LANGUAGE TECH NEWS

ata A Publication of the Language Technology Division of the American Translators Association.

Editor's Note

The first word that comes to my mind while writing this letter from the editor is: volunteer. You know who you are. From the writers to the editorial team, all of you who have donated your time to make this newsletter a reality have made a valuable contribution—not only to the LTD itself, but also to the larger objective of disseminating the kind of knowledge that helps our multi-disciplinary profession evolve. We all have numerous commitments competing for our attention, so a big thank you for your belief and commitment to the LTD!

As the recently elected Assistant Administrator, my number one priority is this newsletter. Michael recently sent out a call for volunteers to help in the newsletter production. I would like to thank those of you who got in touch and would like to encourage those of you who would like to, but have not yet done so. There are different ways you can collaborate, from writing and editing to scouting for articles. My plan is to create a newsletter production flowchart to visually depict the positions available.

To end on a positive note, enjoy this first newsletter of the new administration and please don't hesitate to give us your feedback. We are also departing from former newsletters, not only in terms of format, but also in terms of member interactivity. So contact us! ▢

Emily Tell
Editor & Assistant Administrator

VOL. 4, NO. 1 / JULY 2010

IN THIS ISSUE:

From the Administrator:

Dear LTD Member,

When you read this it will mean that we have published our first newsletter of the year 2010! What might have gone unnoticed, however, is the fact that we have been publishing the very same articles throughout the first half of this year on our website at www.ata-divisions.org/LTD. This is our new model of publication: as we receive articles, we edit and publish them on our website. Once we reach the expected volume, we put the articles together for publication in the LTD Newsletter. It helps keep our volunteers from getting overwhelmed and reduces the stress to maintain a strict publication schedule. This is not something that we particularly care about.

Speaking of care… I do however care about the fact that we are such a strong group within the ATA divisions by membership numbers and I know so little about you! We have only a small subset of members who are subscribers to the mailing list, followers in Twitter and visitors to the LTDnews website. Here at the admin desk I wonder how I can reach you better and understand your interests in this group. Participation on your part would not mean less enjoyment for you—just the opposite. Join us in our mailing list, contact me, write a comment on the blog and/or meet us in Denver 2010 at the General Meeting! That would really help us to make the Language Technology Division an even more interesting offering within the ATA.

Hope to hear from you!
Michael Metzger,
LTD Administrator

Emily Tell and Michael Metzger

Language Tech News
Vol. 4, No. 1 / July 2010

American Translators Association
225 Reinekers Lane, Suite 590
Alexandria, VA 22314
Telephone (703) 683-6100
Fax (703) 683-6122
ata@atanet.org
www.atanet.org

Editor
Emily Tell
emily@tellmealgo.com

Editorial Committee
Barbara Guggemos
Naomi Sutcliffe de Moraes

Contributors to this Issue:
Emily Tell
Patrick Reuter
Rubén de la Fuente
Jon Ritzdorf
Tuomas Kostiainen

Layout:
Cindy Gresham
illustrator@cgresham.com

LTD is the
Language Technology Division
of the American Translators
Association

LTD Administrator:
Michael Metzger
mm9@sbcglobal.net

LTD Assistant Administrator:
Emily Tell
emily@tellmealgo.com

ATA Language Technology Division Annual Meeting

Minutes taken by Emily Tell, Friday, October 30, 2009

Dierk Seeburg called the meeting to order at 4:08 p.m.

The agenda for the 2009 meeting and the minutes from the 2008 meeting were approved. Naomi de Moraes reported on the following activities from the year:

- Some recent newsletter articles were summarized.
- The mailing list was mentioned.
- The 2009 ATA conference speakers as well as language technology-related activities such as tutorials and talks were announced.

Dierk acknowledged the hard work that Naomi put into the LTD during the year and expressed his gratitude.

Elections

Elections were held for the officers of the LTD. The membership acclaimed the election of officers Michael Metzger and Emily Tell as incoming Administrator and Assistant Administrator.

Michael thanked the LTD administration, the newsletter contributors and the mailing list administrators for their work. He expressed his goals for the direction of the LTD for 2010, including:

- Creating new communication channels with the membership such as via Twitter.
- Identifying division issues with the division task force.
- Working closely with the Computers in Translation Committee.
- Opening a dialogue with the machine translation community.

Michael gave the floor to Alon Lavie, the current President of AMTA (the Association for Machine Translation in the Americas) to speak. Alon expressed that one of their goals has been to bring machine translation closer to the translators' community. He announced that their biennial conference will take place immediately following the 2010 ATA Conference in Denver, Colorado, which will give translators the opportunity to understand machine technology through tutorials. He is also encouraging AMTA members to attend the ATA conference to bring the two communities closer together to foster collaboration.

Adjournment

The meeting was adjourned. ▢

AMTA
The Association for Machine Translation in the Americas

AMTA 2010
Oct. 31 – Nov. 4, 2010
The Westin, Tabor Center
Co-located with ATA Conference in Denver, Colorado
more info at: amta2010.amtaweb.org

Panel Presentation on Cloud Computing, Crowdsourcing & Machine Translation

LTD General Meeting, ATA Conference, 2010

By Emily Tell

A lively and energetic panel presentation at the general meeting of the Language Technology Division at the ATA conference addressed the technology trend issues that most bedevil the translation community. Cloud computing, crowdsourcing and machine translation are some of the emerging trends that affect translators in their daily lives, in ways that range from their contracts with clients to the legal and ethical ramifications of using online translation memories. Each speaker gave a five-minute presentation focusing on one example of the phenomenon. Each presentation was followed by a question from the moderator and then a question from the audience. This article will address the first two presentations, since the third presenter, Rosana Wolochwianski, wrote an article about her own presentation. Rosana's article can be found in the next edition of the LTD newsletter under the title, "An Overview of the Impact of Machine Translation on the Professional Translation Community."

Laurie Gerber (left) and Rosana Wolochwianski

I. Cloud Computing

The first speaker, Beatriz Bonnet, President and CEO of Syntes Language Group, spoke about ***cloud computing***, a term used in very different ways, but mostly as a marketing term that denotes processing power in a lot of different servers. This concept, which has been around ever since we have had virtual private servers, is more applicable to companies than individuals in general. However, with the increase in processing power, cloud computing is now a little bit more difficult to define, as it generally takes place on multiple servers. The location of the data, its destination and how it's being backed up is more invisible to people using cloud computing.

In addition to defining ***cloud computing*** (the processing power in a lot of different servers), Beatriz also defined the terms ***service cloud*** and ***translation cloud*** and described how applications such as SaaS (Software-as-a-Service) always reside in the cloud. An example of a translation cloud is what TAUS (Translation Automation Users Society) is doing: a group of companies have come together to form a data repository of translation memories. Such companies as Cisco, Adobe and Microsoft are members of this society, which has created a legal framework for sharing all their translation memories with each other. Paying members of the society have the right to download the translation memories for their own use. The idea is that one company may have the need to translate something already translated by another company, so they avoid "reinventing the wheel." The TAUS Data Association (TDA) shares all this information and has just opened it up to the translator community – not for downloading the translation memories but for consulting and using existing translations. Some issues to look out for are the quality of

the translation memories and the possibility that all the translations might be moved to the cloud, thereby replacing translators. An example of cloud computing outside the translation industry is salesforce.com, which is a product of the Web 2.0 scenario.

How does this trend in what big companies are doing affect the life of freelance translators?

