MT-ECOGUIDE

The practical Guide through the Machine Translation EcoSystem

for

- Developers and Providers of Language Resources and Language Technologies
- Policy and Decision Makers and
- End Users

January 2017

Compiled and edited by the LT Observatory project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 644583.

TABLE OF CONTENTS

Compiled and edited by the LT Observatory project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 644583.

1. INTRODUCTION AND OVERVIEW

Background

We are living in a connected world. Digital technology allows for seamless, ubiquitous communication that brings the world closer together than ever. However, languages are still a major barrier: they hamper transborder eCommerce, social communication and exchange of knowledge, content and services. Language communities without sufficient technological support may become marginalised.

These barriers must be overcome by language technology, like machine translation solutions. Data-driven language technology, especially machine translation is now reaching a critical mass of awareness and performance. With sufficient support, other existing and emerging language technologies can follow in effectiveness and scope and provide solutions to multilingual needs in relation to economic and societal challenges.

What is MT EcoGuide and what are its objectives?

Many of the most important advances in the area of language technologies, in particular machine translation (MT), and language resources have come from Europe. With a considerable body of work in Europe ongoing in the field, MT EcoGuide acts as a pathfinder through the European machine translation ecosystem. It aims at providing:

- practical hints and recommendations for use of MT and language resources and tools for MT use,
- practical hints and recommendations for funding opportunities,
- a concise picture of where the EU stands with regard to language resources for MT,
- an outlook and recommendations geared at European, national and regional policy and decision makers, with emphasis on the CEF and related national and regional services.

Who are the target groups of MT EcoGuide?

MT EcoGuide addresses primarily the following groups:

End users of language technology, especially machine translation: European business, public services, language service providers, and citizens that need solutions for their multilingual needs.

Policy and decision makers who are interested in measures that can be introduced to accelerate the uptake of language technology, especially MT in public services and business, and can support reaching the end result of ubiquitous MT deployment across Europe.

Developers and providers of language technology, especially MT, who are interested in language resources in an operational context, and are looking for funding opportunities for their projects.

Developers and providers of language resources interested in making language resources usable in an operational context for the MT use, and are looking for funding opportunities for their projects.

The colour labels ( ) help different target groups to quickly find the content that is addressed to them.

Compiled and edited by the LT Observatory project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 644583.

How is MT EcoGuide organized?

The Guide has four main parts:

- *Guide to language technologies, language resources and machine translation*
- *Funding opportunities*
- *Recommendations for policy makers*
- *Conclusions*

The Annexes provide supplementary background information and analysis, on which the recommendations for policy and decision makers are based. The following topics are explored in greater detail:

- *Language technology and language resources for machine translation in Europe*
- *Stakeholders and enablers of language technology and language resources*
- *Support for languages with best practices in Europe*

In the Annexes, you will also find:

- abbreviations
- glossary
- list of tools
- references

Note: MT EcoGuide is not intended to be read linearly, but is organized in a way that different target groups will be able to quickly find the content that is addressed to them in the on-line platform. The Guide is organized in modules, cross-linked and color-labeled for an easier integration in the on-line platform.

Compiled and edited by the LT Observatory project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 644583.

2. GUIDE TO LANGUAGE TECHNOLOGIES, LANGUAGE RESOURCES, AND MACHINE TRANSLATION

2.1. FAQS ABOUT LANGUAGE TECHNOLOGIES, LANGUAGE RESOURCES AND MACHINE TRANSLATION

This part of the Guide provides the answers to general frequently asked questions regarding language technologies (LTs), machine translation (MT) and language resources (LRs).

This part is recommended for:

- End users
- Policy/decision makers

This is a good starting point for those end users, and policy and decision makers, who are interested in these topics and may need solutions to their multilingual needs, but are not yet very familiar with language technology, machine translation, and language resources.

FAQ1: What are language technologies (LTs) and the tools/systems they provide?

A: Language technologies comprise computational methods, computer programs and electronic devices that are specialized for analyzing, producing or modifying texts and speech[1]. They are manifold but can be, roughly, divided into three main areas:

- TEXT AND TRANSLATION: From simple spell check to almost fully-automated "language transfer"
- INTERACTION: From spoken human-machine interaction to speech synthesis
- ANALYTICS: analysis of text/speech data to gather deep insights

All of them can be further differentiated into sub-types (and combinations), which may follow different approaches. No wonder there are hybrid LTs, others are or can be combined or bundled.

Like all digital technologies, LTs are 'vehicles' for reaching an ultimate goal. By their very nature, LTs are closely related to content in various written and spoken languages and formats, and communication in a multitude of situations. Language technologies can provide solutions to multilingual needs in relation to economic and societal challenges.

[1] https://www.dfki.de/lt/lt-general.php

Compiled and edited by the LT Observatory project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 644583.

FAQ2: Why are language technologies (LTs) important?

