Towards a MultiLingual Data & Services Infrastructure

Support Action
Grant agreement no.: 610951

D6.2 – Analysis and Specification of Candidate Services
MLi – WP6-D6.2

<table>
<thead>
<tr>
<th>Project Number</th>
<th>FP7-ICT-610951</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual Date</td>
<td>28 April 2016</td>
</tr>
<tr>
<td>Document Author/s</td>
<td>Tilde, DCU</td>
</tr>
<tr>
<td>Version</td>
<td>1.0</td>
</tr>
<tr>
<td>Dissemination level</td>
<td>PU</td>
</tr>
<tr>
<td>Status</td>
<td>Final</td>
</tr>
<tr>
<td>Document approved by</td>
<td>Tilde</td>
</tr>
</tbody>
</table>
### Document Version Control

<table>
<thead>
<tr>
<th>Version</th>
<th>Date</th>
<th>Change Made (and if appropriate reason for change)</th>
<th>Initials of Commentator(s) or Author(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>05.01.2015</td>
<td>Initial fishbone</td>
<td>Tilde</td>
</tr>
<tr>
<td>0.2</td>
<td>01.02.2015</td>
<td>Stable fishbone</td>
<td>Tilde</td>
</tr>
<tr>
<td>0.3</td>
<td>16.03.2015</td>
<td>First draft</td>
<td>Tilde</td>
</tr>
<tr>
<td>0.4</td>
<td>21.08.2015</td>
<td>Input from DCU integrated</td>
<td>DCU</td>
</tr>
<tr>
<td>0.5</td>
<td>23.10.2015</td>
<td>Results from the MLi MT@EC Workshop included</td>
<td>Tilde</td>
</tr>
<tr>
<td>0.6</td>
<td>08.01.2016</td>
<td>Section on e-Commerce included</td>
<td>Tilde</td>
</tr>
<tr>
<td>0.7</td>
<td>19.02.2016</td>
<td>Generic specification for Speech services added</td>
<td>Tilde</td>
</tr>
<tr>
<td>0.8</td>
<td>31.03.2016</td>
<td>Evaluation report on LTi Cloud prototype added</td>
<td>Tilde</td>
</tr>
<tr>
<td>1.0</td>
<td>28.04.2016</td>
<td>Final version</td>
<td>Tilde</td>
</tr>
</tbody>
</table>

### Document Quality Control

<table>
<thead>
<tr>
<th>Version QA</th>
<th>Date</th>
<th>Comments (and if appropriate reason for change)</th>
<th>Initials of QA Person</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>28.04.2016</td>
<td>Final version</td>
<td>Tilde</td>
</tr>
</tbody>
</table>
TABLE OF CONTENTS

EXECUTIVE SUMMARY ............................................................................................................. 6

1. INTRODUCTION ...................................................................................................................... 7

2. LT SERVICES FOR THE LANGUAGE INDUSTRY ................................................................. 9
   2.1. SERVICES IN SUPPORT OF THE ONLINE CONTENT LOCALISATION/GLOBALISATION CHAIN .............................................................................................................................................................. 9
   2.2. TRANSLATION SERVICES FOR COMPANIES OPERATING ACROSS NATIONAL AND LINGUISTIC BOUNDARIES ................................................................................................................................. 11

3. LANGUAGE TECHNOLOGY SERVICES FOR THE FINANCIAL SECTOR ...... 13

4. MT SERVICES FOR E-COMMERCE SECTOR: CASE STUDIES ............................. 15
   4.1. CASE STUDY: ZALANDO .................................................................................................. 15
   4.2. CASE STUDY: ETSY ...................................................................................................... 19
   4.3. COMMON OUTLOOK ..................................................................................................... 25

5. MT SERVICES FOR PUBLIC SECTOR .............................................................................. 26
   5.1. MT IN THE LEGISLATIVE PROCESS ........................................................................... 27
   5.2. TRANSLATION SERVICE FOR THE EU PRESIDENCY ............................................. 29
   5.3. MACHINE TRANSLATION FOR CONNECTING EUROPE FACILITY ..................... 29
   5.4. MT FOR THE PUBLIC SECTOR CASE STUDY: HUGO.LV .................................... 30

6. MT SERVICES FOR CYBER SECURITY AND DEFENCE ........................................... 32

7. MT SOLUTIONS ..................................................................................................................... 34
   7.1. LARGE-SCALE CLOUD-BASED SERVICES .................................................................... 34
   7.2. PROLIFERATION OF MOSES TOOLKIT ......................................................................... 34
   7.3. LONG-TERM SUSTAINABILITY OF MOSES .................................................................... 35
   7.4. COMBINATION OF STATISTICAL AND KNOWLEDGE-BASED METHODS ............... 35
   7.5. STRATEGIES FOR CUSTOMIZED MT SYSTEMS ..................................................... 35
   7.6. DATA PREPARATION SERVICES .................................................................................... 35
   7.7. MT SOLUTIONS AND DEVELOPMENTS ....................................................................... 36

8. GENERIC SPECIFICATION OF ESSENTIAL SERVICE CATEGORIES ............. 42
   8.1. MACHINE TRANSLATION ................................................................................................ 42
   8.2. NATURAL LANGUAGE PROCESSING ............................................................................... 46
   8.3. SPEECH TECHNOLOGIES .............................................................................................. 49
   8.4. OPEN DATA MANAGEMENT FOR PUBLIC AUTOMATED TRANSLATION SERVICES . 51

9. CONCLUSIONS ....................................................................................................................... 54

10. REFERENCES .......................................................................................................................... 56

11. APPENDIX LTI CLOUD EVALUATION REPORT ......................................................... 57
   11.1. INTRODUCTION ............................................................................................................. 58

MLi WP6/D6.2
11.2. Evaluation Approach and Methodology ........................................... 58
11.3. Test Report of the Service Platform ............................................. 59
11.4. Comparison with Other Platforms and Similar Solutions .................. 63
11.5. Conclusions and Suggestions for Future Development ...................... 65
LIST OF FIGURES
Figure 1. Zalando.com landing page................................................................. 15
Figure 2. Search result on English version of Zalando.com for "Men’s Sandals"........ 16
Figure 3. Search result on German version of Zalando.com for “Sandalen”......... 17
Figure 4. Window on Etsy.com prompting users to choose their language.......... 20
Figure 5. Item description for "O’Malley sunglasses" on Etsy.com ...................... 21
Figure 6. Screenshot of item description machine translation into French, offering users
to see original (left) and screenshot of item description in English, offering users to
translate into French (right)........................................................................... 21
Figure 7. Search results for the word “clutch” on Etsy.com ................................ 23
Figure 8. Reference architecture for MT services.............................................. 45
Figure 9. Language tool grouping according their functional area ....................... 47
Figure 10. Request execution of Natural Language Processing tool...................... 48
Figure 11. Speech Services architecture.......................................................... 50
Figure 12. Bitext lifecycle .................................................................................. 52

LIST OF TABLES
Table 0.1 Test of LTI Cloud Broker ................................................................. 59
Table 0.2 LTI broker tests for case sensitivity.................................................. 60
Table 0.3 Test results of LT Store categories .................................................. 61
Table 0.4 Test results of filtering scenarios....................................................... 62
Table 0.5 Test results of section "My LTCs"...................................................... 63
Table 0.6 Comparison to different types of existing API cloud......................... 63
Table 0.7 Suggestions for future work on LTI Cloud ........................................ 65
EXECUTIVE SUMMARY

The MLi support action\(^1\) aims to establish the foundations of a comprehensive *European Multilingual data & services Infrastructure* (MLi). In scope of this task the WP6 aims to identify, characterise and lay the foundations of a number of high-impact innovation driven public-interest services and innovation services geared towards the digital single market.

In the Deliverable 6.1, we identified and reviewed the potential needs and current experience on end-user systems and plug-in components for (a) trans-European public interest services (mainly but not exclusively in the public sector) and (b) cross-national/sectoral services geared towards Europe's digital single market. Findings of D6.1. is used as a starting point to identify the needs, opportunities, and pre-conditions for automated translation and related language technology services.

In this deliverable, we continue analysis and specification of essential service categories for sectors identified in D6.1. We have identified what sort of setups and facilities should be foreseen to operate candidate services.

---

\(^1\) http://mli-project.eu/
MLi WP6/D6.2
1. INTRODUCTION

This report is prepared by Tilde and DCU. DCU prepared the Section 2 which analysis usage of the language technology services in Localisation and Globalization industry. Many of the observations presented in the Section 2 stem from the vantage point of the activities conducted at the ADAPT Centre, Ireland’s global centre of excellence for digital content technology, which over time has developed partnerships with several key international players in the localisation industry. With localisation at its heart, and building on the success of its predecessor CNGL (Centre for Next Generation Localisation), which was founded in 2007, ADAPT’s objective is to conduct world-class research into fundamental technologies to enable better content curation, multilingual discovery, translation and localisation, personalisation and multi-modal interaction across global markets.

Section 3 summarizes assessment of multilingual challenges and innovation work for financial sector. The example solutions that could address the end-user needs in multilingual and actionable information include (1) a web based service consisting of the real time multilingual information about a set of financial products and (2) an authoritative news and customer sentiment feed which provides insight in specific financial products and their providers, in real-time based on the wisdom of the crowd.

Section 4 provides case studies of MT services for e-commerce sector carried out for two major e-commerce providers. Both e-commerce providers see that language technology in general and machine translation in particular bring huge benefits for e-commerce.

As the public sector in Europe faces language barriers every day in Section 5 we focus on MT services for public sector. One area where language barriers to information exchange have a direct impact is legislation. Thus MT can make an effective contribution to the process of harmonisation of national legislations across EU, particularly by aligning many norms and standards of new EU democracies. Positive examples of MT usage include translation service for EU presidency in Latvia and machine translation service hugo.lv developed by Latvian government. Finally, Machine translation for Connecting Europe Facility is analysed.

Section 6 discusses MT services for cyber security and defence. We have identified and discussed a need for a technological toolkit of automated translation, textual pattern and keyword tracking, and text analytic functionality that will automate the monitoring and analysis of content from social and news media.

Section 7 provides overview of workshop organized by MLI that was devoted to possible offerings from the industry and the academic community that can support the MT engine-building factory in CEF. Major trends, solutions and developments in MT are summarized. The workshop demonstrated that there are several large-scale cloud-based MT services of industrial strength available on the market.

Finally, Section 8 provides generic specification of essential service categories – machine translation, natural language processing and speech technologies. Requirements and architecture are presented. This section also provides recommendations for data management to support public automated translation services.

The deliverable has an appendix in which the LTI Cloud evaluation report is included. The evaluation is performed for three main use scenarios.

MLi WP6/D6.2
2. LT SERVICES FOR THE LANGUAGE INDUSTRY

Digital and online content is increasingly a part of everybody’s lives, for multiple purposes, both professional and recreational: to make business decisions, for academic study, to plan holidays, to shop online and for online gaming, to name but a few examples. This diverse online content is consumed in a variety of settings, not only across multiple platforms (on laptop computers, mobile phones, tablets and other portable electronic devices, etc.), but also as part of innumerable everyday situations: while jogging, during a lunch break away from home, driving or commuting to work on public transport, etc. Individual users typically combine a number of these usage scenarios on a regular basis.

Mobile solutions for websites and online services require new, more flexible modes of localised content delivery and consumption and adds further layers of technical complexity to localisation. The online content localisation and globalisation chain must respond to an increasing number of ever more difficult challenges, to make content flexible and adaptable in order to be successfully consumed with positive user experiences in many different scenarios.

Against this background, and within the wider context of the analysis and specification of essential service categories covered in this deliverable, this section provides a high-level analysis and specification of localisation industry needs. To this end, in this section we refrain from going into the minute details and specifics of what products and services are available, but aim to offer a high-level overview of the dominant trends and key needs, to capture the essence of the dynamic evolution that is taking place in this complex and rapidly developing area, without basing our analysis on excessively specific cases. In general, we refrain from mentioning specific companies and products, but we aim to explain the services that can provide successful responses to industry needs not only at the present time, but also for the expected challenges that lie ahead in the short-to medium-term future.

2.1. Services in support of the online content localisation/globalisation chain

Individuals and organisations, including companies, are increasingly seeking to interact with content, services and other people in their own language, according to their own needs, preferences and scenarios of use, across multiple platforms. This situation is creating fundamental challenges concerning how the online content localisation and globalisation chain can dynamically create, process and deliver contents for global usage, which is also suited to users’ expectations, especially in terms of language. This corresponds to a need to enable global digital content that can be generated, discovered, managed, personalised and delivered multimodal (via text, video, image, and/or speech) more effectively, efficiently and with significantly enhanced impact – the importance of doing all this taking into account the users’ languages cut across all these activities.

The notion of the global content value chain was initially introduced as a strategy to unify the complex process of moving multilingual content from its creation stage through to final consumption by end users (Emery et al., 2011). This notion is predicated on the assumption that value can, and indeed should, be added to content as it passes from one stage of the chain to the next, by enriching it to meet the needs of the final users. While a thorough review of all the services that can support this process that is central to the success of the online content localisation and globalisation chain is beyond the scope of this deliverable, in what follows we emphasise the importance of its essential
components that we believe to be of crucial importance, regardless of other contextual variables.

One important service that is all too often dismissed or disregarded when localising online content concerns the cultural dimension of products and services, which goes beyond the language of the accompanying copy. Cultural consultancy services can ensure that the localisation process duly takes into account the socio-cultural norms and requirements of the targeted user population. This can be useful both for products and services that are localised in-house (e.g. by huge multinational corporations) and for localisation projects entrusted to external single- or multi-language vendors in outsourcing. Cultural consultancy ensures that all the elements of the localised product or service are suited to the norms and expectations of the final users, beyond the technical requirements and the language used – this encompasses elements as diverse as images and use of colour palettes in the product's packaging, typical use case scenarios described in accompanying documentation, etc.

2.1.1. Search engine optimisation

Search engine optimisation (SEO) is of increasing importance to the success of online content localisation and globalisation. SEO refers to the process of making a website or online service more visible to search engines, in addition to people and visitors. Many steps contribute to successful SEO, including e.g. careful keyword selection to differentiate oneself from the competition and emphasising unique selling points; the website’s code can also be updated and tailored to become more appealing for search engines, with a view to boosting results positioning, frequently refreshing contents in the interest of keeping them current. The number of incoming links also plays a role in search engines' ranking algorithms, as a way to measure a website’s general importance and authority. When a website or online service is localised, professional and effective SEO into one or more target languages requires dedicated actions and in-depth research on all these fronts, so that the newly created versions occupy the ideal space in the online environment of the target language, that can guarantee the highest visibility and, in case of commercial enterprises, the best exposure to prospective customers, who can be converted into paying buyers.

