusively shown to work.

## [tags]

- The content we need to translate consists mostly of text – but not exclusively. One oddity is formatting changes: how do you represent a change of text color in the middle of a sentence? The other oddity is a consequence of structured content, where text is intertwined with markup like hyperlinks, cross-references, or placeholders that will be substituted with, say, a number when a piece of software runs. CAT tools cope with all this by introducing tags: symbols inside your segments that act like a character in the editor, but look completely different. These creatures are called inline or internal tags.

Then you have formats like XML or HTML that have tags woven into their own DNA. Some of these tags define structure (“this is a headword”, “this is a caption”), always enclosing text from the outside. These are called structural or external tags, and should never show up in your segments. They only do if the XML filter was not configured properly before the import. You can fix that by hiring a good localization engineer.

The analysis output of well-behaved CAT tools has a separate section that shows how many tags the text contains in addition to good old-fashioned characters. This is important, because tags can be a lot of work and really slow you down as you translate.

[TB]

» See [term base](#)

—

## [TBX; TermBase eXchange]

An XML-based standard that allows CAT tools and other software to exchange [terminology](#).

—

## [TC match]

A bit of a schizophrenic creature that cannot completely make up its mind whether it's a match rate or a [segment status](#). It rears its head in the complicated scenario when you need to translate a source [segment](#) that contains tracked changes, which you need to reproduce in the translation too. A TC match is basically an [exact match](#) for the original form of the source segment, pretending those tracked changes were never put in there.

—

» See also [track changes](#)

## [TEP; translation-editing-proofreading]

- A widely used workflow that involves a translator and two different people subsequently reviewing her work, with the aim of ensuring a high-quality translation, and giving feedback for the translator to improve.

## [term base]

- A “database” or a component of **CAT** tools that allows users to store important words/expressions and their equivalents. It saves the hassle of researching the same term twice. It also helps translators adhere to terminology mandated by their clients, or at least stay consistent with themselves. In fact, it’s indispensable for consistency if different people are translating the same large text simultaneously, collaborating **online** from different locations.

Often used interchangeably with glossary, but they’re not quite the same. A glossary is usually just a word list in two languages, while term bases have structure and **metadata** too.



## [term extraction]

A function of advanced CAT tools that looks at new source text or a body of existing translations and extracts important words and expressions. The output typically contains a lot of “false positives,” but it allows a translator to research important terms before starting to translate, include them in a **term base**, and make sure they are then translated both correctly and consistently. —

## [terminology database]

» See **term base** —

## [TM; translation memory]

The idea that initially gave rise to commercial translation technology. Why translate the same sentence twice? The TM is a “database” of **segments** and their translations. TMs quickly evolved to give a hint also for segments that are only similar (see **TM match types**), to allow **concordance** searches, and to support **subsegment leverage**. —

## [TM match types]

- Translation memories are big bags full of source segments and their translations. When a new segment comes along, **CAT** tools rack these bags for segments that were translated before, and return these as matches. If the same sentence is there in the bag, that's an exact match, whose match rate is 100%. It can get even better: if there's a translation of not only the same segment, but the same segment from the exact same **context**, that's a context match (aka "ICE" for in-context exact). If the best there is is the translation of something similar but not identical, that's a **fuzzy** match with a match rate below 100%. Often, high fuzzies are distinguished: these matches only differ in punctuation, capitalization or numbers, and are therefore easier to fix.



i

*Trepverter is the Yiddish way to refer to a witty riposte or comeback you think of only when it is too late to use. Literally, "stair-case words."*

## [TM-driven segmentation]

An advanced function of memoQ that dynamically splits or joins segments during **pre-translation** to get better **TM** matches. It's a simple idea. What if a translator joined two segments before storing the translation in the TM, and now the same two segments show up again? By recognizing this on the spot, the two segments can be joined in the current document too for a perfect match, without human intervention.

## [TMS; translation management system]

Software that helps you manage translations and organize resources. Re: manage, think “these 1500 files must be translated into 25 languages, with 6 translators and 2 reviewers working in parallel for each, making sure that nobody overrides approved translations from the past, and ready by next Monday, with real-time visibility into the project’s progress until then.” Re: organize, think “I must find the right **translation memories** and **term bases** from among the 2000 resources I have around for various clients and language pairs.”

# TERMINOLOGY

① That feeling when...

I know I researched  
this before...



② Yes! It's right there in your terminology database

Is there anything else  
I'm forgetting?



Your glossaries are  
your most valuable  
**ASSET**

③ No need to ask!  
I've highlighted all these  
other terms for you  
already!