A: In wake of globalisation, LTs have become a necessity for facilitating communication across language boundaries – including the marketing of products and services beyond local markets. Within a language community they facilitate the creation of and distribution of public information, literature, as well as private and commercial communication. Thus, LTs have a huge economic and societal impact. The European Digital Single Market can only materialise if cross-border commerce is truly seamless – i.e. every European citizen can sell or buy wherever s/he wants in any language s/he wants to use, without geo-blocking or translation failures. Without this, Europe will still remain a mega-market of many fragmented local markets that are not big enough to compete with big players at global level.

LTs can support society at many levels – provided that they are used to create and apply content that concerns the most urgent demands of society and public administration, for example:

Quickly obtaining accurate information at all levels of government on all topics concerning the citizens of EU, even when traveling or living in another EU country, comprising

- Tourist and traffic information for foreign visitors
- Patient information for travelers with diseases or disabilities
- Product information on allergenic ingredients for comestible goods

Human-machine voice interaction e.g. for

- Visually impaired persons or elderly with reduced mobility
- Patient–doctor interaction in travel or immigrant situations
- Intuitive language learning using audio-visual modules

Data analytics (including social media analytics) for predicting trends (e.g. refugee waves) or for preventing disasters (whether human-made or natural)

FAQ3: How can language technologies (LTs) be obtained and applied?

A: There are several ways of having access to LTs:

- Purchase
- Lease
- Software as a service (SaaS)
- Freely available LTs in the form of free and open source software
- Freely accessible LTs in the form of online LTs which can be freely used under certain conditions

Compiled and edited by the LT Observatory project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 644583.

Depending on the volume of data to be handled, the purpose and complexity of the data, as well as the frequency of need for a particular LT, the respective software may be highly sophisticated. Often, LTs, in particular machine translation systems used in professional context are custom built for the specific "touch and feel" of the client, which includes using specific language resources, for example organization-specific terminology. Ensuring data security and confidentiality also speak for customized solutions. Additional costs may arise for regular system maintenance and support. In addition, it needs trained or professional users for some LTs.

FAQ4: Isn't it sufficient, if language technologies (LTs) cover my language and English?

A: In many technical texts today occur foreign loan words, quotes in foreign language, and special characters in all kind of combinations. Besides, even local markets are becoming more multilingual through migration, worker mobility, etc. Usually, sooner or later LTs tools/systems capable of processing multilingual data will become necessary. In order to avoid costly adaptations or upgrades at a later point in time, it is worthwhile to consider a more adaptable or upgradeable tool/system than needed at the moment.

FAQ5: What are language resources (LRs)?

A: The term language resource (LR) refers to a set of written or spoken data and their descriptions in digital form. They are used for building, improving or evaluating natural language (human language) and speech algorithms or technologies, and increasingly, for machine-learning. They are also widely used in the language industry, translation industry, publishing, international transactions, language and translation studies, etc.

Examples of LRs are (monolingual, bilingual and multilingual) written and spoken corpora, computational lexica, terminology databases, speech collections, etc. Various tools/systems of the LTs support the acquisition, preparation, collection, management, customisation and use of these LRs.[2]

FAQ6: Why are language resources (LRs) important?

A: The same what was said for the LTs, is also true for the LRs. They are among others – depending on the purpose of their use – raw material for LTs development and upgrading, the medium for conveying information and knowledge (if possible in the most efficient and effective way), the content for developing culture and civilizing societies. LRs in combination with the LTs have tremendously changed the user-experience and interactive possibilities of apps, tools and systems, and public media over the years. Thus, LRs – in combination with LTs – have a huge economic and societal impact.

In addition, LRs having also other functions beyond: they can support society in various fields of applications, such as tourist and traffic information, eHealth, persons with disabilities, eLearning, etc.

[2] http://www.elra.info/en/about/what-language-resource/

Compiled and edited by the LT Observatory project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 644583.

FAQ7: Which kinds of users are using language technologies (LTs) and language resources (LRs) intensively?

A: The following main user groups can be identified:

- European LT developers and vendors
- Language service providers (LSPs) that offer services to customers with high text volumes and short turnaround times. Increasing numbers of LSPs are considering MT as a competitive advantage and necessity, often in combination with post-editing. Some of the LSPs are also technology developers or offer system maintenance services.
- Large public and administrative organizations
- Large enterprises with global outreach
- (Academic) research institutions

…not to forget European SMEs and each individual using the broad and growing variety of apps, tools or systems for written or spoken communication.

FAQ8: Are language technologies (LTs) and language resources (LRs) developed only for English?

A: More and more LTs are developed

- With graphical user interfaces (GUIs) in more than one language
- That can process data in more than one language

However, still languages of larger language communities are more represented than others. There are large multi-lingual corpora (e.g. Acquis Communautaire of DGT, see https://ec.europa.eu/jrc/en/language-technologies/dgt-translation-memory) that cover all 552 language pairs of the 24 official European languages. But apart from English, many language resources figure better in multilingual corpora than parallel corpora strictly speaking.

FAQ9: Why do language technologies (LTs) and language resources (LRs) often need to be developed for each language?