While TAUS is what the big companies are doing, Wordfast and Google Translate are creating big translation memories that freelance translators can contribute to. There is a question as to whether freelance translators are willing to share their translations and contribute them to translation memories. The speaker stressed the importance of being careful about sharing and about reusing translation memories to avoid retranslating, because of the potential for violating contracts between translators and direct clients and/or translation agencies. It is easy for translators not to pay attention to the contracts they have signed. If the contracts are well written, there will probably be a clause that says that everything they do is considered "work-for-hire," and therefore they don't own the translations provided. Furthermore, translation contracts normally have clauses stipulating that everything we do is confidential and that the translation is owned by the translation agency or by the translation agency's client. If we are thinking about uploading our translation memories to Wordfast or to any repository in the cloud, it is important to ask ourselves beforehand if we really own them, if we have the right to them. Since our translation memories probably contain a lot of company names. uploading them could be deemed a breach of contract between the translator and the translation agency, and also a breach of contract between the translation agency and its client, which could be a software company which hasn't released its product yet. Also, translators must ask themselves if they are commingling translation memories from a lot of different clients. It may be okay to share the data from one client, but not okay to share the data from the other three clients. The other point is that there are a lot of resources in the cloud that translators can consult—for example, the TAUS translation memories. Other resources such as Translated.net, glossaries and forums are available to translators. Be careful to understand where a translation comes from and don't trust it automatically. Just because there is a lot of information out there, and it was worked on by other people, sometimes we automatically think it is good, and don't think it over. If translators find data from the top 10 Fortune 500 companies, don't assume it's good. Remember that whatever translators get has to come from the right source.

Before uploading our translation memories to any repository in the cloud, it is important to ask ourselves if we really have the right to them.

Relevant links:

www.tausdata.org
www.translationautomation.com
www.translated.net
www.wordfast.net
Google Translate: www.google.com/ig

II. Crowdsourcing

Naomi Bear, the second speaker, is the Director of Microloan Translations and Review at Kiva and she spoke about crowdsourcing and Facebook.

Crowdsourcing is a new term which is defined as the act of taking a task traditionally performed by an employee or a contractor and outsourcing it to a generally unknown large undefined group of people in the form of an open call. This definition is taken from Wikipedia, which is the best example of crowdsourcing. In the case of Wikipedia, a large group of people have come together to iteratively create a body of content. Crowdsourcing is about three years old. It has been around for a long time, however, although in a different form. Sun Microsystems has been doing crowdsourcing for the translation of its open source software for a number of years. Crowdsourcing is a growing trend in the industry, and there are a number of organizations and companies that have begun to experiment with this in the last year or so. In

the case of translation, the definition could be a little bit broader: An undefined generally large group of people that may vary in terms of size or amount of definition. Facebook is the example which is most talked about these days. It is a very specific type of crowdsourcing and it represents just one point on the range. The way it works at Facebook is that there is an open crowd which is undefined. The only limitation is that you have to be a member of Facebook. Once you're a member of Facebook, you can download the application, contribute translations to their user interface, and participate as you wish. The model for Facebook also works in an iterative way. That is, Translator 1 might contribute a translation into French to befriend somebody. Then Translators 2, 3, 4 and 5 might contribute an alternate version of the same translation. These versions would then be reviewed and voted on by the crowd until one of them is approved. The other defining aspect of the Facebook model is that it is translation for very specific segments of the user interface. For example, strings that are short and relatively small in volume – such as 35,000 strings – are being translated into about 65 languages now.

Naomi Baer (left) and Beatrice Bonnet

Other points along the spectrum are the open versus closed crowd and everything in between. There is an open crowd with the unrestrictive Facebook model, but there are other organizations like Asia Online, which is running a project to translate all of Wikipedia into Thai. Asia Online has a system where they initially differentiate their users through some automated testing. Then the users acquire a history, as the paid in-house linguistic reviewers edit the translations to determine which of the contributors of the crowd are more or less reliable.

There are some systems where it's a fully closed crowd. The program being run at Kiva has human testing in which the translators take a traditional translation test. The test is evaluated by humans and either accepted or rejected. If translators are accepted then they have free reign to participate in the translation crowdsourcing program.

The other point on the spectrum that differentiates translation crowdsourcing models today is the review mechanism. There is a perception that no crowdsourced translations are reviewed and that there is 100% reliance on the crowd to identify the best translations, regardless of the crowd's experience and background. In some cases this may work relatively well, as in Facebook, where the elements are relatively small and the crowd is very familiar with that environment and terminology. However, most companies are actually doing some sort of professional review on top of that. So even in a Facebook model, with a voting, ranking system, most companies are paying their linguistic reviewers in top tier languages. So there is an element of professionalism in most translation crowdsourcing today. Organizations and companies that are seeing the value of community involvement still understand the importance of having quality. There is a range and we're still learning which of these models works best.

Why are companies using crowdsourcing for translations and taking the work away from professional translators?

According to Naomi, there is a concern and perception that the motivation behind crowdsourcing is simply to cut translation costs.

This isn't true. A recent industry insights report on crowdsourcing stated that one of the key findings was that companies were not implementing crowdsourcing to reduce their costs. There is a real cost to setting up and maintaining an infrastructure like crowdsourcing, and the community needs to be recruited, motivated, sustained. Quality needs to be monitored. The top motivation for companies is community engagement. For an organization or product that has a passionate user base, this is one more way that users can be involved. It is a way for them to participate, stay engaged and define the terminology or the atmosphere that they're working in.

Emily Tell is the CEO of Tellmealgo Translations, Inc. and Assistant Administrator of the ATA Language Technology Division. She can be reached at emily@tellmealgo.com

There is also the issue of market reach. Many crowdsourcing translation programs for software products, for example, are creating translations in a vast number of languages that would not be supported by paid translation. Companies can begin to access new markets in some cases, test the translations, and see if they are viable. Then perhaps they will make a case for professional support.

On the flip side of reach there is the question of access. Going back to the Asia Online project, project participants talk a lot about the fact that internet users in the Asia Pacific region are expected to reach 47% of all internet users by the year 2013. At the same time, only 13.8% of internet content is in those languages, and non-CJK content is only 0.03%. There is real information poverty, and some of these crowdsourcing programs are providing access to information. At the same time that they are enabling people in information-impoverished markets to access all kinds of paid and unpaid content online, they're also making it possible for organizations and companies to reach these markets to sell their products and services.

Another reason why translation crowdsourcing is taking place is time to market. Unlike the traditional sequential localization model, there is a lot faster turnaround. An additional reason is that we can translate more. There is a lot of additional content being translated that organizations would not budget for. Even organizations that have large translation budgets for their traditional projects and user documentation are using the translation crowdsourcing model. There is one case of a software company that pays for professional translation of all of their products, but also has a large base of user-generated development documentation that they would not translate professionally. Their user base has asked for access to this content in languages other than English and they have begun to support a crowdsourcing model.

So we're not seeing crowdsourcing taking away very much translation of existing content, but are really beginning to see an expansion of the pie in terms of what is being translated and which languages are being translated. In addition, we are also seeing possible new roles for professional translators, as paid reviewers of community translated content.

Relevant Links:

www.kiva.org
www.facebook.com
www.asiaonline.net
www.wikipedia.org

Conclusion

The panel presentation addressed very important issues affecting translators today. Big companies are using cloud computing to create repositories of data located in invisible online servers so that they do not have to invest in new translations and can share existing ones. On an individual level, translators need to be aware of the legal and ethical consequences of sharing translation memories, as there may be contracts in place with their clients regarding the ownership of these translation memories. The phenomenon of crowdsourcing is also a source of concern for professional translators, who may feel that work is being taken away from them in the interest of developing online communities around a brand or website like in the case of Facebook. Translators need not hop on the wagon or sit in silence. Rather, they can observe how these trends are unfolding. With this awareness, they can make educated decisions based on fact, not illusion. ▢

A Multi-Dimensional Approach to Productivity and Automation:

The Beetext Flow Leadership

By Patrick Reuter

Note: The material in this article is presented for informational purposes only. It is not intended as an endorsement of a product or service.

