KantanAnalytics.

Build Better Business Models for *Machine Translation*

KantanMT.com

Executive Summary

KantanAnalytics™ is a newly developed, Machine Translation (MT) quality scoring technology that it is anticipated to have a considerable impact on the localization industry. It will not only give Project Managers more control and flexibility when scheduling and costing translation projects, but it will also facilitate the increased efficiencies and cost savings of localization projects and workflows.

As an increasing number of businesses target online sales to increase revenues, demand for localized content will continue to grow. The Common Sense Advisory has indicated that the localization industry will reach approx. €29 billion ($40bn) by 2014, and TechNavio predicts the Global Machine Translation market will increase 18.05% Compound Annual Growth Rate (CAGR) between 2012 and 2016.

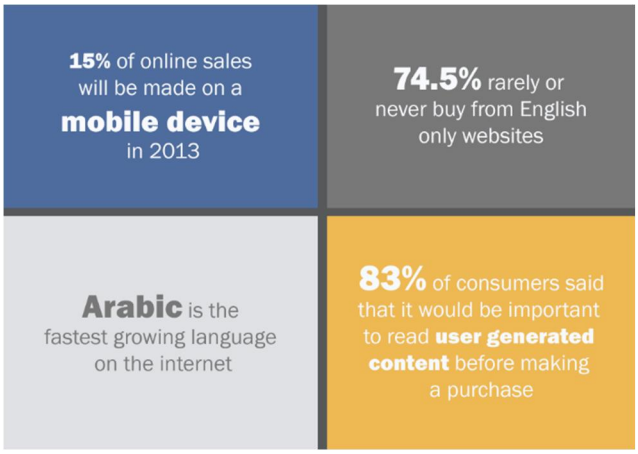
Machine Translation, which is defined by the Oxford English Dictionary as; “translation carried out by a computer” and in the Collins English Dictionary as; “the production of text in one natural language from that in another by means of computer procedures”, can provide comprehensive, high speed and high quality translations. However, measuring MT quality, and integrating MT into localization workflows has long been a challenging undertaking for Language Service Providers (LSPs) interested in using Machine Translation. Inadequate MT Quality Evaluation (QE) Metrics have made it difficult to confidently cost and schedule MT related projects, leading to the greatest challenge facing LSPs; how to develop suitable business models for costing and scheduling scalable MT projects. The ability to develop and apply such a model would give LSPs a powerful competitive advantage in leveraging MT and Translation Memory (TM) technology.

KantanMT, a cloud-based implementation of MOSES Statistical Machine Translation (SMT) technology offers members an easy to navigate platform for building customized MT engines. KantanMT takes on the challenges faced by LSPs today and creates innovative industry solutions unmatched by any other MT service provider. The KantanMT platform effortlessly scales to generate a high-quality, low-cost MT system, which gives members the ability to build competitive business models through the use of its unique KantanAnalytics technology. Using KantanAnalytics, LSPs can accurately evaluate Machine Translation quality, fairly price Machine Translation, precisely schedule projects, and take on challenges such as estimating productivity calculations, predicting scalability, optimising language asset integration and forecasting human resources management. Perhaps most importantly, it allows LSPs to confidently predict a Return on Investment (ROI) for their MT investment strategies.

