LetsMT!

Platform for Online Sharing of Training Data and Building User Tailored MT

www.letsmt.eu/

Project no. 250456

Deliverable D1.1
Requirements Analysis

Version No. 1.0

30/06/2010

Document Information

Deliverable number:	D1.1
Deliverable title:	Requirements analysis
Due date of deliverable according to DoW:	30/06/2010
Actual submission date of deliverable:	30/06/2010
Main Author(s):	UCPH
Participants:	UCPH, Tilde, UPP, SEM, MOR
Reviewer	Aivars Berzins (Tilde), Jörg Tiedemann (UPP)
Workpackage:	WP1
Workpackage title:	LetsMT! Platform and infrastructure
Workpackage leader:	Tilde
Dissemination Level:	PU
Version:	V1.0
Keywords:	Users, localisation, requirements, CAT, translation tasks, financial news, machine translation, metadata, formats

History of Versions

Version	Date	Status	Name of the Author (Partner)	Contributions	Description/Approval Level
0.1	17/06/2010	Initial draft	Lina Henriksen Claus Povlsen Lene Offersgaard (UCPH) David Filip (MOR)	Initial draft completed	Initial draft for review
1.0	30/06/2010	Final version	Lina Henriksen Claus Povlsen Lene Offersgaard (UCPH) David Filip (MOR) Bram Stalknecht (SEM)	Final version completed	Ready for submission to EC

EXECUTIVE SUMMARY

This present report represents the work done in Task 1.1 in WP 1. The main aim is to provide a requirements analysis in order to establish a safe basis for an initial functional specification of the LetsMT! translation platform.

The evidence that is used in this requirements analysis is collected from two sources. An internal source in terms of two project partners' detailed description of use scenarios and an external view on user requirements provided by collecting information from a wide range of end users. This information is provided by making interviews with the end users based on an interview template.

One important observation in this report is that the input from the external partners in comparison with the feedback from the end users in broad terms point to the same translational needs.

Table of Contents

Abbreviations

Abbreviation	Term/definition
LetsMT!	Platform for Online Sharing of Training Data and Building User Tailored MT
API	Application programming interface
BLEU	Bilingual Evaluation Understudy
CAT	Computer aided translation
CRM	Customer Relationship Management
CSV	comma-separated values
ERP	Enterprise resource planning
GUI	graphical user interface
IPR	Intellectual property rights
Locale	Market with specific language, legal, cultural etc. needs. Locale is typically the same or smaller than a country, such as DE-DE or FR-CA, but can be also larger, such as ES-LA, which is rather a useful abstraction motivated by economies of scale than a real locale.
L10N	Localization - Creation of locale specific versions of products, documentation, and support materials. Translation is typically an important part of L10N process.
LSP	Language service provider
METEOR	Automatic Metric for MT Evaluation
MT	Machine translation
OLAP	Online analytical processing
SOV language	Languages with word order: Subject-Object-Verb
TBX	Term Base eXchange
TDA	TAUS Data Association
TER	Translation Edit Rate
TMX	Translation Memory eXchange format
TM	Translation memory
XLIFF	XML Localisation Interchange File Format

1. Requirements analysis – Purpose and Objectives

The target audience of the LetsMT! platform includes many different types of users which makes it essential to ensure that the different elements of the LetsMT! platform accommodate many different types of user requirements. Therefore some crucial project objectives involve great emphasis on LetsMT! system usability and suitability.

In order to achieve these objectives an approach involving an external as well as an internal perspective on user requirements has been adopted. This report thus concerns an analysis of user requirements identified among potential end users from many different types of organizations (external view) together with detailed descriptions of a LetsMT! use scenario prepared by a project partner (internal view). The approach involved collection of information that creates an overview of possible user translation tasks, workflows, general contexts, requests for MT system functionalities, etc.

This report will first provide an overview of the approach adopted for collection of user requirements among external end users followed by a classification of these end users in different user groups (section 2). The classification of users in user groups will be used in section 3 to describe the typical user requirements of each user group.

Section 4 deals with detailed descriptions of 2 use scenarios providing an internal perspective from project partners. These partners have long term experience within the fields of localization/ translation and financial news.

Section 5 will give an overview of main interview results and will thus provide a use scenario from an external view. In section 5 we will also compare user requirements as seen from an internal view with requirements as seen from an external view.

2. User requirements survey - approach

The survey of user requirements as perceived from an external perspective is based on interviews. These interviews were conducted by project partners, and interviewees (also called respondents) are potential MT users from many different types of organizations in the various partner countries (for more information about interviewees see section 2.2).

Interviews were selected as a preferable approach for collection of user requirements to ensure some flexibility in the selection of apposite questions in relation to a particular interviewee. This approach also allowed the interviewer to use his/her intuition and sensitivity to obtain accurate, relevant and sufficient information in the many different interview contexts.

Interviews were conducted as a series of *open questions* (open questions encourage interviewees to think and reflect; they will often result in personal opinions and will give some control of the interview to the respondent). This approach was considered important for harvesting as much information as possible and also for embracing the “unexpected” information.

2.1. Interview template

A template of interview questions was created (attachment A). This template covers a catalogue of questions where different subsets have been relevant in different interview contexts. The template constituted the backbone of all interviews and ensured comparable and relatively homogeneous data.

Interview questions of the template are divided in two groups. One group concerns the contexts of the organization and the interviewee. The objective of this question category was to provide a rough picture of the settings in which the organization/interviewee operates as this insight is essential in a system development context. The other group concerns the potential users’ expected requirements to the different elements of the LetsMT! platform.

2.1.1 Interviewee's background information

In order to develop a well designed, high quality and easy-to-use system, the LetsMT! development team needs an overview of user types with respect to job profile, job tasks and technical competencies together with aspects related to general working conditions, availability of different CAT tools and some overall specifications of general translation tasks. Therefore, a series of questions have been developed to describe the interviewee organization.

This group of questions includes a simple classification of the interviewee's organization in 4 different categories: *localization/translation agency, organization with multilingual translation needs, research institution* or *news agency* (for more information about these categories see section 2.2).In order to grasp the interviewee's perspective, short general descriptions of the interviewee's job profile and job tasks are also included.

Subsequent questions concern a specification of CAT tools (and other tools) employed in the organization together with an outline of the organization's experience with these CAT tools. This is followed by a description of the organization's translation tasks. This description includes information about domains, language pairs and translation volume as well as some information about the organization's stored text resources.

Closing questions of this group focus on the localization/translation workflow of the particular organization and specify intellectual property rights of text resources stored in the organization.

2.1.2 Interviewee's LetsMT! platform requirements

The overall purpose of this part of the interview template is to identify the criteria and functionalities that define a good MT system. Questions for example pertain to upload of corpora, the organization's willingness to share data, metadata requirements and preferred feedback facilities.

The template also covers questions concerning the translation widget and browser plug-in. The last questions deal with accessibility to different kinds of resources stored in the LetsMT! platform and are especially directed towards researchers, students and teachers.

2.2. Classification of interviewed organizations

The total number of interviews conducted by the partners is 43. These interviews have been divided into 4 groups based on the organizations' primary business activities and in line with the goals of the LetsMT! system. One LetsMT! goal concerns efficiency increase of localization and translation work performed by industry professionals and organizations with multilingual translation needs. Another goal concerns a free online translation service for global business and financial news. Furthermore, it is expected that academic users as for example researchers, teachers and students will be interested in access to (and exchange of) linguistic resources of the LetsMT! system to further other SMT related research work.

Accordingly, the 4 user groups are the following:

- Localization/translation companies
- Organizations with multilingual translation needs
- Companies distributing business and financial news
- Users in research and university communities

The below table indicates names of interviewed organizations within each user group. The partners have conducted 21 interviews with *localization/translation* agencies, 10 interviews with *organizations with multilingual translation needs*, 6 interviews with *companies distributing business and financial news* and 6 interviews with *research organizations*. The table also shows names of partner organizations conducting each interview.

User Groups	Interviewee Organizations	Interviewers
Localization/Translation Agencies	Freelance translator	FFZG
	Freelance translator	FFZG
	Lancon	FFZG
	Lancon	FFZG
	Oversætterhuset	UCPH
	Oversættelsescentret, UCPH	UCPH
	Inter-Set A/S	UCPH
	CLS Communications	UCPH
	TILDE	TILDE
	SDI Media	TILDE
	Skrivanek Latvia Ltd.	TILDE
	TILDE	TILDE
	TILDE	TILDE
	TILDE	TILDE
	Moravia IT	MOR
	Moravia IT	MOR
	Moravia IT	MOR
	Moravia IT	MOR
	Amesto Translations AB	UUP
	Oneliner Language & eBusiness Solutions bvba	UUP
	ESTeam	UUP
Organizations with multilingual translation needs	Novo Nordisk A/S	UCPH
	Novo Nordisk A/S Region Danmark	UCPH
	Communication department of a bank	UCPH
	An international bank	UCPH
	National Library of Latvia	TILDE
	European Centre for Disease Prevention and Control (ECDC)	TILDE
	EU institution	TILDE
	Latvenergo	TILDE
	EU institution	MOR
	EU institution	MOR
Companies distributing business and financial news	Schultz Information International A/S	UCPH
	Infopaq	UCPH
	Dow Jones	SEM
	The Times	SEM
	Nomura International	SEM
	Thomson Reuters	SEM

Users in research and university community	Faculty of Humanities and Social Sciences, Zagreb	FFZG
	University of Copenhagen	UCPH
	University of Latvia	TILDE
	Ventspils University College	TILDE
	Research institution	TILDE
	Linköping University	UUP

3. Survey results of the four groups of user profiles

This section describes survey results of the 4 different user groups (attachment B). The answers of each user group will be reported in accordance with the structure of the question template, i.e. subsections of below user group sections refer to overall question categories in the interview template.

Survey result descriptions of the 4 user groups in section 3 will generally not be exhaustive, but will instead try to emphasize main points (attachment B provides a full version of all answers). In section 5 we will for selected topics provide overall conclusions on the basis of all user groups. These overall conclusions will include diagrams reflecting an exhaustive enumeration of all answer types.

3.1. Localization and Translation organizations

This section describes survey results of the 4 different user groups (attachment B). The answers of each user group will be reported in accordance with the structure of the question template, i.e. subsections of below user group sections refer to overall question categories in the interview template.

Survey result descriptions of the 4 user groups in section 3 will generally not be exhaustive, but will instead try to emphasize main points (attachment B provides a full version of all answers). In section 5 we will for selected topics provide overall conclusions on the basis of all user groups. These overall conclusions will include diagrams reflecting an exhaustive enumeration of all answer types.

For topics mentioned in this section results of an enumeration of interview answers are often stated in brackets as actual numbers.

3.1.1 Use of CAT tools

Not surprisingly, in comparison with the other groups in this investigation, this group of interviewees has shown great interest and willingness to express their views on the quality of the CAT tools that they are using in their daily work. A majority of the respondents belongs to the decision level of their organization, in that 14 of the interviewees describe their job profiles as being either manager or administrator. Professional translators are thus less represented in this survey. A fact that can be said to have both cons and pros. It could be perceived as a drawback if the respondents' practical knowledge is very limited. On the other hand, it is an advantage being given responses from a representative part of the organization's decision level. Especially the answers concerning copyright and data sharing can be judged as being both precise and exhaustive.