This applies equally to businesses and services of all types, from established brands to new entrants trying to carve a market niche for themselves, regardless of market saturation and of the competition that is already there: if potential users and customers in a new locale cannot find a product or service, they will never know about it, let alone use it, even if it has been made available in their language and is perfectly adapted from a cultural point of view. One important service supporting online content localisation and globalisation efforts accompanying effective SEO is web analytics, which consists in the measurement, collection, analysis and reporting of web data; this ongoing monitoring activity is aimed at managing and optimising the web presence of a product or service in the interest of serving (potential) customers in the best possible way. This is particularly important, for instance, when visitors to a localised e-commerce website regularly fail to finalise a purchase, or when an online service has gained visibility through effective SEO, but does not attract paying subscribers. Far from being just a mechanism measuring web traffic, web analytics is vital for business and market research, as well as to evaluate the effectiveness of a localised website or online service. Essentially, web analytics can help businesses track the localisation return-on-investment for the languages and locales that they have targeted, flagging up potential areas where performance remains below expectations, prompting corrective actions.
2.1.2. Quality assurance

In addition, the importance of rigorous testing and quality assurance of localised products and services cannot be underestimated: this involves both technical and linguistic aspects, and it may be performed in-house and/or by dedicated LSPs in outsourcing (as noted above for cultural consultancy services). Such tests should involve in-country validation or acceptance testing by local subsidiaries, distributors or customers. Troubleshooting activities in this phase benefit from cross-version coordination, to ensure that issues detected in one language version are also addressed in others, if/as relevant and appropriate. In many (but not all) cases, all of the above actions are geared towards the attainment of simship (short for 'simultaneous shipment', cf. Gaspari, 2005), which occurs when vendors release and launch their new products at the same time in several locales, possibly in all the new markets they wish to enter at the same time. This makes it clear how important and challenging the successful management of all these complex tasks is, also to make sure that they do not end up causing delays or unnecessarily inflating the budgets allocated to localisation projects, especially when multiple languages are involved – it is normally imperative to minimise time-to-market, which is the period of time elapsing between the moment in which the development of a product is technically finished, thus making it functionally ready to be marketed in a new locale, and its actual release to customers, i.e. when it is finally available for purchase through commercial channels. This is very important because of the competitive advantage it presents in gaining, or maintaining, market shares.

Once the localised products or services have been tested and finally released to the markets in question, an essential service in support of the online content localisation and globalisation chain is local product marketing, support and feedback. After the localised product has entered the marketing and support phase of the specific country or region, businesses need to develop effective customer support services, e.g. to gather usage data on errors and for after-sale customer help, along with mechanisms to report customer feedback and user wish lists to the main central development and marketing teams if/when appropriate.

2.2. Translation services for companies operating across national and linguistic boundaries

It is easy to see how the services described above in support of the online content localisation and globalisation chain are, at least potentially, relevant to a wide range of players in the industry. Clearly, depending on their size and ambitions, the budgets allocated by the diverse players to specific localisation projects will dictate to what extent they can address the above steps, and other related ones, to position their localised products and services in the markets in which they wish to operate. In this section we zoom in on an underlying requirement of all localisation projects, namely the translation needs of companies operating across national and linguistic boundaries. Similarly, to our approach above, we describe general trends and dominant situations, without delving into the transient details of product and company names.

By now there is consensus in the translation industry on the essential software tools whose use benefits translation services, in particular computer-assisted tools such as terminology management systems and translation memories; these can be used alongside localisation toolkits and platforms, global content management systems and translation quality assurance software to ensure an optimal translation service. A thorough review of these software tools and their related workflows is beyond the scope of this document.
of this deliverable, but we focus on the analysis and specification of the machine translation (MT) solutions that can serve the needs of companies operating across national and linguistic boundaries. Despite resistance from some quarters of the professional translation community, MT is increasingly popular in connection with localisation, and MT systems can be used alongside the other, by now commonplace, translation technologies mentioned above.

One growing application of MT is to translate user-generated content. From the companies’ perspective, the use of MT is helpful for sentiment analysis, e.g. to monitor comments on products and services on user forums, social media, etc., which can be helpful for managing brand reputation online. Essentially, sentiment analysis consists of the automatic identification and classification of users’ opinions on a particular topic, and it is particularly challenging to mine opinions on brands and/or products from potentially unpredictable and unstructured sources such as user forums or social media platforms. In such scenarios, there is a need to ensure that MT can cope with ill-formed content, including the features that are typical of online communication (e.g. emoticons, abbreviations, etc.), so that MT output is good enough for subsequent applications, such as meaningful sentiment analysis that can help the company take action where problems are identified, or monitor the positioning of its online reputation.

An interesting growth area for MT linked to localisation concerns cloud-based services, where, from the point of view of the end users, MT engines made available on the cloud remove the overheads caused by maintaining the required hardware infrastructure and lessen the problems of the technical expertise that is needed to develop cost-effective MT-based solutions. Cloud-based MT systems are typically tailored and customised to the needs of corporate users, on the basis of their preferred domains and text types for the language pairs of interest to them. The customisation of MT systems can be achieved by using corporate linguistic assets and resources provided by the clients themselves, but can also be supported by additional data harvested by the companies offering these MT-based services, to improve translation quality. In this way, the MT output can adopt the preferred terminology and style of the company in question, to obtain a consistent corporate voice in the targeted locale(s). Two MT companies specialised in cloud-based MT and customisation services are KantanMT and Iconic Translation Machines, which were reviewed in Section 7, respectively, of MLi’s deliverable D6.1 on “Characterisation and recommendations for pilot services”.

Since cloud-based MT solutions and engine customisation do not necessarily guarantee perfect MT quality, two related areas that are often investigated in connection with MT-based translation services for companies operating multilingually across countries are controlled language and post-editing. When single sourcing with approved vocabulary and restricted grammatical and syntactic rules is implemented, there may be benefits not only in terms of enhanced MT quality, but also simplified content for non-native speakers, that can be easier to understand (e.g. reduced ambiguity and sentence complexity in user and maintenance manuals, instructions on how to assemble a piece of furniture, etc.). With regard to MT post-editing, even in the face of resistance from some translators, LSPs are increasingly offering this service for a growing number of language pairs, and this trend seems set to continue for the foreseeable future. Even if the quality of raw MT output is far from perfect, the adoption of controlled language and the integration of post-editing into the translation workflow of companies needing translation services can help turn MT into a viable solution.
3. LANGUAGE TECHNOLOGY SERVICES FOR THE FINANCIAL SECTOR

This section summarizes assessment of multilingual challenges and innovation work for the financial sector that was carried out in EU cooperation activities such as the Eurostars SAFE project. The financial service industry is one of the largest industries in Europe, with an annual turnover of €600 billion and dealing with multiple languages. According to Gartner the financial service industry process 20,000 Exabyte and is among the highest growth sectors in the European service industry. It faces huge changes in usage of multilingual data and new European Union legislation following the financial crisis that harmonised regulation for investment services across the member states.

There are approximately 140,000 adviser firms in Europe, employing 810,000 staff in total, of whom 250,000 are advisers. 40% of investment and protection products are sold through financial advisers, with annual revenue estimated at €77.5 billion (€15.2 billion from investment business, €12.2 billion from general insurance and €50.1 billion from mortgages).

In the domain of financial services targeting consumers the balance of knowledge between the buyers and suppliers is extremely uneven. Suppliers of financial products specifically design their offers to tempt consumers who, due to the extremely complex nature of the financial markets are unable to assess the risk involved.

The main problem is the lack of balance in information between the suppliers and the consumers of financial products. Over the last year, financial products in general have become more and more complex to the extent that even experts have difficulty in recognising the risk involved in the use of specific products. Whereas suppliers (banks, insurance agencies) are well positioned to inform themselves on the intricacies of product offers, consumers often have to depend on the suppliers to provide them with the information they need and at the time it is needed. This effectively leads to an information monopoly that is undesirable for the consumer and since it undermines market trust, in the long term, also undesirable for the suppliers.

It is not difficult to find recent examples of developments in the financial world that lead to severe consumer uncertainty on a European scale: SNS bank nationalisation in The Netherlands, Central bank failure in Cyprus, bankruptcy in Greece, record unemployment in Spain etc.

Current product information systems are one-dimensional, often in English-only, and require human processing for content analysis, which is enormously cumbersome.

There is no authoritative information source which provides insight in specific financial products and their providers, in real-time, accessible to all European language communities, based on the trusted sources and wisdom of the crowd.

For many consumers, English is not their first language. In Europe alone, there are dozens of ‘smaller’ languages such as Polish, Dutch, Latvian, Swedish, and larger languages such as French, German. In order to capture product related information from different media and sources, these languages need to be instantly accessible. To gain quick access, machine translation is the most suitable technology. While machine translation is not as accurate as human translation, it can be done much quicker and on a much larger scale than human translation.

As noted above, the volume of financial information on the web is great. In order to process, filter and interpret all of this unstructured information, semantic analysis is the MLi WP6/D6.2
key technology. Semantic analysis has a distinct benefit over statistical sentiment classification as it captures the “meaning” of the messages at a much higher resolution. This technology has proven to be capable of extracting the sentiment of large amounts of data within a short time (for example in the financial sector, where news is processed for traders within milliseconds) and is a necessity in determining customer’s evaluation of new products very quickly. In addition, sentiment analysis is a powerful technique for determining the “wisdom of the crowd” by assessing the public sentiment regarding specific topics and aggregating the results from many news messages. It is particularly useful for the establishing qualitative linguistic identifiers (good / bad) concerning concrete objects (products or services) and it lends itself well to intuitive visualisations (positive, up, green vs negative, down, bad).

In order to capture the information of social media, web crawling tools are needed. These range from keyword search tools to advanced web mining technology. Web crawlers scan web sites in a systematic way to collect information. The data collected by crawlers is raw, and not only contains the information presented to a web user, but also contains information regarding the format and structure of the web site. After the data is collected by an automated crawler, it should be restructured in order to be automatically processed. This crawling is done in a language-independent way, in order to quickly determine relevant sources and subjects for machine translation and subsequent semantic analysis.

In financial services there will be a growing need for multilingual consumer driven data analytics. Real time analytics as well as in-depth analytics for trend discovery will grow according to IDC from 1,2 billion in 2011 to 6,8 billion € per annum in 2016 in Europe. According to Gartner, financial service analytics will grow to 1,8 billion per year for big data analytics. The highest grow will come from social media analytics according to IBM Watson and Cisco.

Example solutions that could address the end-user needs in multilingual and actionable information:

- a web based service consisting of the real time multilingual information about a set of financial products. This service needs to be accessible through a dedicated web site providing a clear visualization of data and as a generic plug-in for third party apps and websites and social media platforms (twitter, Facebook).
- an authoritative news and customer sentiment feed which provides insight in specific financial products and their providers, in real-time based on the wisdom of the crowd. This feed can use social media publications of both users and independent experts as expressed on e.g. twitter, Facebook and blogs (WordPress, Tumblr blogger).

The reputation risk management market is focused on investment banks and insurance companies. These companies valued the content coming from such service as consumer feedback of the findings of their financial products and how new financial product are received in the market. Customers in the target market are banks, insurance companies and fund management companies.

The investment & consulting services in the financial adviser industry is another target group for multilingual tools. Over 50% of the population rank financial advisers as one of their top three most trusted sources of advice about money matters. As such, financial advisers represent a leading force in the maintenance of a competitive and dynamic retail financial services market.
4. MT SERVICES FOR E-COMMERCE SECTOR: CASE STUDIES

To get an in-depth view in the multilingual challenges and needs of e-commerce providers, we carried out a study of two major e-commerce providers – Zalando and Etsy. As part of this study we carried out interviews with Andreas Antrup, VP Data at Zalando, and Raymond Flournoy, Localization and Translation Group Product Manager at Etsy.

This section is focused on these two case studies that in our opinion are representative for the overall e-commerce sector. See separate MLi deliverable MLi e-Commerce Report on the generalized analysis and recommendations on how language technologies can help e-commerce players to overcome the language barriers and reaching multilingual markets.

4.1. Case Study: Zalando

Zalando is Europe’s leading online fashion platform founded in 2008 and based in Berlin, Germany. The site began as an online outlook for purchasing shoes, but has now expanded to fashion retail more broadly, with over 150,000 product choices. The site is online in 15 European markets and has over 130 million visits per month, leading to 2.2 EUR billion in net sales in 2014.

The company has approximately eight thousand employees in Germany, of whom 3,500 are based at three locations in Berlin (tech, management/market teams, and fashion). All three Zalando warehouses are located in Germany, therefore all goods are shipped from a central location.

Zalando maintains separate sites (i.e. web shops) for fifteen countries: Germany, Netherlands, France, Spain, Great Britain, Austria, Switzerland, Poland, Belgium, Norway, Finland, Spain, Luxembourg, Denmark, and Sweden.

The national internet based shops feature local language and pricing, a localized assortment as well as local payment methods. The aim is to address the needs of audiences in the different countries. Because of this setup – a central headquarters and
“virtual” shops in fifteen countries – Zalando actually engages in a higher percentage of cross-border trade than the EU average.

This arrangement also comes with its own challenges. It is complex to set up logistics for fifteen countries rather than one, and to integrate different service providers. Different countries also have diverse payment landscapes: in some countries customers prefer cash, in others credit and in still others bank transfers. Zalando currently offers over 20 localized payment options for selected countries. In addition, Zalando communicates the different lead and delivery times to customers in various countries.

4.1.1. Language barriers in Zalando e-commerce activities

In order to serve its customer base in 15 countries, Zalando has created highly localized front ends. According to its 2014 Annual Report, Zalando offers “local websites tailored to domestic needs, including country specific assortment, local online catalogue sorting, and visual merchandizing and marketing campaigns.”

Zalando admits to facing language complexities in its e-commerce activities. According to Andreas Antrup, VP Data at Zalando, these complexities are divided into two categories: the “dictionary question” and the “translation question.”

To resolve the translation question, the company localizes its web activities using a team of in-house human translators. To deal more in-depth with the language issues in its fifteen markets, Zalando also employees native speakers at its central offices in Berlin. These native speakers are tasked with responding to customer requests, engaging audiences in social media, etc.

