Pangeanic's Do-It-Yourself Machine Translation: User Empowerment and User-Driven MT Processing

[Pangeanic の DIY 機械翻訳：管理権限をユーザーに委ねた ユーザー主導型機械翻訳の全貌]

Elia Yuste, Manuel Herranz, & Alexandre Helle
eyuste/mherranz/ahelle@pangea.com.mt

Pangeanic / B.I Europa
PangeaMT Technologies Division
Valencia, Spain

A.-L. Lagarda, M. García, J. Pla-Civera, M. Blasco, A. Morellá, & J. Mallach
alagarda@iti.upv.es
Institute of Computer Technology (ITI)
Universidad Politécnica de Valencia (UPV)
Valencia, Spain

1. Introduction

This paper reports on how Pangeanic's machine translation (MT) offering, PangeaMT[1], has evolved to include technical components that allow the user to perform MT-related actions that were in the exclusive hands of the MT provider in the past, namely engine creation and updating or retraining. We explain why we thought this was a necessity among our user-base in the translation industry and enterprise market; in other words, why and how we decided to set the DIY[2] footprint in the translation automation arena. The components of a **PangeaMT DIY** solution as released in late Spring 2011 are presented. This information is framed taking into consideration the full picture of PangeaMT technologies and their business models and services until 2012/Q1. Given PangeaMT's innovation-driven slant, we then revamped the concept of MT-DIYing and decided to make it not only available for PangeaMT DIY self-hosted solution clients, like in 2011, but also for users working in the so-called new **SaaS Power** mode. This is a/n r/evolutionary MT SaaS[3] type, whereby users not only enjoy the typical PangeaMT benefit of machine translating as much and as often as needed but they are empowered to retrain their engines online and fuzz-free. This represents a major breakthrough in the translation industry, particularly in the case of organizations and language service providers (LSP) that are really keen to be in full command of all their MT processes, not only machine translating, in a transparent and efficient fashion but yet cannot afford the time or the little to medium range investment to host a DIY solution themselves. Nonetheless, those multilingual content production agents, handling highly confidential data, will continue to be interested in self-hosting their MT solution and growing their own MT ecosystem. Having them in mind, the PangeaMT DIY offering first released in 2011 can be delivered in an integrative, intuitive and user-layered platform called **PangeaMT Full Power** since June 2012. This article concludes with a reflection on how we think user

[1] www.pangea.com.mt
[2] DIY stands for Do-It-Yourself. In this context, it refers to empowering the user of a machine translation (MT) system to go beyond the activity of an MT query or request via a command line or web panel, and be able to manage engine training data and engines, updating them when need be, without having to resort to the MT provider.
[3] SaaS stands for Software as a Service.

intervention leads to user empowerment, an increased intake of MT technology and, ultimately, to overall MT technology advancement.

2. Birth of the PangeaMT DIY Concept

Pangeanic/B.I Europa (Pangeanic for short, an associate member of B.I Corporation in Japan) is a Spain-based language service provider (LSP) with close operational links in Asia and a machine translation division called PangeaMT. In Yuste et al. (2011) we provide an introduction on Pangeanic's transformation from a translation agency into a global LSP and a statistical machine translation solution customizer. This is worth reading if you are interested in finding out more about the early days and the philosophy of PangeaMT and its evolved connection to Moses[4], i.e. PangeaMT has built in several modules and pre- / post-processing scripts on top of Moses, as well as interfaces and retraining procedures in order to develop an industry-tested alternative[5] to off-the-shelf, more rigid MT products.

The fact that Pangeanic wears two hats, that of an LSP and of an MT provider put the company in a privileged position from the very beginning to contemplate different MT technology deployment scenarios. We listened not only to our internal users, i.e. translators, reviewers/QA checkers and project managers, but also to our clients, who interestingly enough come both from the enterprise market and the international language service industry. While the former usually play the role of buyers and hardly own a full localization department with a link to MT or other HLT[6] areas (unless we talk about cutting-edge IT and Internet related players who may even have experience in MT development), the latter feel the urge to adopt MT technology in their processes but, depending on the LSP size and background, they may still be somewhat confused as to the MT paradigm or product that allows them to translate more for less.

While multilingual enterprises are deploying MT for internal communication and information dissemination, and lately even in combination with business intelligence analytical tasks, particularly of rather more volatile content such as UGC[7], the LSP primarily looks at MT as a multilingual technical content production-enhancement tool and a means to react against the ever-increasing demand for publication quality translations in squeezed TAT[8] and drastically constrained budgeting conditions. Therefore, what motivates these two types of MT user/customer segments is cost-effectiveness and ROI[9] even if their deployment scenarios and goals may sometimes differ.

Another aspect they have in common is a demand for rapid and user-focused system's performance, and this may go well beyond the actual step of machine translating. Enterprise market buyers are keen to elicit an MT request from a web interface or an API for integration in other applications, and this has to be run as quickly and smoothly as possible. If they have an in-house localization department, they will usually perform a black-box

[4] www.statmt.org/moses

[5] See Yuste et al. (2010) for more information.

[6] HLT stands for Human Language Technology.

[7] UGC stands for user-generated content.

[8] TAT stands for Turn Around Times.

[9] ROI stands for Return On Investment.

evaluation of outputs across their languages of use and from several MT developers until they make up their mind about what MT provider(s) to choose. Real engine customization, previous in-production experience in the same specialization domain, scalability and robustness will be determining factors for success. Moreover, buyers with a keen interest in MT and a view to intervene in MT processes, although not willing to reinvent the wheel or adapt the Moses toolkit to their needs, are becoming really appreciative of MT solutions that allow for **rapid engine customization** and in particular, **routines for updating engines at will**.

LSPs have been confronted with the challenge of using MT without having technical/computational linguistics background and, on many occasions, the resources to invest in MT technology self-development. However, a growing number of LSPs are getting bored of being just 'MT output-only handlers.' Instead, as it happens with MT users from the buyers' segment, they are becoming more and more interested in learning about MT and playing a role in MT processes, such as engine updating, particularly if this is made easy to them. Even in the case of LSPs that were more experienced with MT, we noticed a clear dependence from the MT provider. This reminded us of Pangeanic's own dependence from our programmers whenever we felt that an engine needed updating or retraining.