The answers concerning the use of TMs reflect the fact that the commercial SDL Trados translation platform is a dominant player in the market. Twelve of the LSPs have responded that they in varying degrees use the translation memory imbedded in the SDL Trados package. Other TM tools such as Transit NXT and Acros are also mentioned in the survey result, but they are as implied above outnumbered by SDL Trados in this context.

Another significant feature is the respondents' use of old and to some extent outdated TM tools. MS Helium and MS Locstudion, are examples of localisation tools that are not commercially available any longer. IBM translation manager, abandoned in 2002, is another example of an old TM system still in use. An additional observation one can make, is that 6 of the interviewees do not use any kind of TM.

Only 7 of the respondents reply that they use fully automatic MT systems in their translational practice. The MT systems that are used vary a lot ranging from SMT systems (Language Weaver and Asian Online) to more traditional rule-based systems such as Systran and Promt. In short, fully automatic systems are not a dominant CAT tool amongst the respondents.

Fifteen of the interviewees use some kind of terminology tools. Bearing in mind that there was a widespread use of the TM tool imbedded in SDL Trados cf. above, it does not come as a surprise that 7 of the respondents use Multiterm (also a part of the SDL Trados package).

If the link between Multiterm and TM in SDL Trados is understood as an API (application programming interface), 3 of the organizations respond that they benefit from this facility.

The respondents' assessments of the various CAT tools differ profoundly. One organization states that the tool is easy to use, while others think that the tools are too complicated to use. An advantage of the tools emphasized in the survey is that by using the tool you ensure translational consistency and makes it possible to gather all relevant information in one place.

Since 20 out of 21 organizations in the LSP group are business agencies, efficiency in terms of low labour cost is an important parameter. By reducing labour costs, these agencies therefore value CAT-tools highly.

Regarding disadvantages three issues are mentioned by more respondents. TMs are criticized for containing too many minor errors. Furthermore, it is pointed out that some versions of SDL Trados do not have an adequate file format treatment. The other area is dealing with lack of functionality. In some of the TMs used, it is not possible to use the fuzzy match facility. Finally, the CAT-tools are criticized for lack of interoperability making it troublesome to port the resources from one application to another. It would for instance, be advantageous if it would be possible, in a parallel session, to use both TMs and MTs in one's translational work.

Concerning technical problems using CAT tools the opinions vary greatly. Many respondents have not faced any problems while others express that they are often confronted with technical problems using language technology in their translational work. The latter group report on more problems. To mention a few, inadequate handling of file formats, installation problems, compatibility problems (lack of backwards compatibility), and lack of a defined standard of TMX (SDL Trados is, for instance, using its own TMX dialect).

'On-line language resources' are used by 14 of the respondents. A wide range of 'on-line language resources is listed which reflects the fact that resources useful for minor languages are most often found domestically. Seven of the respondents state that they use 'Other relevant tools'. Again no general pattern in terms of that more than one respondent uses a particular tool, can be found.

3.1.2 Localization/translation workflow

Based on the feedback concerning the translation pipeline issue, it can be concluded that each localisation organization has defined and implemented its own company specific translation pipeline. The larger companies have often implemented their own software to facilitate their translation pipeline, where the smaller companies, in general, use standard software such as the MS-office products.

Not surprisingly, the respondents primarily use the mainstream browsers, Internet Explorer (13) and Firefox (13). Other browsers mentioned are, IE8 (3), Chrome (2), Opera (1), and Safari (1).

3.1.3 Translation tasks

Many subject domains are mentioned by the respondents. The most frequently translated domains are: IT (8), law (legislation) (8), Economics (finance) (5) and medicine (2). How to classify and categorise subject domains has been disputed for many years and no agreement has been reached, meaning that the information above concerning subject domains should be interpreted with caution.

Concerning file formats of translation texts, the respondents answer that they use and handle many file formats in their translation work. The most frequently mentioned are, MS Word (13), HTML (5), XML (5), txt (if that can be considered as being a format?), TMX (4), excel (4). But as mentioned above, the list of file formats mentioned is very long.

In order to make replies concerning translation volume comparable, running words have been chosen as the unit of measurement. In order to reach this common unit, it has been assumed that a translator translates 5 pages a day, a month consists of 25 working days, and a page contains 300 words. Twelve of the respondents have given estimates of their translation volume. The scale in terms of translation volume spans from 810,000 words a year to 23,400,000 translated words.

A huge amount of language pairs are represented in the answers. One distinctive feature is that besides the large languages (Russian, English, Spanish, French, etc.), the minor languages spoken by the states involved in this project are represented. Meaning that if Denmark had not been a part of LetsMT! it would probably not have been included in the list of translated language pairs.

Based on the replies concerning the respondents' use of CAT tools, it is a bit surprising that (full) manual translation is done as much as TM based translation. The replies with respect to MT based translation reflect the observation that MT systems are not frequently used by the respondents. Thus only 1 LSP organization employs MT as the primary translation method. It should be added, that some confusion about how human involvement in the translation process should be understood. It seems as if some organizations count the human post-editing efforts as human translation. This confusion could explain the 'over-weight' of human translations.

As expected very few LSP organizations can give substantial feedback to MT language pairs.

3.1.4 Text resources

The information given by the respondents on text resources in the organizations is quite scarce. Several of the respondents have not replied, some do not know the exact number, but assess that the resources are huge.

Revising data is a way to ensure that resources remain of a high quality. Seven of the interviewees have not answered. The remaining replies are distributed more or less equally on, once a year, twice a year, and often.

The language pairs in the text resources can be considered as a subset of the 'Translation language pairs'. Meaning that e.g. Croatian is not represented in the set of text resources since the Croatian translators do not use TMs in their translation work. Otherwise the same pattern recognised under 'Translation language pairs' can be observed here.

All the respondents have answered that they do not possess any parallel corpora besides texts in the TMs.

Most of the respondents (14) which has answered this question, inform that the most important criterion in connection with structuring their text resources is the customer (9). Subject domain (5) is also considered important in this context.

Since almost all of the parallel texts are stored in TMs, the overall dominant file format is TMX (11).

Most of the organizations that have described their translation workflow (12) have a fine-grained distribution of work. The most significant feature in the replies is the division of labour between data administration (file conversion, tag-editing, and upload) and the translation task as one put it, 'the translator should concentrate only on translation'. If the respondents use a data management tool it is most often developed in-house.

A clear majority of the interviewees that have replied to which segmentation type that is typical in their organization answers 'at sentence level' (13).

In the organizations that use CAT-tools the translator's access to TMs differs dependent on the translation work. If more translators work on the same translation task, they often share the TMs via internal networks.

In those organizations that have division of labour between data administration and the translation task, an appointed project manager often carries out the data administration.

In cases where high translation quality is required most of the data validation is carried out be humans (10), but also automatic validation is used. Especially when the customer has required a high quality translation, the validation is carried out by humans.

3.1.5 IPR

The feedback from the interviews about Intellectual Property Rights can be divided into two groups. One that says that allocation of IPRs depends on the contract made with the client (5) and another group of organizations that gives the IPRs to the customer/client (6).

3.1.6 Definition of a good MT system

The interviewees mention many relevant criteria that should be met in order to implement an ideal MT system. The most frequently criteria mentioned are, translation quality (7) and interoperability (4), i.e. being able either to integrate an application in the MT system in question or the other way around to integrate the MT system into another translation platform.

3.1.7 Upload of parallel corpora

In connection with upload of files most of the respondents prefer or at least can live with TMX (9) although TXT is also mentioned a couple of times. The MS-word format is also mentioned. Most of the interviewees that have answered this question (11 out of 15) about data management when uploading, regard this as a very useful functionality.

Regarding the idea of sharing data, 6 of the respondents haven't answered (they have no resources), six have said *no*, and nine have said *yes with reservations*. In this group, you will find answers like, *probably* and *maybe*. But, first and foremost, the willingness to share data depends on the costumers since they often own the IPRs.

The replies concerning the need for an alignment tool is connected with the small amount of 'parallel texts not in TMs' which the organizations possess, cf. above. They show an interest, but they can't really see how they can benefit from such a facility.

Not surprisingly, the organizations, in general, are not willing (or are not allowed to) make their parallel resources public. Only two of the respondents think that setting up a user group for SMT would solve the problem concerning the unwillingness to share data. The organizations show no interest in making their parallel corpora accessible for research purposes. As expected more organizations would find it acceptable to agree on access only for the organizations (5) - but again, you can observe a major group that does not reply or answers, *no idea*.

3.1.8 LetsMT! resource metadata

The replies on which kind of metadata that the organizations would consider useful reflect the different ways their translation work is organised. Some organizations do not (or do almost not) use CAT tools and therefore they do not reply. Other organizations translate specific text types, such as movie subtitles, and therefore have different views on the importance of the various metadata. In general, however, all respondents are positive to having access to as many metadata as possible.

3.1.9 Feedback

More than half of the organizations (11) consider it very useful to be able to give feedback to parallel texts. Several of those who have replied (5 out of 14) think it is good idea to rate resources, especially the rating of specific data owners and entire resources.

3.1.10 Configuration

About half of the interviewees (11) think it is a good idea to be told how much text volume that normally is required to reach acceptable SMT based translation results. Access to information about system configurations is also considered useful (7). The possibility of specifying parameters for your own system is even more wanted by the respondents (11).

3.1.11 Website for translation

Not surprisingly, the feedback regarding the LetsMT! website for translation is more or less a copy of what the organizations replied to the questions about their translation tasks.

3.1.12 Translation widget

In general, the Localisation and Translation organizations do not show great interest in having a translation widget included into websites of business and financial news. The most frequently formulated argument against having such a facility, is that they are LSPs and therefore in contrast to such a widget provide high quality translations.

3.1.13 Browser plug-in

Most of the respondents (7 out of 14) answer *yes* concerning whether speed is an important parameter in automatic translation. It seems, however, as if they do not relate their answers to the performance of the browser plug-in but to translation systems in general. Concerning file formats and language pairs it is more or less a copy of the answers in connection with respondents' description of their translation tasks.

3.1.14 Integration with other tools

As one could expect the respondents want to have access to an SMT system from the CAT tools they are using. They want the SMT system to be available from all commercial systems on the market, specifically of course from all the TMs, (cf. above) that they use in their translational work.

3.1.15 Access to resources in LetsMT!

Localisation and Translation organizations, in general, do not express any needs to do research by having direct access to text resources and by working with the trained SMT systems.

3.1.16 Conclusion

The assumption or preconception that the translation industry with respect to integrate innovative language technology is somewhat reluctant, seems to find support in this survey.

First, not all organizations are using CAT-tools in their translation work. Secondly, very few of the respondents in this survey are using or have thought of integrating SMT systems in their translation pipeline.