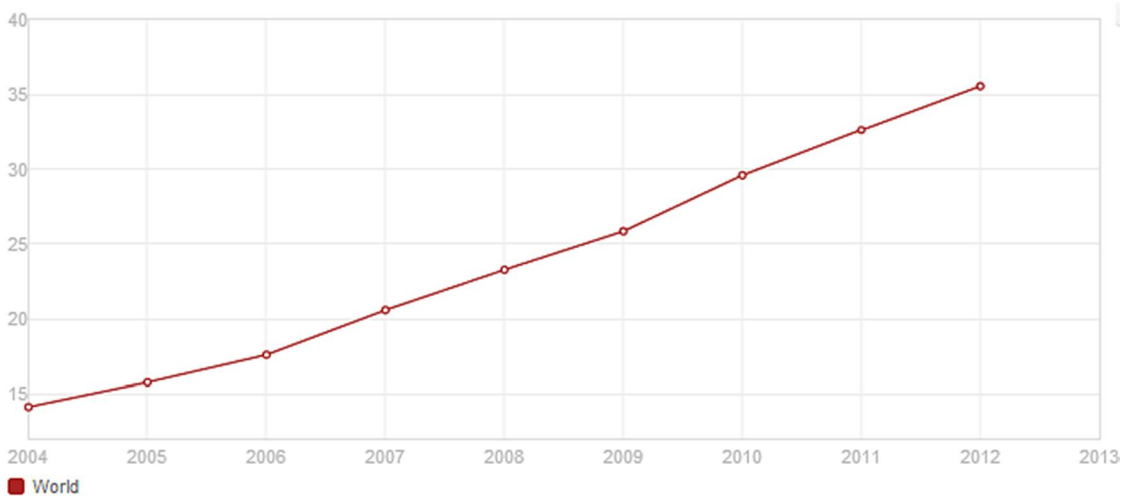
Evolution of Translation Services

Publishing multilingual content is a standard part of any company's communication strategy when operating - or planning to operate - globally. Companies that succeed in capturing international markets are those that speak to their audience in the languages of their 'locales'. However, the challenge in meeting the volume of content that must be translated is growing exponentially and companies find it ever more difficult to keep pace with this 'explosion of content'.

Online Consumer Power



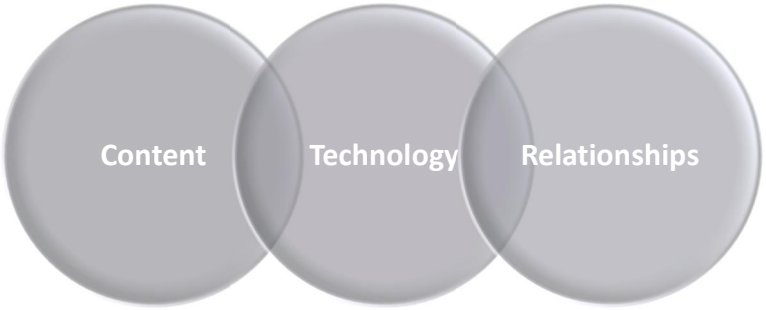
One of the major challenges for companies operating globally is converting sales prospects in different language regions. Forrester research identified that **42% of European online shoppers only shop online in their native language**, 34% of Canadians prefer a French language site, and in Quebec the preference for shopping on French sites is 64%. These preferences continue to be replicated, as more online consumers emerge from nations such as Brazil and China. **Increasing demands for fully localized websites** and information will continue to have a marked influence on the localization industry, as consumer demand for own language information grows.



Global Internet users (per 100 people), Source: International Telecommunication Union, World Telecommunication/ICT Development Report and database, and World Bank estimates.

Translation Industry Growth

The localization industry is one of the fastest growing global industries. The Common Sense Advisory Board estimates that the global 'language services market' is growing at an annual rate of 7.4% and that by 2014 it will be worth €29 billion ($40bn). The natural language processing sector is larger still, with 21.1% CAGR, growing from €2.8 billion in 2013 to €7.4 billion in 2018, and the market for Machine Translation is estimated to grow from €1.2 billion in 2012 to €5.7 billion in 2019, according to research carried out by Markets and Markets. TechNavio explains that innovation in the localization industry is being driven by a variety of factors including:



Automated Translation

Automated translation was first brought into the public domain in the 1950s, but it is only in the last few decades that MT research has sharpened its focus. A major turning point for modern research in MT technology came during the 1990s, and was driven by the rapidly increasing internet usage for commercial transactions, the common accessibility of computers in society, and the incredible computing power of the average computer.

This renewed drive to develop a powerful, accessible Machine Translation model has meant that today, Machine Translation not only can be used both as a tool to **facilitate professional translators**, but it also provides a stand-alone tool capable of **translating large volumes of data quickly**.

The greater levels of computational resources, and the development of the **open source MOSES SMT system,** have made a great impact on the **quality of text now provided by Machine Translation** systems.