When I come across people who are interested in learning about translation management systems (TMS) at professional conferences, I am often met with puzzled looks. It has become evident to me that most people cannot fully understand the basic concepts around TMS and the place a Software as a Service (SaaS) product has in their work universe. In this article, I will introduce readers to the basic TMS concepts, and the benefits of Beetext Flow (http://beetext.com/flow_saas.php) as it relates to those concepts. Let us start by examining the most common questions, and what is currently considered widely accepted knowledge about TMS.

The acronym soup: Is TMS a form of CMS (Content Management System)? And how are TMSs connected to TMs (Translation Memories) and MT (Machine Translation)?

From a technical and practical dimension, there is a connection between these technologies, insofar as they are a part of the language services infrastructure. A CMS allows for effective management, storage and publishing of content, including multilingual content. A TMS enables users to automate many essential projects, programs, and back office tasks, as well as specific productivity (*workflow*) tasks.

Makers of TMS have to use aliases to try and differentiate their products from those of their competitors, and make them stand out in the acronym soup.

Translation memories are a part of the workflow for those users who employ them, but they are *linguistic resources* that can be integrated into the workflow. MT is another resource available to the language services market and which, like the TMs, can become a part of the workflow. Simply put, all of these components may be integrated into an Enterprise Management System (EMS) or any number of "acronymic" business management systems. Neither CMS, nor translation memory programs, nor MT allows users to execute and track the milestones associated with a project. A TMS does.

TMS by many aliases

Many software products fall under the general definition of Translation Management Systems, but the use of several combinations of adjectives and nouns to describe available TMS products also contributes to the confusion and the puzzled looks in the exhibit hall. Everyone who works in the development of TMS programs is guilty of calling it by a different name, and yet every TMS program is different. Therefore, makers of TMS have to use aliases to try and differentiate their products from those of their competitors, and make them stand out in the acronym soup.

In some cases, a TMS is also considered a part of a larger system, such as a CMS or a Global Information Management (GIM) system. The bottom line is that there are many ways to label a TMS program, product or set of programs, but the ultimate goal is to streamline and automate tasks that take place both concurrently and sequentially, and that are essential components of a business process.

Suppliers, contractors and corporate translation departments all face a challenge in trying to understand—let alone choose—the technology offered by a "workflow management platform" (which is how Beetext Flow is defined), "business management software," "productivity solution," "management information system," "translation workflow automation," "translation office management tool," and many others.

Users then have to make a decision based not just on traditional business-minded criteria like budgets and return on investment (ROI), but also on how closely the TMS they are evaluating meets their current and planned business needs. These are not the only con-

cerns a TMS should address, though. The criteria for choosing a TMS are as diverse as its potential users, and the best way to address this diversity is to allow for a reasonable degree of customization.

In the next few paragraphs we examine how Beetext Flow conforms to the general idea of a TMS and goes beyond it to adapt to what customers want and need.

Solution, migration, integration... consternation?

One of the main concerns often voiced by those contemplating the acquisition of an off-the-shelf TMS is migration of their current data. When customers have tracked their projects using a database or even spreadsheets, data migration is possible with Beetext Flow. Beetext works with its clients to transfer data to Flow, where it is then available in a central location.

In the end, providing a workflow management platform or solution is all about the customer's needs.

Integration with back office programs is another frequent concern, and to the question "Can this TMS be integrated with X?" the short answer is, it *depends*. Integration with certain programs, such as translation memory applications, makes all the sense in the world, and Flow is currently integrated with two of them (MemoQ, and LogitermWeb). Integration with XTM is in progress. Flow is also compatible with a few other applications like Trados and SDLX (see http://www.beetext.com/flow.php for more information).

Integration with other software products, such as accounting systems, is an important and delicate consideration for a software company like Beetext, and it only makes sense when there are standards, and when "true code collaboration" between applications is possible within reason. The challenge is that every program has different complexities, and such complexities are compounded by an application's level of compliance with standards or lack thereof.

Surveys and research have shown that complexity is the last thing a workflow management solution user desires. To keep things simple while still providing what can be called a complete TMS, Beetext Flow offers connectivity to all major CMS, QuickBooks and other applications, but it lets customers decide if they wish to integrate Flow with other programs. The integration process is one of careful evaluation, planning and execution, and Beetext does take the time to examine all angles in collaboration with its clients.

Transforming business models: Doing three times as much business in a third of the time

When Beetext started back in 2002, the company adapted its definition of Beetext Flow, the first TMS available as SaaS, to whatever their clients wanted it to be—because in the end, providing a workflow management platform or solution is all about the customer's needs.

From the start, Flow users have provided Beetext with input that has helped to augment and at the same time optimize the program's capabilities. As a result, eight years later the robust workflow platform has been refined and improved. Many of the most complex but common tasks related to document translation, software and web localization are further automated in Beetext's newest program, Flow MMX, which will be available for implementation soon. This is a fast application that enables agencies, contractors and corporate or government translation departments to double and even triple their productivity (thereby augmenting their profits).

While a chapter could be written about each module in the Flow MMX program, for the sake of space in this newsletter let us highlight a few key features:

1. Business around the clock, and around the world: Flow MMX enables 24/7 operation for suppliers of any size. A secure, intuitive, and simple wiki-style dashboard allows users to quickly start, track or close a project (see Figure 1, next page). The dashboard is a portal that is accessible to users in accordance with their defined privileges. With appropriate privileges, clients, project managers, contractors, and others can execute a step within a business or project cycle, obtain a report, or benefit from any of the myriad

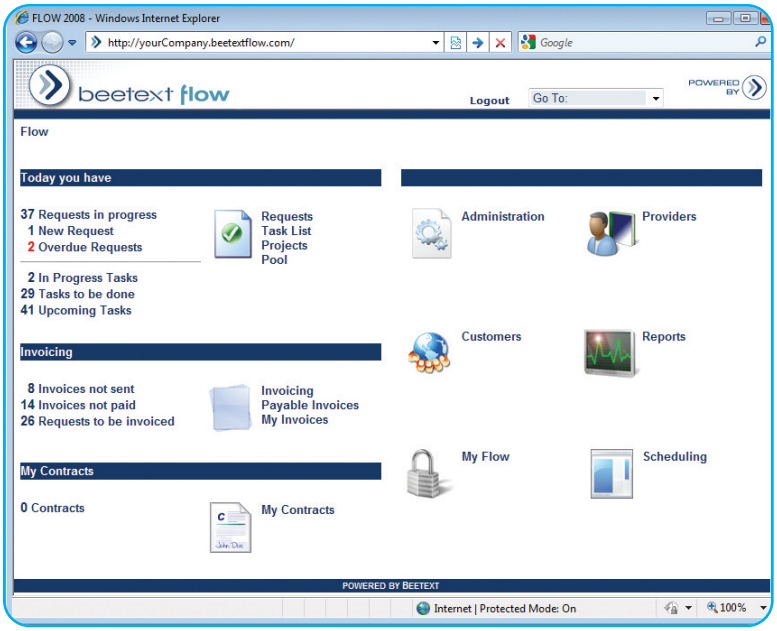


Figure 1: The Beetext Flow dashboard

functionalities available.