Most of the respondents are using TMs, especially SDL Trados is used, but also outdated TMs and TM platforms are used. As a consequence of this fact, the parallel texts in the organizations' possession consist first of all of the translation memories that they have developed. No other types of parallel texts seem to be part of the organizations' data repositories. The latter observation explains why so few of the respondents find it useful to have access to an online sentence alignment tool. The many translation memories the respondents report that they have in their possession, seem at first glance to be perfect seen from the LetsMT! point of view. Translation memories are thus tailored as SMT training data. So in principle the sharing idea of training data and henceforth SMT systems seem to be feasible. Unfortunately, the respondents' motivation to share data and systems with each other is very limited. Part of this lack of motivation is due to contractual obligations towards the customers. Should it be the case that some of the respondents' translation memories are allowed to uploaded to the LetsMT! platform, then the only use of these data will be to develop customized SMT systems - exclusively based on the organization's own data.

3.2. Organizations with multilingual translation needs

The number of respondents in this category is 10 distributed over six public and four private organizations. All the interviewees have managing job profiles, meaning that in this group of respondents professional translators are not represented.

3.2.1 Use of CAT tools

In general, Organizations with multilingual translation needs use CAT tools more infrequently in their translational work than the LSP group. Four out of 10 use SDL Trados, two organizations use MT systems, five use some kind of terminological tool and finally five out of the ten respondents use on-line services (eg. www.letonika.lv). Other CAT tools that are mentioned are TVT (Text Verification Tool and Tilde's Birojs.

An opinion about the use of CAT tools is conveyed by only one respondent so no common pattern can be derived from the answers. The advantages mentioned are that use of CAT tools ensures consistent translation and 'is overcoming the language barrier'. One of the interviewees has experienced that when the TMs are getting bigger mistakes become more frequent . The technical problems mentioned are installation difficulties (Trados Studio), handling of file formats, and insufficient translation quality for some languages.

Seven of the respondents reply that they use Internet Explorer as their standard browser.

3.2.2 Translation tasks

Many subject domains are mentioned by the respondents. The most frequently translated domains are law (legislation/patents) (4), medicine (3), and finance (2).

The respondents use and treat many file formats in their translation work. The most frequently mentioned are, MS Word (7), PDF (6) XML (4)In order to make the replies translation volume comparable, running words have been chosen as the unit of measurement. In order to reach this measure unit, it has been assumed that a translator translates five pages a day, a month consists of 25 working days, and a page contains 300 words. Five of the respondents have estimated their translation volume. The scale in terms of translation volume spans from 120,000 words a year to 600,000,000 translated words. The translation language pair consists of all the official EU languages (both directions), Norwegian to and from Danish, Swedish, Finnish and English, and from Latvian to Russian.

Only 5 of the interviewees have estimated the ratio between TM, MT, and human translation. A clear majority of these respondents states that all translation is carried out by humans. The union set of language pairs in which MT was involved consisted of all of the languages represented in EU until 2004.

3.2.3 Text resources

The size of the organizations in this group varies profoundly, which is supported by the fact that one organization has 6,5 million words in its possession while another organization has 6 billion words in 23 languages. In short, the amount of text resources in this group of organizations is huge.

Revising your data is a way to ensure that your resources remain of a high quality. Four respondents inform that revision takes place often while three say that revision is done rather rare. The coverage in terms of language pairs of text resources is the same as the 'Translation language pairs'. Three of the respondents reply that their organizations have access to huge numbers of text resources not in TMs. Two of these respondents refer to Acquis Communautaire (the body of EU law) as their text source and claim that these resources are sentence aligned. The most important criteria in connection with structuring their text resources are, version (date, year) (4). Subject domain (2), customer (2), and title/product name (2). The file formats of stored text resources are, TMX (4), MS-Word (2), PDF (2), and XML (1).

In broad terms, the document flow is typically taken care of in three consequtive steps. First, a department in the organization receives the document to be translated and then it is either outsourced (2) or sent to the in house translation department (3). After having been translated, the documents are then finally returned to the management department. No specific data management tool in order to detect corrupted data, is used. Five of the respondents state that the translations made are validated by at least one additional person. Based on the replies about data administration, it seems as if no distinction is made between 'Reciept of text' and 'Data administration'. All the respondents are conducting manual proof-reading in their validation of data.

3.2.4 IPR

The feedback from the interviewees reveals that all the organizations own their data resources.

3.2.5 Definition of a good MT system

The most frequently criteria mentioned here are, translation quality (4) and interoperability (2), and time (i.e. cost) saving (2).

3.2.6 Upload of parallel corpora

In connection with upload of files most of the respondents prefer that the file formats are the ones represented in the Microsoft Office package. Tmx is also mentioned. Only two of the respondents think it would be useful to be able to detect corrupted data automatically.

Regarding the idea of sharing data, the replies can be divided into three groups, *yes* (3 including already public data), *no* (4), and *unsettled* (2).

Three of the respondents would appreciate an online alignment tool. The already available data on the internet can of course be used for any purpose (2). Only one respondent is willing to provide data for any purpose. An additional respondent is willing to share data in a user group for data. One more is included in the category of 'Research only'

3.2.7 LetsMT! resource metadata

A majority of the respondents are positive to having access to as many metadata as possible at least no one rejects any of the metadata categories

3.2.8 Feedback

Two respondents think that giving feedback to specific text resources would be subjective – can the assessment be trusted. Four of the interviewee find it useful or very useful. Three of the respondents would like to be able to rate specific text resources

3.2.9 Configuration

Six of the respondents speak in favour of being given information about the text volume for training a reliable SMT system. Four respondents would to have access to information about the system configurations used to the development of the SMT system. Two doubt that there will be enough time to get acquainted with the detail of such a system. Four of the interviewees would like be able to configure or to make experiment with the available SMT system.

3.2.10 Website for translation

Not surprisingly, the feedback regarding the LetsMT! website for translation is more or less a copy of what the organisations replied to the questions about their translation tasks.

3.2.11 Translation widget

The replies from the group ’Organizations with multilingual translation needs’ with respect to having translations from specific websites of business and financial news, are distributed as follows, *mayby/no answer* (4), *yes* (3), and *no* (3). The lack of interest from some of the organizations is due to worry about the translation quality and that other sections in their organization are translating these text types. The domains and the language pairs are to a large extent a copy of the answers in connection with the organizations’ translation tasks.

3.2.12 Browser plug-in

Most of the respondents’ replies on whether speed is an important parameter in automatic translation is, *no* (5). Concerning file formats and language pairs, the answers resemble the ones given to the questions about their translation tasks

3.2.13 Integration with other tools

The answers are distributed as follows, not answered (5), n/a (2), yes to Microsoft Office package (2), SDL Trados (1).

3.2.14 Conclusion

The group, ’Organizations with multilingual translation needs’ uses in comparison with the LSPs less CAT tools. The most used tool by the respondents is SDL Trados.

Unlike the LSPs, this group of organizations does not offer translation services. As a consequence of this fact, the obstacles in terms of unwillingness to share parallel data with other organizations are less noticeable. More organizations in this group state that they see no hindrance for uploading their data in a common repository.

It should, however, be noted that for two of the respondents, the data they find sharable are already available on the internet (Acquis Communautaire).

3.3. News agencies

This category comprises 6 agencies/companies offering news and other types of information. Four interviews were conducted by the Dutch partner with interviewees from *Dow Jones*, *The Times*, *Nomura International* and *Thomson Reuters*. The remaining two interviews were conducted by the Danish partner with interviewees from *Schultz Information* and *Infopaq International A/S.*

Job profiles of interviewees include an editor/translator, an executive director and some business reporters. Job tasks involved in the interviewees’ work are for example quality assurance, administrative tasks, development of user specifications and customer relationship innovation.

Some of the interview questions have not been applicable or relevant for a number of the respondents in the *News agency* group. Therefore in the following description of interviewee responses, the obtained number of answers is mentioned for all questions where only few interviewees have given an answer.

3.3.1 Use of CAT tools

None of the organizations in this group employs translation memory systems. One organization uses Language Weaver (a statistical machine translation system) and two organizations store their terminology; one organization stores terminology in the i-term system (a regular term coding tool) and the other in Factiva (a business information and research tool including information management features). Factiva has a web service API potentially enabling interaction with the LetsMT! platform. The use of online language resources is very limited in the group as a whole. The very limited use of CAT tools in this group also entails that the organizations have very limited CAT tool experience to report from.

Two interviewees employ Firefox and Internet Explorer, the rest have not answered this question.

3.3.2 Translation tasks

All questions describing translation tasks are answered by three interviewees.

Translation domains reported by interviewees include law, medicine, newspaper material and economic news. The language pairs of translation tasks usually embrace English as either source or target language and the other language is Czech, Croatian, Danish, Dutch, German, Polish, Slovakian or Swedish. Reported translation volumes are very diverse. One interviewee translates around 100,000 words per year, another translates around 1.8 million words per year. Mentioned file formats of translation texts are Microsoft Word, Excel and PowerPoint, Adobe PDF, html and xml.

3.3.3 Text resources

All questions about text resources stored in the interviewee's organization are answered by two (in one instance three) interviewees.

Regarding size of text resources, one interviewee reports storage of 30-40,000 sentences within the medical domain, the other reports storage of 500-1200 words per news item, and the third interviewee says that no texts are stored.

One interviewee mentions that the stored 30-40,000 sentences are often revised, that source and target sentences are stored in an xml-format and aligned, and that English is the target language while source languages cover more than 40 different languages.

As indicated by respondents of this user group it seems that text resources may be rather scarce. In attachment C is included a description of this challenge together with suggestions for solutions. This description represents an internal perspective.

3.3.4 Localization/translation workflow

The questions in this section describing the organization's workflow are answered by two interviewees.

Both interviewees have data management tools (one an Oracle application, the other unspecified) and the translation process involves only the translator.

3.3.5 IPR

This question is answered by two interviewees. Both have restricted IPR of text resources.

3.3.6 Definition of a good MT system

This question is answered by two interviewees.

Both mention translation quality as the most important issue. Other issues mentioned are the price of the system together with integration features.

3.3.7 Upload of parallel corpora

Questions concerning requested and desired LetsMT! system facilities for upload of parallel corpora are answered by two interviewees.

Both interviewees suggest xml as the file format for upload data; one thinks that a data management tool in LetsMT! would be a good idea, the other considers this irrelevant. Both interviewees foresee problems in connection with data sharing, but one is willing to consider data sharing of stored data for research purposes.

3.3.8 LetsMT! resource metadata

This section of questions covering metadata requirements for corpora in the LetsMT! platform are answered by two interviewees.

Both interviewees are interested in all metadata types mentioned in the question template (language pair, source language identifier, domain, text type, data owner, data provider, upload date, text production date, alignment type – for more information see attachment A).

3.3.9 Feedback

Questions about feedback/rating facilities in the LetsMT! system are answered by two interviewees.

Both interviewees would be interested in feedback and rating facilities.

3.3.10 Configuration

Two interviewees answered questions concerning access to system configuration information.

One interviewee would be interested in such information; the other would not as this type of information was estimated as too technical.

3.3.11 Website for translation

Questions about the LetsMT! website for translation are answered by 4 interviewees.

The organizations will typically be interested in the following domains: law, medicine, education in general, news paper articles and economic news. Language pairs always include English together with Czech, Croatian, Danish, Dutch, Polish and Slovakian. Typical file formats to be translated via the website are xml, Microsoft Word, regular txt-format, djnml and rdms.

3.3.12 Translation widget

Questions about the translation widget are answered by 5 interviewees.