About MOSES:

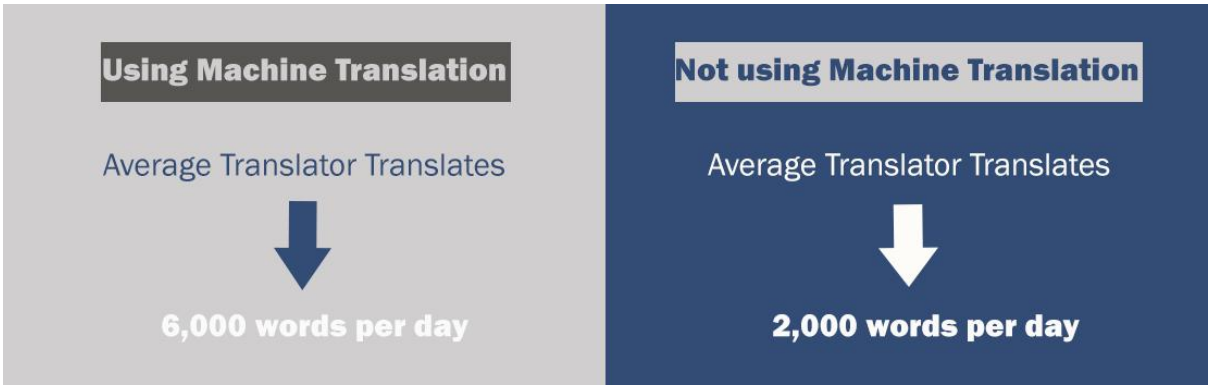
MOSES is an open source Statistical Machine Translation (SMT) system developed and maintained by the MOSES community and now available to potential end users everywhere. SMT systems use algorithms to select the best possible translations based how frequently a pattern occurs. MOSES is now the preferred MT system in the localization industry because of its flexibility and speed.

Content's Relationship with MT

The steady increase in online multilingual content is providing an abundance of language resources that can be optimized in the training of SMT engines. The growing number of mobile device users, is also generating quantum leaps in 'Big Data' creation. Cisco has estimated that this global mobile data traffic will have a 66% CAGR between 2012 and 2017, reaching approximately 11.2 exabytes (11.2 billion gigabytes) of mobile data per month. The growth in content, and the need to make this content multilingual is good news for LSPs. However, the industry still faces monumental challenges when it comes to pricing Automated Translation services, including Statistical Machine Translation. Costing and scheduling translation jobs using error typologies and/or industry standard metrics is not practical and can lead to misestimation. The localization industry needs a system of quality analyses that they can trust and that fits with existing business and pricing models.

The content explosion has made it unfeasible for organizations to include traditional 'word rates' in their localization budgets. According to Forrester, an average professional translator can translate approx. 2,000 words per day, with machine-aided translation this figure increases to approx. 6,000 words on average, equal to a 40% cost reduction.

Real-time customized MT systems are the only solution for translating large volumes of content on-demand. Maintaining a competitive advantage in global economic markets requires measuring productivity gains and adding value. Costs do not necessarily need to be lowered, only provide more 'value' for the same price.



Technology

The need for segment-by-segment Quality Estimation (QE) and TM integration will drive the future of MT technology. LSPs will experience improvements in productivity and ROI through a more efficient and streamlined translation process. As anticipated by TAUS, MT technology will dominate the new "convergence era", where translation and LSP services are available through any application, and on any device.