2. Security is not a concern: Most people who voice fears about a SaaS model lack enough information. There is no reason to fear that data will be lost or compromised when you use Flow. Think about it – when you use your cellular phone, a land line, a fax line, email, online banking, you are entrusting your data transmission to a means you have no control over. Using a SaaS program, i.e., a program "in the cloud" where transactions are virtual and stored remotely but securely, is akin to all of those other modes of data management and transmission we now take for granted. Beetext assures customers that their data is secure when they use Flow SaaS, because every measure has been taken to protect it. Customers also have the choice of installing Flow on a local server, but savings are greatly increased when the SaaS model is adopted. More on that on the next item...

3. Overhead and expenses on information technology (IT) are minimized: A SaaS model implies that somebody else is taking care of all the maintenance and support for your application. And because the idea behind a workflow management platform is to ease and automate as many production, client, and vendor management tasks as possible, the savings on IT are significant.

4. Efficiency, end to end: Clients set up projects, project managers become business managers, and vendors manage themselves. These are some of the features language services suppliers, corporate translation departments and contractors can look forward to with Flow:

- Intuitive, easy to use web-based portals specific to clients, project managers and linguists
- On-the-fly workflow models for project management, or previously created templates for repetitive projects
- Automatic notification to appropriate parties of new requests, new tasks, completed tasks, completed projects, and delivered projects
- Simple, clear review of the status of project tasks, per file or per project, in real time
- Files collected in a central location, with version control
- Simple management of clients and contractors, with detailed reports
- Shortened turnaround times, increasing productivity while controlling costs
- Scalability (some LSPs outsource project management to a certain extent, therefore having an "elastic" PM team)

Solving the project management conundrum

A California translation agency had been searching for a suitable solution to their lack of a project management system for two years. They had evaluated several software programs, and found that none of them accomplished every single wish they had on what was a really long list. That changed when the agency's CEO came across Beetext Flow at an ATA conference. After examining the program thoroughly, the agency decided that Flow offered what they needed: It could streamline processes and avoid common production efficiency problems, and could be customized to their way of doing business. As a bonus, Flow was less than one-fifth the cost of the other workflow programs they had evaluated.

Before Flow, this American agency was managing projects entirely via email, tying up

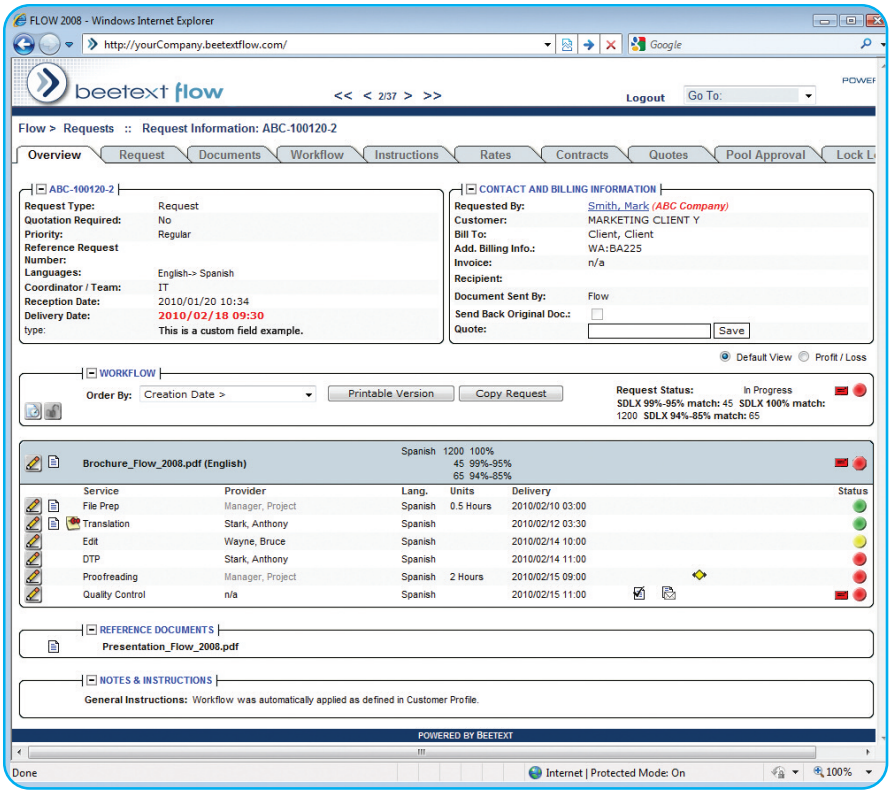


Figure 2: A bird's eye view of the status of a project. Every item can be expanded for further detail.

precious time and resources while manually preparing quotes, sending and receiving files, and performing a host of other tedious tasks. By implementing Flow, the company was able to automate the quoting, setup, management, and tracking of projects, and also start working with remote project managers (see Figure 2). The change in infrastructure resulted in a fast ROI, further savings, productivity gains and accelerated growth for the company.

Like the translation agency described above, other language services suppliers can change their business model with technology, all for the better, and in a matter of just a few months. Participants in a survey conducted in 2009 among members of an association of translation companies placed Beetext Flow as the top workflow solution for agencies (please contact Beetext at info@beetext.com for more information about this survey).

Ending the mires at corporate and government translation departments

Corporations, particularly large international companies with internal translation and localization departments, frequently find that the systems used to manage other parts of the company, mostly legacy systems, are ill-suited for the idiosyncrasies of the language trade. Government is no different.

In both the public and private sectors many a dedicated engineer, terminologist, linguist, and project manager has had to adapt his or her work to a generic project management application. Frustration and yearning for change are common topics of conversation by the water cooler until someone decides to try an application specifically built to automate and manage translation and localization processes such as Flow.

With implementation of Flow, productivity sometimes increases so much that some members of the staff become redundant, while others have to account for how they are actually spending the time that used to be consumed by manual data entry, email messages, phone calls, and all manner of back-and-forth communication to even get a project started. In these cases too, the right tool changes the business model for the better.

The contractor as a mini-company

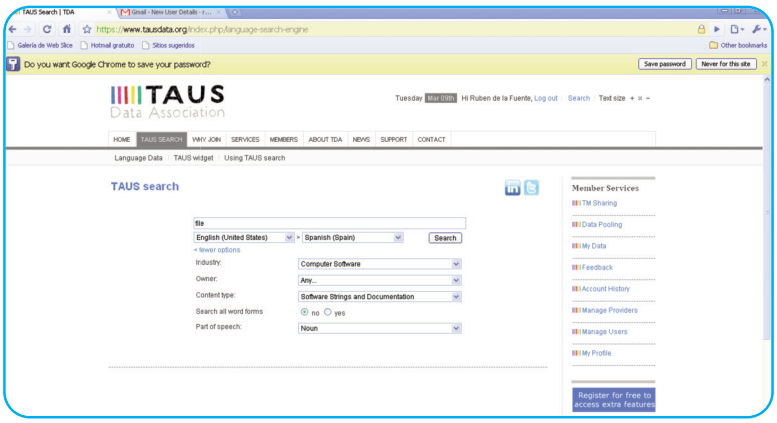
Usually confined to a couple of well-known operating systems and whatever is developed for them, contractors (translators, editors, DTP providers and their cousins, the localization testers) find themselves with few affordable choices to manage their projects, particularly quotes and invoices. A SaaS version of Flow for contractors allows them to enjoy just those functionalities that make sense for them, for a very reasonable price. Contractors can thus devote proper attention to the tasks they are actually paid to do, and minimize the time spent on administrative tasks nobody compensates them for.