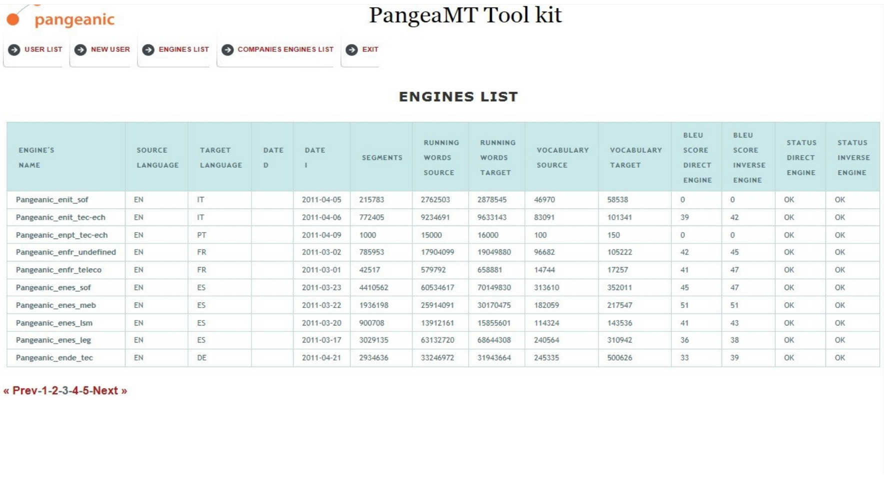
With a view to fostering even more accessibility to **flexible and customizable MT solutions** as well as further **independence from MT providers for engine experimentation and retraining**, Pangeanic officially released the **PangeaMT SMT DIY** Solution for E-FIGS[10] right before the *Localization World Conference* in Barcelona in June 2011. Its main asset consisted of a trouble-free automatic engine training routine, so there was no need to knock on the MT provider's door any more and also pay for every engine updating. Since then PangeaMT DIY, being a self-hosted solution, has proved to be appealing to enterprise and LSP MT users alike, particularly in deployment scenarios where **data confidentiality** is at stake.

3. PangeaMT DIY as released in 2011

PangeaMT SMT DIY included a **training data set structure**, similar to that of an FTP window, where users would simply leave their bilingual aligned files in TMX 1.4b format. Thanks to their associated automatic training routines, engine training and updating could be done automatically, either on a time basis or incrementally (after 50Mb, 100Mb, 300Mb of additional data are detected – a setting that could be determined by the client).

This was paired with a **Control Panel** that showed an engine listing table, with a focus on engine status and automatic quality metrics, such as BLEU for the direct and reverse translation directions, as shown in Fig. 1. PangeaMT DIY was then the first solution on the market to incorporate informative engine status charts with quantitative and qualitative data available on a 24/7 basis.

[10] English into/from French, Italian, German, and Spanish. Offering first this language bundle, then some combinations thereof, other languages, and linguistic families, according to clients' specifications and needs (e.g. Brazilian/Iberian Portuguese, Scandinavian languages, etc.).



ENGINES LIST

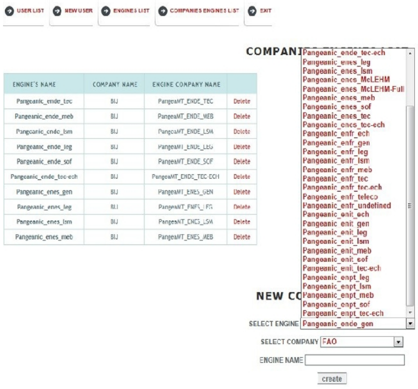
ENGINE'S NAME	SOURCE LANGUAGE	TARGET LANGUAGE	DATE D	DATE I	SEGMENTS	RUNNING WORDS SOURCE	RUNNING WORDS TARGET	VOCABULARY SOURCE	VOCABULARY TARGET	BLEU SCORE DIRECT ENGINE	BLEU SCORE INVERSE ENGINE	STATUS DIRECT ENGINE	STATUS INVERSE ENGINE
Pangeanic_enit_sof	EN	IT		2011-04-05	215783	2762503	2878545	46970	58538	0	0	OK	OK
Pangeanic_enit_tec-ech	EN	IT		2011-04-06	772405	9234691	9633143	83091	101341	39	42	OK	OK
Pangeanic_enpt_tec-ech	EN	PT		2011-04-09	1000	15000	16000	100	150	0	0	OK	OK
Pangeanic_enfr_undefined	EN	FR		2011-03-02	785953	17904099	19049880	96682	105222	42	45	OK	OK
Pangeanic_enfr_teleco	EN	FR		2011-03-01	42517	579792	658881	14744	17257	41	47	OK	OK
Pangeanic_enes_sof	EN	ES		2011-03-23	4410562	60534617	70149830	313610	352011	45	47	OK	OK
Pangeanic_enes_meb	EN	ES		2011-03-22	1936198	25914091	30170475	182059	217547	51	51	OK	OK
Pangeanic_enes_lsm	EN	ES		2011-03-20	900708	13912161	15855601	114324	143536	41	43	OK	OK
Pangeanic_enes_leg	EN	ES		2011-03-17	3029135	63132720	68644308	240564	310942	36	38	OK	OK
Pangeanic_ende_tec	EN	DE		2011-04-21	2934636	33246972	31943664	245335	500626	33	39	OK	OK

« Prev-1-2-3-4-5-Next »

Fig. 1 *Control Panel: Example of Engine List with Statistics*

No computational linguistics knowledge was necessary to navigate through the Control Panel or to perform an engine training task. For instance, just by knowing for sure which bilingual data assets or TMs to throw in a given structure data folder, a new engine or an existing engine could undergo a training process - as simple as that. As soon as the (re)training was completed, the statistical information about the engine would be made available on the Control Panel's Engine List section. No other MT offering had encompassed automatic training routines while displaying their effects on the engines immediately.

The Control Panel also had a basic user activity tracking display and allowed for unlimited users' creation. Any newly created or updated engine could then be assigned to a certain user or company's section in an intuitive application within the Control Panel, just like this:



COMPANI

ENGINE'S NAME	COMPANY NAME	ENGINE COMPANY NAME	
Pangeanic_ende_tec	BIJ	PangeaMT_ENDE_TEC	Delete
Pangeanic_ende_meb	BIJ	PangeaMT_ENDE_MEB	Delete
Pangeanic_ende_lsm	BIJ	PangeaMT_ENDE_LSM	Delete
Pangeanic_ende_leg	BIJ	PangeaMT_ENDE_LEG	Delete
Pangeanic_ende_sof	BIJ	PangeaMT_ENDE_SOF	Delete
Pangeanic_ende_tec-ech	BIJ	PangeaMT_ENDE_TEC-ECH	Delete
Pangeanic_enes_gen	BIJ	PangeaMT_ENES_GEN	Delete
Pangeanic_enes_leg	BIJ	PangeaMT_ENES_LEG	Delete
Pangeanic_enes_lsm	BIJ	PangeaMT_ENES_LSM	Delete
Pangeanic_enes_meb	BIJ	PangeaMT_ENES_MEB	Delete