Three interviewees think a translation widget would provide quicker updates and generally added value. The other two organizations are not interested in a translation widget. Language pairs relevant in terms of a translation widget always include English together with Czech, Croatian, Danish, Dutch, Polish and Slovakian. Domains relevant in connection with this translation service are law, medicine, education in general and economic news.

3.3.13 Browser plug-in

This section of questions about a browser plug-in, which provides instant translation of web pages, is answered by three interviewees.

Interviewees are generally not particularly interested in a browser plug-in. All mention xml as a typical file format.

3.3.14 Integration with other tools

This question is answered by two interviewees.

They would appreciate integration with i-term, ORACLE, Microsoft Word and other text editors.

3.3.15 Access to resources in LetsMT!

These questions are answered by two interviewees.

One interviewee is not interested in access to LetsMT! resources, the other thinks additional knowledge is always desirable.

3.3.16 Conclusions

Six interviewees form the basis of the *News agency* group, and half of the interview questions are answered by only two interviewees. As mentioned above some of the questions were not applicable to a number of the interviewees in this group; probably because some of the interviewees have rather small translation volumes. The number of answers within the *News agency* group is of course an inadequate basis for generalization of user requirements..

Having mentioned that, it can be noted that this user group does not have much CAT tool experience, their text resource volumes are relatively small and they do not necessarily have the IPR of text resources. Regarding requirements to the LetsMT! platform, the preferred file format for upload data is xml and data sharing could possibly turn out to be a difficult issue. Features of the LetsMT! platform that will be interesting for this user group are the *website for translation* and possibly the *translation widget*. The *browser plug-in* does not seem to capture that much interest.

3.4. Research organizations

This category comprises 6 research organizations. One interview was conducted by the Croatian partner with an interviewee from *the University of Zagreb, Faculty of Humanities and Social Sciences;* another interview was conducted by the Danish partner with an interviewee from the *University of Copenhagen, Faculty of Hunmanities* and a third interview was conducted by the Swedish partner with an interviewee from *Linköping University*. The remaining three interviews were conducted by the Latvian partner with interviewees from *University of Latvia, Ventspils University College* and another research institution not mentioned by name.

Job profiles of interviewees include researchers, academic users and teachers. Job tasks involved in the interviewees' work are for example teaching, research, translation and use of CAT tools.

3.4.1 Use of CAT tools

Three organizations in this group employ TRADOS translation memory system and one of these organizations employs the MultiTerm terminology tool as well. One interviewee mentions machine translation systems as Moses, ITG, Berkeley Aligner, GIZA++, Phrasal, HIERO and Joshua, all of them for research purposes; another interviewee mentions Google translate. Online language resources used by interviewees are for example termnet.lv, EU pages, www.letonika.lv, other online term databases and dictionaries. Browsers used in interviewee organizations are Firefox and Internet Explorer.

Mentioned advantages in connection with use of CAT tools are speed, history tracking, standard compliance and one-stop-shopping. None of the interviewees have experienced technical problems with CAT tools.

3.4.2 Translation tasks

Translation domains reported by interviewees include law, science, business, economics, transportation, environment and domains within the general vocabulary. Language pairs of translation tasks often include English as either source or target language and the other language belongs to a long list. Reported translation volumes are usually small as translation as a product is not a core activity for the research institutions. Mentioned file formats of translation texts are Microsoft Word, Excel and PowerPoint, Adobe PDF, html and rtf.

3.4.3 Text resources

Research institutions usually have rather limited text resources and revisions are therefore also very limited. File formats of text resources are Microsoft Word, Adobe PDF and TMX. Language pairs of text resources stored in the interviewees' organizations are: Latvian/English, English/Latvian, Latvian/German, German/Latvian, French/Latvian, Latvian/French and English/Swedish. (This question only received responses from Latvian and Swedish interviewees).

3.4.4 IPR

This question is answered by two interviewees. One states that IPR usually belongs to the university itself, the other reports that IPR belongs to the particular client.

3.4.5 Definition of a good MT system

Suggestions for criteria that define a good MT system include high translation quality, usability, high speed, linking to online resources, fluency and modularity (to ensure transparency of component strengths and weaknesses).

3.4.6 Upload of parallel corpora

Interviewees suggest different file formats for upload of data. One interviewee is satisfied as long as standards are observed; others prefer Microsoft Word, Adobe PDF and TMX while one of the interviewees specifies that TMX is not a preferred format.

Two interviewees expect that they will be willing to share data for research purposes while the remaining interviewees are either reluctant or have not answered this question.

Interviewees from research organizations would generally appreciate if the LetsMT! platform included an online alignment tool.

3.4.7 LetsMT! resource metadata

Interviewees are generally interested in the possibility to use metadata, but not necessarily all the metadata types mentioned in the question template (language pair, source language identifier, domain, text type, data owner, data provider, upload date, text production date, alignment type – for more information see attachment A).

3.4.8 Feedback

Half of the interviewees would be interested in feedback and rating facilities, the others would not.

3.4.9 Configuration

The organizations would appreciate access to information about learning curves for training of an SMT system. A couple of the interviewees would also appreciate access to system configuration information and configuration options.

3.4.10 Website for translation

The organizations will typically be interested in the following domains: law, business, transportation, environment and economics. Language pairs always include English, Latvian, Russian and Swedish. Typical file formats to be translated via the website are html, Microsoft Word, Adobe PDF, rtf and regular txt-format.

3.4.11 Translation widget

The organizations are generally not very interested in a translation widget.

3.4.12 Browser plug-in

Interviewees agree that translation speed is a very important issue; they agree that html is the preferred file format and a large number of language pairs are suggested.

3.4.13 Integration with other tools

One interviewee suggests integration with SDL TRADOS.

3.4.14 Access to resources in LetsMT!

One researcher mentions that he would prefer to work with simple formats: raw text, alignments in GIZA++ or ACL format. He thinks sufficient resources are available for some domains in European languages, but would welcome publicly available resources for SOV languages, in more domains and language pairs where none of the languages are English. Another researcher mentions that he would like to use the LetsMT! resources for student research work.

Regarding access to trained SMT systems for research purposes, one interviewee answers that he is interested in new techniques for realignment, reordering, decoding or new language modeling techniques. If LetsMT! could provide interesting datasets with pre-computed baseline results, e.g. MOSES with near-optimal parameter settings, that would be a really nice feature for research purposes.

To the question whether interviewees would appreciate access to the sets of training data used to develop the language models, three answers that it would be useful.

3.4.15 Conclusions

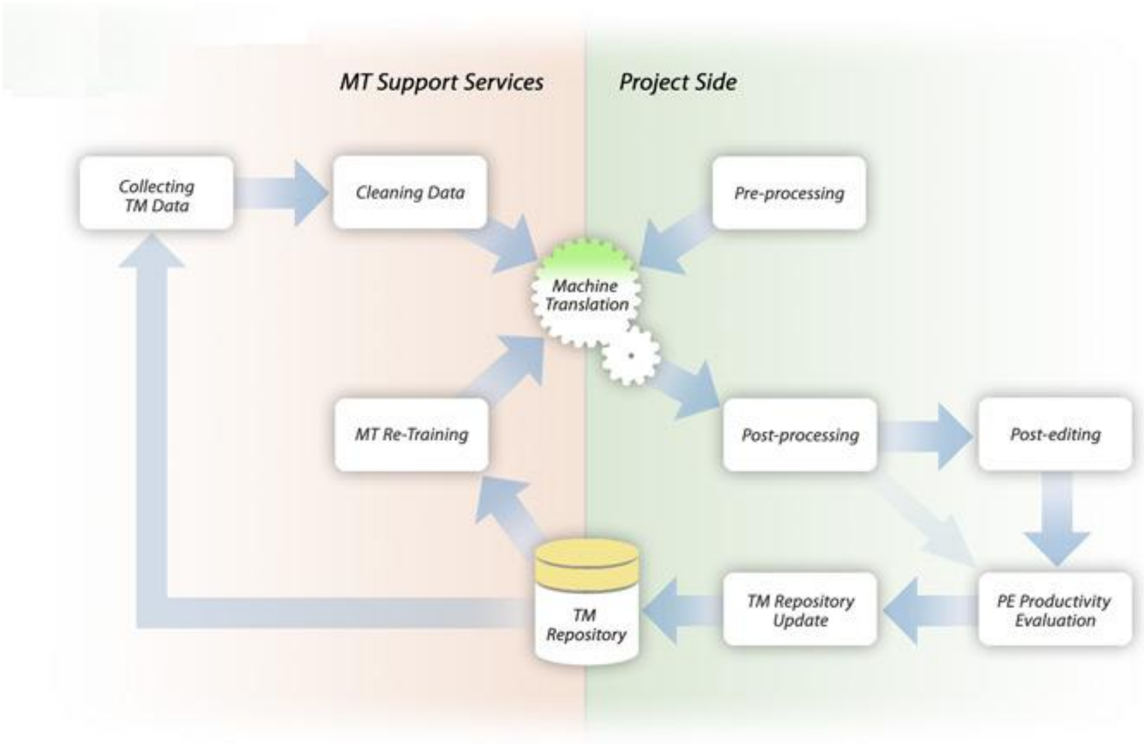
Six interviewees form the basis of the *Research organization* group. Interviewees of this user group generally employ CAT tools together with other tools, and they do not experience technical problems in this respect. The organizations' text resource volumes are relatively small and they do not necessarily have the IPR of text resources. Regarding requirements to the LetsMT! platform, interviewees have suggestions for overall success criteria which include usability, fluency, modularity and linking to other online resources. A preferred file format for upload data does not exist as different opinions are presented. Data sharing could turn out to be a difficult issue, but some are prepared to share with other researchers. Some of the researchers would appreciate access to LetsMT! resources for research purposes; especially access to the sets of training data used to develop the language models would be interesting.

4. Descriptions of LetsMT! Use Scenarios

This section concerns descriptions of two usage scenarios as seen from an internal perspective together with recommendations. Topics discussed in this section are for example collection of training data, MT training, pre- and post-processing and post-editing.

4.1. Localization Use Scenario

This is a description of the Localization MT usage scenario for D1.1. The description follows the below diagram:



Process of using MT in L10N consists of the following sub-processes. The subprocesses are not analyzed and modelled. A high level description is given with important caveats that will have impact on LetsMT! Functional Specification.

4.1.1 Collecting Training Data (Collecting TM Data)

In localization (L10N) TM (CAT) technology has been ubiquitous since late 20th century. Hence, unlike other industries, L10N typically has well aligned legacy bilingual corpora.

However there is a big legal challenge. TMs that resulted from CAT projects were generally produced as work for hire and are generally considered ownership of customers who ordered the translations. Using TMs for MT is not the main legal challenge because in MT training you are not using the customer property directly. You are rather using statistical and syntactic patterns that are derived from the original content but constitute separate IP. This is considered fair use and does not construe an IPR violation.

However, due to other contract patterns common in L10N confidentiality is another frequent issue. TMs are often classified as confidential materials not to be used.