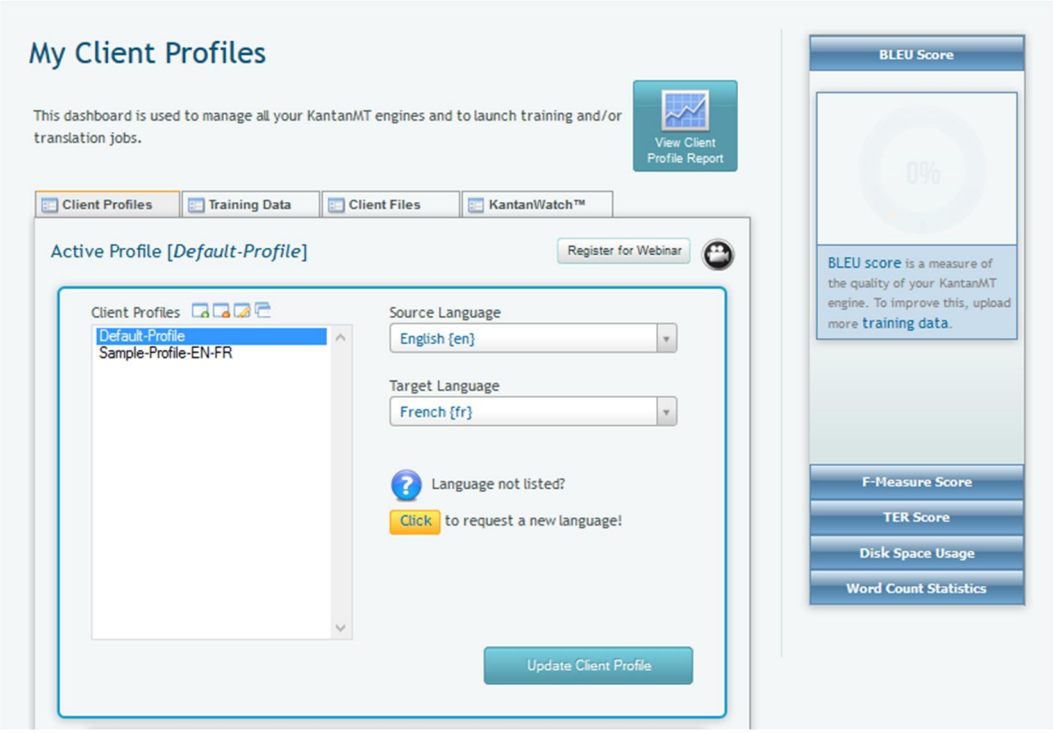
Convergence era:
Translation on demand,
on all devices.

Relationships – Community and Business

LSPs are increasingly leveraging crowdsourcing communities (non-professional and volunteer translators) for translating more informal User Generated Content (UGC), which has been created en masse by internet users, particularly through social media. The creation and translation of this content is being used to foster business and community relationships through online groups and forums. LSPs are also realizing the value of partnering with technology companies and research centres to improve workflow integration.

What is KantanMT?

KantanMT is a cloud-based implementation of MOSES SMT technology that offers members an intuitive, easy to navigate platform for building customized MT engines. KantanMT effortlessly scales to generate a high-quality, low-cost MT platform for LSPs, by leveraging the power and flexibility of the cloud.



Members can build domain specific engines for individual clients with a guarantee that all data is fully encrypted and hosted on a secure Amazon web server. The KantanMT Application Programming Interface (API) enables members to fully integrate MT into their existing Workflows, and by using KantanAnalytics segment level quality analysis members can accurately evaluate a project's cost and schedule.

KantanAnalytics is a revolutionary new technology, which will transform the localization industry. It generates QE scores on a segment-by-segment basis for translations generated by a KantanMT engine. KantanAnalytics takes a granular approach to segment level QE analysis and will become the cornerstone for creating new methods to predict post-editing effort and project management costs.

The scores will help localization Project Managers prepare accurate schedules for MT projects creating transparent and streamlined workflows for their clients.

KantanAnalytics provides solutions to Industry Challenges

Machine Translation adoption is not without its challenges. Some of the more common problems are listed below along with solutions developed by the KantanMT team.

Quality Estimation

According to a benchmarking study conducted by TAUS in 2011, current QE models are very inflexible. QE metrics do not take into consideration different content types, context or communicative function, and the 'one size fits all' approach is not applicable to real world business applications, rendering even the strictest error-based evaluation models inadequate. Time consuming manual evaluation exercises are only practical when carried out on small samples, but selecting samples may also prove inefficient if the sample is not representative of the translated text.

Solution: The KantanAnalytics analysis feature generates a QE score for each segment of a translation. The segments are assigned an individual quality score as a percentage value, similar to the 'fuzzy match' scoring method used for generating TM match values. This is a flexible QE method giving accurate individual quality scores. Members can use KantanWatch™, BLEU, TER and F-measure scores to show the engine's overall quality level during the training or development stage, knowing the engine's quality level provides a benchmark. KantanAnalytics can then be used to analyse the quality of each segment generated by a KantanMT engine. Segments with TM matches lower than 85% are put through the MT engine to generate a translation.