Pangeanic_ende_tec-ech
Pangeanic_enes_leg
Pangeanic_enes_lsm
Pangeanic_enes_McLEHM
Pangeanic_enes_McLEHM-Full
Pangeanic_enes_meb
Pangeanic_enes_sof
Pangeanic_enes_tec
Pangeanic_enes_tec-ech
Pangeanic_enfr_ech
Pangeanic_enfr_gen
Pangeanic_enfr_leg
Pangeanic_enfr_lsm
Pangeanic_enfr_meb
Pangeanic_enfr_tec
Pangeanic_enfr_tec-ech
Pangeanic_enfr_teleco
Pangeanic_enfr_undefined
Pangeanic_enit_ech
Pangeanic_enit_gen
Pangeanic_enit_leg
Pangeanic_enit_lsm
Pangeanic_enit_meb
Pangeanic_enit_sof
Pangeanic_enit_tec-ech
Pangeanic_enpt_leg
Pangeanic_enpt_lsm
Pangeanic_enpt_meb
Pangeanic_enpt_sof
Pangeanic_enpt_tec-ech

NEW CO

SELECT ENGINE: Pangeanic_ende_gen
SELECT COMPANY: FAO
ENGINE NAME:
create

Fig. 2 *Control Panel: Example of Assignment of an Engine to a Company (or section, department)*

As far as MT requesting and user-application interaction are concerned, there was an intuitive Google-like **web-based translation panel** and, most interesting, a **%MT file translation feature** whereby pre-translated XLIFF or SDL Trados TTX files could be translated after a certain match percent had been leveraged[11] from the translation memory by the user. This is fully customizable by project, so one chooses exactly at what pre-translation level MT will come in, not touching any material leveraged from the TM. Other useful feature available in the translation panel was the **Glossary** TXT file upload that helped fix terminology and preferred expressions, such as *untranslatables* or DNTs[12] in the background prior to the MT output – by the way, quite a historic request to SMT developers.

3.1 Perception of PangeaMT DIY Technology in the industry: Benefits and *side effects*?

MT competing players that are not willing to be as open-minded and **user-empowering** as Pangeanic prefer to have an immovably enshrined focus on providing solutions that are output-driven[13] and not **tool-driven**. Coming also from the Language Service Provision sphere, we at Pangeanic consider that this may be counterproductive for the industry. Allowing for new MT user-focused paradigms and easing typically user-obscure MT processes, enabling interested users to retrain engines or reconvert their TM assets into *fuel* for their engines, can only lead to a wider and more motivated adoption of the technology. Doing this does not mean though that Pangeanic is offering DIY MT tools without taking care of the necessary steps and tasks that go in a real MT customization, as some may have tried to argue.

On the contrary, Pangeanic also offers **Data Consultancy** as part and parcel of every first-time PangeaMT custom engine creation, independent from the mode in which it will be made available to the user: in a PangeaMT SaaS mode or as a self-hosted solution, or from whether the user will be just a translator or an organization interested in performing DIY tasks, too. Clients, whose existing bilingual data may be unclean or scarce for engine training purposes, are particularly appreciative of this type of service. Depending on the condition and size of the starting training corpus, a customization may then encompass the training of one than more engines, with the project resulting in at least two or three engines, ranging from one trained with the client's TM material only, plus others having been trained with other suitable datasets.

[11] PangeaMT's TMX / XLIFF /TTX parsers facilitate the user's preference for any CAT tool as post-editing environment (a TMX or XLIFF editor can be used, and even Tag Editor, which will identify pre-translations as coming from MT! and stop at each segment. This allows users to leverage their TMs as well.

[12] DNT is a localization industry acronym that stands for "do-not-translate" a given word or expression, usually due to product marketing and corporate identity related issues. A typical example is the brand name Apple. For a computational linguist, a DNT is somewhat similar to the NLP concept of *named entity*.

[13] These MT companies rather stick to traditional word-based pricing business models, imposing limits about the number or weight of MT translated word on the user.

When a PangeaMT DIY customer gets hold of their solution, much effort has thus gone into analyzing their training data, seasoning it with any necessary advanced filtering/cleaning[14] pre-processing techniques and additional translational assets from public and reliable sources[15] and repositories. We regard this as a kind of necessary *planting-the-seeds* operation to ensure that if the client later opts to work on their own in a PangeaMT DIY framework, they can then operate intuitively and freely, yet harvesting more than adequate MT offspring, in terms of new and self-retrained engines and their output. It is only then that true MT DIYing can come into play, with minimum intervention from Pangeanic as a MT provider. Yet the PangeaMT team remains at the client's disposal for support or training purposes at any time after the PangeaMT DIY solution has been made available.

The essence of MT DIYing, also as initiated by Pangeanic, has been discussed in detail in the article put together by Simpkins (2012), which targets a readership interested in the hottest localization industry topics.

4. Going Full Power – The PangeaMT DIY Concept Gets Revamped in Spring'12

MT DIY features of engine self-training and updating as well as round-the-clock engine status information access were at the core of the so-called PangeaMT SMT DIY offering launched in Localization World Barcelona in June 2011. These DIY components were easy to use and truly well-received by testers and clients. However, we wondered if we could even go one step forward and make all these DIY features accessible from a single spot. This would in principle facilitate and speed up remote installation in the case of self-hosted solutions. PangeaMT DIY components would have to remain fully functional but also further interlinked and viewable from a platform, which would become highly versatile on the basis of user profiles, that is, the ones using it.

The new **PangeaMT Platform** released in 2012/Q2 is the result of listening to our own needs as an MT-proficient LSP and those of our MT clients with a strong desire to implement customized MT without MT request limits, and with full tracking of all MT process actions and top-notch DIY capabilities – all without having to log in or open several different applications!