In the “dictionary question”, Zalando sees a greater challenge. Within the Zalando web shops, products are grouped into various categories and subsections. For example, “shoes” have a subsection “sandals,” which are grouped into “beach shoes,” “casual sandals,” and “flip flops.”

These terms are not just used for categorizing the products, but are also “tags” that designate certain attributes. This metadata must be consistent on different sites. This is where the dictionary question presents a challenge.

MLi WP6/D6.2
According to Antrup, the problem is two-fold. First, the right expressions must be found for each country. The challenges lie in the fact that categories differ between languages. The reasons for the differences in categories across languages is culturally specific, as well as directly linked to the purchasing habits of the buyers in each country. Products from sellers must fit into the categories in the native languages, meeting the demand of customers as well as the culturally specific categories.

The challenge for Zalando is to transfer the keyword searches for different countries – to link the keywords between languages – so that customers can more accurately search for goods. This would also help in SEO, a foundational aspect of online marketing. Linking keywords across languages would help Zalando to steer its marketing spend more efficiently, as well as to target customers more accurately.

The lack of interoperable search terms is challenging the implementation and is also preventing unified learning across countries.

According to Antrup, in e-commerce, the focus of differentiation is moving away from a wide range of products for good value to curation and personalization – toward inspiring the customers with personalized offers.

The type of personalization today is mostly very crude and simple, filtering down to what the customer has recently engaged with. This is called the “filter bubble.” In the future, personalization is going to move toward a field where it recommends and opens new experiences for the new customer.

This is only possible if the retailer has a deep understanding of (1) its merchandise and (2) the preferences of its customers. Both can only be achieved if the retailer has a very clear, granular understanding of the metadata surrounding customers with its articles.

For instance, if Zalando knows from metadata that (1) a customer is looking at a particular type of shoe in Italy, (2) the shoe has a certain kind of material such as bio leather, (3) this is currently a popular trend among urbanites and (4) this person is in Paris, then Zalando can deduce that the customer is following this trend. As a result, Zalando can display to the customer the most popular brands within this particular trend.
This example shows how metadata can lead to an educational experience about customers in a particular market. But this is only possible if the retailer has fine grained, granular understanding of the metadata surrounding the merchandise and the customer.

However, it takes a lot of efforts to get this understanding. Metadata must be collected in different language spheres. Because all of the data is different – the data is specific to the local market – retailers must have a localized approach to metadata. This is an area where Antrup sees that language technology – and automation – can help.

The solution is to have a trustworthy corpus of linked terms in various domains. This can then be used by the search engines to classify products, making for better product recommendations.

4.1.2. What could the EU do to help Zalando?

As a solution to the aforementioned problem Zalando sees a need to create a corpus of identity links for commercially relevant terms, in various domains.

By creating such a pan-European corpus, e-commerce sites would be able to utilize SEM and customer targeting more effectively. This would provide a platform for learning and steering across languages.

The short-term advantages are currently too small to justify the large investment. However, global players such as Baidu (from China) can afford the investment, thus they threaten smaller EU e-commerce sites.

If such a corpus were in place, according to Andreas Antrup, this would help to generate a “level digital playing field.” Antrup adds: “At the present time, biased advantage is given to companies coming from bigger markets – from the U.S. and China – which have the scale and the data to break through these problems. The smaller-scale EU companies don’t have this capability.

4.1.3. Possibilities for automation and Language Technology tools

At the present time, the majority of customer contact at Zalando works with highly standardized responses. In general, the field of customer correspondence shows high potential for automation.

However, Zalando deliberately decided against automation in customer service due to possible quality issues. According to Antrup, tools like MT must reach a certain level of quality before it is considered.

The decision to employ language technology tools would be made by the Tech (IT) department at Zalando.

Terminology management tools

Antrup says that Zalando would benefit greatly from a pan-European database of linked terms in the fashion domain. This would help Zalando to solve the aforementioned “dictionary question,” achieve consistency in metadata, as well as help Zalando to more accurately engage in SEM and create a more personalized shopping experience for buyers in its markets.

Machine translation

Andreas Antrup sees that MT could be used to speed up the localization process. This would include translation of product descriptions, interfaces and customer contacts.
Zalando sees two options for using MT: either building MT systems in-house, based on open-source software or outsourcing to third-party vendors.

The decision would be based on the availability of open-source software and/or the specific offer of the third-party vendor.

The languages that Zalando would be interested in are German to Dutch, French, Spanish, English, Polish, Norwegian, Finnish, Danish and Swedish.

Text analysis

According to Andreas Antrup, a major potential for LT at Zalando is text analysis. Zalando receives huge volumes of customer feedback and reviews. This offers great opportunities to learn about customers and products. Customer feedback about sizing is particularly valuable for Zalando.

However, locating the valuable data is a problem, as employees must manually scan the reams of feedback data. It is therefore difficult to integrate the valuable comments into internal reports.

Therefore, Andreas Antrup sees the potential for integrating text analysis tools into Zalando’s workflow. These would help Zalando to more accurately locate valuable comments by extracting the necessary data (e.g. sizing information) out of the text.

Andreas Antrup also sees possibility for sentiment analysis tools, to gauge customers’ opinions about the products on offer.

4.2. Case Study: Etsy

Founded in 2005, Etsy is a “double-sided marketplace” e-commerce platform, which means that Etsy provides a venue—a platform—for sellers to sell their wares to buyers. In particular, Etsy specializes in three categories: handmade, vintage, and craft supplies. The platform has 1.5M active sellers, 21.7M active buyers, and $1.93 billion gross merchandise sales in 2014.

Etsy is based in Brooklyn, New York.

4.2.1. Language barriers in Etsy e-commerce activities

The platform has localized sites for the English-speaking markets—UK, Canada, and Australia—as well as localized sites for Germany, France, Italy, Spain, and the Netherlands. A couple more sites are currently in Beta (in particular, Japanese). These are Etsy’s strategic target markets.

Although Etsy only has sites for these specific regions, where texts are localized for those regions, Etsy has buyers and sellers in every country on the planet.

This creates a tension, as Etsy can’t afford to localize its site in every single language. But they do know that people are posting in the forums in languages other than those listed. Etsy also knows that people are having “conversations” on the platform in languages other than those listed.
Supporting that user-generated content is often a different issue than Etsy’s official site and app localization. One answer is machine translation.

### 4.2.2. Machine Translation at Etsy

According to Flournoy, Etsy’s decision to implement machine translation was a “blue-sky initiative.” Originally, a project manager had proposed this as an experimental project to hook up machine translation with user listings.

That initial work was launched two years ago, November of 2013. Since then, Flournoy has worked on developing the integration as well as expanding the integration points.

For user listings—i.e., actual descriptions of items that people are selling—Etsy gives users tremendous latitude in terms of how they format and describe their items. Etsy tries to be very hands-off.
Figure 5. Item description for “O’Malley sunglasses” on Etsy.com

But Etsy has a structure for people to translate their items themselves. For example, if a seller in France speaks French, German, and English, they can post translations in all three languages. These translations are then indexed. Human translation is always the first choice.

But for people who don’t have the ability to translate themselves, Etsy provides machine translation as fall-back. Currently, if your language setting is set to English, French, or German, and you visit a listing originally published in another language, you will automatically see that translated into your language (with a link to see the original). Users can toggle back and forth between the translation and the original.

Figure 6. Screenshot of item description machine translation into French, offering users to see original (left) and screenshot of item description in English, offering users to machine translate into French (right)
For Italian and Spanish speakers, if you visit a listing—if it hasn’t been visited before—you will see the option to translate it. So it will be translated dynamically.

In short, English, French, and German content is translated by default. Italian and Spanish content is translated on demand.

In addition to user listings, Etsy also provides machine translation for reviews, comments, forum posts, and “convos”—the Etsy internal messaging system.

Etsy runs language detection. And if Etsy detects that the detected language does not match the “view language,” then a dynamic translation is offered.

The MT technology used by Etsy is Bing. Etsy is currently working on customizing its own engines with Microsoft Translator Hub. Although Etsy is currently just using the generic engines. The language detection software used at Etsy is also the Bing software.

Generally speaking, Flournoy says that Microsoft are “very easy to work with” on an enterprise level. They are very responsive and the pricing is good. This was the original impulse to work with Microsoft.

Flournoy has led a quality test to compare Bing against four other providers for one language pair. Bing came out on top. That was good data for Etsy, confirming that they did not have to rush into moving to other providers.

Microsoft also features self-training. Though in Etsy’s case, customization was not launched. However, Google does not feature self-training. So from Etsy’s point of view, this was a negative. Though Etsy has looked at other providers that have self-trainers. So this is something that Etsy considers to be a plus, and something they want to keep access to.

Why not built MT in-house? In Flournoy’s previous job, at Adobe, the developers tried to work with Moses in a fairly small team. At the time, Moses as a package was, according to Flournoy, “not ready for prime time”—a lot of aspects in terms of load balancing, the API, and the interface were either not available or not developed. Adobe didn’t have the bandwidth to create these engines.

At Etsy, building MT in-house is not something they have explored yet, as they have other priorities, such as getting the MT integration cleaned up. Though Flournoy could see that, if the self-serve options like Moses continue to develop the way they have been, it could be something they could look into. But this doesn’t come for free: you require people in-house, highly specialized people in-house. If you balance those salaries against the licensing costs, as well as the advantages of the flexibility of licensing, it’s not a clear win for building it in-house.

Etsy has two developers/engineers for machine translation. But the overall Localization and Translation Team has 6 developers, one designer, two product managers, and two project program managers who manage a team of translators (who are all outside contractors).

4.2.3. Challenges in application of machine translation

Challenge: Customization

According to Flournoy, Etsy uses a certain language that is not necessarily the default interpretation of languages on Bing.
A classic example: the word “clutch.” Generally Bing translates the word “clutch” by its automotive meaning, as in “part of a transmission.” But on Etsy, clutch is—99.98% of the time—referring to a woman’s handbag or purse. This is the kind of language manipulation that Etsy wants to do.

Etsy generally translates user-generated content. In addition, there are certain forms that people use when they post listings—the type of language they use. Therefore, according to Flournoy, doing self-training—i.e., customization—would help Etsy steer toward these structural directions. But Flournoy says that Etsy is mostly concerned about vocabulary at this point.

Etsy is still in the process of figuring out the best way to customize engines. They had a first, fairly large “first pass” at trying to build a customized engine. The path they took was as follows: Etsy has terminology lists that are maintained internally and are used for navigation and SEO. In addition, Etsy regularly collects examples of terms that are not being translated correctly. Etsy has also done some things to try to collect terms automatically that might not be translated correctly.

In the customization process, Etsy used some terms from term lists maintained for navigation and SEO; some from term lists maintained for translators in order to instruct them to do official translations; others from errors reported by users or employees over the course of time. These made up a long series of term lists. Etsy also did light analysis of trying to pull common two-word terms that weren’t already in the lists.

Etsy took these terms—which were manually curated—and hand-filtered them to a smaller list of terms (about 300). After determining that these were the terms that Etsy wanted to concentrate on, Etsy automatically pulled actual sentences from its listings—actual real live user-generated content. These sentences were then professionally translated. This then became the training corpus.

According to Flournoy, Bing does have a dictionary feature. Though his understanding is that is applied fairly brutally—i.e., it is a brute-force “search and replace.” So Etsy didn’t necessarily want that. Etsy’s hope was that, by translating actual data by hand and then using this as training data in Microsoft Translator Hub, they would get “a bit subtler integration of these new terms.”
The result, according to Flournoy, was that on the terms themselves performance improved. However, Etsy noted that the accuracy overall—on other test sets—declined. This is the issue that Etsy is currently grappling with: how to use Microsoft Translator Hub so that they can improve the quality on those specific terms, but also so that the performance of the engine is not affected too much by that fairly small training set. This is what Etsy is working on right now.

*Challenge: linked ontologies*

The terminology lists maintained by Etsy for professional translators—to help them known the “official translation” of terms like “listing” or “returns”—are linked across languages, and are maintained that way. The terms used for navigation are also maintained in all languages. This has a cross-language view. This forms an ontology, a hierarchy. But there are places where in certain languages that hierarchy doesn’t work, where the languages affect the hierarchy. So in those cases Etsy has a cross-language view.

But as far as MT goes, Etsy has been taking this challenge “language by language.”

*Challenge: Unit conversion*

There is a difference between what one could objectively call a “good translation” or a “correct translation” and what is useful for a visitor to the site. A prime example is unit conversion. This is something that Etsy has struggled with. According to Flournoy, the Bing engine is struggling with this too. In some place you might put in a sentence with the word “miles,” and the translation might have the word kilometers, but will not necessarily have done anything with the number. Or vice versa: it might translate “miles” into the word “miles” in the other language, although the speaker might not have any concept of what a mile is.

For Etsy, this is a problem. If a word like foot is translated to the word foot in the target language, you could make the argument that there is an objective correctness there. But it is still not great usability. This has been an issue for Etsy, and it is not a straightforward thing to fix. Bing and Google cannot intelligently recognize “four feet” and translate that into the equivalent in meters. This would be the translation with highest utility. This is currently not available and it is not straight-forward for someone like Etsy to implement.

*Results of MT use*

Etsy can't get into specific numbers about the results of MT integration in a business sense—i.e., has it led to increase traffic or transaction. But Flournoy does say that, when he started in commercializing MT, commerce was the area considered “off limits” for MT. There was the idea that MT could have good enough quality for certain things, like web browsing or internet chat. But commerce requires accuracy and fidelity at a level that people have to be able to trust. There was a belief that MT will never reach that level.

According to Flournoy, his experience at Etsy has shown that MT is at that level. Maybe not for all languages, but Etsy is definitely seeing that, for people viewing machine translated listings, the purchase rate is encouraging. The idea that MT is automatically not trustworthy or not useful for a commercial transaction—Flournoy believes that now this is false. He had known that anecdotally before. But he is now seeing this with general users as well.

Flournoy says: “We can look at our metrics and group visitors by whether they saw a machine translated listing, a listing that was hand-translated, or a listing that was in their
native language natively. We have this categorization into various buckets. Somewhat surprisingly, the purchase rate for machine translation is higher than I would have expected a few years ago.”

4.2.4. Further plans

Post-editing functionality
The listing process is currently an online process with form fills. Etsy’s philosophy is to try to allow the sellers maximum flexibility in how they format and describe items. So an extension of that philosophy would be to roll out a post-editing function for machine translated texts in a user’s own webshop.

Searches
Etsy uses machine translation to try to expand user queries. But MT is used to improve search results. According to Flournoy, right now that is a fairly “naïve integration.” Etsy knows from user feedback that users want this to be more accurate and dependable. Looking at results they know that they can do better.