# A TERM BASE IS WAY MORE THAN AN EXCEL SHEET!

1

## METADATA



Just a fancy way of saying:

For each term, you know which client it is for, where you found it, etc.

2

## SPEED



CAT tools have auto-complete. If you've stored a term, you never need to type it out again.

3

## SUGGESTIONS



No need to remember and search.

Your CAT tool finds and highlights every term in the sentence you are translating.

4

## VALUE



You can share it with your customer. They will love it and hire you again!

## [TMX; Translation Memory eXchange]

- An XML-based format to, well, exchange translation memories. The adoption of this standard was a crucial step in the industry towards **interoperability**, and at this point virtually all tools support it.

## [track changes]

- Many regulated industries (like pharma) are required by law to track every change they make to crucial documents, such as the usage instructions and side effects of a medicament. Not only that, but when they sell to multiple markets, they must reproduce all these changes in translated materials too. As a translator or **LSP**, the only way to achieve this without losing your sanity and/or getting sued out of your profits is if your **CAT** tool has special functions to both cope with change-tracked documents and preserve the benefits of **TMs**, **term bases**, **QA** and everything else.
  - » *See also* **TC match**

## [translation unit]

In a **CAT** tool, you translate documents **segment** by segment. Once you store the translation of a source segment in your **TM**, the two together, plus some **metadata** like “who translated this and when,” are bundled up and transmogrified into a translation unit. —

## [trash in, trash out]

Imagine you store a nonsense translation in your **TM**. When you receive an updated document, **pre-translation** picks it up as a **perfect match**, and you don't even get to see it. If you train an **MT** engine with this data, it will produce nonsense translations. Once trash gets into the system, it perpetuates itself. How do you avoid that? Through **QA**, through **TEP**, through separating **working and master TMs**, and other similar efforts. All of which are only possible if you use a **CAT** tool that allows you centralize your resources, define the right processes, and eliminate error-prone manual steps. —

## [two-column Excel]

» See **multilingual Excel** —



## [Unicode]

- According to Wikipedia, Unicode is a computing industry standard for the consistent encoding, representation, and handling of text expressed in most of the world's writing systems. According to me, Unicode is the best thing that has happened since sliced bread. It means if you write a text with your own language's special characters, that text can be read by people anywhere in the world, using any gadget with a CPU and a display. Even (usually) in Excel. Nonetheless, to prove that our world is not the best of all possible worlds, you must keep in mind that while Unicode doesn't support Klingon, it does have a character for the handgun emoji.

## [UTF-8]

- » [See Unicode](#)





## [vendor]

In our industry, a person or a business that offers translation services to other persons or businesses. —

## [view]

Since CAT tools are apparently great fans of deconstructivism and start their day by tearing text into chunks called segments, you might as well max this out by slicing and dicing the living daylight out of those poor segments. As in: “I have just turned this User Guide into 1300 segments and pre-translated them. Now give me those segments that have no TM match, occur at least twice, and have the words ‘squinting squirrels’ in them. Also, show me each segment only once, and order them alphabetically.” That is the kind of thing that views allow you to do. —

# WHAT IS A TRANSLATION MEMORY?

## 1 SENTENCES KEEP REPEATING



a few pages later



zZZ



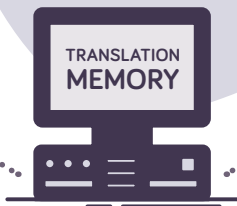
or a few months later

## 2 You don't want to translate them differently

You don't want to translate them **twice**

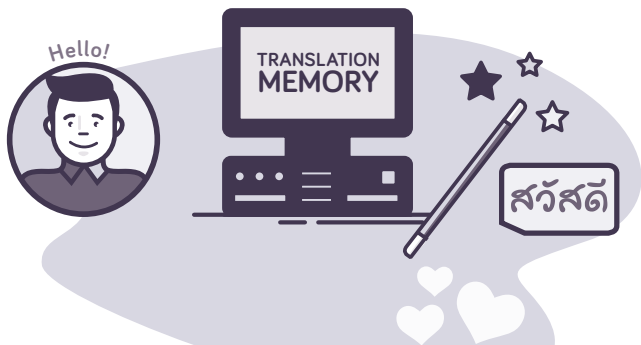


## 3 The translation memory remembers everything you translate



4

## THE NEXT TIME



**BUT WAIT IT  
GETS BETTER**

The translation memory also gives you sentences that are only **similar**

and shows you all occurrences of a particular expression

**FUZZY  
MATCHES**

**CONCORDANCE**





## [web editor]

- A component of **CAT** tools that allows translators and reviewers in an **online project** to work from a browser, without installing software on their own computer. A web editor is to traditional desktop tools as Google Docs is to Word, except advanced CAT tools offer both options (even within the same project) and don't force you to choose between two incompatible companies.