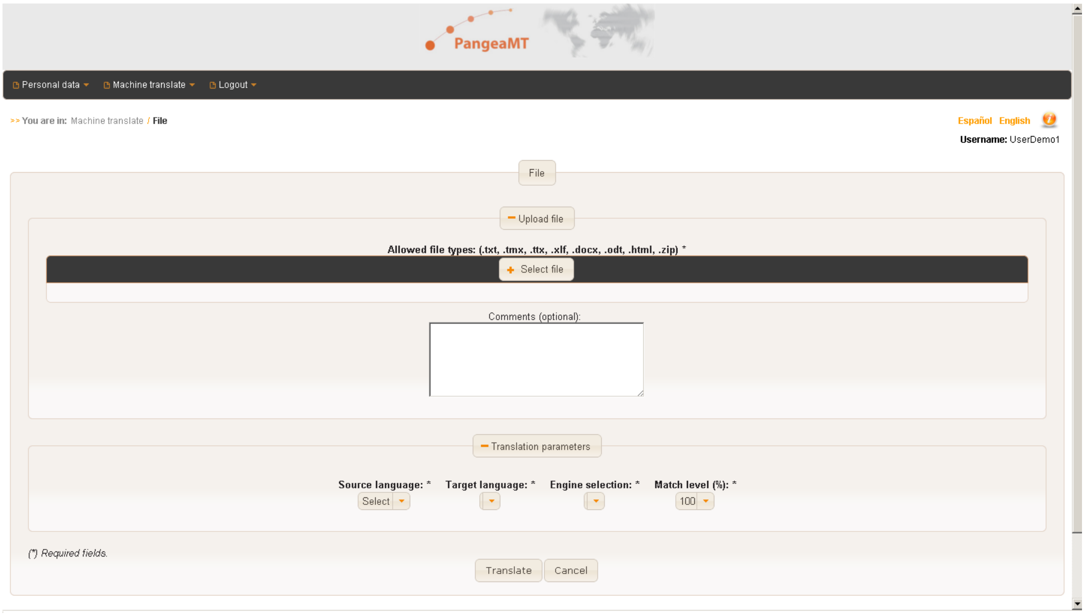
Most importantly, engine training and updating will also be available to users under the **SaaS Power** framework. This is definitely Pangeanic's 2012 innovation with regard to PangeaMT, as discussed in the next section. All in all, users demanding the PangeaMT DIY solution as borne in 2011 are now happy to find even more features that guarantee their independence from the MT provider and a full control over engines and data, hence the new name **PangeaMT Full Power**.

[14] Automatic cleaning routines in the PangeaMT pre-processing module that may be applied for automatically detecting suspicious or unclean translation units and extracting them from the training corpus, that is, the bilingual files or TMX files that the client provides the PangeaMT team for a custom engine creation. If interested, the extracted suspicious units can be handed in for later verification by the user. Suspicious units are those which contain spelling errors or punctuation/numerical inconsistencies, or in which the source language segment is very similar or is identical to that of the target language. The client commissioning the custom solution should be aware of the fact that applying these filtering methods has a pruning, or cleaning effect, therefore leaving their data clean and free of segments that may represent a hurdle to a statistical based translation output. However, the clean training set usually gets smaller in size.

[15] An example of this is TAUS – see www.translationautomation.com/

Integrative in nature – all components that made up the PangeaMT DIY offering of 2011 are made available from a sort of one-stop spot, that is, a password-protected, web-based, unified interface with enhanced MT process informative functionalities that offer more or less options on the basis of a clear-cut user profiling.

Let us know explore those different profiles from a bottom-up approach. Fig. 3 shows the landing page of someone logged in as a translator in the new PangeaMT platform, where no further advanced features are offered. This would be the platform's basic or level 1 user profile.



>> You are in: Machine translate / File

Username: UserDemo1

Allowed file types: (.txt, .tmx, .ttx, .xlf, .docx, .odt, .html, .zip) *

Comments (optional):

Source language: * Select

Target language: *

Engine selection: *

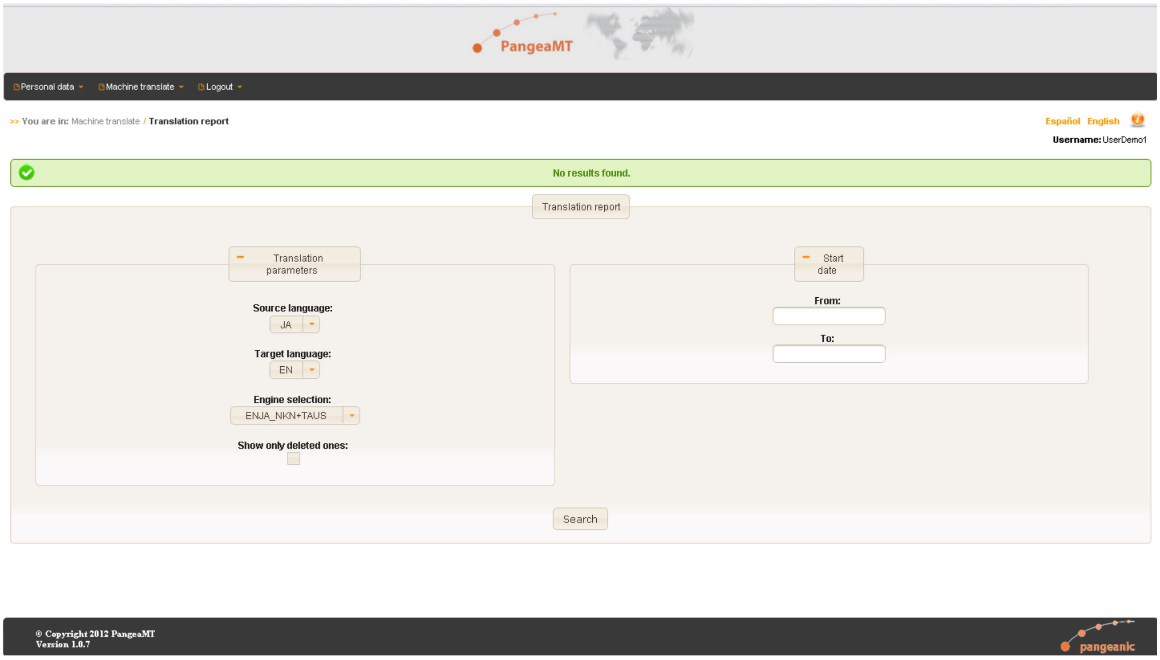
Match level (%): * 100

(*) Required fields.

Fig. 3 *Translator-only Profile: Detail of Machine Translate > File Upload function window view*

In particular, this figure shows how to upload a file or a number of files (subsequently or in an archived .zip file) to request machine translating. Worth noting are the Match level (%) function as one of the required translation parameters as well as the possibility to upload a Glossary (further down in the same window) at the time of requesting MT. The translator will be able to make as many MT requests or upload as many files for MT as necessary, without any limit whatsoever.

Every translator's activity gets logged in the so-called **Translation report**, available from the Machine translate pull-down menu as well. The translator can then query about and get access to any MT jobs of them by specifying a given timeframe. These jobs exclusively and not the ones requested by any other user get listed in inversed chronological order. This is why our translator does not get to see any MT jobs from the selected JP into EN engine, as s/he has not performed any MT request with this engine yet. As shown in Fig. 4, there are no results to see here. The Translation report functionality can be really useful, as all kinds of administrative information and the machine translated files from the jobs listed are accessible here, too. The person logged as company's administrator, enjoying advanced DIY features and unrestricted data access rights, will be the only one accessing every translator's job and related info.