4.3. Common outlook
Both e-commerce providers see that language technology in general and machine translation in particular bring huge benefits for e-commerce. Most importantly, you can associate an ROI with it; it generates money. That makes the investment that much easier to justify for corporations.

According to Zalando, the European market is subscale. This puts European e-commerce platforms at a disadvantage compared to competitors particularly from the US or China. There are especially high threads for smaller European companies that don’t have the necessary financial and human resources to compete against enterprises from other markets.

This challenge can be overcome by public investment – for example, in the type of infrastructure envisaged by the MLi project – or by the joined forces of private companies in the European Union.
5. MT SERVICES FOR PUBLIC SECTOR

The public sector in Europe faces language barriers every day. Language barriers impede the free flow of information between public administration institutions, inter-parliamentary cooperation, and the capacity for civil engagement, particularly for speakers of less resourced languages.

Language barriers to information exchange have a direct impact on legislation. Due to these barriers, citizens encounter obstacles to pan-European engagement in the law-making process; they are unable to debate and discuss European law-making in a communal manner. In addition, national parliaments are unable to share information about national legislation, leading to a fragmentation in the legislative process. The quality of law-making in Europe could be improved by offering the public sector MT solution as a tool for overcoming fragmentation, allowing national parliaments to reach the objectives of inter-parliamentary cooperation in the EU².

Likewise, the exchange of information across borders for public administration institutions in Europe is severely affected by language barriers. A study conducted for the European Commission has recognized “language problems” as a key obstacle that thwarts cooperation between law enforcement institutions in EU member states, highlighting that “the language barrier is one of the biggest hindrances in cross-border information exchange.”³ Recent assessments show that 66% of public administration portals provide a translation in non-national languages; however, in 12% of these, only a small portion of the pages is translated⁴. For public administration a MT solution can open up the flow of inter-institutional data as well as between the public administration and society at large.

In its efforts to engage civil society, the public sector encounters the problem of language barriers to communication. Studies show that 9 in 10 Internet users in the EU say that, when given a choice of languages, they always visited a website in their own language and only 39% of European Internet users use other languages than their mother tongue to communicate online.⁶ Nowhere is this more apparent than for smaller languages, which are underrepresented online, hindering the full engagement of civil society in EU affairs.⁵

Each of the public sector’s key roles – legislation, public administration, and civil society engagement – has a wide array of differing needs today, ranging from confidentiality and data security to public access through civil engagement platforms. To meet all of these needs, there is necessity to provide machine translation services that satisfy each of their individual requirements. Through its applicability to this broad spectrum of public sector needs, MT solution needs to be a truly wide-ranging tool for enabling the public sector

---

M Li WP6/D6.2
as a whole, helping it to crack the language barrier and promote the spread of e-democracy across Europe.

To further provide an optimal MT solution for the public sector, meeting the sector’s wide range of needs, it needs to be available in two distinct formats. First, an MT solution should be cloud-based, allowing for instantaneous translation of publicly available information and online texts. This is ideal for the needs of civil society engagement, which demands MT tools that can be widely accessed by the general public, as well as for the needs of legislation, which often requires sharing information with citizens.

Second, MT platform deployable on the local hardware infrastructure of public institution. In this format, the MT solution will work from a piece of actual computing hardware (as opposed to working as a cloud-based service), which institutions will store on their own premises, guaranteeing 100% data security and confidentiality. This deployable solution is ideal for the needs of many public administrations in Europe, which analyse and process information that is highly confidential, as well as for legislation, which often requires translation of documentation that is sensitive. The solution would be administered solely by the institution for which it was designed.

MT adaptation by the public sector presents its own challenges, such as domain adaptation and small language quality. For the public sector, MT quality requirements are often high. The public sector’s specific needs for MT solutions have not been satisfied by current MT solutions. For instance, the public sector has a need to adjust MT systems to particular language pairs and domains, e.g. specific topic, language style preferences, terminology, taxonomies, and named entities. Another important requirement is more widespread availability of MT: the public sector has a need for instant text translation, document translation, web page content translation, and MT integration in information systems and work processes. In order to meet these needs, a MT solution must be stable and highly scalable to process a high amount of translation tasks. Finally, the public sector has a need for highly secure MT solutions that could be used for translation of even classified information.

5.1. MT in the legislative process

The omnipresence of Internet and rapid development of myriads of online services have triggered transformation of society in general and the public sector in particular. Technology has opened up access to information like never before, empowering citizens to participate more fully in the political and legislative life of their country. Machine translation is a prime example of this power. MT can provide easy access to political and administrative documents and information in citizens’ native language. This can make an effective contribution to the process of harmonisation of national legislations across EU, particularly by aligning many norms and standards of new EU democracies.

To this end, MT can enable cross-border communication by providing automated translation of legislative documents, meaning that a parliament can access legislative proceedings from other countries. MT can also allow all European citizens to actively join pan-European discussions in their native languages, by proving a MT add-on or plug-in for e-democracy platforms, allowing readers to machine-translate other users’ entries.

Parliaments can use high-quality machine translation to perform the following actions:

- Translate information and different types of working documents that support the legislative drafting and legislative process at large.

MLi WP6/D6.2
- Translate background information relevant to parliamentary business (news, web, articles etc.).
- Translate incoming correspondence, e-mails in particular, to understand the general context and to define what parts need to be translated by translation service.
- Provide the possibility for non-native speakers to access information and legal documents, even draft laws and adopted laws.
- Get access to the legislative process and documents on the websites of other national parliaments.
- Translate the information on the parliament website.

For the development of civil society, it is extremely important that the work of the national parliament is democratic, transparent, and controllable. ICT provides an array of tools to follow parliamentary activities.

The MT solution can be integrated into parliament’s information system and legislative workflow process, which increases the productivity of parliamentary staff and makes members’ work more effective, providing all interested parties with necessary timely information about the content of the draft and the pace of legislative process. This system allows users to follow-up on all stages of the drafting of bills, responding to the needs of MPs and parliamentary staff, as well of the civil society. Machine translation can open multilingual access to all types of information of such information system, e.g.:

- Bills, laws, legislative acts, and related documents;
- History of each law and legislative act;
- History of Standing Committee action on laws and legislative acts;
- Inquiries from the Members of Saeima and responses to them.

All such documents are usually made available to the general public online and machine translation will make them accessible to both domestic and pan-European multilingual communities. Constituents, inhabitants, and non-governmental organisations may monitor the progress of legislation and interact with the relevant Standing Committee. As a result, such multilingual system makes the legislative and political process more transparent and accessible for everyone.

Currently, only a limited number of documents can be translated due to the budget constraints, posing the risk of making decisions based on inadequate information. By providing high-quality instant MT of legislative documents – including analytical reports, studies, laws and normative documents of other countries – MT will have a substantial impact on improving the legislative process in member states and foster the continued flourishing of democracy and the rule of law in parliaments throughout Europe.

In addition, the timeliness of access is a crucial factor for public involvement and social dialogue, therefore making information available in real-time on the Internet makes access to the parliament virtually open to everyone. Plenary agenda and agendas of Standing Committee meetings are usually published ahead of time with the links to full text drafts and amendments; online video and audio streaming are just few of the opportunities on the parliament website.

MT widget could be integrated in the parliament webpage, which usually includes the following information:

- Legislative information management system

MLi WP6/D6.2
• Interactive plenary agenda, current and future activities of the Parliament
• Websites, calendar and agenda of the Standing Committees
• Parliamentary publications and archive (minutes of the Parliament sittings and voting results)
• Information about the Members of Parliament

Particular consideration needs to be paid to analysing the legal aspects of providing machine-translated information on an official parliamentary website. Involvement of legal experts to prepare the necessary disclaimers and usage guidelines to avoid the possible legal consequences of reproducing or otherwise using imprecise translations is necessary.

The MT solution could also be used to satisfy cross-country information requests from national parliaments in Europe and other regions of the world. Due to the limited human translation capacity, members and staff from other national parliaments are not able to follow adequately the processes and legislative proceedings of the respective national parliament. Parliament staff usually receives many requests from other legislatures for information about policies and activities and particular drafts or laws, and it takes lot of work and time to prepare responses. Similar limitations and problems are faced by most of the parliaments and state institutions, especially from the “under resourced languages” group.

5.2. Translation service for the EU presidency

A vivid example of a machine translation application supporting the European Union policy process is the machine translation tool developed for the EU Presidency in Latvia. The EU Presidency Translator was a machine translation tool for the 2015 Presidency of the Council of the European Union. It was specially designed for use by international journalists and EU Member State delegates attending official events during the 2015 EU Presidency. With the translation tool, journalists and delegates could follow EU Presidency events in the Latvian media, read local websites and blogs, translate press materials and other news sources not available in English, and access government documents for general research purposes.

The EU Presidency Translator provided instant translation of texts, documents, and websites from Latvian to English and vice versa. The translation tool is available for download as a desktop and mobile app. Two freestanding translation kiosks were set up at the EU Council Presidency headquarters in the Latvian National Library.

This is first time that a machine translation tool has been used at such a high-level event to enable multilingual communication. The EU Presidency Translator provides the world’s highest quality automated translation quality for the Latvian language and is part of the Latvian e-gov machine translation service. The machine translation technology powering the EU Presidency Translator can be applied to other EU languages as well.

5.3. Machine Translation for Connecting Europe Facility

The CEF (Connecting Europe Facility) programme for building and deploying infrastructures has the broad goal of facilitating pan-European access to a range of public online services (e.g. health, justice, procurement). A specific Automated Translation platform (CEF.AT) will be set up as one of the core building blocks to support these pan-

6 hugo.lv/translate2015
MLi WP6/D6.2
European services. As mentioned in the CEF guidelines, the CEF.AT platform will comprise “machine translation engines and specialised language resources including the necessary tools and programming interfaces needed to operate the pan-European digital services in a multilingual environment”.

The CEF.AT will build on the Moses-based MT@EC service developed and operated by the Directorate-General for Translation. MT@EC is a stable and mature solution for MT needs within EU institutions which is showing steady quality improvements. The system is being used heavily and is a good starting point for CEF.AT.

Currently, all training data is collected together to build one Moses-based system per language pair. Accordingly, there will be some challenges with scaling up to more users and new domains, both in terms of infrastructure and obtaining more training data. These include the need to cover more inbound and direct machine translations requiring more and broader training data together with increased robustness.

Domain adaptation is a major challenge to support DSIs that need specific terminology, but for which little parallel data is available. The MT@EC team is interested in novel ways to increase coverage, e.g. by extracting parallel data from Web Crawls, or exploiting comparable corpora. Moving to a cloud-based infrastructure is also being explored to adapt to high capacity demands for almost instant translation.

The Directorate-General for Translation (DGT) is working together with the Directorate General for Communication Networks, Content and Technology (DG CONNECT) for the deployment of the CEF.AT platform, focusing essentially on the MT service part and on implementing an “engines factory” focusing on domain adaptation and on the automatic engine creation. In this context DGT is keen to investigate cloud-based solutions, as well as domain adaptation and customization techniques and technologies that have been developed as intermediate or end results of EU-funded projects and/or solutions that are available on the market.

5.4. MT for the public sector case study: Hugo.LV

To provide multilingual communication in the public sector, the Latvian government has developed the machine translation service Hugo.lv administered by the Culture Information Systems Centre, which operates under the Ministry of Culture of the Republic of Latvia, the system provides instant translation of texts, documents, and websites from Latvian to English (and vice versa) and from Latvian to Russian.

The service is freely available online as a public web interface, offering widespread multilingual access to information originally published in Latvian. Hugo.lv is also integrated into e-service platforms – including the official public services website Latvija.lv – so that residents, citizens, and visitors can receive e-services in multiple languages.

As a result, the Latvian public sector can now communicate in multiple languages and share crucial information about government policies, legislation, and services with audiences not only in Latvia, but also across the world. This communication is particularly important for a small language like Latvian, with only 2 million speakers worldwide.

At the present time, the machine translation service Hugo.lv is specially tailored to general-domain texts and legislative acts. This means that the service can currently be used to translate general descriptions of e-services and other e-gov texts, as well as
more specialized legislative documents such as normative acts, legal proceedings, legislation, official regulations, and laws.

In the next phase of development for Hugo.lv, the Culture Information Systems Centre plans to further adapt the Latvian public sector’s machine translation service for cultural heritage data. This would mean that the system could be used to provide instant translations of crucial cultural information like museum catalogues, library catalogues, archives, and metadata.

Latvia currently has nearly a million items in the Joint Catalogue of the National Holdings of Museums. Only a small fraction of these items – less than 10,000 – are translated into English. Therefore, adaptation of Hugo.lv to the cultural domain, and the integration of Hugo.lv into the Joint Catalogue, would allow Latvian museums to share descriptions of their joint holdings with vast audiences around the world.

The process for adapting the machine translation service for the cultural domain entails incorporating glossaries from museums, libraries, and other memory institutions; integrating specific cultural terminology; and training the service with large volumes of data from the cultural domain. The result would provide high-quality translation of cultural heritage data from Latvian to English (and vice versa) and from Latvian to Russian.
6. MT SERVICES FOR CYBER SECURITY AND DEFENCE

The cyber-space is increasingly becoming a battlefield. It is widely exploited to target, harass or in other ways influence individuals, organisations and states. One of the platforms used for these attacks is social media, which due to its ability to spread information in high volumes and at speed, can cause dangerous and devastating effects. Cyber-attacks in the form of false and aggressive messages by fake or artificially created users ("troll armies", botnets etc.) are conducted to further the aims of different actors.

Taking into account the increasing dependence on technology and on the Internet, it is important to advance the efforts to confront the wide range of cyber threats targeting defence agencies on a daily basis. The growing sophistication of cyber-attacks makes the protection of communications and information systems (CIS) an urgent task.

To identify and prevent cyber-threats it is necessary to monitor news and social media content which can be predictive of and help deter hostile, and or terroristic cyber-attacks and counter the impact of the use of the internet for terrorism. It can also help to understand the sentiments of the public and serve as an early warning for the possible threats.

There is a need for technical solution to monitor cyberspace in order to identify and evaluate potential threats specifically through the monitoring of multilingual content in social media. There is a need to implement a technological toolkit of automated translation, textual pattern and keyword tracking, and text analytic functionality that will automate the monitoring and analysis of content from social and news media. A monitoring system producing analytic output will improve capacity to monitor potential threats, and will serve as an early warning, improve situational awareness and provide the ability to formulate response based on analysis.