Pricing

There is no standardized pricing model for Machine Translation, making it very difficult to set standard prices for clients. Traditional volume-based pricing models based on word rates, which are adjusted according to the project cannot be applied, and hourly rates, or task-based pricing are also inefficient for pricing and costing MT projects.

Solution: Using KantanAnalytics, Project Managers can measure post-editing effort in the same manner as TM matches. The costs can then be calculated according to existing business models used for pricing TM matches. KantanAnalytics data can be easily downloaded to an Excel file, so Project Managers can calculate costs and schedules easily and efficiently, without having to manually add data into a spreadsheet. Costing can be accurately calculated for both the client and LSP, improving everyone's bottom line.

Scheduling

Localization Project Managers usually work on a number of localization projects simultaneously, so planning and scheduling can be challenging, particularly when calculating a project's time-to-market. Content and quality of the language assets including human translators, MT systems and TMs are all factors contributing to the time taken per project.

Solution: Project Managers can manage a projects time-to-market more effectively with an automated QE system in place. The KantanAnalytics segment score can dictate how post-editing work is divided. More experienced and/or domain specialized post-editors can often work through lower scoring segments faster than less experienced post-editors.

Scalability Challenge

Scaling to meet real-time translation demand requirements, is not possible using traditional translation methods. Traditional approaches have limits on the volume of translations that can be produced per hour and per day, as even expert translators using CAT tools have a maximum daily translation output. This can be particularly problematic for Project Managers when a 'sim-ship' or simultaneous shipment release of a product is planned.

Solution: KantanAnalytics can help Project Managers prioritise segments for translation. SMT engines process vast quantities of data quickly – this output can be assigned a quality score at segment level meaning trained post-editors can work more efficiently and faster depending on client requirements. 'Fit-for-purpose' translation requirements can be completed very quickly.

Human Resources Challenge

Localization Project Managers require excellent organization and strategy planning skills to manage teams of translators, and technical, linguistic and quality control specialists, both internally and externally to the organization. The management of these resources can be one of the greatest challenges for a LSP, and a costly one also.

Solution: TAUS has carried out studies, which prove that post-editing Machine Translation output is more productive than starting a human translation from scratch. Experienced post-editors who have a higher level of technical or domain knowledge will be more productive. When deciding on post-editing strategies, Project Managers can improve productivity by assigning post-editing tasks based on KantanAnalytics QE scores relative to their technical or domain knowledge.

Calculating Improved Productivity and ROI

Project Managers often use a variety of production metrics to measure the progress of each project and ROI. Application of these metrics can be challenging, each project is different and there is no industry standard for measuring Machine Translation output.

Solution: The KantanAnalytics QE scoring feature can be used as a tool to measure production metrics such as the percentage of text re-use (100% matches, fuzzy matches etc.) to improve operational throughput and efficiency. This also makes it more straightforward for LSPs to evaluate the productivity for post editors.

Language Asset Integration Challenge

Translation memories are an integral part of the translation workflow. However, they can only be productive when there is a similar segment stored in the TM. If software is not integrated effectively it makes it more difficult to generate centralized feedback.

Solution: Integration of both MT and TM technologies takes advantage of both the high quality and extensive translation memories clients build up over time. These language assets are either translated or quality checked by humans and make excellent client specific training data that can be used to train highly customized engines. KantanAnalytics lets the user know exactly how many segment matches above 85% came from the TM, and how many segments are translated through the MT engine.

KantanAnalytics enables Project Managers to:

- Develop a transparent framework for accurate cost and schedule estimations, resulting in a more efficient cost management structure during all stages of a project.
- Develop a tiered cost model based on KantanAnalytics QE scores, similar to the traditional TM costing model, so costs can be easily calculated even when integrating both MT and TM technologies.
- Predict project schedules by identifying, which segments require the most post-editing based on the client requirements. Higher scoring segments are proven to take less time to post-edit.
- Distribute post-editing jobs appropriately based on the post-editor's level of expertise.
- Control and manage quality performance of the engine, higher KantanAnalytics scores indicate a better quality engine. Engines with a high distribution of low-scores can be retrained or replaced with better more domain specific engines.