>> You are in: Machine translate / Translation report

Username: UserDemo1

No results found.

Source language: JA

Target language: EN

Engine selection: ENJA_NKN+TAUS

Show only deleted ones: □

From:

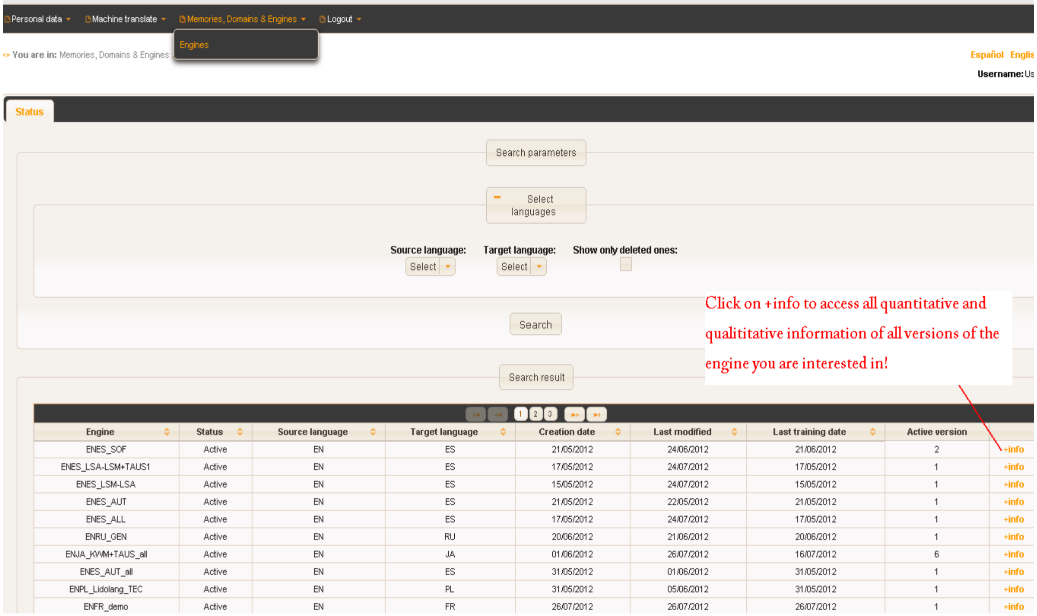
To:

© Copyright 2012 PangeaMT
Version 1.0.7

Fig. 4 *Translator-only Profile: Detail of Translation Report – No results found (no jobs requested by this user with this engine yet)*

The next level user profile, level 2, could be appropriate to a project manager (PM) interested in having the same system's functionalities as the translator but also more administrative information about the custom engines belonging to the company. This user profile in the new platform is benefitting from the same engine control information available in the Control Panel available within the first PangeaMT DIY suite of 2011 (see previous section, particularly Fig. 1). This information is now accessible through the so-called **Memories, Domains, Engines** pull-down menu, particularly in the Engines section, which is the only one partly available for this type of user. When going to Memories, Domains, Engines > Engines, the user sees the engine chart shown in Fig. 5. Apart from some admin info, such as *creation date* or *last training date*, the PM gets to know about the status of an engine, for instance, whether it is *Active*, *Inactive*, in *Training*, etc. This is essential for day-to-day translation project management purposes.

When the PM is keen to know more about a particular engine, clicking on the *+info* button will lead to a really informative page, containing all statistical and administrative details about the various versions of that engine. See an extract of this information in Fig. 6 on next page. Right at the bottom of the displayed information, the user can read about all the bilingual files making up the Corpus List, that is, the files that have served as a training set for that engine version. Any corpus file can get selected and accessed for a quick view, if interested. The same transparency principle applies to the engine training process itself. Should the user be keen to know more about how this happened, there is an engine training log file for both directions of the engine version, direct and inverse[16].



» You are in: Memories, Domains & Engines

Username: Us

Source language: Select

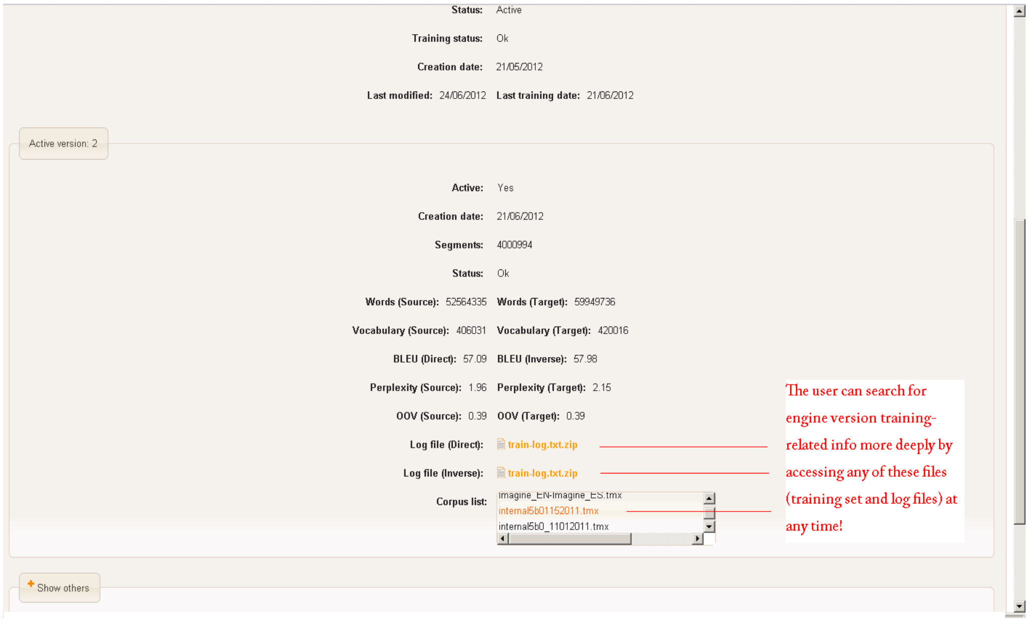
Target language: Select

Show only deleted ones: □

Click on +info to access all quantitative and qualititative information of all versions of the engine you are interested in!