The basic components and use scenario is to monitor social media and news content for characteristics of manipulation across multiple information streams and languages through automated translation to English, so that the output from different sources and languages, can be pooled for broader analysis and may be used for situational awareness by defence agencies and the member states. The output will be analysed using statistical algorithms, pattern, keyword and phrase detection, and other analysis as required.

High quality automated translation of large flows of cyber information is a key technology in data mining and the analysis value chain for monitoring large quantities of information. Fast, quality automated translation can provide near real time ability to improve analysis and expedite investigations and facilitate regional diplomacy efforts, communication with counterparts, and support the formulation of informed response. The benefits of these translation capabilities will be very significantly enhanced by coupling them to social media analytics technologies that will detect rapid changes in activity on specific sites/topics, similarity of content/new postings across sites, adversarial language, claims contradicting claims on trusted sites. The solution needs to be scalable to include other agencies both domestic and foreign with similar requirements.

The long term impact is to develop and strengthen the ability to anticipate, prepare for and disrupt hostile acts in cyberspace as well as strengthening international security. The solution should be scalable for a growing number of languages, data capacity, and analytic complexity as required. The amount of data being disseminated in cyberspace
can only be captured and analysed using cutting edge big data analysis technologies to support further human analysis, decision-making, and response formulation.

The private sector is a key player in cyberspace, and technological innovations and expertise from the private sector are crucial to enable defence agencies to mount an effective cyber defence. 

7 http://www.nato.int/cps/en/natohq/topics_78170.htm
MLi WP6/D6.2
7. MT SOLUTIONS

At the request of the EC, MLi organized a workshop on September 22, 2015 in Luxembourg to present the offerings from the industry and the academic community that can support the MT engine-building factory in CEF. The workshop focused in particular on the following areas:

- Deployed solutions to domain adaptation
- Discovering domain-specific terminology/conventions in parallel, comparable, and monolingual corpora
- Hybrid combinations of statistical models with linguistic and domain-specific knowledge
- Distributed SW Architectures for scalable MT solutions
- Cloud-Based MT solutions (not necessarily statistical)
- Shrinkwrapped installable MT solutions
- Setting up/accessing cloud-based engines in industry

The workshop demonstrated that there are several large-scale cloud-based MT services of industrial strength available on the market. Iconic, KantanMT and TildeMT showed their MT platforms on the cloud that fully automate the training and running of MT engines based on the data provided by users or offered on the platform’s data repository.

Taking into account the importance and wide use of the Moses toolkit it is critical to ensure its continuous support, development and sustainability. Hieu Hoang from NYU (Abu Dhabi) presented a preliminary plan to establish the Moses Foundation, to ensure the long-term sustainability and continuous development of the toolkit.

It was emphasized that a large amount of high-quality in-domain training data is very important for MT. Several initiatives were discussed in this regard: ELRC project (presented by Andrejs Vasiļjevs), Language Data Provision from Web Crawls (Philipp Koehn), Acquisition of domain-specific MT-related resources from the Web (Prokopis Prokopidis) and Multilingual Data Value Chain for CEF.AT (David Lewis). Industry players offer their services in data collection, processing and cleaning as well as different tools such as web crawlers and tools for collecting comparable data.

The following sections details the trends and technologies presented at the workshop.

7.1. Large-scale cloud-based services

The workshop demonstrated that there are several large-scale cloud-based MT services of industrial strength available on the market. Iconic, KantanMT and TildeMT showed their MT platforms on the cloud that fully automate the training and running of MT engines based on the data provided by users or offered on the platform’s data repository.

7.2. Proliferation of Moses Toolkit

Most of the MT solutions discussed at the workshop use statistical approaches (SMT). One exception is the Apertium MT platform presented by Prompsit. Although Apertium is a rule-based system, although in different practical applications Prompsit often combines it with SMT models.

All the existing SMT solutions (Iconic, KantanMT, TildeMT, CasmaCAT, PangeaMT, Falcon and Taeyou) that were discussed are based on the Moses toolkit. The Moses toolkit is also used by MT@EC. That shows the significance and maturity of the Moses
toolkit. An alternative statistical method is under development by the Modern MT project, but this new technology also builds upon the Moses legacy.

The extensive experience and sophisticated technologies of MT companies that help to maximize the efficiency of the Moses toolkit could be very helpful to support MT@EC to address the massive requirements of CEF.AT, in order to cover the translation needs of the public sector.

7.3. Long-term sustainability of Moses

Taking into account the importance and wide use of the Moses toolkit it is critical to ensure its continuous support, development and sustainability. Hieu Hoang from NYU (Abu Dhabi) presented a preliminary plan to establish the Moses Foundation, to ensure the long-term sustainability and continuous development of the toolkit.

The aim is that the Foundation will be a non-profit organization funded by the subscription model. It will take care of the continuous development of the Moses toolkit to make it faster, smaller, easier to use, supporting different platforms and providing day-to-day maintenance. The subscriptions will help to ensure that Moses developers in the Foundation receive competitive salaries.

7.4. Combination of statistical and knowledge-based methods

The necessity of further language pairs and domain-specific components in addition to pure SMT methods was discussed. All commercial MT solutions presented in the workshop contain some additional features. The main features that were discussed included:

- linguistic components, e.g. part-of-speech taggers, stemmers, sentence breakers, syntactic parsers, multi-word entity and named-entity recognizers;
- special treatment of numbers, dates, URLs and other non-translatable tokens;
- domain-specific terminology;
- pre- and post-processing of translation data;
- automatic post-editing.

7.5. Strategies for customized MT systems

The workshop participants discussed various strategies for MT domain adaptation. The proposed methods are related to in-domain data collection (including data selection from large corpora) and MT training using this in-domain data. The other option was to tailor existing MT systems by adding additional data and terminology. The specific treatment of terminology for morphologically rich languages was also discussed.

7.6. Data preparation services

It was emphasized that a large amount of high-quality in-domain training data is very important for MT. Several initiatives were discussed in this regard:

- ELRC project (presented by Andrejs Vasiljevs)
- Language Data Provision from Web Crawls (Philipp Koehn)
- Acquisition of domain-specific MT-related resources from the Web (Prokopis Prokopidis)
- Multilingual Data Value Chain for CEF.AT (David Lewis)
Industry players offer their services in data collection, processing and cleaning as well as different tools such as web crawlers and tools for collecting comparable data.

### 7.7. MT solutions and developments

The workshop started with brief presentations setting the scene of the latest activities of the European Commission related to the workshop theme.

#### 7.7.1. CEF AT DSI

Kimmo Rossi presented the CEF programme and CEF Automated Translation digital service infrastructure (CEF.AT DSI), as well as current and future actions in their implementation. The Commission plans to scale up CEF.AT to production for "selected DSIs" (following the requirements analysis stage). Cloud capacity will be procured to run CEF.AT in/from the cloud. Preparations will be made for a facility that will automate MT domain adaptation. Language resource collection will be extended by setting up a language resource repository.

#### 7.7.2. MT@EC

Andreas Eisele presented current developments of MT@EC and new requirements to scale it up for CEF AT.

MT@EC has developed a stable and mature solution for MT needs within EU institutions which is showing steady quality improvements. The system is being used heavily and is a good starting point for CEF.AT.

Currently, all training data is collected together to build one Moses-based system per language pair. Accordingly, there will be some challenges with scaling up to more users and new domains, both in terms of infrastructure and obtaining more training data. These include the need to cover more inbound and direct machine translations requiring more and broader training data together with increased robustness. Domain adaptation is a major challenge to support DSIs that need specific terminology, but for which little parallel data is available. The MT@EC team is interested in novel ways to increase coverage, e.g. by extracting parallel data from Web Crawler, or exploiting comparable corpora. Moving to a cloud-based infrastructure is also being explored to adapt to high capacity demands for almost instant translation.

The MT@EC team is curious to learn from experts in the MT community about solutions they may suggest.

#### 7.7.3. ELRC

Andrejs Vasiljevs presented the European Language Resource Coordination (ELRC) project which is part of the CEF.AT implementation process. ELRC aims to establish a permanent Language Resource Coordination mechanism that will feed the CEF Automated Translation DSI with relevant LRs in all official languages of the EU and CEF Associated countries, in order to improve the quality, coverage and performance of MT systems and solutions in the context of current and future CEF digital services.

ELRC objectives are to improve the availability and simplification of access to LRs, with a particular focus on those LRs held by public institutions across Europe. ELRC acts as an observatory for LRs as it identifies, catalogues and documents LRs that can be used for MT. ELRC helps to raise awareness among public administration bodies regarding the value of LRs, and to clarify legal and technical issues related to LRs.
7.7.4. Moses Foundation

Hieu Hoang from NYU (Abu Dhabi) emphasized the broad adoption of the Moses Toolkit. It is widely used by numerous MT service providers (e.g. KantanMT, Tilde, Iconic, Capita, Systran, Safaba, Lucy Software, CrossLang, and many others) and EU research and innovation projects (Falcon, HimL, ModernMT, TraMOOC, QT21, Cracker).

He presented the plan to establish the Moses Foundation, to ensure long-term sustainability and continuous development of the toolkit. It will be a non-profit UK-based Community-Interest Company, which will take care of coordinating the further development of Moses, preparing new releases and supporting its subscribers who will regularly pay a subscription fee.

7.7.5. CASMACAT Home Edition

Philipp Koehn presented the Computer-Aided Translation Workbench CASMACAT, which is an outcome from the FP7 project with the same name.

CASMACAT Home Edition is provided as a preinstalled virtual machine for use by translators on their home computers. CASMACAT Home Edition includes all the software integrated into a workbench. It also integrates the Moses and Thot training pipelines, which allows the MT systems to be trained on the translators’ own data. This solution is targeted to users who do not want to provide their training data to service providers due to confidentiality or privacy concerns.

7.7.6. Language Data Provision from Web Crawls

Philipp Koehn proposed to use massive web crawling initiatives such as Common Crawl to collect additional very large corpora to train SMT models of the size of Google Translate and Bing Translator. This could be especially useful for collecting monolingual data for creating very large language models, which has shown to be effective in SMT.

In order to extract parallel data, the University of Edinburgh has developed state-of-the-art methods for the alignment of parallel documents that outperform traditional methods. Current work is focused on an advanced data cleaning tool that identifies different types of noise, including machine-translated text.

7.7.7. Acquisition of domain-specific MT-related resources from the Web

Prokopidis presented the ILSP Focused Crawler that can perform both monolingual and bilingual crawling to collect MT training data from the Web. The required input from the user consists of a list of seed URLs pointing to relevant web pages and a list of terms that describe a topic.

ILSP-FC integrates modules for text normalization, language identification, document clean-up, text classification and identification of bitexts (documents that are translations of each other). If the user does not provide a list of terms, the software can be used as a general crawler. The crawler is a modular open-source tool available from: http://nlp.ilsp.gr/redmine/projects/ilsp-fc/.

7.7.8. Modern MT

Ulrich Germann from the University of Edinburgh presented ongoing work from the EU-funded H2020 project MMT (www.modernmt.eu).

The main ambition of the MMT project is the creation of an MT infrastructure that, once fuelled with enough data, works without any further training iterations and translates MLi WP6/D6.2
sentences in a domain-sensitive way, without any need for prior adaptation nor definition of the domain. The MMT infrastructure eliminates model training and domain adaptation altogether as separate steps in setting up a system, and exploits model scores that are computed on-the-fly from massively distributed database look-ups, and biased by the input context.

The internal release of Version 1.0 (confidential) is scheduled for November 2015 and will feature massively reduced build time even for very large systems (hours, not days), fast domain adaptation (ca. 1 min. overhead per document), fast integration of new data (e.g. from the CAT data cycle), and a stable MT backbone (Moses for now). A full open-source release is planned for November 2017.

7.7.9. Multilingual Data Value Chain for CEF.AT: Interoperability Plan

Dave Lewis from the ADAPT Centre presented some of the key results of the EU FP7 project FALCON (falcon-project.eu). This project integrates incremental MT engine training with an integrated chain of best-of-breed, commercial cloud-based localisation tools consisting of website translation from EasyLing, Translation Management from XTM International and Terminology Management from Interverbum Technology. The FALCON solution uses both the Moses and cdec frameworks in a novel way to massively reduce incremental MT training time. It is integrated fully into the XTM CAT tool, and has recently been tested in-house by the DGT.

It also integrates Automated Term Extraction (ATE), forced decoding by the MT engine of validated term translation, leverage of external lexical-conceptual resources in the form of BabelNet, active curation and in-project capture of parallel text and its annotation with terms using linked data.

7.7.10. Tauyou

Diego Bartolomé presented the tauyou platform, providing in-domain adaptation mechanisms, especially where little data exists. Tauyou has substantial experience in building automatic post-editing rules to enhance MT output (hybrid combinations of statistical models with linguistic knowledge). Tauyou can run on the cloud, sometimes making use of the rule-based platform Apertium.

Tauyou provides integration with the CAT tools XTM, Memsource, SDL Trados, memoQ, and Wordbee. More than 700 million post-edited words were translated on the platform in 2014.

7.7.11. Prompsit

Mikel Forcada and Gema Ramírez-Sánchez presented the activities of Prompsit Language Engineering in the area of language technologies. The company was created in 2006 as a spin-off from the Universitat d’Alacant, responding to the commercial demand generated by the Apertium free/open-source MT platform.

Apertium-based MT systems offer excellent translation quality for related language pairs (e.g., Spanish-Portuguese), are easily installed (e.g. Apertium-Caffeine or the Apertium app for Android) and may be easily customized. In particular, Apertium offers excellent format management for plain text, HTML, ODF, and OOXML (.docx), LATEX, QuarkXPress Tags and custom XML formats, featuring synthetic sentence-boundary insertion (for example, for section headings) to increase the quality of the MT output.

MLi WP6/D6.2
Prompsit is currently developing Spidextor, a hybrid platform that combines two open-source components: Spiderling – a monolingual text crawler, and Bitextor – a bilingual text crawler. Prompsit also develops Reverso Context (context.reverso.net/translation) – a bilingual concordancer or translation spotter.

Prompsit AltLang (www.altlang.net/home) is a rule-based approach to same-language translation, e.g. to convert texts from American English to British English, or from Brazilian Portuguese to European Portuguese. This may be used to make new data available for training and domain adaptation, but also to adapt input text to a variety which conforms to a specific MT engine.

Prompsit LAUG-Moses is a complete MT solution consisting of a pipeline around the Moses engine which uses linguistic components from Apertium (morphological analysis, POS tagging, multiword lexical unit and named-entity detection and tagging, for pre- and post-processing) to improve capitalization, treatment of numbers, dates, and other fixed expressions.