Recommendations

- It is very important to have safe legal framework for collecting TM data. Collaboration with data-sharing organizations that have such a legal framework is highly advisable.
- Functionally, API integration with legally safe data owners such TDA should be developed.
- Functionally, LetsMT! Must be able to track legal metadata of training materials
- Most important formats for data collection in L10N scenario are TMX and XLIFF
- TBX, csv, and tab-delimited are important for collecting terminology data

4.1.2 Cleaning data

L10N data typically contain loads of code. The code goes in two varieties meta-segment and in-line. Whereas meta-segment mark-up is easily filtered out the in-line elements constitute a cleaning challenge. MT training platforms are typically unable to deal with un-natural language code in the training phase. Throwing-out data with mark-up would render the engines virtually unusable in L10N scenarios.

Variables are the toughest part of in-line mark-up. Typical formatting tags can be filtered out relatively easily without distorting the general meaning of segments. But variables are usually placeholders for core-information. Therefore we cannot just filter out placeholders. In case no working substitution algorithm is found these segments must be thrown away not to pollute the training corpus.

Factoids such as figures, dates, product names, currency symbols etc. need to have special handling. We know that MSR have developed a proprietary solution for dealing with factoids, the algorithms are however unknown.

Recommendations

- Develop automated easily configurable cleaning scripts for most common standard formats such as TMX, XLIFF and TBX.
- Develop configurable scripts for dealing with variables and factoids.
- Pay attention to L10N specific standard formats (TMX, XLIFF, TBX, csv etc.) during data definition.
    - The data definition spec should be robust, considering major flavours of standard formats.

4.1.3 MT training an re-training

MT in L10N can only be successful if highly specific engines are developed. The MT training capability should be exposed to a technically skilled end user (production coordinator in a L10N company or department) through simple API and GUI.

The end user performing data based training should have integrated access to automated evaluation results for diagnostic purposes.

The training capability should allow for forking engines, such as in the following case

General IT

➔ CRM -> [forks for clients] -> [forks for product lines]

➔ ERP -> [forks for clients] -> [forks for product lines]

➔ Security -> [forks for clients] -> [forks for product lines]

➔ HW -> [forks for clients] -> [forks for product lines]

➔ Telecoms -> [forks for clients] -> [forks for product lines]

➔ Etc.

Medical

➔ Devices -> [cardio, blood, etc.] -> [..] -> []

➔ Pharma -> [cancer, trials, etc.] -> [..]

➔ Healthcare -> [..] ->..

➔ Etc.

Automotive

Etc.

Incremental retraining based on post-edited data (TM repository) should be exposed as a standard functionality, through simple API and GUI. It should be possible to perform incremental retraining by up to 2 mill new words within minutes or up to a couple of hours rather than days.

Recommendations

- Expose training and retraining capability through simple API and GUI
- Allow hierarchical management (ordered by parent-child relationship, by domain inclusion etc.) of trained engines
    - GUI and API exposure of the forked engines in a clear tree structure (based on the above hierarchical relationships)
    - Simple querying and filtering to avoid navigation through complex trees
    - Multitenant
- Advanced access management (both data and SMT models) – at least two dimensional (role and group based)
    - Role can be for instance customer, engineer, PM, admin etc.; groups can contain users in different roles and can themselves be ordered by inclusion. Groups can either be mutually invisible, or have different levels of trust among them.
- Advanced meta-data management, including legal

4.1.4 Pre-processing and Post-processing

Assuming that current MT platform cannot process in-line mark-up as part of training, and hence cannot learn how to place the mark-up as part of the training, it is nevertheless vital to ensure mark-up preservation throughout the L10N process in spite of inclusion of an MT step. All CAT integrations must ensure seamless and lossless roundtrip of mark-up and metadata. Tools that produce recalcitrant flavours of formats should be explicitly excluded and their exclusion publicized. Special attention must be paid to variables and factoids as discussed under “Cleaning data”.

So the production pre-processing must be able to recognize factoids and process them according to strict rules. It must strip or replace all mark-up but store it to attempt reapplication during post-processing.

Recommendations

- Process factoids through separate rules
- Strip mark-up from source -> Reapply mark-up on target
    - This applies to meta-markup and formatting in-line mark-up
- Replace mark-up indicating placeholders in source -> Reapply placeholder mark-up on target
    - This applies to content placeholders only
- Alternatively: process placeholders through separate rules (similar to factoids)
- All pre- and post-processing steps must be automated and end-user-configurable

4.1.5 Machine Translation

Optimise performance. It should be capable of translating roughly millions of words within minutes. This can be a challenge in a massive multitenant environment. This function should be transparent in the L10N scenario. The user should see informative progress bars for all operations taking more than 3 seconds.

Recommendation

- Transparent and quick. Progress bars.

4.1.6 Evaluation

Automated and human evaluation is very important for managing the life-cycle of specific trained engines.

Major automated evaluation metrics – BLEU, TER (including language specific where available), METEOR (including language specific where available) should be built in or at least integrated as quick and reliable third party web services. These metrics must be available in real time for samples of up to 100K words to allow for hill-climbing monitoring during training. The end-user MT trainer should be able to rapidly assess the efficiency of added training data with respect to the test set.

Ideally the GUI should provide a human evaluation editing grid that would support common and custom human evaluation models (categories of errors). The GUI should allow for upload of human reference sample, and showing it in the editing grid with along with source and MT candidate.

It should be possible to store automated and human evaluation results for later reference and analysis, ideally through an OLAP cube. It should be possible to track the quality of the engine based on added training material, new content in production, renewed human evaluation, ongoing editing distance etc.

Recommendations

- Built in TM systems automated metrics tested for performance
  - BLEU, TER (including lang specific), METEOR (including lang specific)
  - Allow for storing and managing results
- Human evaluation grid
  - Display source, MT candidate and human reference
  - Allow for upload of custom error category models
  - Allow for storing and managing results
- Analytics services (OLAP)

4.1.7 Post-editing

Post-editing is very important in L10N MT scenarios. It is important even in scenarios were raw output of the **final** trained engine is intended for publication. In the raw publishing scenarios the MT quality typically must be higher than in classical post-editing (MT-human hybrid) scenarios. You can only achieve good enough quality in raw-output-publishing scenarios through human feedback, i.e. post-editing and incremental retraining.

In human quality publishing scenarios integration with CAT tools is a must, because only segments that do not have good TM matches (based on a configurable threshold) are typically sent to MT. Translator/Post-editor typically edits the TM and MT suggestions at the same time. Therefore it is critical for post-editing scenarios to integrate MT suggestions in major CAT tools, such as Trados, Worldserver, etc. Generic XLIFF integration prototype should help build solutions for other XLIFF capable CAT solutions such as OmegaT, memoQ, Heartsome etc. The seamlessness and user friendliness of any such integration will depend on the capability of the target tool to receive, carry and display MT suggestions along with its own TM matches.

Post-edited translations should be automatically stored in a TM repository (ideally as increment to the original training data) to be used from time to time for MT retraining.

Ideally, the end-quality post-edited strings should feedback directly into the MT's retraining capability.

Recommendations

- Develop a generic XLIFF round-tripping prototype
- Develop a generic TMX round-tripping prototype
- Develop a couple of specific CAT tool integrations, full roundtrip with correct and transparent display of MT suggestions
- Automate storing of post-edited strings directly into TM repository
- Configurable trigger for retraining

4.2. Financial News Use Scenario

The LetsMT! consortium has identified press releases to be an ideal source for LetsMT! data, specifically for the online translation service of business and financial news. International press releases are almost invariably distributed in English to the international business community. However, as many companies registered in non native English countries are required to release important business news in their native language, a large percentage of business press releases are available in two languages.

4.2.1 Collecting Training Data

To be useful for the LetsMT! project, the parallel corpus should ideally consist of about 1 million words in both languages.

To estimate the size of the available parallel corpus in the targeted small languages (Dutch, Swedish, Danish, Czech and Polish) the following holds.

A major businesses provider's newsfeed ranges between 5000 and 10000 messages per day resulting in 2 to 4 million messages per year. However, only a fraction of these are press releases and for that reason likely to be available in multiple languages. Commercial news providers archive of "significant development", consisting of relevant press releases translated to English consists of about 30.000 messages per year. This corresponds to at least 30.000 original press releases.

The following table provides an overview over the last three years:

Year	2006	2007	2008
Number of significant press releases	23.413	30.476	30.932

The majority of European companies are required to submit press releases in their native (non-English) language. Therefore many press releases in the commercial news feeds are English translation of the original press release, either made by the provider of the feed or by the releasing company.

If the translation was done by the provider of the news feed, the original language message is also available in the newsfeed. As stated before, this will be a minority of the news messages. Unfortunately exact figures are unspecified, but an analysis of the language of the news messages provides the information in Figure 1.

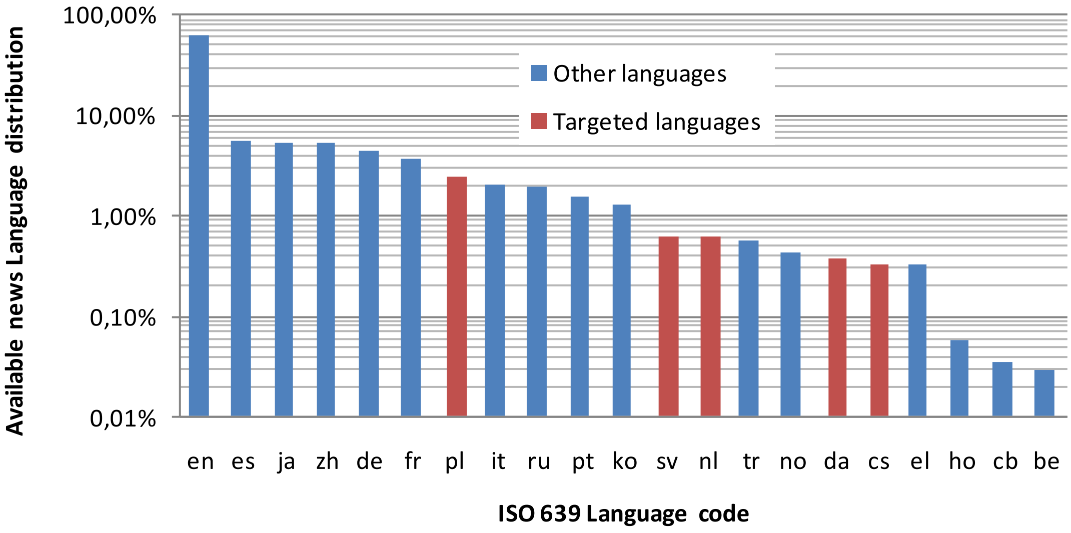


Figure 1: Schematic language distribution of targeted languages in business news streams from major commercial news providers.

For the targeted languages the following corpus sizes are expected:

Language	Dutch	Swedish	Polish	Danish	Czech
Number of press releases per year (approx).	11K	11K	43K	6.8K	6.0K
Yearly corpus size in words (max.)	2.3M	2.3M	8.7M	1.4M	1.2M

For the majority of cases, the translation was done by the company themselves, before submitting the press release to the news provider. In this case either the company dissemination (their website) or the local national press release agency provides the original language press release. Listed companies normally provide an overview of their press releases from their corporate websites. National press agencies also support access to the releases made through their service. Full texts are available only for a fee to registered users. In addition also archived data is commercially available. As timing is critical for press release data, the tie stamp of the message is a reliable indicator by which the original language document can be matched to the English translation.