Engine	Status	Source language	Target language	Creation date	Last modified	Last training date	Active version	
ENES_SOF	Active	EN	ES	21/05/2012	24/06/2012	21/06/2012	2	+info
ENES_LSA-LSM+TAUS1	Active	EN	ES	17/05/2012	24/07/2012	17/05/2012	1	+info
ENES_LSM-LSA	Active	EN	ES	15/05/2012	24/07/2012	15/05/2012	1	+info
ENES_AUT	Active	EN	ES	21/05/2012	22/05/2012	21/05/2012	1	+info
ENES_ALL	Active	EN	ES	17/05/2012	24/07/2012	17/05/2012	1	+info
ENRU_GEN	Active	EN	RU	20/06/2012	21/06/2012	20/06/2012	1	+info
ENJA_KWM+TAUS_all	Active	EN	JA	01/06/2012	26/07/2012	16/07/2012	6	+info
ENES_AUT_all	Active	EN	ES	31/05/2012	01/06/2012	31/05/2012	1	+info
ENPL_Lidolang_TEC	Active	EN	PL	31/05/2012	05/06/2012	31/05/2012	1	+info
ENFR_demo	Active	EN	FR	26/07/2012	26/07/2012	26/07/2012	1	+info

Fig. 5 *Memories, Domains & Engines – Engine List: Status (as seen by a level 2 user profile, e.g. a PM)*



Status: Active

Training status: Ok

Creation date: 21/05/2012

Last modified: 24/06/2012 Last training date: 21/06/2012

Active: Yes

Creation date: 21/06/2012

Segments: 4000994

Status: Ok

Words (Source): 52564335 Words (Target): 59949736

Vocabulary (Source): 406031 Vocabulary (Target): 420016

BLEU (Direct): 57.09 BLEU (Inverse): 57.98

Perplexity (Source): 1.96 Perplexity (Target): 2.15

OOV (Source): 0.39 OOV (Target): 0.39

Log file (Direct): train-log.txt.zip

Log file (Inverse): train-log.txt.zip

Corpus list: Imagine_EN-Imagine_ES.tmx
internal5b01152011.tmx
internal5b0_11012011.tmx

The user can search for engine version training-related info more deeply by accessing any of these files (training set and log files) at any time!

Fig. 6 *+info: Detail of Qualitative and Quantitative Info about an engine version of interest*

[16] A into B language direction and the opposite one. PangeaMT engines are always bi-directional, containing in fact two training models corresponding to the two translation directions, direct and inverse.

The information displayed here gets updated constantly and is available on a 24/7 basis. Fig. 6 shows a detail of statistical information about an engine version, containing e.g. the BLEU[17] scores of both corresponding translation models. The page can display this type of information for all versions of an engine at the user's will. This is particularly of use if a PM wishes to know how a translation engine has improved across the different versions, that is, after one or several retrainings, just by looking at the BLEU figures along the different engine versions. In particular, this represents an enhancement with regard to the first PangeaMT DIY solution, where the Control Panel was capable of showing this kind of information but only about the last active version.

An advanced/more senior translation PM or the person(s) designated in the company to act as MT manager needs to have a more advanced user profile in the PangeaMT platform, i.e. they are the ones who will be able to enjoy real DIY functionalities. Logging onto the system as this so-to-speak level 3 user profile, the **Memories, Domains & Engines** pull-down menu is now fully active in all its sections.

Let us now explore what this user profile can do! The **Memories** subsection has two tabs available: *Search* and *Upload*. Intuitively enough, the *Search* tab is resorted to in order to search for any translation memories (TMs) that have been previously uploaded in the PangeaMT data repository, which contains all bilingual translation assets in **TMX 1.4b** format and ready to use for engine training purposes. The search resulting window will display the memories the user is after according to the search parameters previously selected, such as language direction, uploading user, or time frame. The list of displayed memories includes administrative and statistical information, and, as it happens in the case of the Engine Status section discussed above, the user can access the files themselves, by clicking on the Filename highlighted in orange. At the end of every TM line in the table, there is a clickable button called **Edit**, which not only provides extensive information on the TM file itself but allows for a fairly powerful functionality, namely that of associating this TM file to an existing domain. This drag-and-drop easy operation comes handy if our advanced user thinks that it makes sense that this TM file, belonging to X domain, gets also activated in another domain.

The **Memories** *Upload* tab allows the user to upload as many memories as they see fit. This can be done one by one or in a zip archived file. The most important thing is that the user remembers that the system should hold all TMs in the TMX 1.4b standard format. Fig. 7 shows the looks of this tab. As soon as the memory asset(s) get(s) uploaded in the system, this new material is searchable from the Memories *Search* tab described above.

The next section in the **Memories, Domains & Engines** pull-down menu is **Domains**. Natural Language Processing (NLP) systems and applications that are specifically created in or adapted to a restricted domain, that is, a specialized area of knowledge, tend to perform better than those handling or processing open-domain content and theoretically speaking at least, they are less dependent on a vast amount of data. Modeling and training in the PangeaMT framework, as opposed to well-known generalist counterparts such as Google Translate or Bing, has focused on restricted domains as the company itself has a language service provision tradition to serve industry vertical clients that are highly specialized.

[17] Information resulting from other qualitative metrics may be included as per client's request. BLEU stands for Bilingual Evaluation Understudy. An explanation of BLEU is included in the PangeaMT Platform's online help. For further details, please refer to Papineni (2002) or read the entry on Wikipedia (en.wikipedia.org/wiki/BLEU).