## [word count]

- » See **analysis**

## [working, master and reference TM]

Keeping stuff organized is an age-old challenge. If you don't get it right, you end up with **trash in, trash out**. One way to stay on top of data within a translation project is to designate one **TM** as the master (translations coming from there get precedence over others); another one as the working TM (new, as-yet unrevised translations get stored there, keeping the master pristine); and the rest as reference (to fill in the gaps that the master does not cover).



i

*According to Urban Dictionary, "Going there was a complete WOMBAT" should be understood as reference to wasted money, brains and time.*



## [XLIFF]

- Short for XML Localization Interchange File Format, a standard maintained by OASIS. In practice, it is a file format that allows translatable text to be passed between tools in a source/target, bilingual form.

## [XML; eXtensible Markup Language]

- An extremely versatile format for storing structured information in files that are readable by machines and not completely unreadable by humans. A tremendous amount of content that gets translated comes from XML files, particularly if the content's source is a **CMS** system. You generally don't need to understand the dirty details unless you are a **localization engineer**, in which case you are a wizard who knows all about **file format filters** and don't need this glossary anyway.

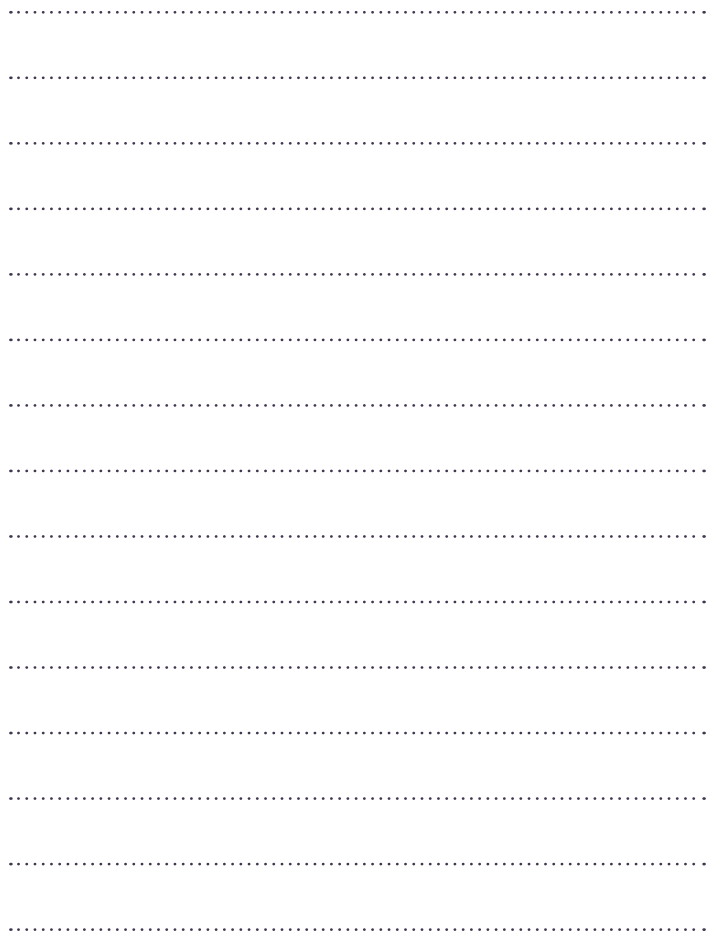
## [X-translate]

So you're half done translating a file, with tons of **comments** in there and segments in all imaginable **statuses**. At this point your client calls you and says, "Hey, our editors have been busy, we have a revised source file, we actually need you translate that instead of the one from last week." If your **CAT** tool has X-translate, this is not a problem. The function compares the original source document with the updated source document, going **segment** by segment, and recreates your work, including comments, ignored **QA** warnings, segment statuses and all the rest. Whatever changed in the source text is left empty, so you can catch up with the changes and continue where you left off.



**i**

*The letters "ough" can be pronounced 9 different ways in English.*





**SOURCE:** [www.jealousmarkup.xyz/texts/cat-tool-glossary](http://www.jealousmarkup.xyz/texts/cat-tool-glossary)

**EDITOR:** Ágnes Gázsó

**DESIGN AND PRODUCTION:** 823 studio kft. & Gelbert ECO Print Kft.

© Gábor Ugray