All in all, Beetext Flow is as versatile a tool as its users want it to be. Although it was built specifically for the translation industry and as such is already a very complete system, it is constantly validated and improved by the users' experience. Flow strips the complexities and opens a portal of ease, speed and efficiency. ▢

Wealth of Multilingual Reference Material at Your Finger Tips

By Rubén de la Fuente

There is a more or less recent trend in companies and institutions to foster cooperation and development by granting open access to the public to very valuable resources: open courseware from prestigious institutions such as MIT, or APIs and source code for third party application development from the likes of Apple or Google. The translation and localization industry is no exception to that trend and a wealth of multilingual reference material has been made publicly available in the past few years, which translators can leverage from in order to speed up their work. The aim of this article is to provide and comment on a selection of very useful multilingual reference resources, namely Taus Search from TAUS, MyMemory from Translated.Net, the Acquis communautaire and the DGT Translation memory from the European Union, OPUS (Open Source Parallel Corpus) and Bitextor, an open source tool to produce translation memories from multilingual web sites.



Taus search

http://www.tausdata.org/index.php/language-search-engine

TAUS (Translation Automation User Society) has undertaken to create a huge shared pool of multilingual resources (translation memories, glossaries) so as to meet a constantly increasing demand for translation that traditional models can no longer cover: as a result of Internet improvements leading to Web 2.0, more and more content intended for wider global audiences is generated daily. TAUS brings together companies (e.g. Microsoft, Adobe, Google), LSPs (e.g. Lionbridge, WeLocalize), language technology developers (e.g. Language Weaver, ProMT) and universities (e.g. Universitat de Valencia, University of Leeds) in a combined effort to produce a pool of resources that will provide the basis for growth and innovation within the localization industry and ultimately make it possible to meet the demand for translation.

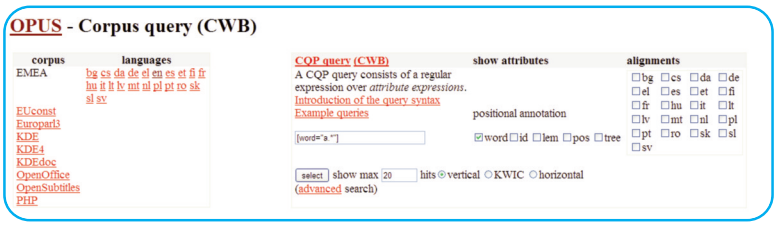
Since one of the aims of TAUS is to reduce translation costs (or, as a representative from one of its members put it: "Why pay for a translation that has been done already somewhere else?"), it has been seen as a threat by some translators. However, TAUS can also be good news for translators, since it grants free access to these multilingual resources through Taus search, a web service that allows you to search in the existing translations of TAUS members for a selected language pair. Place your mouse over the i icon in any of the search results if you want to know the origin of that translation, including content type (software, help, documentation) and data owner (company using that translation). Registration is free and gives you access to advanced features (such as filtering by industry, data owner or content type) and allows you to download a widget that will let you run Taus search from your desktop.

MyMemory

http://mymemory.translated.net

MyMemory is a huge translation memory (nearly 300 million segments) put together by Translated.net through web alignment, corpora like the EU corpus and contributions from translators. Registration is free. Once registered, you can upload source documents and download a translation memory in TMX

format with segments relevant for those documents. Also, you can choose if you want machine translation to be applied for sentences with no matches and have the resulting translations included in the TM. Source documents are kept confidential and not used to train any machine translation engine. You can contribute to this project by deleting wrong alignments, rating existing translations and adding new ones.



OPUS

http://urd.let.rug.nl/tiedeman/OPUS

OPUS is a collection of multilingual parallel corpora and tools compiled by Jörg Tiedemann from the University of Groningen. This collection includes the following corpora: OpenSubtitles (subtitles from OpenSubtitles.org, covering 18900 movies in 59 languages), EMEA (1500 documents related to medicinal products translated into 22 languages for the European Medicines Agency), KDE, OpenOffice.

Apart from the corpora, OPUS also provides a set of tools to handle the corpora (align, browse, convert). The most interesting tool for translators is the Corpus Workbench, an interface that can be used as a general query engine for all corpora included in OPUS. The corpora themselves or a sample of them can also be downloaded either to have them as reference or to use them for training machine translation engines or other natural language processing purposes.

Acquis Communautaire and DGT Translation Memory

http://langtech.jrc.it/DGT-TM.html

The is the entire body of European legislation: treaties, regulations and directives adopted by the EU, plus rulings from the European Court of Justice. Legislation must be translated into the official languages of the Member States, which results in the biggest parallel corpus in existence. Pairwise alignments of the corpus (231 language pairs) can be downloaded at http://wt.jrc.it/lt/Acquis/JRCAcquis.3.0/alignments/index.html. Below is a table with global figures (number of words and documents) for the AC.

The DGT Translation Memory is a shorter version of the AC, but probably more interesting for the translator: instead of complete texts aligned, translation units in TMX format are supplied. Split into 12 different volumes, the DGT TM can be downloaded at http://wt.jrc.it/lt/Acquis/DGT_TU_1.0/data. An extraction tool called TMXtract is also supplied, so users can retrieve segments for the desired language pairs (NB: the extraction tool considers English as the source language).

Bitextor

http://bitextor.sourceforge.net

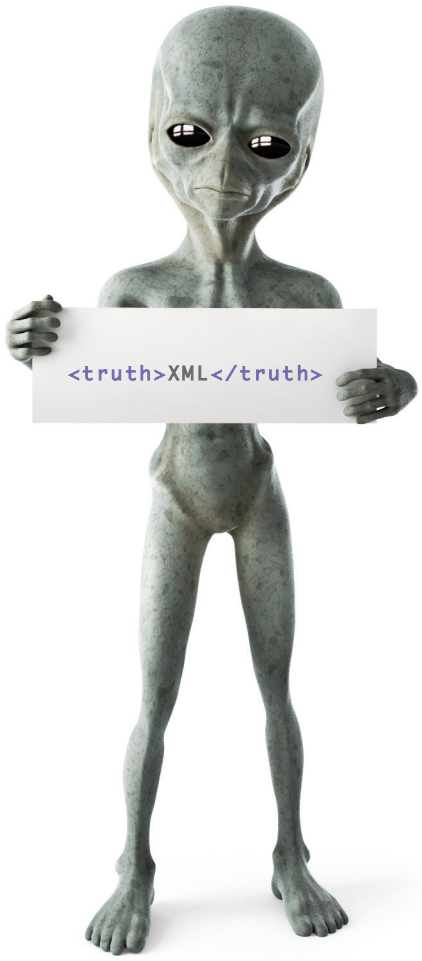
Bitextor is an open source application, developed by Universitat d'Alacant's research group Transducens, which allows you to download an entire web site, analyze its pages to see if any have the same content written in different languages and if so, produce a TMX file. Bitextor can be downloaded at http://bitextor.sourceforge.net.

If we are to meet the increasing demand for translation mentioned above, a boost to our productivity is needed. I believe the resources described here can be very useful in that sense, since they provide access to extensive reference material from trustworthy resources (international institutions, major companies) in just a few clicks. ▢

Rubén de la Fuente holds a BA in Translation and Interpretation from the University of Granada, Spain. He has worked in the localization industry for nearly 10 years, as a translator, reviewer and project manager. Currently, he works as the in-house Spanish linguist for PayPal Spain. He also lectures in localization tools and processes at Universidad Alfonso X El Sabio and gives localization seminars and workshops frequently. He likes sharing knowledge through social media like twitter and LinkedIn. He can be contacted at rubo@wordbonds.es.