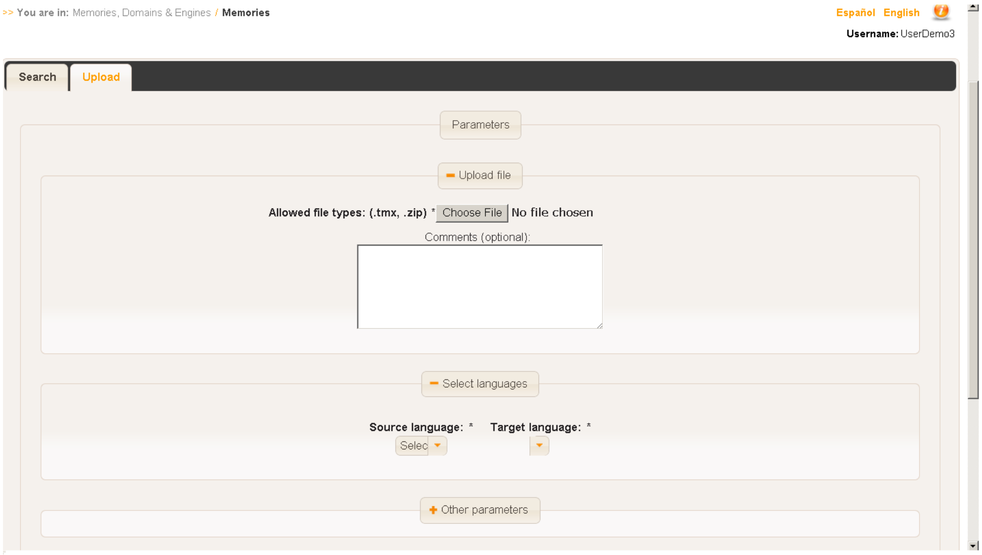


Fig. 7 *Detail of Memories, Domains & Engines > Memories > Upload*

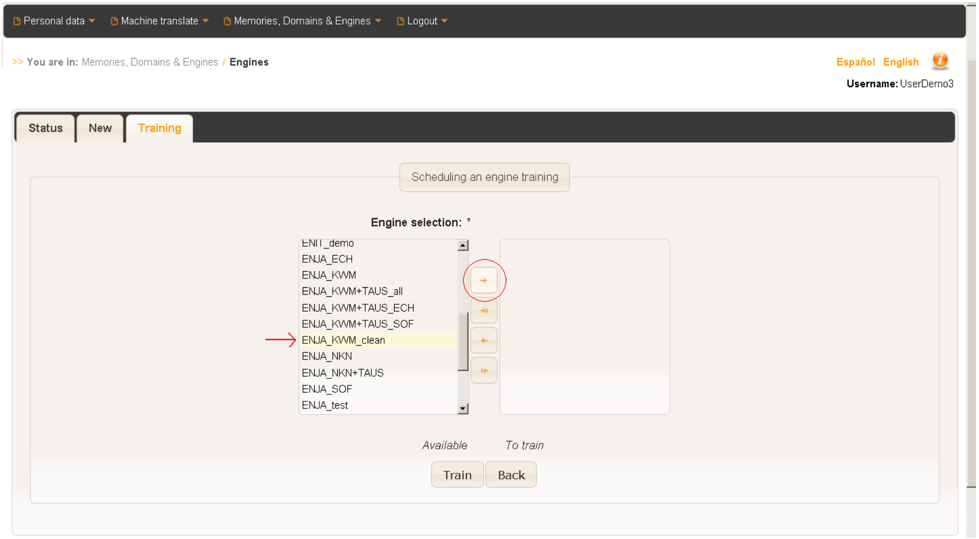
Therefore, in our context, the concept of *domain* is always intertwined with a given industry sector, such as automotive, banking, biotechnology, tourism or renewable energies, to name a few, and the corporate language specificity[18] of the client commissioning the custom MT solution. PangeaMT custom solutions are thus in-domain and client-specific. Some of the domains, meaning therefore *client-driven domains*, we have tackled pose inherent challenges, such as data scarcity or training data showing text types that are closer to natural language than controlled, technical language. This is one of the reasons why the notion of domain in PangeaMT had to be flexible, in that we realized that some customizations, if not aiming to be completely open-domain, would necessarily have to be kind of mixed-domain and even contain supporting bilingual data coming from trusted sources belonging to a different sector to that of our client. Filtering data as well as picking & mixing data from different domains to experiment and generate better-performing custom engines are common operations. How to facilitate this in an intuitive fashion in the new PangeaMT Platform has been achieved through this menu, whereby memories are held, searchable and uploadable, and then get associated with domains in a flexible and traceable step prior to launching an engine re/training.

The **Memories** *Domain* section has two tabs, *Domain Management* to search for domains already available in the system, and *Domain Creation* to create a new domain, which depending on the user's context, it could well be a sub-domain to designate the content of a corporate division, a product line, etc. An illustrative example of this would be a multi-site pharmacy industry client that purchases the **PangeaMT Full Power** solution hosted at their HQ to machine translate their content in 20+ languages originating from four divisions across the globe. They would manage their TMs and assign them to domains that reflect their four divisions, dealing respectively with nine medical areas, such as cancer, tropical diseases, AIDS/HIV, etc. They could create domains on the basis of their geographical divisions or their medicine coverage areas.

[18] Mainly in terms of style, language register, terminology including name entities (NE) or untranslatables/DNTs as called in the translation industry, etc.

Domains are interpreted somehow more traditionally across our translation agency client-base in that they tend to name their newly created domains using industry sectors names and conventions, such as automotive or AUT. However, given the system's flexibility to designate domains and engines according to one's needs, the data available and the envisaged scope of the engine, many LSP companies name their domains per client or sector and client, e.g. AUT or AUT_Mitsubishi. This comes handy when the LSP wishes to launch the training of a sort of **supra-engine** of a given domain, encompassing the memories belonging to domains that in reality are client data specifications.

This example leads to the most powerful DIY function of the new platform – that of the **Memories, Domains & Engines** pull-down menu called **Engines**. This section has three tabs, *Status*, *New* and *Training*. While the former two are rather of admin nature, to search for and learn all about the status of all existing engines and to create a new engine respectively, the *Training* tab empowers the user to train an existing or a new engine – that is really the biggest DIY takeaway of the new PangeaMT Platform in its Full Power flavor (or SaaS Power, as we will see next). There are many reasons that may motivate the action of training or retraining an engine. While retraining usually results from the user's wish to expand and improve the output quality of an engine, such ease-of-use to create a new engine opens up lots of possibilities in a day-to-day MT-enhanced localization business.



>> You are in: Memories, Domains & Engines / Engines

Username: UserDemo3

Engine selection: *

ENIT_demo
ENJA_ECH
ENJA_KWM
ENJA_KWM+TAUS_all
ENJA_KWM+TAUS_ECH
ENJA_KWM+TAUS_SOF
ENJA_KWM_clean
ENJA_NKN
ENJA_NKN+TAUS
ENJA_SOF
ENJA_test

Fig. 8 *Memories, Domains & Engines > Engines >Scheduling the training of a new engine*

Fig. 8 shows the actual moment of dragging and dropping the designation of a new domain, ENJA_KWM_clean, to the right-hand side before pressing the *Train* button that elicits the training of a new engine. ENJA_KWM as well as other mixed-data ENJA_KWM engines existed already. ENJA_KWM_clean now contained the memories provided by the client[19] that had just undergone an advanced filtering (cleaning) pre-processing as part of our ongoing MT customization for this Japanese LSP. In a matter of hours this new engine got ready and automatically viewable to the client, accessing the new PangeaMT platform remotely in a SaaS mode. In just no

[19] KWM is the short form of a client of ours.

time they could access their existing engines and this new one, and then assess on their own if this cleaning process had had a positive impact on the output quality by submitting a translation job to the whole range of engines and evaluating results contrastively.

5. DIY Concept Extended: SaaS Power

In the process of revamping the PangeaMT DIY concept, originally thinking mainly of those clients willing to buy a self-hosted PangeaMT DIY solution, we came up with the idea of an integrative platform as explained in Section 4. In parallel to this, we soon realized that this platform should internally be powerful, secure and scalable enough to allow us, as Pangeanic, to bring to market highly innovative SaaS versions of PangeaMT that would not have been conceived without the inception of the platform.