7.7.12. Iconic Translation Machines

John Tinsley presented Iconic Translation Machines – a spin-out company of Dublin City University based on the legacy of the EU-funded PLuTO project on the development of MT for the patent domain.

Iconic Ensemble Architecture for MT allows users to lower the barrier for entry to adopting MT by making high-quality solutions instantly available and easy to use, primarily through a process of customisation and domain adaptation.

For domain-adapted MT systems, Iconic has developed a set of processes to adapt the training process for the domains and content being translated. For each Iconic domain, a unique training process is constructed. Techniques include data selection from out-of-domain data using the Modified Moore-Lewis algorithm with Vocabulary Saturation filtering and development set selection using perplexity and MML, among other approaches.

Information Retrieval ranking metrics are used to select candidate terms from parallel, comparable or monolingual corpora.

Iconic’s Ensemble Architecture is designed to facilitate hybrid combinations of statistical models with linguistic and domain-specific knowledge. Techniques include incorporating syntactic knowledge and syntactic parsing in the Hierarchical Phrase-Based SMT system, pre-ordering techniques (e.g. for English-Japanese) and combining SMT techniques with translating domain terms using translation templates.

Iconic has developed and deployed a scalable and secure cloud-based architecture with APIs. Iconic has also developed virtual machine-based installable solutions and has installed MT solutions at client sites.

The primary mode of interaction for Iconic customers is through access to cloud-based engines. Iconic has developed a diverse range of API access methods and integrations for common Translation Management Systems (TMS) as well as CAT tools. Iconic has also developed a web interface for clients with limited technical infrastructure and expertise in order to complete batch translations.
7.7.13. PangeaMT

Manuel Herranz presented the PangeaMT platform, a Moses-based hybrid MT platform that provides complete automation of engine creation based on training data domain selection, tuning automation and automation of data cleaning.

PangeaMT has a DIY interface for easy training of MT engines. It includes facilities for the management of companies, users, training data and several formats, engines and translations. It supports temporary users with a limit on the amount of words they can translate, and also restricts particular computers/IP domains.

Engines can be used through a web interface or through a REST-based API. The API can also be accessed from browsers by using Cross Origin Resource Sharing (CORS) with AJAX or by using JSON-P if the API needs to be accessed from a browser that does not support CORS.

Plugins are provided for major commercial CAT tools (making use of the API): MemoQ, Trados Studio and older versions of Trados and MemSource.

“External” engines can be integrated into PangeaMT, e.g. engines trained with other systems (Thot, Apertium, rule-based systems, or APIs which call other systems like Google Translate or Bing Translator).

7.7.14. TildeMT / LetsMT

Andrejs Vasiljevs and Raivis Skadiņš presented TildeMT – a scalable cloud-based “factory” for SMT. It provides automated training of SMT systems from specified collections of data. TildeMT integrates a repository of parallel, monolingual and terminology data with 3.7 billion parallel segments and 7.9 billion monolingual segments. It provides fully automated services for data collection, MT generation, customization and running of a variety of user-tailored MT systems. TildeMT is based on the Moses toolkit and LetsMT technologies, resulting from cooperation with the Universities of Edinburgh, Copenhagen, Uppsala and Zagreb in the ICT-PSP project.

The platform can be hosted on the cloud or on a local infrastructure. About 800 SMT systems are trained on TildeMT. The platform serves 2.5 million translation requests per month.

Domain adaptation and customization is provided by a combination of generic and domain-specific data models, integration of language-specific modules, terminology integration and specific processing of formatting tags. Language-specific modules include tokenization and sentence breaking, part-of-speech tagging, stemming, inflection-aware dictionaries, terminology glossaries and named entity glossaries, and special treatment of dates, numbers, IDs, placeholders, etc.

To better serve complex languages, it uses factored models and morphology-aware language models, as well as reordering limits to prevent broken dependencies.

Terminology integration is performed through pre-processing of the content prior to decoding, identifying terms and finding translations in a term dictionary. For morphologically rich languages, term surface forms are acquired through surface form generation or corpus look-up techniques. Term translation candidates are ranked based on language transfer models or simple frequency analysis.
7.7.15. KantanMT.com

Tony O’Dowd presented KantanMT.com, a SaaS-based statistical MT platform that enables the KantanMT community to develop and manage customised MT engines in the cloud in over 750 language combinations, integrating into localization workflows and web applications.

With over 480 servers delivering MT services, KantanMT.com serves various international business sectors, including Localisation Service Providers, Automotive, Governmental, Medical, Finance, Banking and eCommerce sectors.

The KantanMT.com platform uses distributed software architecture on the cloud. It uses a Moses-based hybrid SMT approach integrating MT, TM and rules. It also includes custom Named Entity Recognition technology.

KantanMT.com also includes automatic quality evaluation and analysis and provides predictive quality estimation in an SMT environment. A REST-based API is provided for all KantanMT services.

The KantanMT.com platform hosts over 7,600 SMT engines, stores over 120 billion words of bilingual training data and has processed over 1.7 billion words of translation in 2015.

KantanMT.com uses 680 Moses Servers, 40 primary Web Servers and 9 Sentinel Servers deployed on 3 Data Centres (Level 3 SFT), achieving average request latency time of 1,500ms.

7.8. Conclusions from the Workshop

The majority of MT solutions used in production are based on European-based MT platforms. It is clear that Europe can compete with US tools such as Bing Translator and Google Translate. All of the systems presented at the Workshop are cutting-edge state-of-the-art solutions to real problems of interest to both industry and academia; they work now, they are robust, and solve hard problems in real time for a range of clients across many industry sectors.

The Workshop demonstrated that both of the main issues underpinning the planning process of CEF.AT – access to a range of pan-European services via MT, and porting the MT@EC engines to the cloud – have real prospects of being solved for a range of language pairs across several vertical sectors, which will significantly benefit all of us as citizens of Europe.
8. GENERIC SPECIFICATION OF ESSENTIAL SERVICE CATEGORIES

8.1. Machine Translation

In this section we describe a possible reference solution and technical architecture for a public sector machine translation platform. It is inspired by the technical architecture of Latvian e-government machine translation infrastructure. It is based on Moses technologies for the demands of dynamically scalable real-time applications. The solution supports multiple input formats, maintain tag and formatting integrity for translating documents. Integration in services is enabled through API, translation widgets, and browser add-ons.

8.1.1. Requirements

The requirements specified in this section are based both on user needs analysis in the MLi Work Package 6.

The requirements include training the system on specific terms for legislative acts, public services, and e-government content; and adapting the system to the functional needs of the public sector, such as integration into e-government systems and services.

Given that the public sector covers a wide range of different tasks and shares a myriad of responsibilities, the MT system had to be capable of translating texts in a variety of domains: from public service information and enterprise registry data to museum databases and library catalogues. This would ensure that the MT system could be used by various institutions.

Public sector MT service should support different use scenarios, including:

- widget in government websites, so web administrators can machine translate content in CMS and then perform light in-place post-editing, and visitors can machine translate Latvian-language content into Russian or English;
- translation API, so that ministry programmers can choose where and how to include machine translation function in government sites;
- website translator, so that government employees can translate websites using systems tailor-made for public sector content (e.g., legislation).

The most substantial use scenario is an online public MT interface – a website accessible from any computer connected to the Internet. On the site, visitors could machine translate phrases and sentences. They would also have the option to upload and translate entire documents in several formats: DOCX, PDF, or PPT. The system would be open to any user. But it would have particular benefit for employees in the public sector, who can use the interface to translate documents, e-mails, PowerPoint presentations, or a range of other texts.

Due to the specific needs of the public sector, however, all information entered on the site would be translated securely, guaranteeing the confidentiality of sensitive data. The MT system would be hosted in an on premise appliance at a public sector institution, further reinforcing the security of data. This means that the interface can also be used to translate highly confidential public sector documents, such as diplomatic communiques, confidential Defence sector information, draft legislation, etc.

MT adaptation by the public sector presents its own challenges, however, such as domain adaptation and small language quality. For the public sector, MT quality requirements are often high. The public sector’s specific needs for MT solutions have not
been satisfied by current MT solutions. For instance, the public sector has a need to adjust MT systems to particular language pairs and domains, e.g. specific topic, language style preferences, terminology, taxonomies, and named entities.

Another important requirement is more widespread availability of MT: the public sector has a need for instant text translation, document translation, web page content translation, and MT integration in information systems and work processes.

In order to meet these needs, a MT solution must be stable and highly scalable to process a high amount of translation tasks. Finally, the public sector has a need for highly secure MT solutions that could be used for translation of even classified information.

8.1.2. Reference architecture

In this section we propose reference architecture for SMT training and decoding based on the open source toolkit Moses. Moses includes the essential components needed to pre-process data and to train language and translation models. SMT training is automated using the Moses experiment management system.

Although this proposal may serve as a guidance in specifying MT solutions, by no means does it excludes other technical approaches based on Moses or other open or proprietary technologies.

The proposed solution has a multitier architecture (Figure 8). It has (i) an interface layer implementing the user interface and APIs with external systems; (ii) an application logic layer for the system logic, (iii) a data storage layer consisting of file and database storage and (iv) a high performance computing (HPC) cluster. The MT system should perform various time and resource consuming tasks; these tasks are defined by the application logic and the data storage and are sent to the HPC cluster for execution.

The interface layer provides interfaces between the system and external users - both human users and machine users like websites and third party solutions. Human users can access the system through web browsers by using web page interface. External systems such as web browser plug-ins can access MT services through a public API.

The public API is available through both REST/JSON and SOAP protocol web services. An HTTPS protocol is used to ensure secure user authentication and secure data transfer.

The application logic layer contains a set of modules responsible for the main functionality or logic of the system. It receives queries and commands from the interface layer and prepares answers or performs tasks using the data storage and the HPC cluster. This layer contains several modules such as the Resource Repository Adapter, the User Manager, the SMT Training Manager etc. The interface layer accesses the application logic layer through both REST/JSON and SOAP protocol web services. The same protocols are used for communication between modules in the application logic layer.

The data is stored in one central Resource Repository (RR). As training data may change (for example, grow), the RR is based on a version-controlled file system (currently we use SVN as the backend system). A key-value store is used to keep metadata and statistics about training data and trained SMT systems. Modules from the application logic layer and HPC cluster access RR through a REST-based web service interface.

A High Performance Computing Cluster is used to execute many different computationally heavy data processing tasks – SMT training and running, corpora
processing and converting etc. Modules from the application logic and data storage layers create jobs and send them to HPC cluster to execute. HPC cluster is responsible for accepting, scheduling, dispatching, and managing the remote and distributed execution of large numbers of standalone, parallel or interactive jobs. It also manages and schedules the allocation of distributed resources such as processors, memory and disk space. The HPC cluster can be based on the Oracle Grid Engine (SGE) or similar technologies.

For adaptability to different backbone infrastructures, it is advisable to support heterogeneous hardware infrastructure. Both services for Linux platform (e.g. Moses tools) and Microsoft Windows platform (e.g. Web server or application logic services) may co-exist in this architecture.

The system hardware architecture is designed to be highly scalable. The system contains several machines with both continuous and on-demand availability. Continuous availability machines are used to run the core frontend and backend services, and HPC grid master that guarantee stable system functioning. On-demand availability machines are used (i) to scale up the system by adding more computing power to training, translation and data import services (HPC cluster nodes), (ii) to increase performance of frontend and backend server instances.
To ensure secure and confidential processing of data solution can be hosted on the local infrastructures of the client institutions.

Many European languages are highly inflected posing a major challenge for statistical MT. An efficient approach to address this challenge is application of factored translation technique (Koehn and Hoang, 2007; Bojar, 2009) devised for morphologically rich languages. Standard SMT systems treat each word form as an individual unit, regarding different inflected forms of the same word as completely unrelated tokens, leading to high data sparseness for morphologically rich languages. Factored models try to alleviate this issue by analysing the data with part-of-speech taggers and annotating each word form with its lemma and part-of-speech tag. This additional information is then provided to the SMT system, allowing it to draw better generalizations over the data. Such approaches may be used both for the translation model and for the language model, as appropriate.
8.2. Natural language processing

Humans express natural language differently in written and verbal forms. Language technology tools can be classified and grouped into several groups. Grouping can be done based on data type being processed by language tools as well as functionality needed by the use case. META-NET White Paper Series group language tools into six groups: speech recognition, speech synthesis, grammar analysis, semantic analysis, text generation and machine translation.

A natural language processing task usually is split into smaller sequential tasks reflecting different linguistic levels. For example, to in order to syntactically analyse a text it must first be tokenized, then each word analysed morphologically, and finally the text must be disambiguated in order to get a correct sense of every word. To ensure this workflow usually tools are chained sequentially and often input for one tool is the output of the previous tool.

A unified platform for language processing tools faces challenges that must be resolved or improve on the data processing workflow. A big challenge is various data formats. Language tools developed in different countries and for different languages have not been controlled by any regulations and not have not been developed based on any standard (also because some language tools were created earlier than any standard). So this Platform for language tools must:

- Differentiate output formats that are based on the selected tool and the selected language must convert into an intermediate format or broker format. Thus the platform must provide the possibility to attach data conversion between any tool format and the broker format.
- Provide a possibility for the client to submit data and retrieve results in their own format (e.g., in most common data formats based on client’s technological platform). This will force faster (thus also with lower costs) integration of platform tools into the client’s data processing workflow.

Different language tools often have been developed for different technological platforms, with different dependencies and different management of system resources. So the platform architecture must be flexible on supporting different tools. The solution must not be limited to certain platform, e.g. include only tools developed for JAVA platform.

For one certain functional task (e.g. text tokenization) there can be numerous tools, each with support of the set of a certain language. Each tool can be written in a different programming language, support different input/output formats. But from the clients’ perspective it should be a single tool with a single API, one unified language processing service with the support of certain input/output data formats, with list of language support and some other parameters. Such tool grouping also improves monetization and API use control possibilities.
The Platform for natural language tools should provide a public web page and a machine-enabled API using the latest web trends and technologies to reach more potential clients. Clients have different environments, different knowledge and competence levels therefore natural language tools must be presented as simply and friendly as possible. Following components are recommended to satisfy different clients:

- A catalogue of available natural language services with search, filtering and listing capabilities;
- Details about each Service:
  - List of real tools, their vendors and links to the related documentation and algorithms
  - List of supported data input/output formats
  - Clean and easy to understand pricing model
- Trial or demo web form:
  - The potential client can fill the form with some values and see the result (IT specialist, researcher, business manager)
  - The client can fill the form with some values and get request and response examples to help integrating the language tool into its environment (developer, researcher, data specialist)
  - The client can automatically generate code snippets for the most common programming languages (for developers)
- Download of code libraries for the most common environments to enable the use of natural language tools for non-developers.