KantanAnalytics

KantanAnalytics technology assigns a QE score as a percentage value for each machine translated segment in a KantanMT engine. The technology predicts the score a human translator would likely assign as to the utility of the translation, and the higher the score, the better the quality and consequently the less effort required to post-edit it.

KantanAnalytics Report

The KantanAnalytics project management report can be easily generated from the member's dashboard. The report can either be viewed directly, or downloaded as a Microsoft Excel file and includes; segment-by-segment QE scores, word, character, placeholder and tag counts.

1	KantanAnalytics Summary Report for job no [**22671**]			Date	28/11/2013		
2							
3	Summary						
4		**Segments**	**Words**		**Characters**	**Tags**	**Placeholders**
5	**Total**	17	316		1,857	0	0
6	**Matched**	16	311	98.40%	1,824	0	0
7	KantanAnalytics						
8		**Matched**	**Words**	**Matched [%]**	**Characters**	**Tags**	**Placeholders**
9	**100%**	9	229	72.50%	1,349	0	0
10	**90-99%**	0	0	0%	0	0	0
11	**85-89%**	0	0	0%	0	0	0
12	**70-84%**	3	7	2.20%	41	0	0
13	**55-69%**	3	56	17.70%	326	0	0
14	**40-54%**	1	19	6%	108	0	0
15	**Others**	1	5	1.60%	33	0	0
16	**Totals**	**17**	**316**		**1,857**	**0**	**0**
17	Profile scores						
18			**Word Count**	**Unique WC**	**BLEU**	**TER**	**F-Measure**
19	**Default-Profile**		407,649	17,552	69.70%	80.70%	28%
20							

KantanAnalytics™ Excel Report

KantanAnalytics™ dashboard report

The dashboard report results are graphically represented at the top of the report. The first graph illustrates the quantity of matches above 85% from the Translation Memory using TotalRecall™ technology, as a percentage of the file to be translated. The second graph, also represented as a percentage, shows matches with a value less than 85%. These are put through the customized KantanMT engine. Then a third graph, represented as a bar chart shows the KantanAnalytics QE scores in 10% increments. This data is also listed below the graphs in numerical form.

KantanMT Technology

KantanMT Technology is based on the MOSES Statistical Machine Translation (SMT) phrase-based decoder. The decoder uses translation tables to find equivalent target words or phrases from words or phrases in the source language. Using the basic decoder, KantanMT developed a new core processing and finishing technology designed to dramatically improve the speed, processing accuracy and reliability of the original MOSES system.

Key Characteristics:

- Introduced **Unicode character encoding** to use with topographically different language pairs, like Asian languages, as the MOSES decoder built from a European perspective only supports plain ASCII files.
- Incorporates the building of custom file parsers using **GENTRY technology** in the industry standard XML file formats. GENTRY is a script driven technology and therefore, it is possible to create rule files. These rule files, created in a simple text editor can instruct your KantanMT engine on what to translate, and a XML file parser can be completed in a matter of minutes.
- Uses PEX technology to automate many manual or repetitive post-editing tasks, this technology can be customized for client specific styles or formatting.
- Operates completely on the cloud, hosted by Amazon Web Services (AWS).

For more information on KantanMT Technologies please go to: www.kantanmt.com, or contact Niamh Lacy, niamhl@kantanmt.com.

CNGL Centre for Global Intelligent Content

KantanAnalytics™ was co-developed with the CNGL Centre for Global Intelligent Content (Dublin City University, Ireland). CNGL is a collaborative academia-industry research centre dedicated to the development of advanced global content processing technologies. The CNGL research centre combines the expertise of its more than 150 investigators at four universities: Trinity College Dublin (lead institute), Dublin City University, University College Dublin and University of Limerick, as well as its industry partners, to produce technologies with significant economic and societal impact. More information can be found at www.cngl.ie.