In essence, up to the end of 2011, our existing SaaS clients could already enjoy the benefits of a typical PangeaMT solution and community-oriented client policies: namely, unlimited MT requests, open translation industry standard and interoperability fostering through PangeaMT TMX and XLIFF parsers and generators, competitive opt-out MT SaaS subscriptions including initial training and support, gradual adoption of the PangeaMT technology across language domains through custom engine piloting, and last but not least, total 24/7 engine access and traceability through the Control Panel as described in Section 3 above. However, our PangeaMT SaaS users were still dependent on us, as MT service providers, for some essential MT tasks, such as managing domain-specific and their own training data and retraining engines, which used to come at a cost, even if moderate.

The new PangeaMT platform would ideally have to allow any SaaS user interested in remote engine training to do so online. This has been possible in early 2012/Q2 and is, in our modest opinion, PangeaMT's breakthrough of the year 2012. The so-called **PangeaMT SaaS Power** offers all DIY functionalities online, corresponding to the level 3 or the most advanced user profile described in Section 5. In other words, when logging in, a PangeaMT SaaS Power user can see **Memories, Domains & Engines** menu completely activated and enjoy all powerful functionalities therein.

The PangeaMT platform managed by the PangeaMT team to cater for all levels of PangeaMT SaaS users, including Saas Power ones, works in dedicated physical and cloud servers and makes use of a number of sophisticated technical features in the background to ensure optimum and secure performance of MT work processes, such as heavy machine translating and engine training, being requested simultaneously by a significant number of users scattered worldwide. PangeaMT SaaS Power users may rest assured that their MT jobs, their TM assets associated to domains of their choice and their custom engines, which they can self update, are available to them round the clock and, most importantly, exclusively.

6. Conclusion

While the PangeaMT technology was borne out of Pangeanic's translation automation needs and expectations as an LSP, we soon decided to market it as custom, fully-tailored in-domain and client-specific engines. PangeaMT technology has evolved to go far beyond the custom engine creation and output-only offering portrayed by our competitors. We wished to set apart from other MT providers that impose word-based or user number limitations to their clients in the deployment of those engines. But was that the only means to empower our users? How about letting them schedule an automatic engine training without our intervention once their engine was first trained by us? Since 2011, users of a self-hosted PangeaMT DIY solution make the most out of their domain- or client-specific bi-texts by mixing and experimenting with these data sets for unlimited in-domain or mixed domain SMT engines (e.g. safety/insurance and automotive, software and life sciences, etc.) and automatic engine retraining or updating.

The new PangeaMT platform presented here, released in 2012, has allowed us to become even more customer-focused and much more versatile in the array of MT-related services and solutions that we can now provide. Thanks to this year's revamping of the 2011 PangeaMT DIY concept in the form of an integrative, user profile-driven and multi-function platform, which can be installed at the client's end or accessible remotely on the Web, MT processes that were previously in the exclusive hands of the MT provider can now be elicited by the user when need be. This is so far the maximum exponent of our mission, i.e. to democratize MT – lowering or even removing access barriers to MT.

The PangeaMT team has strived to ensure that an organization that gets interested in MT but is a newcomer to the field can adopt the technology in an intuitive fashion, yet powerfully and independently from us as much as possible – and when that makes sense, that is, once they get acquainted with essential preparatory steps that we do gladly take care of, such as Data Consultancy, team-wide and management demonstrations and custom development project piloting.

Acknowledgements

The authors wish to thank their ever-growing worldwide and heterogeneous client-base for their invaluable feedback, particularly w.r.t. their different DIY platform and API related needs. The PangeaMT team would like to express their gratitude for the awarded European Union funds under the FEDER program managed by Valencia's local government IMPIVA in order to advance our Statistical Machine Translation technologies (award numbers: IMIDTA/2010/1016, IMIDTA/2011/777 and IMEXPB/2011/1). The collaborating UPV team feels indebted to Spain's Ministry of Science and Innovation (MICINN) and their own University's Research Council for awarding them the grants TIN2009-14511 and UPV/2009/2851 respectively.

Appendix: Last Word on Technical Matters

The PangeaMT Full Power platform is also functionally linked to the PangeaMT API, which allows for easy integration of a PangeaMT customization in computer-assisted localization and multilingual content consumption workflows and applications. Depending on whether you are the localization department of a corporation, an LSP or an organization in need of MT as an enabling technology to accomplish other multilingual technology tasks, you will need direct access to the Platform to query and retrain engines yourself or your own technology applications will be calling in your custom PangeaMT engines via API to get content translated automatically.

For those interested in knowing what is in a PangeaMT solution from inside out, particularly of interest to those clients in need of a self-hosted Full Power solution, PangeaMT works through the virtualization software VirtualBox Virtual Machine (VM). These prospects should get in touch for the latest information on the VM version in use, as well as the recommended technical specifications of the machine where the VM will be run, at the time of commissioning the solution. User's system requirements, such as compatible browsers, recommended resolutions and the like, will be of interest to both self-hosted and SaaS customers, and up-to-date information can also be provided at any time. While SaaS users do not have to bother about hosting technicalities, a natural concern that may arise in them is that of security, at the level of both data and engine handling. Having partnered with European secure hosting leaders, such as the prized company Strato, PangeaMT can ensure a high level of security at all times.

Should you require further details on technical matters, please do not hesitate to contact Alexandre Helle by e-mail. For a cost-free consultation on commercial aspects of a real-world PangeaMT customization as well as cross-company technology collaborations and integrations, kindly get in touch with either Elia Yuste or Manuel Herranz.

References

Papineni, K. et al. (2002) BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proc. of the 20th ACL.* Association for Computational Linguistics. Issue: July, pp. 311-318.

Simpkins, A. (2012) Do-It-Yourself MT. In *Multilingual*. Multilingual Computing Inc. Issue: July/August 2012, # 129 Vol. 23 Issue 5, pp. 44-44. www.multilingual.com/articleDetail.php?id=1946

Yuste, E. et al. (2011) Going Hybrid: Pangeanic's and Toshiba's First Steps Toward ENJP MT Hybridization. In AAMT Journal. Issue 50: December 2011, pages: 33-39. ISSN 1883-1818.

Yuste, E. et al. (2010) PangeaMT – putting standards to work... well. In *Proceedings of The Ninth Conference of the Association for Machine Translation in the Americas – AMTA 2010*, Denver. Available at: amta2010.amtaweb.org/AMTA/papers/4-04-HerranzYusteEtal.pdf