A minimal initial estimate of the obtainable corpus for the Dutch language are 40-50 releases per working day with an average of 200 words per release, which yields about 2.3 M words per year. Significantly larger estimate holds for Polish whereas Swedish is similarly represented. Yearly releases for Danish and Czech languages are expected to be smaller, but still exceeding the 1M words threshold. In addition, by using multiple year archives, the required corpus size of 1 million words is easily surpassed.

4.2.2 Recommendations for Financial News Use Scenario

Following assumptions should be taken under consideration:

- computer readable archive or stream
- relevant news (business/finance related press releases)
- accurate timestamps

- accurate company codes (ISIN, stock ticker, RIC or similar)
- sufficient target language coverage (see above)
    - English – Dutch
    - English – Polish
    - English – Swedish
    - English – Danish
    - English – Czech
- minimum total corpus size 1million words

5. Conclusion

In this section we will elaborate on main findings of interviews which as mentioned represent the external perspective in terms of user requirements (attachment C). All conclusions are supported by diagrams reflecting the distribution of particular answer types within the user groups. Where appropriate, we will compare main findings of external user groups with findings and recommendations of the usage scenarios of section 4 which represent the internal perspective. Finally, we will emphasize the most significant recommendations based on the internal as well as the external perspective.

5.1. Summary of requirements for upload and handling of parallel corpora – external use scenario

Generally, the interviewee organizations are relatively willing to and interested in sharing their text resources via the LetsMT! platform. The below diagram (figure1) reflects answers of all respondents including those that did not find this question applicable or relevant (for example because they do not have any data at the moment). Of all interviewees 44% are to some extent willing to share data even if half of them are reluctant, mostly because of IPR issues. Only 16% of all interviewees are completely dismissive of the sharing idea.

Another aspect closely related to the organizations' ability to share data, is IPR. Figure 2 shows all answers, including the blanks. Interviewee organizations without IPR of text resources can of course not commit themselves to the sharing idea. This means that a larger subset of the organizations might be able to share data if they get the opportunity to solve IPR issues with their customers or clients.

Many different file formats are relevant in connection with upload data. The most popular file format mentioned is TMX (46%), but also Microsoft Word (15%) and Adobe PDF (10%) are popular, see figure 3.

Metadata on corpora uploaded in the LetsMT! system can be used to select the most appropriate training material for MT training. Figure 4 shows to which extent interviewees prefer to have specific metadata information available in the selection process. All the "No answers" are returned by the same two interviewees. Apart from these two respondents, all respondents prefer to have the metadata categories listed in the diagram below. Other metadata are also suggested by different respondents, this list include:

- Usage Counts
- Availability (Corpus can/cannot be used for MT training)
- Contact Person at LSP
- MT usage history: where has the segment previously been used for training
- Version

- Tokenization
- Encoding

All the suggested nine metadata categories (figure 4) should therefore be specified for uploaded parallel corpora.

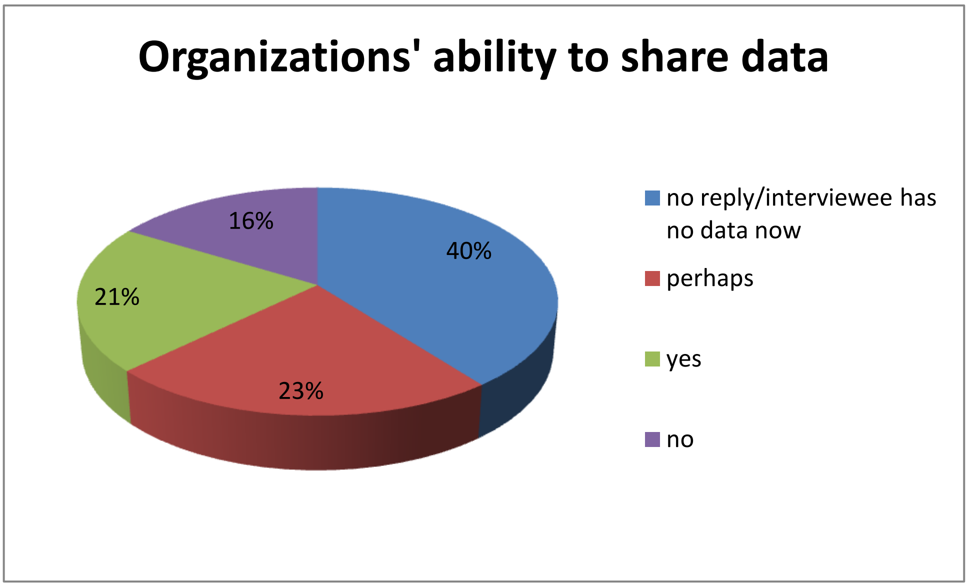


Figure 2 Data sharing

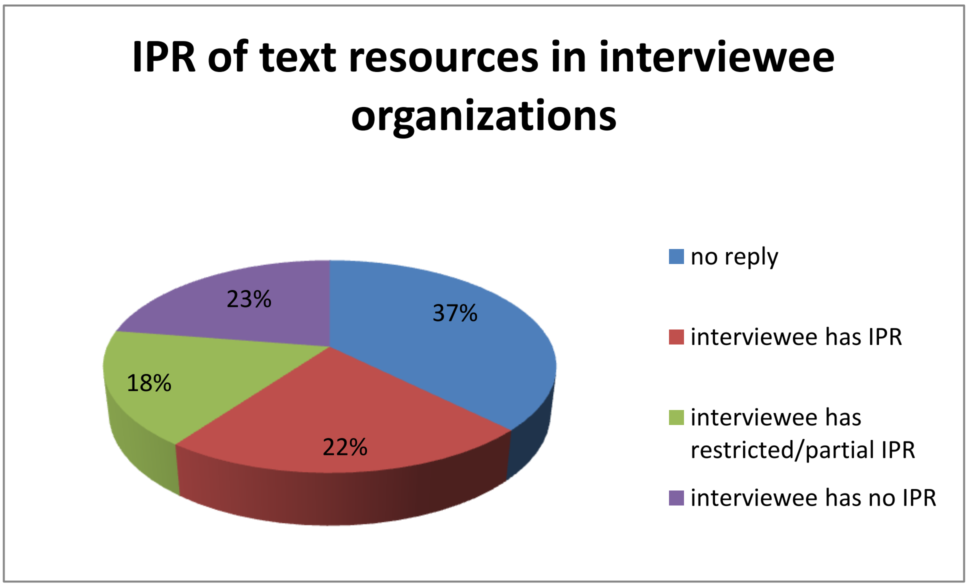


Figure 3 IPR

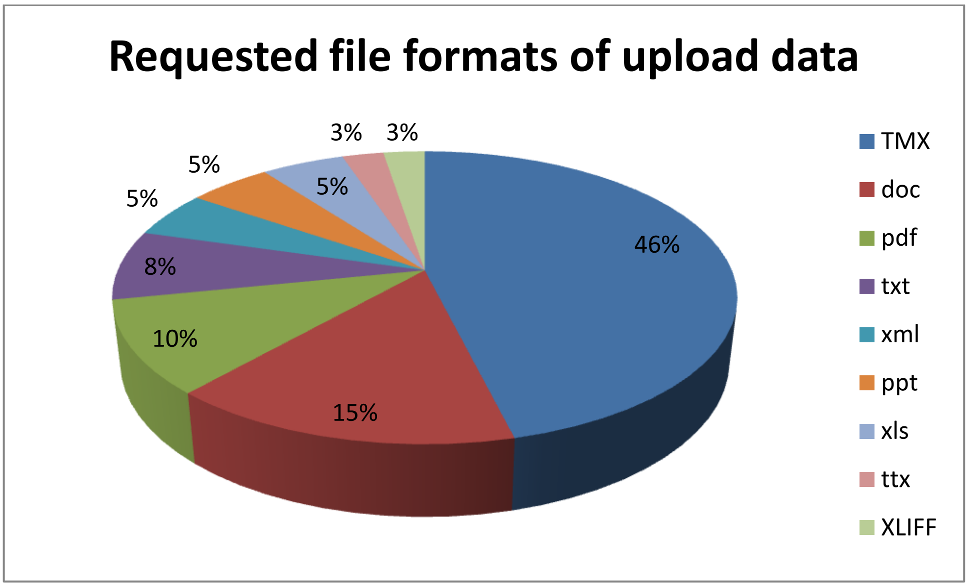


Figure 4 File formats – upload data

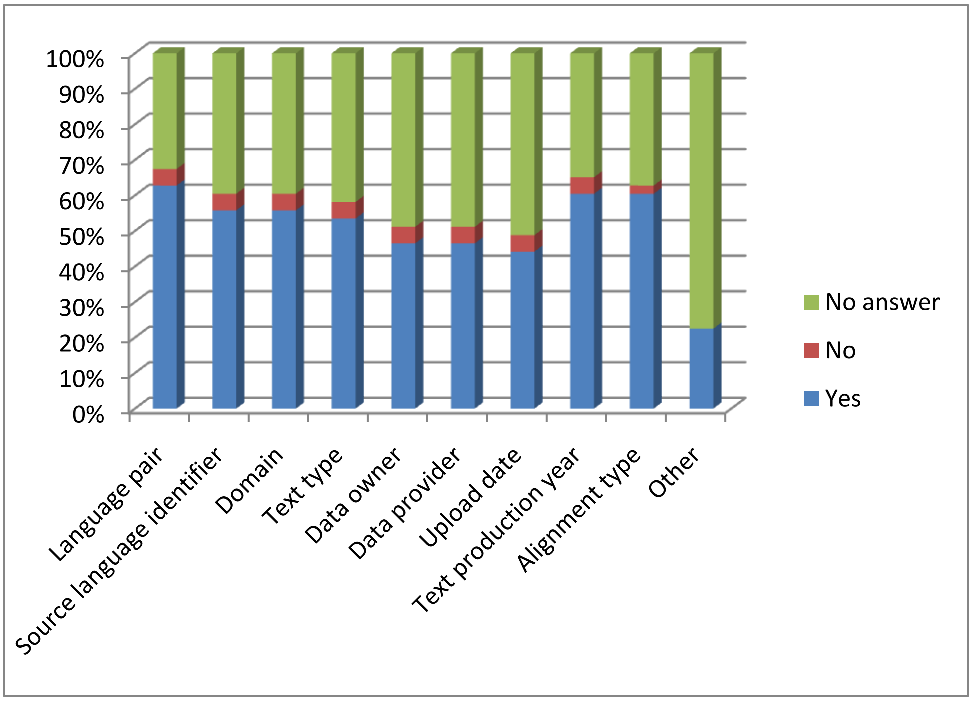


Figure 5 Required metadata for parallel corpora

5.1.1 External scenario compared with the internal scenario

Section 4 of this report describing the internal view also points out that IPR issues may be difficult and that successful data sharing requires collaboration with organizations that have control of IPR. Concerning formats for data collection, it is considered that the most important ones are TMX and XLIFF which is - as regards the TMX format - in line with requests of the external user groups.

5.1.2 Recommendations for upload and handling of parallel corpora

It can be concluded that the LetsMT! system design should give high priority to IPR issues. The LetsMT! platform must be able to handle IPR issues so that users are able to upload and access their own data without having IPR violated - and when possible we should generally focus on collaboration with organizations that have IPR of data collections.