"X" Files

By Jon Ritzdorf



iStockphoto

The "X" files have returned. Not the popular FOX network show of the 1990's, but rather the file format known as "XML." Since the introduction of the standard over a decade ago, XML has begun to crop up in almost every industry. In fact, without being aware of it, you have probably already had "close encounters" with what I lightheartedly refer to as "X" files. Have you ever received a file created in the latest version of Microsoft Office (Office 2007)? Pay close attention to the file extensions. All of the new Office 2007 file formats ending in doc**x** (Microsoft Word), .ppt**x** (Microsoft PowerPoint) and .xls**x** (Microsoft Excel) are XML derivatives of the original .doc, .ppt , and .xls formats used since Office '95. Additionally, if you're an "open source" or Apple devotee, you may already know that XML has been the default file format of both Sun's OpenOffice and Apple's iWork suite since their inceptions. Popular due to its infinite flexiblity, XML is a chameleon of sorts and can represent any type of content for any purpose. As a translator, you will inevitably come face to face with the format sooner or later. Even if you think you've never seen one, you've likely encountered "X" files before and never realized it: To quote the "X Files" television series, "The truth is out there"…in fact it's probably lurking among your documents right now.

"The truth is out there"...in fact it's probably lurking among your documents right now.

"X" Factor

Why was XML created and why has it gained such prominence today? XML offers a simple way to structure complex data that can be shared without restrictions between various parties and across various mediums. XML-based formats are used to author a variety of content, such as databases, software, electronic office documents and even hardcopy manuals or CAD drawings for print.

The "X" in XML stands for e**X**tensible, indicating the fact that its usage is not restricted to any particular subject area, medium of communication, or application. The "ML" in XML stands for "markup language." With a basic understanding of another markup language such as HTML, XML is easily read. However, XML is not restricted to internet usage or building web pages.

When introducing newcomers to XML, I regularly equate the flexibility of the format to a standard deck of playing cards. When preparing for a card game, you are held to a highly flexible set of pre-defined constraints by the deck of cards. A standard 52-card deck contains only 4 suits and 13 cards per suit; however, there are multiple games at your disposal given those general guidelines. Regardless of the game you decide to play, the constraints of the deck are always present.

The XML standard works much the same way. XML follows a set of constraints defined by the overall XML standard, yet from that,

```
<?xml version="1.0" encoding="utf-8"?>
<!DOCTYPE article PUBLIC "-//OASIS//DTD DocBook V4.1//EN">
<article>
  <articleinfo>
    <title>An Example Article</title>
    <author>
      <firstname>Your first name</firstname>
      <surname>Your surname</surname>
      <affiliation>
        <address><email>foo@example.com</email></address>
      </affiliation>
    </author>
    <copyright>
      <year>2000</year>
      <holder>Copyright string here</holder>
    </copyright>
    <abstract>
      <para>If your article has an abstract then it should go here.</para>
    </abstract>
  </articleinfo>
  <sect1>
    <title>My First Section</title>
    <para>This is the first section in my article.</para>
    <sect2>
      <title>My First Sub-Section</title>
      <para>This is the first sub-section in my article.</para>
    </sect2>
  </sect1>
</article>
```

Figure 1

```
<?xml version='1.0'?>
<Pet>
  <name>Bibob</name>
  <gender>Male</gender>
  <fedBy>Jon</fedBy>
  <colors>
    <primary>Black and Tan</primary>
    <secondary>Silver</secondary>
  </colors>
  <order>Mammal</order>
  <species>Dog</species>
  <breed>Dachshund</breed>
  <furstyle>Short-haired</furstyle>
  <favoriteFood>Anything</favoriteFood>
  <likes>Eating</likes>
  <dislikes>Walking</dislikes>
  <notes>Old and stinky...but has a lot of personality</notes>
</Pet>
```

Figure 2

thousands of XML "games" can be created. Like poker, bridge, or solitaire, each XML-based document has a unique method of "play" and is defined by certain characteristics, such as the intended output (a document, a webpage, a help system, a marketing brochure), the purpose (financial reporting, chemical structure descriptions, or FDA regulatory submissions), or simply the need to structure data in an organized manner.

Figure 1 shows a section from a "DocBook" XML file defined by the OASIS standards committee for technical documentation.

Note how the structure of the document is clearly denoted with markup tags indicating the author (`<author>`), titles and subtitles (`<title>`), and paragraphs (`<para>`). However, XML presents no restrictions to content or purposing. I can just as easily describe my dog using customized XML markup as in Figure 2. Both usages are perfectly valid.

Brand "X"

Due to its flexible nature, XML is useful for any number of functions. Even if you are not "technical," as a translator, it is important that you can identify the "X" files common to your respective areas of specialty. Although your particular domain(s) may not adopt an XML standard, at the minimum you should be familiar with the various XML vocabularies, such as TMX, TBX and XLIFF that are part of translation/localization industry standards.

The following is a list of some of the most common domains and subject specialists that use XML markup vocabularies:

XML in Office Productivity Suites

- OpenDocument XML: *Sun Microsystems* Open Office
- Office Open XML: *Microsoft* Office 2007
- iWork XML: *Apple* iWork

The oft-encountered word processing, spreadsheet, presentation and graph/diagram files from these office productivity application suites are all composed of XML that has been compressed in an archive. The files must first be decompressed (a.k.a. "unzipped") before the XML is accessible.

XML in Technical Communication ("Techcom")

- DITA
- DocBook

Used to create technical manuals, help systems, online help and knowledge bases for medical, science, technical and industrial/manufacturing domains from a single set of XML source content. Technical writers and user assistance authors are the primary users of these standards. Using a so-called "multi-

channel" publishing model they can write once (in XML) and publish to a variety of outputs.

XML in Software, Web 2.0 Applications & Websites, and Telecommunications

- XHTML
- RSS
- SOAP
- WML
- VoiceXML
- UML
- RESX and XAML - *Microsoft* .NET
- XUL - *Mozilla*

Without a doubt, the IT industry is the largest adopter of XML for everything including websites, wikis, blogs, internet protocols, web-based applications, telecom, software and user interfaces. This is only a small, representative sample of the more common XML vocabularies for these subjects. Many of these are encountered in daily web browsing and when using a mobile device.

XML in Marketing Communications ("Marcom")

- INX and IDML - *Adobe* InDesign
- SVG

With the exception of web programmers, XML has been slow to catch on among marketers, graphic designers and desktop publishers since the technical nature of the standard is not easily embraced by those with more visual modes of thinking. That said, the benefits of moving to a multichannel publishing model can be as significant for graphic communicators as it is for technical communicators. Recognizing this, the World Wide Web Consortium (W3C) and private companies that dominate the industry (such as Adobe and Quark) constantly push new ideas for XML adoption forward in the marketing industry. XML savvy marketing firms can now create automated publishing workflows though InDesign's customized XML export, and an increasing amount of 2-dimensional imagery for the web is authored in SVG as popular web browsers add support for the format.

XML in other Industries

- (*E-Learning*) SCORM, LOM
- (*Financial*) XBRL
- (*Life Sciences/Pharma*) SPL
- (*Manufacturing*) CAMX
- (*Translation*) TMX, XLIFF, TBX, SRX, GMX

For a full list of the main XML markup languages in use today, check out Wikipedia:

http://en.wikipedia.org/wiki/List_of_XML_markup_languages

If you happen to specialize in one or more of the areas listed above, proactively becoming familiar with the standard(s) that relate to your interests is recommended. Soon, even literary translators will not be able to escape the grip of "X" files, as almost every eBook format is written in XML.