MLi WP6/D6.2
- Extended documentation of API and tools.
- API methods

The following workflow should be executed for every Language technology service API request:

- Authentication. Each client must be able to authenticate both with login/password based approach (suggested also 3rd party authenticators like Google\textsuperscript{8}, Twitter\textsuperscript{9}, Facebook Login\textsuperscript{10}, and LinkedIn\textsuperscript{11}) as well as client-API related authentication like Basic authentication\textsuperscript{12} or other token/key based system.

---

\textsuperscript{8} https://accounts.google.com/
\textsuperscript{9} https://dev.twitter.com/web/sign-in
\textsuperscript{10} https://developers.facebook.com/docs/facebook-login
\textsuperscript{11} https://developer.linkedin.com/docs/signin-with-linkedin
\textsuperscript{12} http://www.ietf.org/rfc/rfc2617.txt
MLi WP6/D6.2
• Service and data format routing. The system should choose relevant real language tool based on the client’s selected language and other parameters. Also must detect submitted content format, select corresponding data conversion component to covert data directly or via broker format to tool’s supported format.
• Authorization. At this stage must ensure if the client has access to the requested service depending on tools pricing model, clients selected data plan, already used amount in current period as well must check if not exceeding client’s set costs constraint.
• Language tool exploitation. Invoke real tool implementation by providing data in its supported format. Every language tool should be treated as a “black box” solution to enable simpler extending the list of available tools.
• Service use count. Statistics registration – used amount, request registration, used later for analytics and invoice generation.
• Serving the result. Result must convert to client requested data format directly or via broker format from tool’s supported format.

New language tools are developed all the time, some are improved and regularly must install updates therefore Platform for natural language tools must provide reasonable approach of adding, updating and removing the tools.

Also scalability of the Platform itself and every tool is very important. As every tool can be differently developed and with very deferent quality levels, scalability strategy could be considered for each tool differently.

The above described approach enables the client to create a chain of language services to reach a certain goal. Using a central broker data format allows seamless integration of previously incompatible tools into one workflow. The platform should also provide that workflow generation and execution on a server thus opening doors to language processing tools for a wider range of potential clients.

8.3. Speech technologies

The Speech services platform should be a multilingual cloud-based service for automatic speech recognition (ASR) and text to speech synthesis (TTS). It should provide both end-user and API interfaces. The end-user interfaces should allow users to use speech technology in the web page, dictation software and mobile applications. The API should allow developers to integrate speech technology in web pages, mobile apps, and other information systems.

There should be several speech applications for end-users. Users should be able to dictate text in office applications or to upload audio file and get back its transcription, and vice versa, they should be able to enter text and get it pronounced. Another application for end-users would be voice enabled text input methods (for SMS and any other app that allows to enter text) in mobile phones allowing users to dictate text they enter and hear in voice messages they receive.

Typical examples of solutions that should use the Speech services though API are web pages, audio and video indexing solutions, virtual assistants, chat bots, voice controlled devices (including cars and devices that operate in cars), dictation systems for doctors, journalists and other professionals, audio books, automatic subtitling etc.

In addition to the general purpose ASR, the platform should allow users to customize speech recognition and synthesis for different needs. It should be possible to set up and
adjust speech technologies for special text types, for example, medical texts, special keyword recognition, address or name recognition etc.

The Speech Services should address all typical issues every developer faces targeting European market (or smaller regional markets) with the software products that uses speech technologies. Questions they ask are: Is there speech recogniser and synthesiser for particular language? Can I integrate them in my software? Is there API? Can I adapt it to my specific needs? Can I get also tools for other languages? Can I get all languages from the same provider with the same API? The proposed infrastructure should advance the speech technology so that the answer to all these questions should be – YES.

![Speech Services Architecture](image)

**Figure 11. Speech Services architecture**

Having Speech Services platform in place, various products and services based on it can be enabled.

Examples of end-user (home users, individuals, professionals) products and services:

- **Mobile Apps** for voice enabled text input. Users should use the app to input text in any mobile app, just like they can dictate English text using the microphone icon on default keyboard in mobile devices. But this app should support Lithuanian, Latvian and other languages relevant to the market. It is planned that this app should be free, and it should serve as the technology demonstrator to build user awareness of the technology and to promote other products and services.

- **Audio file transcription service.** Users should upload their audio recordings (lectures, interviews etc.) to this webpage and should get transcriptions in various formats (texts, documents, audio synchronised texts, subtitles etc.), which they can edit online and download later. There should be limited amount of audio that can be transcribed for free to demonstrate the service, the authorized users should be able to use also engines that are customized for them. Users should be charged depending on amount of audio data they are transcribing. For users who would like to process huge amounts of audio files, there should be an option to purchase the extra processing power, to process the files in reasonable time.
• Dictation system for PC. Users should use the system to dictate texts in a separate Microsoft Windows application. There should be a free version of it with limited amount of text that can be dictated to demonstrate the technology and make user awareness. Users who should exceed the free usage limit should be encouraged to use the professional dictation system.

• Professional dictation system for PC. Users should use the system to dictate texts (documents, e-mails etc.). The dictation system should be integrated with Microsoft Office. The authorized users should be able to use also engines that are customized for them.

The platform should provide TTS and ASR APIs that allow developers to integrate both speech technologies in their information systems. The API should support all EU official languages and should allow to use both general TTS/ASR engines and customized engines.

The customized TTS and ASR engines should be available for system developers through API calls and in applications for professional users. The customization may also include development for specific API, specific integration solutions or user applications. TTS and ASR customization options should depend on language.

8.4. Open Data Management for Public Automated Translation Services

Language data plays an essential role in European Multilingual infrastructure. LT community representatives13 are currently working under W3C umbrella on recommendations for data management to support public automated translation services. In this section we cite the key parts of the current results of this work that is still in progress14.

The recommendations are based on interoperability experience and knowhow from various projects and initiatives: the LT-Web project; the MONNET project; the FALCON project; the LIDER project; the QTLaunchpad project and the interoperability research at CNGL.

The aim of the recommendations is to inform the development of public automated translation services about requirements for achieving interoperability in data management, which is a key aspect of realizing such services. The document should provide answers to questions like:

• What type of interoperability requirements are relevant for public automated translation services?
• What goals can be achieved by following these requirements?
• How does a reference model for data management in automated translation services look like? To this end, the document adopts a data management lifecycle approach to structuring the requirements of different stakeholders.

More broadly these recommendations also aim at formulating data interoperability requirements for automated translation services in general: the document should be relevant for all such services.

13 Dave Lewis (CNGL at Trinity College Dublin), Felix Sasaki (DFKI / W3C Fellow), Asun Gomez-Perez( Universidad Politécnica de Madrid, Serge Gladkoff (Logrus / GALA CRISP)
14 https://www.w3.org/International/its/wiki/Open_Data_Management_for_Public_Automated_Translation_Services
The recommendation is based on the following model data management for automated translation. The model assumes three main groups of data to be taken into account:

- The content being translated with assistance of automated translation.
- The bitext used in training automated translation components.
- The lexical-conceptual resources associated with the source and target that assist in the consistent use and translation of terms.

The Reference Model identifies some high-level activities which are judged important to the data management related to automated translation, and which should therefore be subject to data interoperability recommendations. This set of activities are not intended to be fully comprehensive, as many variations of these process chains will potentially use automated translation, however they are sufficient to highlighting the main interoperability issues to be addressed.

These requirement recommendations focus on the data management lifecycles identified in the reference model. They should be complemented by additional interoperability recommendation on the services required to implement these activities.

Requirement are marked as mandatory (M) or optional (O).

8.4.1. General Data Management Requirements

1. (M) Public automated translation services should use open standards for representing content, metadata and annotations. “Open” means that the standards are available on a royalty-free basis, to maximize re-use and adoption.
2. (M) All data used and content processed by public automated translation services should be annotated with license information in a machine readable format.
3. Where licensing terms permit, content and data should be published on the web and referenceable via a unique URI.
4. (O) To enable content and data to be used as a resource in the generation of automated translation engines they should be:
- Annotated with common machine readable meta-data to allow them to be automatically indexed and discovered. This meta-data should conform to a profile of the DCAT meta-data.
- Provided with a persistent URL.
5. (M) It should be possible for third parties (general public, individual experts and language service providers alike, as well as automated language services) to submit error, QA or corrective annotations to published data, provided it is presented in a common format, with metadata conformant to one of the commonly accepted and documented universal error typologies, and/or appropriate quality metrics.
6. (O) The status of submitted error, QA or corrective annotation, as it is considered for integration by the curator of the data set, should be made available with reference to the original submissions and in a common format.

8.4.2. **Bitext Data Management Requirements**

1. (O) If content that is translated with assistance of public automated translation services is published on the web together with the source content, then the corresponding bitext aligned at a segment level should also be published. A different license to the source and target language documents may be used.
2. (O) Bitext data may be discovered via a web-based API that can return specific segment bitext data selected via query parameters. Parameters may include:
   - Source and target languages.
   - Presence of terms or phrases.
   - Translation provenance meta-data including: identification of the automated translation component used and its operational parameters; characteristics of post-editors; characteristics of the post-edits (edit type, edit distance, time to postedit); characteristics of the QA (quality assurance) method applied (parameters and assessment guidelines); QA annotations; and annotation links to specific lexical-conceptual resources.

8.4.3. **Lexical-Conceptual Data Management Requirements**

1. (O) If content is translated with assistance of the public automated translation in combination with lexical-conceptual data and is published on the web together with the source content, then the source and/or target content should also be available in a form that annotates the relevant terms or multi-word units with the lexical-conceptual concepts used.
2. (O) Lexical-conceptual data may be discovered via a web-based API that can return lexical and conceptual data selected via query parameters. Parameters may include:
   - Terms or multi-word units in source or target languages.
   - Contextual information for the terms or multi-word units
   - Conceptual restrictions upon which to filter results
   - Lexical restrictions upon which to filter results
   - Provenance meta-data of the lexical-conceptual data sought, including: source of the data; process by which the data was created; current status of data still under curation.
9. CONCLUSIONS

This deliverable concentrated on deep analysis and specification of key services that are important for the language industry, enterprise sector, public sector (eGovernment), security and Digital Single Market, and facilitates crossing the language barriers. The analysis and specification of essential service categories demonstrated, that the language technologies, especially automatic translation services, are important and requested by different sectors.

The exchange of information across borders for public administration institutions in Europe is severely affected by language barriers. For public sector MT solution needs to be a truly wide-ranging tool for enabling the public sector as a whole, helping it to crack the language barrier and promote the spread of e-democracy across Europe. To provide an optimal MT solution for the public sector, it needs to be available in two distinct formats. First, an MT solution is cloud-based, allowing for instantaneous translation of publicly available information and online texts. This is ideal for the needs of civil society engagement, which demands MT tools that can be widely accessed by the general public, as well as for the needs of legislation, which often requires sharing information with citizens. Second, an MT appliance or hardware-based MT solution. In this format, the MT solution will work from a piece of actual computing hardware, which institutions will store on their own premises, guaranteeing 100% data security and confidentiality.

Machine Translation service has been already successfully used during Latvia’s EU presidency. The EU Presidency Translator provided instant translation of texts, documents, and websites from Latvian to English and vice versa. The machine translation technology powering the EU Presidency Translator can be applied to other EU languages as well. The Latvian public sector can also communicate in multiple languages and share crucial information about government policies, legislation, and services with audiences not only in Latvia, but also across the world. The machine translation service Hugo.lv is specially tailored to general-domain texts and legislative acts - the service can be used to translate general descriptions of e-services and other e-gov texts, as well as more specialized legislative documents such as legal proceedings, legislation, official regulations, and laws.

The NATO Strategic Communications Centre of Excellence (NATO StratCom COE) has identified a need to monitor news and social media content which can be predictive of and help detect hostile, and or terrorist cyber-attacks and counter the impact of the use of the internet for terrorism. It can also help to understand the sentiments of the public and serve as an early warning for the possible threats. There is a need to implement a technological toolkit of automated translation, textual pattern and keyword tracking, and text analytic functionality that will automate the monitoring and analysis of content from social and news media. High quality automated translation of large flows of cyber information is a key technology in data mining and the analysis value chain for monitoring large quantities of information. Fast, quality automated translation can provide near real time ability to improve analysis and expedite investigations and facilitate regional diplomacy efforts, communication with counterparts, and support the formulation of informed response.

In finance sector there is a need for platform that incorporates the latest technologies for optimal multilingual sentiment analysis providing real-time access to “the wisdom of the crowd”. The combination of machine translation and automatic sentiment analysis has proven to be effective, as semantic sentiment analysis appear to be
relatively insensitive to grammatical translation errors. This requires innovations on semantic analysis, machine translation and social data sourcing.

For companies working in e-commerce the challenge is to transfer the keyword searches for different countries – to link the keywords between languages. The lack of interoperable search terms is challenging the implementation. The solution is to have a trustworthy corpus of linked terms in various domains. This can then be used by the search engines to classify products, making for better product recommendations.

In e-commerce a major potential for LT is text analysis that can be used for customer feedback analysis locate valuable comments by extracting the necessary data. Already now some companies see that there are huge benefits of MT for e-commerce. MT could be used to speed up the localization process - translation of product descriptions, interfaces and customer contacts. When human translation is not available, machine translation can be provided as fall-back. In this case users can see the originally published text and machine translation. Machine translation is also used for translation of reviews, comments and forum posts. For companies operating across national and linguistic boundaries the use of MT is helpful for sentiment analysis, e.g. to monitor comments on products and services on user forums, social media, etc., which can be helpful to manage brand reputation online.

Two related areas that are often investigated in connection with MT-based translation services for companies operating multilingually across countries are controlled language and post-editing.

By now there is consensus in the translation industry on the essential software tools whose use benefits translation services in support of localisation projects, in particular computer-assisted tools such as terminology management systems and translation memories; these can be used alongside localisation toolkits and platforms, global content management systems and translation quality assurance software to ensure an optimal translation service.

For services in support of the online content localisation/globalisation chain the cultural dimension of products and services and the search engine optimisation (SEO) is very important. An interesting growth area for MT linked to localisation concerns cloud-based services, where MT engines made available on the cloud remove the overheads caused by maintaining the required hardware infrastructure and lessen the problems of the technical expertise that is needed to develop cost-effective MT-based solutions. Cloud-based MT systems are typically tailored and customised to the needs of corporate users, on the basis of their preferred domains and text types for the language pairs of interest to them.