Requested file formats for upload data are primarily TMX and doc formats. Metadata requirements are language pair, source language identifier, domain, text type, data owner, data provider, upload date, text production year and alignment type.

5.2. Summary of requirements for Website for translation – external use scenario

Many different file formats are suggested for the website for translation (figure 5). Microsoft Word (29%), Adobe PDF (18%), xml (11%) and html (11%) are the most popular ones, but requests are more evenly distributed than in connection with upload data.

The domains most frequently requested for the translation website are law (21%), finances (16%), medicine (14%) and IT (9%) (figure 6).

With respect to language pairs requested by the users, it can be noted that all 23 European languages are mentioned and in all combinations. In addition, Norwegian is mentioned in combinations with Danish, Swedish and German; Russian is mentioned in combinations with English, Latvian, Estonian and French (for a full specification of language pairs see attachment C).

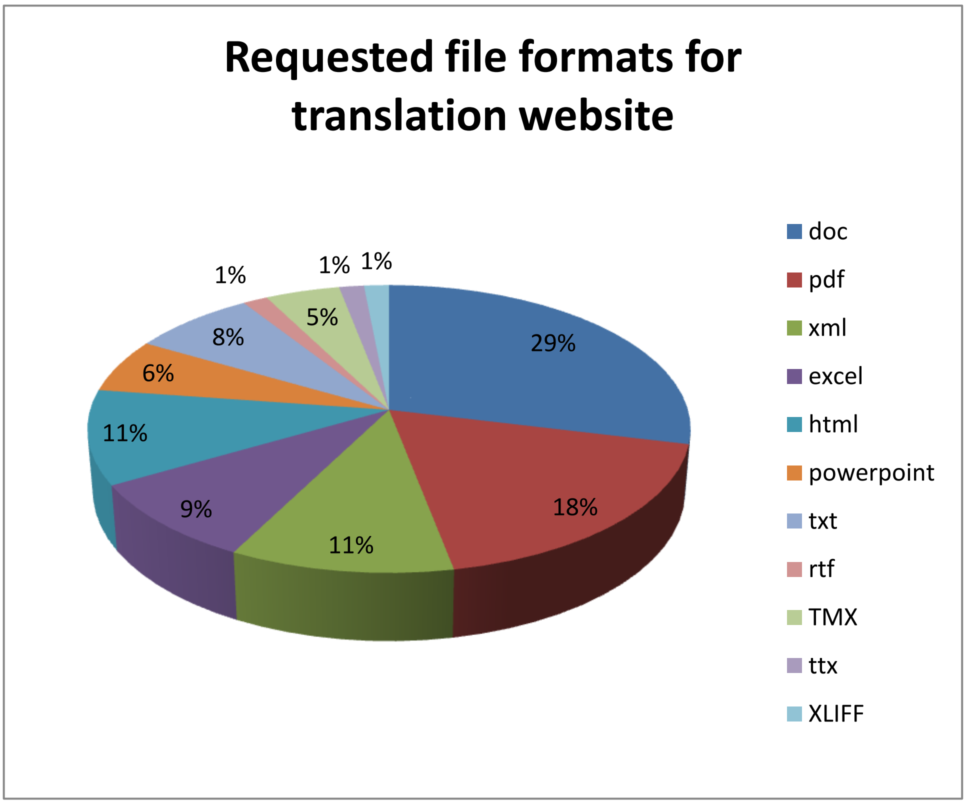


Figure 6 File formats – translation website

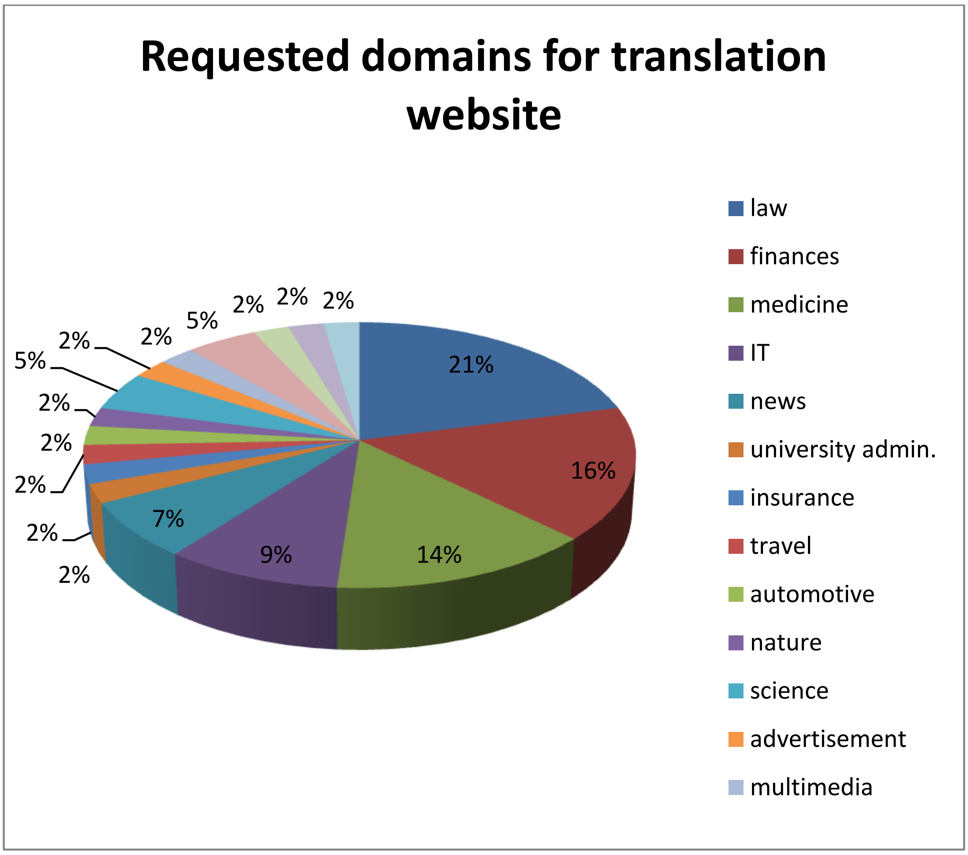


Figure 7 Domains – translation website

5.2.1 Recommendations for Website for translation

File format requests are rather evenly distributed over several different formats. The LetsMT! system should at least be able to handle doc and PDF, but together these formats constitute less half of the requests and other formats as xml and html could therefore also be considered.

The domains most frequently requested are law, finances and medicine, even if the users are generally interested in many different domains. Therefore it is important that the LetsMT! system is able to handle many different domains - and in connection with data collection domains as law, finances and medicine should get a high priority.

As regards languages to be covered by the system, the project will focus on small languages of the EU and a few other languages. The users' requests concentrate on these languages, but also in some combinations with languages as English, German, Russian, Norwegian and French.

5.3. Summary of requirements for Translation widget – external use scenario

Nearly half of all interviewee organizations are potentially interested in working with the translation widget (figure 7).

Requested domains for the translation widget are primarily IT (18%), law (13%), medicine (13%), finance (13%), education (9%) and science (9%) (figure 8).

Also in relation to the translation widget all 23 European languages are mentioned and in many combinations. In addition languages as Russian, Arabic, Norwegian and Chinese are also mentioned (for a full specification of language pairs see attachment C).

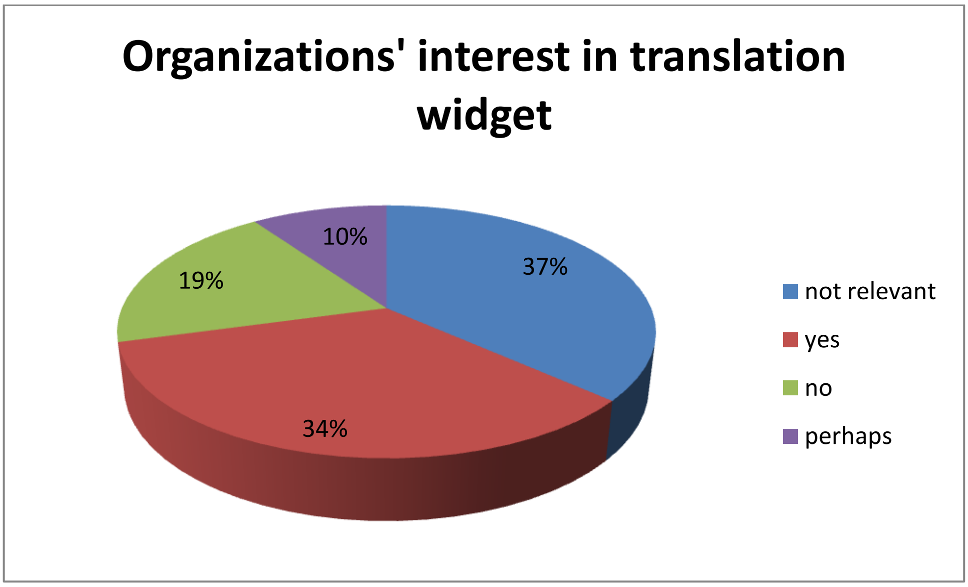


Figure 8 Interest in translation widget

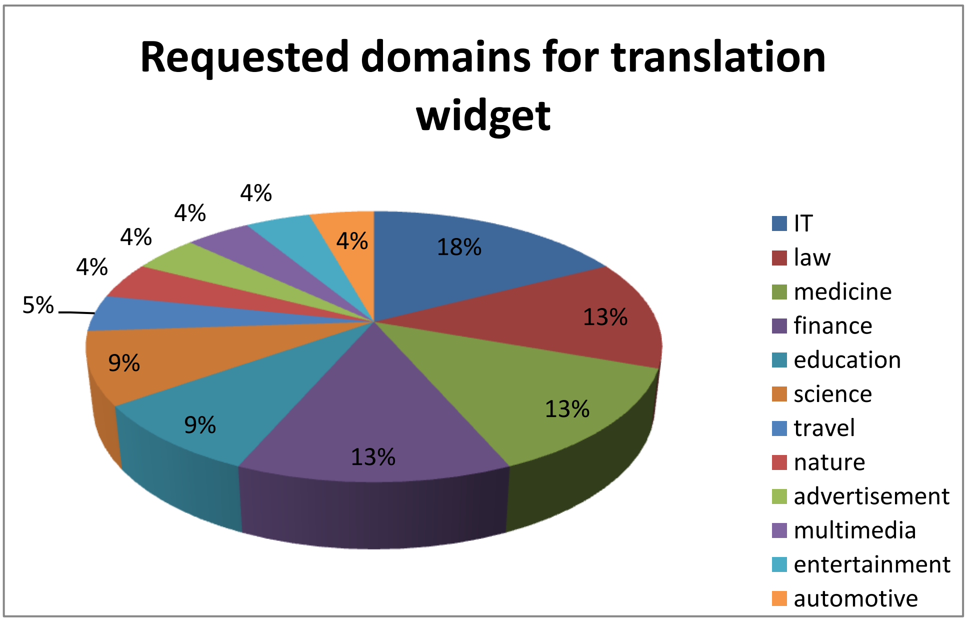


Figure 9 Domains – translation widget

5.3.1 Recommendations for Translation widget

Many users are potentially interested in the translation widget and will especially use it in connection with translation jobs outside their own specific domains. This probably means that the widget will eventually have to cover many different domains.