"X" Marks the Spot

Here are some tips to identify an "X" file:

1. Look at the file extension. Although many types of "X" files come with a generic ".xml" label, others give away their secret by a single "X" in the extension. Notable examples include:

- .inx - *Adobe* InDesign XML
- .wrx - *WordPress* XML export
- .xlf/.xliff - XLIFF
- .resx - *Microsoft* .NET resource file

2. Open the file in a text editor. If the file remains a mystery, open it in any simple text

Jon Ritzdorf serves as the Globalization Architect for Acclaro Inc. and has spent the last decade in a variety of technical positions within the localization industry. Jon is also an Adjunct Professor of translation technology at New York University and the Monterey Institute of International Studies in addition to offering private instruction. He can be contacted at jon@ritzdorfacademy.com.

editor (like "Notepad" (Windows), "TextEdit" (Mac) or "Emacs" (Linux)) and follow these steps:

a. First, make sure you have an XML file. The first line should start with `<?xml`

b. Look for the document type declaration. Declarations are usually in the first few lines of code and are always marked with an exclamation point followed by `DOCTYPE`. The information that follows might offer some hints. In the example below, we can see that the DOCTYPE indicates a DITA file.

c. Identity the root tag (the first tag after the declaration). It can often tell you something about the standard used like `<book>` for DocBook or `<svg>` for SVG or `<vxml>` for VoiceXML. It is possible to look up the root element on the web to see if you get any hints. Again, in the example below, the root tag identifies the XML file type as a DITA document.

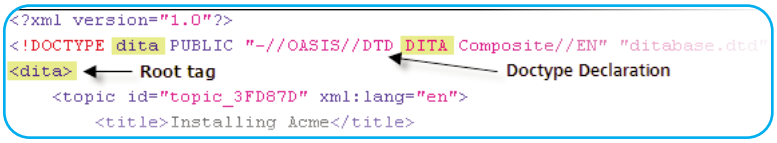


Figure 3

d. Examine the namespace information. This is also usually in the first few lines of code and will be marked with `xmlns`. The namespace will identify information regarding the XML vocabularies being used in the document. (Figure 3)

3. Check with your favorite CAT. Filters in leading CAT tools recognize some of the more common XML standards (XHTML, DocBook and DITA) and can automatically identify your "X" file in many cases. That said, many XML files are not recognized, and others may be customized to the point that they cannot be read correctly by the out-of-the-box parsers. In these cases, definition of a custom filter in your preferred CAT tool is required to process the code and extract the translatable text. Consult with your client to confirm that the filter you create is capturing only the text they want to translate.

Conclusion

Since XML can represent most any type of content created on a computer, it is quite likely that you will come across an "X" file at some point in the future. What type you will encounter is anyone's guess; human innovation is the real X factor at work. Traditional methods of publishing, data dissemination and translation technology are moving rapidly towards specialized XML markup languages. For all who plan to tackle translation projects, a background in the basics of website markup languages such as HTML and XHTML is highly recommended to maintain relevance and competitiveness. Both can be used as a gateway to further understanding the more complex XML-based foundations, which are rapidly becoming the de facto standards for anything authored on a computer. ▢

We invite you to visit our redesigned website at www.ata-divisions.org/LTD. A round of applause and a big thank you goes to our LTD Webmaster Michael Wahlster!

Trados Tip

by Tuomas Kostiainen

My Migration to Trados Studio 2009

I feel like I can't really cover any other topic in this column until I have written something about the new SDL Trados Studio 2009 that was launched in June 2009. Even though the much promoted new features sounded very promising, the actual launch didn't go that well. There were problems with licensing and the general functionality of the software. Unfortunately, that killed the excitement for many translators. For example, I wasn't able to use it because it crashed with such frequency that I was getting worried about my own well-being. However, the licensing issues were quickly fixed and the excitement returned with the arrival of Service Pack 1 (SP1) in October.

Since SP1 came out, I have used Trados Studio for about 95% of my translation work (and gone through the SDL Trados Studio Trainer Training) and I have to say that I'm quite happy with it. It certainly hasn't been like "dancing on roses" all the time, as we say in Finnish, but there are so many major improvements that they definitely outweigh the problems and shortcomings I have encountered.

In this article I want to give you a general picture of the pros and cons of migrating into Studio based on my own personal experiences. You can read more about my experiences on my Trados Studio blog at http://tradoshelp.wordpress.com. In order to save space and make this article more useful, I have linked most of the features I mention to the online Trados Studio Help[1] instead of explaining how to access these features. Let me start with a few main features that I really like.

1 The possibility of using multiple translation memories[2] at the same time and in different capacities (update, lookup, concordance). I have always encouraged translators to use one big TM and label their translation units with attribute fields. However, with Studio it's not necessarily that important anymore. It still doesn't make sense to have scores of little memories, but at least you have the flexibility to use, for example, additional client-specific memories or really large, more reference-type memories without having to import them to your own main TM. And the fact that you can designate each TM as an update, lookup or concordance TM, is like icing on the cake.

2 The AutoSuggest[3] feature brings a whole new level of leveraging and automation to the translation process. With the help of AutoSuggest dictionaries, which you can create from large TMs, you get phrase-level match suggestions that are automatically presented to you in a small list that opens up where you are typing. Comparing this to an "autosuggest" feature found in many cell phones does not do justice to this new feature. This is much more than that. I have to admit that when I first read about this feature I thought it would be too cumbersome to use in practice, but that's not the case at all. It might take a short while to get used to it and to learn to use it efficiently, but then it becomes a part of your normal typing routine. However, if you are a really quick typist this might not be such a big deal for you.

In addition to the AutoSuggest dictionary material, the suggestions also include termbase matches from your active MultiTerm termbases and AutoText[4] matches from your autotext entries. AutoText is also a new feature and it's similar to the AutoText feature in Word. The only complaint I have is the case-sensitivity of AutoSuggest.

3 The third feature that I really like is the very much streamlined and improved interface with MultiTerm (Figure 1, next page): you can see all terminology hits at once, and adding new terms to the termbase on the fly is very easy and fast.

4 The new editor interface[5] and the underlying SDLXLIFF file format. This of course affects several features but one important and practical benefit is that you can work in an "almost tag-free" environment without needing to use TagEditor for non-Word files or to worry about the annoying font and formatting changes that often happened when translating in Word environment. I admit that the table format has its drawbacks, but I still think it works much better overall.

1 Trados Studio Help: http://producthelp.sdl.com/SDL%20Trados%20Studio/client_en/SDL_Trados_Studio_Help.htm#welcome.htm

2 Translation Memory: http://producthelp.sdl.com/SDL%20Trados%20Studio/client_en/Ref/O-T/TM/Translation_Memory.htm

3 AutoSuggest: http://producthelp.sdl.com/SDL%20Trados%20Studio/client_en/Ref/U-Z/Options/Overview_Options_AutoSuggest.htm

4 AutoText: http://producthelp.sdl.com/SDL%20Trados%20Studio/client_en/Ref/U-Z/Options/Options_AutoSuggest_AutoText.htm

5 New editor interface: http://producthelp.sdl.com/SDL%20Trados%20Studio/client_en/Screen_Layout/Editor_View/SbS_EditorCom.htm