To enable European industry and public sector to benefit from the digital content and to get access to valuable information that is accumulated in it, innovative and intelligent use of technologies are needed to revolutionise society, business, and science in the future.
10. REFERENCES


11. APPENDIX

LTI CLOUD EVALUATION REPORT

LIST OF TABLES

Table 0.1 Test of LTI Cloud Broker 59
Table 0.2 LTI broker tests for case sensitivity 60
Table 0.3 Test results of LT Store categories 61
Table 0.4 Test results of filtering scenarios 62
Table 0.5 Test results of section "My LTCs" 63
Table 0.6 Comparison to different types of existing API cloud 63
Table 0.7 Suggestions for future work on LTI Cloud 65

LIST OF ABBREVIATIONS

API Application Programming Interface
CORS HTTP access control
cURL a computer software project providing a library and command-line tool for transferring data
GUI Graphical User Interface
HATEOAS Hypermedia as the Engine of Application State, a constraint of the REST application architecture
HTML HyperText Markup Language, the standard markup language used to create web pages.
HTTP Hypertext Transfer Protocol, an application protocol
HTTPS HTTP Secure, a protocol for secure communication over a computer network
IS Information System
IT Information Technology
11.1. Introduction

This document is an evaluation report for the prototype of LTI Cloud. Prototype was developed by MLi consortium under leadership of Esteam as part of WP4 activities. Prototype has implemented the basic functionality to provide the essential look-and-feel impression about the potential and features of language service aggregator for both for language technology providers and language technology users and potentially byers.

11.2. Evaluation approach and methodology

For the evaluation purposes three main user scenarios were detected and used in action:

1. Users seeking for language services;
2. Developer looking to implement/use service from the cloud;
3. Developers registering its own language service into the cloud;
4. IT specialist looking for solutions for innovation for its workflows and existing processes and data value chain possible improvements.

As this is the first prototype, not all functionality are implemented and/or working properly, therefore during this testing phase also evaluated specification and possible platform

15 http://lticloud.eu

MLi WP6/D6.2
improvements. Also existing other available platforms were analysing and compared to LTI language cloud possibilities.

Trying out different user scenarios, both lticloud.eu provided GUI and machine interface were tested.

All functionality provided by prototype was tried using latest most popular web browsers and HTTP request generating software.

### 11.3. Test report of the Service Platform

The LTI Cloud is the one-stop-shop platform for making available, discovering, assembling, testing and prototyping language technology components, this test plan to outline the test strategy and overall test approach for LTI Cloud web page.

#### 11.3.1. LTI Cloud Broker

The LTC broker evaluates the availability of the desired LTC and maps the LTI Cloud standardized function call(s) to the LTC SaaS’ syntax. It then forwards the reformatted request to the LTC SaaS end Calling a Detect Language LTC to identify the language of a string.

**Example of request**:  

curl --header "X-Lti-Token: GNTwsD3aDDWY63" --data "q=Leben%20wie%20Gott%20in%20Frankreich"  
https://api.lticloud.eu/nlp/com.detectlanguage.ws/detect_language.json

<table>
<thead>
<tr>
<th>Sample data</th>
<th>Response</th>
<th>Detected language</th>
</tr>
</thead>
<tbody>
<tr>
<td>curl.exe -k curl --header &quot;X-Lti-Token: IndividualApiToken&quot; --data &quot;q=Leben&quot; <a href="https://api.lticloud.eu/nlp/com.detectlanguage.ws/detect_language.json">https://api.lticloud.eu/nlp/com.detectlanguage.ws/detect_language.json</a></td>
<td>{&quot;status&quot;:200,&quot;response_time&quot;:0.40926334,&quot;size&quot;:85,&quot;data&quot;:[{&quot;language&quot;:&quot;de&quot;,&quot;score&quot;:0.01}]}</td>
<td>German</td>
</tr>
</tbody>
</table>

---

16 Used Curl library for testing, https://curl.haxx.se

17 IndividualApiToken- available on itcloud.eu website Profile section, for registered user

MLi WP6/D6.2
### Sample data

<table>
<thead>
<tr>
<th>Sample data</th>
<th>Response</th>
<th>Detected language</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>ectlanguage.ws/detect_language.json</code></td>
<td><code>{&quot;status&quot;:200,&quot;response_time&quot;:0.408103182,&quot;size&quot;:88,&quot;data&quot;:[{&quot;language&quot;:&quot;en&quot;,&quot;score&quot;:0.01}]}</code></td>
<td>English</td>
</tr>
<tr>
<td><code>curl.exe -k curl --header &quot;X-Lti-Token: IndividualApiToken&quot; --data &quot;q=Computer&quot; https://api.lticloud.eu/nlp/com.detectlanguage.ws/detect_language.json</code></td>
<td><code>{&quot;status&quot;:200,&quot;response_time&quot;:0.414627734,&quot;size&quot;:98,&quot;data&quot;:[{&quot;language&quot;:&quot;fr&quot;,&quot;score&quot;:0.01}]}</code></td>
<td>French</td>
</tr>
<tr>
<td><code>curl.exe -k curl --header &quot;X-Lti-Token: IndividualApiToken&quot; --data &quot;q=Rappel concernant &quot; https://api.lticloud.eu/nlp/com.detectlanguage.ws/detect_language.json</code></td>
<td><code>{&quot;status&quot;:200,&quot;response_time&quot;:0.411291754,&quot;size&quot;:94,&quot;data&quot;:[{&quot;language&quot;:&quot;sv&quot;,&quot;score&quot;:0.01}]}</code></td>
<td>Swedish</td>
</tr>
</tbody>
</table>

*Tests for case sensitive situations.* Language detection and returned data is correct, response is identical.

| Table 11.2 LTI broker tests for case sensitivity |

MLi WP6/D6.2
**Sample data**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>{&quot;status&quot;:200,&quot;response_time&quot;:0.4162139,&quot;size&quot;:88,&quot;data&quot;:[&quot;language&quot;:&quot;en&quot;,&quot;score&quot;:0.01]}</td>
<td>English</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>{&quot;status&quot;:200,&quot;response_time&quot;:0.410170157,&quot;size&quot;:88,&quot;data&quot;:[&quot;language&quot;:&quot;en&quot;,&quot;score&quot;:0.01]}</td>
<td>English</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>{&quot;status&quot;:200,&quot;response_time&quot;:0.414917798,&quot;size&quot;:88,&quot;data&quot;:[&quot;language&quot;:&quot;en&quot;,&quot;score&quot;:0.01]}</td>
<td>English</td>
<td></td>
</tr>
</tbody>
</table>

**11.3.2. LT Store**

In **Shop** section there is a possibility to select different categories and chose required language tools.

Category selecting selects tools which meets category. Filter works, test passed. Filter works in My LTCs and Shop categories.

| **Table 11.3 Test results of LT Store categories** |
|---|---|---|
| Available selections /Filtration | Pass/Fail/Not Available |
| Human Machine Dialog | Tools not available |
| Knowledge Management | Pass |

MLi WP6/D6.2
| Natural language Generation | Pass |
| Natural language Processing | Pass |
| Search | Pass |
| Text Mining | Pass |
| Translation | Pass |
| Voice | Pass |

Filter by language. Filter is working accordingly. Test passed.

<table>
<thead>
<tr>
<th>Available selections</th>
<th>Tools available</th>
<th>Pass/Fail/Not Available</th>
</tr>
</thead>
<tbody>
<tr>
<td>Translation + English language</td>
<td>Tilde Translation; Document Translation</td>
<td>Pass</td>
</tr>
<tr>
<td>Translation + Afrikaans language</td>
<td>Tools not available</td>
<td>Pass</td>
</tr>
</tbody>
</table>

**Recommendation:** improve filtering by adding filter option by language pairs not only one language. If filter criteria is not met, recommendation is to show user information, by now, nothing is displayed.

Sort by release date- ascending/ descending. Test passed.

Add to cart option allows to try/buy different language tools developed by leading language technology companies. Selection is available only for registered users.

**My LTCs** category displays user added language tools from **Shop** category.

MLi WP6/D6.2
Table 11.5 Test results of section “My LTCs”

<table>
<thead>
<tr>
<th>Operation</th>
<th>Expected result</th>
<th>Pass/Fail</th>
</tr>
</thead>
<tbody>
<tr>
<td>Add language tool to shopping cart</td>
<td>Tool is available in My LTCs list</td>
<td>Pass</td>
</tr>
<tr>
<td>Add same language tool to shopping cart</td>
<td>Tool is displayed once</td>
<td>Pass</td>
</tr>
<tr>
<td>Add different language tool to shopping cart</td>
<td>Two tools available in My LTCs list</td>
<td>Pass</td>
</tr>
</tbody>
</table>

**Recommendation:** add option to remove tool from My LTC’s list.

**Solutions** category is not fully developed. Recommendation complete development of section and perform the testing.

11.4. **Comparison with other platforms and similar solutions**

Currently there are plenty of platforms - but mostly they are not intended for one-stop language technology market place open for various vendors including competitors.

Some different kinds of services that could be stated as competitor platforms are listed below.

Table 11.6 Comparison to different types of existing API cloud

<table>
<thead>
<tr>
<th>Type</th>
<th>Description</th>
<th>Comparison to LTICloud.eu</th>
</tr>
</thead>
<tbody>
<tr>
<td>Electronical catalogues</td>
<td>Platforms that provides possibility to register your language service for different catalogues. Language tool seekers can use search and filtering capabilities to find potential technology providers and retrieve contact info or other useful</td>
<td>Fully satisfies scenarios for IT professionals seeking for certain language service, for research of possibilities to improve company’s workflows, but often later arrives problems - licencing issues, availability of services, sustainability, scalability and also - often such services not available for real implementation or missing</td>
</tr>
<tr>
<td>Type</td>
<td>Description</td>
<td>Comparison to LTIcloud.eu</td>
</tr>
<tr>
<td>----------------------</td>
<td>-----------------------------------------------------------------------------</td>
<td>------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td></td>
<td>information. For example Meta share(^{18}), UIMA tools(^{19}).</td>
<td>monetising capabilities. Many cases there are need for multiple language tools for one value chain and for various languages, such tools provides different vendors. It can take a lot of time and human resources to sign different agreement and make different service integrations.</td>
</tr>
<tr>
<td>Monetising platforms</td>
<td>Platforms that provide service-in-the-middle or proxy, that forwards service byers requests to language technology provider, providing traffic counting, restriction, marketing and pricing strategies. For example, apigee.com(^{20}), Mashape marketplace(^{21}).</td>
<td>These platforms mostly are too general, language tools not a lot available there. Such platforms solve long negotiations about agreement and pricing but still have problems with different API calls and formats.</td>
</tr>
<tr>
<td>Dedicated platforms</td>
<td>Cloud services that provide specific purpose platforms. For example FREME(^{22}) platform is open platform for service users and also technology provider from other side, and the platform is intended for data enrichment purposes. Other specialized platform is platform</td>
<td>Doesn’t provide general language tools for any of client’s environments. Such platform could be client of LTI Cloud - use any general language technology in data enrichment value chain.</td>
</tr>
</tbody>
</table>

---

\(^{18}\) [http://www.meta-net.eu/](http://www.meta-net.eu/)

\(^{19}\) [https://uima.apache.org](https://uima.apache.org)


\(^{21}\) [https://market.mashape.com/](https://market.mashape.com/)

\(^{22}\) [http://www.freme-project.eu/](http://www.freme-project.eu/)
<table>
<thead>
<tr>
<th>Type</th>
<th>Description</th>
<th>Comparison to LTIcloud.eu</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>created during Panacea project, this allows creation of workflows of different language tools using one specific Java based tool Taverna.</td>
<td>This platform is closed for other technology companies and do not provide any capability to register a new services. Such clouds are restricted to its company products and services they provide.</td>
</tr>
<tr>
<td>Propriety platforms</td>
<td>Some technology companies have created their own API selling platform with full monetising and API consuming trials and detailed documentation. Systran.io is an example.</td>
<td></td>
</tr>
</tbody>
</table>

### 11.5. Conclusions and Suggestions for Future Development

The platform has a strong potential to successfully support different business scenarios:

1. Use platform as commercial product, take some share of price paid by customers.

2. Provide new business capabilities by providing access and fast integration of language tools for companies.

3. Provide new selling channel for language technology companies

To make platform widely used and popular - it need also real services integrated - providing solid coverage of different technology across different languages, domains and other parameters.

Implemented features in lticloud.eu are working properly. To become as market leader for one-stop language technology market place, future general platform improvements are grouped by user scenarios.

**Table 11.7 Suggestions for future work on LTI Cloud**

---


24 Apache Taverna is an open source software tool for designing and executing workflows.

25 [https://platform.systran.net](https://platform.systran.net)
<table>
<thead>
<tr>
<th>Scenario</th>
<th>Suggestion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Registering new language service</td>
<td>As platform acts as proxy, it should provide seamless existing LT API integration. Current prototype not fully covers various API technologies and platforms.</td>
</tr>
<tr>
<td></td>
<td>Improve usability of LTC configuration possibilities, do not restrict to certain language list.</td>
</tr>
<tr>
<td>Monetisation</td>
<td>Prototype do not implement full commercialisation scenario - request control, view used service statistic, view statistic about owned LTCs, calculation of costs, charging and invoicing.</td>
</tr>
<tr>
<td>Data format abstraction</td>
<td>Specification of full service covers significant problem of creation data processing workflows – every language tool can have different data format for input/output. It is important to have one central data format that allows chaining different language tools seamlessly. This is not implemented in the prototype, but would be a big gain for solution creators.</td>
</tr>
<tr>
<td>LTC integration into company workflows</td>
<td>Trial web forms that provides API exploring by filling out sample form would be recommended. Developers coming from different industries with different competencies could try out LTC’s API and get examples how integrate this service into its company IS.</td>
</tr>
<tr>
<td></td>
<td>Use best practices and create API according standards’ requirement both for data formats and API consuming. Improve REST pattern for API calls.</td>
</tr>
<tr>
<td></td>
<td>Provide libraries for common programming languages that allows clients consume API from client’s native environment.</td>
</tr>
<tr>
<td></td>
<td>Additional support for JavaScript based solutions, allowing different cross-domain consumption technologies – JSONP or CORS support.</td>
</tr>
</tbody>
</table>

MLi WP6/D6.2