As regards languages to be covered by the system, the project will focus on small languages of the EU and a few other languages. The users' requests concentrate on these languages, but also in some combinations with languages as English, German and Russian.

5.4. Summary of requirements for Browser plug-in

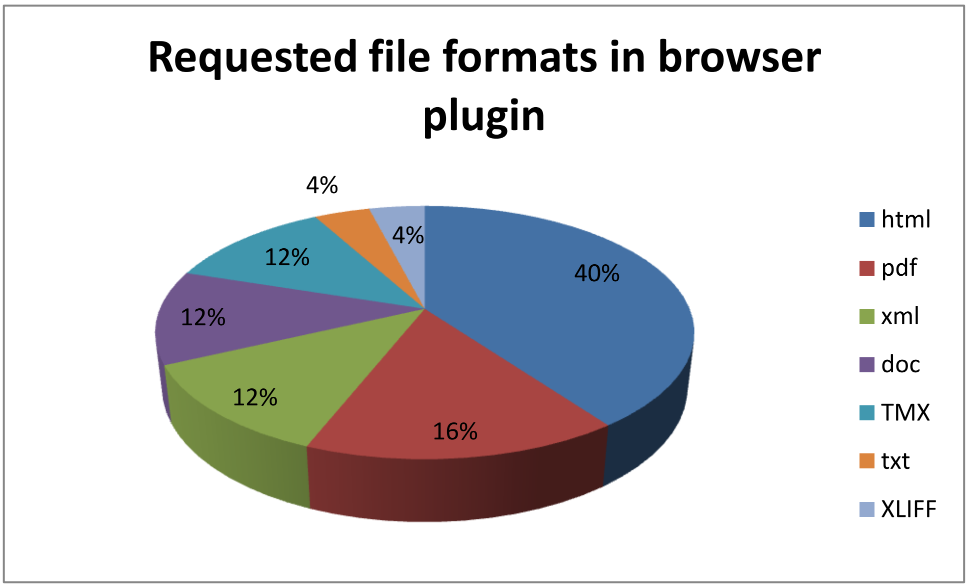


Figure 10 File formats – browser plug-in

Requested file formats for the browser plugin are primarily html (40%), but also pdf (16%), xml (12%), TMX (12%) and doc (12%) are requested formats (figure 9).

Also in relation to the browser plugin all 23 European languages are mentioned and in many combinations. In addition Norwegian is mentioned in combinations with Danish and Swedish and Russian is mentioned in combinations with English, German and Latvian (for a full specification of language pairs see attachment C).

5.4.1 Recommendations for Browser plug-in

Requested file formats for the browser plug-in are primarily html and pdf.

As regards languages to be covered by the system, the project will focus on small languages of the EU and a few other languages. The users' requests concentrate on these languages, but also in some combinations with languages as English, German and Russian.

5.5. Summary of requirements for integration with other tools – external use scenario

Nearly half of all interviewee organizations employ some version of the Trados system (nearly one third does not employ a TM system (figure 10)). The popularity of the Trados system also means that integration of the LetsMT! platform with the Trados system has a high priority among interviewees (29%). Many also think that integration with CAT tools in general is important (21%) and Microsoft Office also gets a relatively high priority (14%) (figure 11).

Figure 12 illustrates MT systems employed in the interviewee organizations today. 67% of interviewees do not use an MT system, 12% employs Google Translate and Language Weaver, Moses and MS Machine Translation Workbench are among other used MT systems. This interview result indicates types of MT systems generally used today and shows that MT systems are in use, but still to a limited degree.

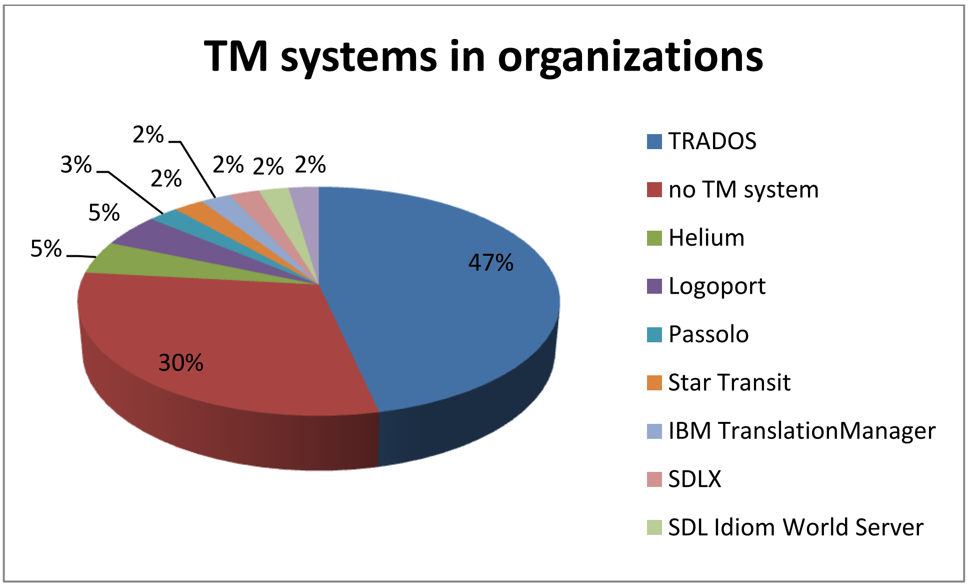


Figure 11 TM systems

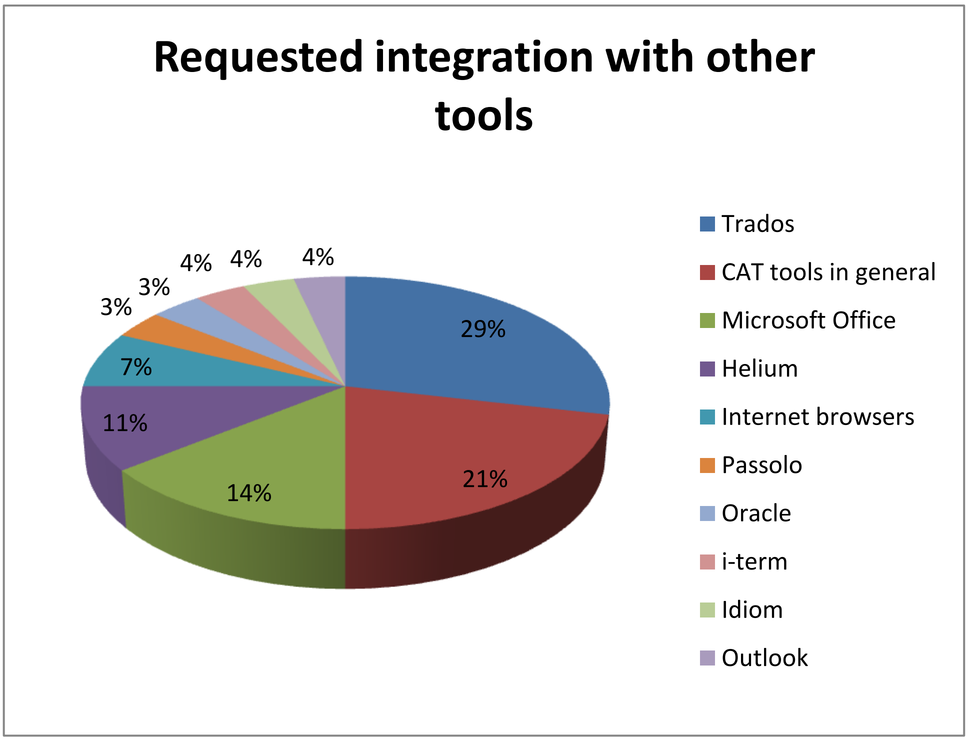


Figure 12 Integration

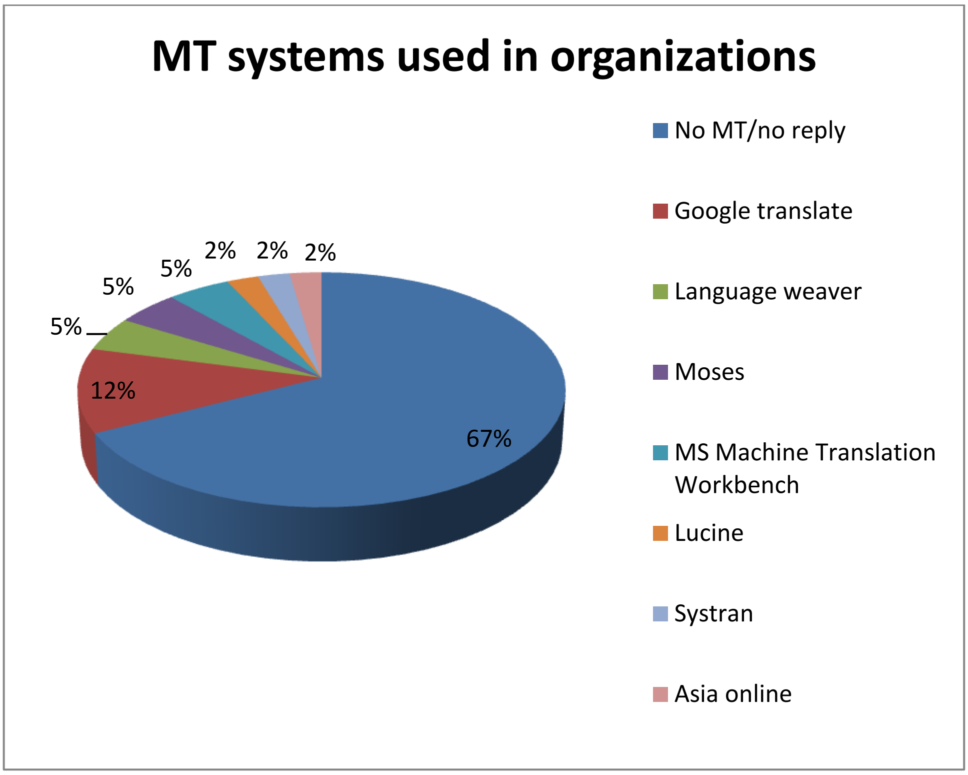


Figure 12 MT systems used

5.5.1 External scenario compared with the internal scenario

Recommendations of section 4 are in line with above requests of user groups and emphasize that integration with CAT tools is necessary as only segments with no/bad TM matches are sent to MT. In section 4 suggestions are that LetsMT! should at least be integrated with systems as Trados and Worldserver. It is also pointed out that usability of the integration depends on seamlessness and the capability of the system to receive and display MT suggestions along with TM matches.

5.5.2 Recommendations for integration with other tools

The LetsMT! platform must have seamless integration with Trados and possibly with Microsoft Office. Some users also mention that they are interested in integration with "CAT tools in general".

6. List of Attachments

6.1. Attachment A: Interview template

See external file D1_1_AttachmentA

6.2. Attachment B: Interview Answers for the four different user groups

See external file D1_1_AttachmentB

6.3. Attachment C: All Interview Answers

See external file D1_1_AttachmentC

See section 7b: Specification of language pairs for Website for translation
See section 8c: Specification of language pairs for Translation Widget

See section 8c: Specification of language pairs for Browser Plugin