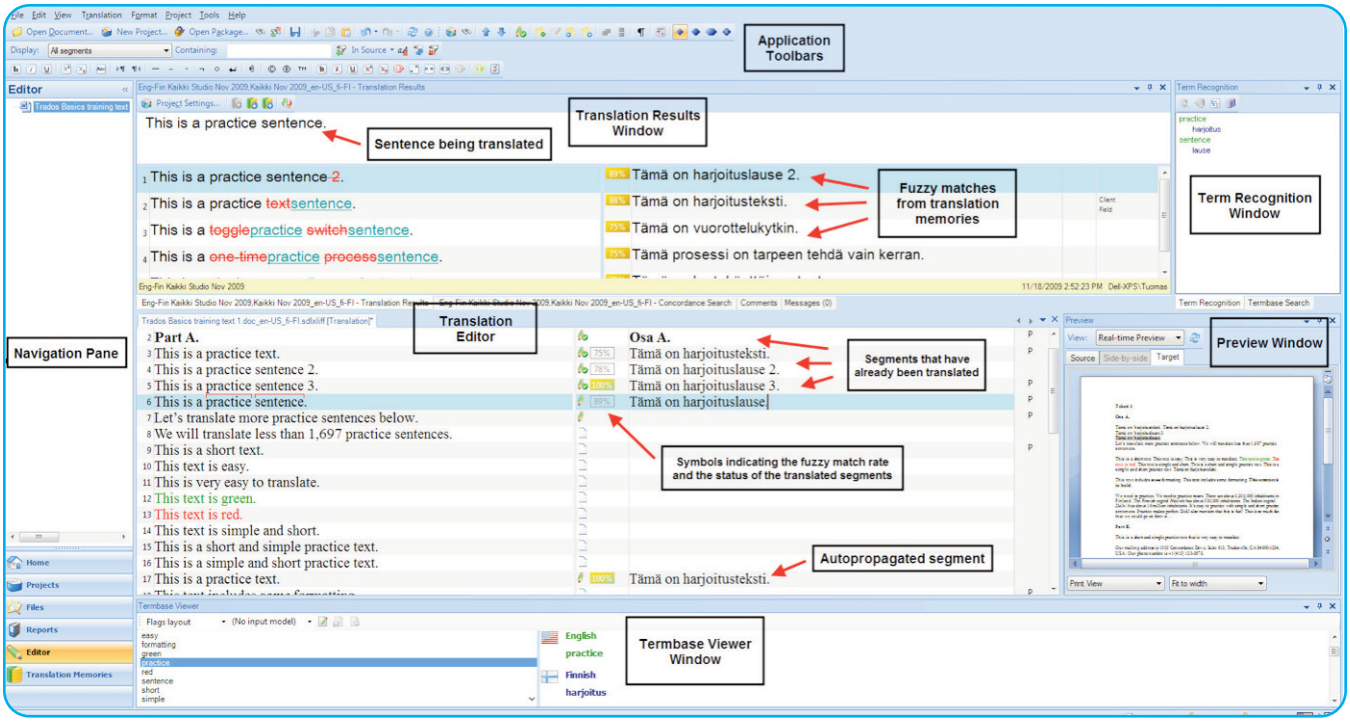


Figure 1. Cathedras suffragarit zothecas. Adlaudabilis ossifragi insectat syrtes, etiam saburre aegre libere iocari quadrupei. Pessimus lascivius matrimonii cel.

5 In addition to the previous four features that I consider major improvements, there are several smaller-scale features that I find really useful. For example, with display filtering[6] you can filter the segments displayed in the Editor view, based on their translation status and source/target content. It also allows you to display duplicate segments and segments that have new translated content. This is great when you want to make sure that you have translated a certain term consistently in the document or want to proofread only segments that include new material.

6 Another smaller but nice feature is the real-time progress counter[7] at the lower right hand corner of the editor window. It shows you real-time word counts (and percentages) for non-translated, draft and translated part of the document. However, I find it strange that it shows only the percentage value by default, and in order to see the word count, you have to change the setting every time the program starts. With all the customization that's possible in Studio, you would hope that this would be customizable too.

7 Since I mentioned the customization, there are plenty of opportunities to get creative, such as the flexible interface layout[8] (you can move, close and resize windows any way you prefer), fully customizable keyboard shortcuts[9], display color settings, fuzzy match difference formatting, adjustable fuzzy bands[10], and six interface language options.

Before I run out of space, I better start covering

some of the problems as well. Some of these are functional problems and some others are features that are missing or don't work in optimal manner. Hopefully, these will be solved by the next service pack or version.

Backwards compatibility is an issue that concerns many Trados Studio users. Possible scenarios range from "no problem" to "impossible" and can be almost anything in-between. To be able to save any bilingual translated file as a TagEditor TTX file would be a major improvement in this regard. However, note that, as a translator, you don't have to wait for your clients to upgrade to Studio before you do it yourself. In most project flow situations, you can easily translate the material in Studio and then provide the translations and

6 Display filtering: http://producthelp.sdl.com/SDL%20Trados%20Studio/client_en/Screen_Layout/Menus_Toolbars/Filters_Toolbar.htm

7 Real-time progress counter: http://producthelp.sdl.com/SDL%20Trados%20Studio/client_en/Ref/A-G/Display_Settings.htm

8 Flexible interface layout: http://producthelp.sdl.com/SDL%20Trados%20Studio/client_en/Screen_Layout/Personalizing_the_Views.htm

9 Keyboard shortcuts: http://producthelp.sdl.com/SDL%20Trados%20Studio/client_en/Setting_Preferences/SpecifyKSs.htm

10 Adjustable fuzzy bands: http://producthelp.sdl.com/SDL%20Trados%20Studio/client_en/Ref/A-G/Batch_Processing/Batch_Processing_Fuzzy_Bands.htm

TM to your client in acceptable formats. Unfortunately, this is not true in every case and it does make sense to figure this out before starting the project. The "SDL Trados Studio Migration Guide[11]" and the recent white paper "TTX Compatibility Guide for SDL Trados Studio 2009 Users[12]" have more info about various scenarios, and are worth studying.

Other problems or missing features that I have encountered include:

- Inability to edit source segments, for example to delete incorrectly placed hard returns or correct typos.
- Unreliable fuzzy match values[13] at least at the lower end of the usable spectrum, below 70%.
- The way source segment differences[14] in fuzzy matches are indicated using the "track changes" type of markings instead of the highlighting used in Trados 2007—this makes it much harder to read and compare the fuzzy matches.
- Strict language variant setting, which does not allow you to use, for example, UK English and US English TMs in the same project.
- The new Hunspell-based spelling checker does not work for Finnish.

You can read more about Studio shortcomings and requested new features at the ideas.sdl.com site. It actually makes very interesting reading to see what users would like to see in their "dream version" of Trados. I have also covered some additional pros and cons on my Trados Studio blog[15].

Of course, one major hurdle for many Trados 2007 users is to learn how to use the new version. It does take some extra effort, but if/when you decide to upgrade, the potential benefits are certainly worth investing a few hours to learn these features. As a trainer, much too often I see translators who have just limped along for years using only the very most basic features of Trados, and not using even those very efficiently. If you feel that you can't learn it by yourself, then get help from workshops, online courses or private trainers to get a good—and happier—start with Studio. ▢

Tuomas Kostiainen is an SDL Trados-approved Trados trainer. He has several years of experience in using Trados as a freelance English to Finnish translator and as a trainer. He has provided training for hundreds of translators and project managers in the US, as well as internationally, and given Trados workshops and presentations for several translator organizations. Contact: tuomas@jps.net, www.finntranslations.com.

11 SDL Trados Studio Migration Guide: http://talisma.sdl.com/display/2n/articleDirect/index.aspx?aid=2783&r=0.06186742

12 TTX Compatibility Guide for SDL Trados Studio 2009 Users: http://snurl.com/ti5ed

13 Unreliable fuzzy match values: http://tradoshelp.wordpress.com/2010/01/03/92/

14 Source segment differences: http://producthelp.sdl.com/SDL%20Trados%20Studio/client_en/Ref/U-Z/Options/Options_TMWindow.htm

15 Trados Studio blog: http://tradoshelp.wordpress.com/

Register for the Mailing List

If you haven't already done so, be sure to subscribe to the LTD mailing list. Go to the Division's website (http://www.ata-divisions.org/LTD/) and click on "LTD Mailinglist." Our listmaster, Katrin Rippel, can't wait to hear from you!