A: The technical issue here is, whether tools/systems of the LTs – including GUI design, web content management systems etc. – are developed under an 'internationalization' approach, which from the point of view of the localization industry means the process of designing a software application in such a way that it can be adapted and seamlessly integrated to various languages and regions without engineering changes.

The content issue here is, whether content has been developed – possibly assisted by LTs developed under the internationalization approach – in such a way that it is interoperable with other kinds of content and easily exchanged as well as – if necessary – translated into other languages. The goal is for original content (or software) and digital services to be international ready. Deployment of

Compiled and edited by the LT Observatory project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 644583.

internationalization must be considered strategically at the beginning of content development, not after original content is already developed.

LRs are the fuel of most modern state-of-the-art LTs: they are used for building, improving or evaluating LTs, and increasingly, for machine-learning. This means that they are crucial for the development of LTs: insufficient amounts of data limit the quality of the language technology system in question. If LRs are not available for deployment in LTs, they need to be prepared – this task is usually time consuming and labour intensive.

In practice, syntax and morphology of the different languages put forward different challenges to translation engines. Therefore, while a machine translation system works well e.g. between English and Latvian, it may give very bad results for English and German. Or as a representative of Facebook recently put it: The challenges are: informal languages that change continuously; low resource languages; morphology; and: build, deploy and maintain so many machine translation systems (Facebook works in 50 languages = approx. 2000 language pairs).

FAQ10: Why do we need more language technologies (LTs) and language resources (LRs)?

A: ICTs, especially mobile technologies, are striving to reach more and more customers – which implies targeting smaller language communities. This triggers an ever-increasing need for LTs capable of processing an ever-increasing number of languages and LRs developed under the perspective of extended interoperability, which means taking into account multilingualism (covering also cultural diversity), multimodality and multimedia, eInclusion and eAccessibility, multi-channel presentations, which have to be considered at the earliest stage of the software design process and data modelling (including the definition of metadata). At the same time, this trend is supported by a decrease of cost for developing the appropriate LTs.

In terms of language resources for LT, in particular machine translation, many LRs exist in Europe and for European languages, but very few are available for commercial use. There are also huge gaps in terms of language and domain coverage. Creation and sharing of new LRs in difficult due to European copyright, privacy and data protection rules. Spoken language resources are becoming more and more important for human-machine interaction (industrial or home robots, automatic cars etc.) and must not be forgotten when mentioning LRs, although they are not part of the investigation of this Guide.

> *Policy makers can play a decisive role as enablers in supporting the development or adaptation of the right LTs and the development of appropriate LRs. In order to speed up the societal and economic development of their constituencies, public services should consider opening up their LR collections, for example in the scope of the action European Language Resource Coordination (ELRC, see http://lr-coordination.eu/) unless there are constraints based on privacy or confidentiality.*

FAQ11: What is machine translation (MT) and how does it work?

A: "Machine translation (MT) is the application of computers to the task of translating texts from one natural language to another. Today a reasonable number of systems are available which produce output which, if not perfect, is of sufficient quality to be useful in a number of specific domains."[3]

[3] http://www.eamt.org/mt.php

Compiled and edited by the LT Observatory project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 644583.

Increasingly, MT is also used for speech translation. MT can be subdivided according to types or approaches:

- Fully automatic machine translation, human-assisted machine translation, machine-assisted human translation;

- Depending on the technological approach:
  - statistical MT: utilizes statistical translation models generated from the analysis of bilingual and monolingual training data and rely on large quantities of these types of data;
  - rule-based MT: utilizes linguistic rules covering morphological, semantic, and syntactic regularities and lexical items and maps them from source to target languages;
  - hybrid MT: takes advantage of statistical and rule-based MT and combines them;
  - neural MT: utilizes neural networks, which are trained by deep learning techniques;
  - adaptive MT: enables use of available data sources to create customized and personalized MT engine for each user, adapted to user domain, style, and feedback.

There are various stages involved in producing an MT system from training data. Whichever type of MT approach is used, the key to creating a good system is lots of (quality) data. A human may optionally edit the output of MT to improve it. This process is known as post-editing.

FAQ12: Why should you consider machine translation (MT)?

A: Many people use machine translation (MT) offered freely through the Internet for something like 'informative raw translation'. In professional application, however, MT covers several processes. Machine translation in professional context offers organizations, such as global enterprises and public administrations, significant translation efficiencies, especially when translation volumes are high.

- It speeds up time-to-market by greatly reducing the length of the translation cycle.
- It increases knowledge sharing internally by making it easier for staff in local markets to access company documents and information in their native language.
- It enhances customer experience and staff effectiveness alike by making real time and on-demand translation a 'regular' part of their business process.

Thus the question in view of increasing volumes of texts to be evaluated, analysed, translated or used otherwise could be: "How to make most effective and efficient use of existing MT so that it becomes part of the value chain?"

FAQ13: What organizations would be interested to introduce or use machine translation (MT) for what purpose?

A: MT is applied in different industries and public sector for different purposes: mainly to address the huge demand for large-scale and/or real-time translation, taking advantage of its speed and low cost. It serves as productivity tool to scale up the availability of multilingual content and to translate content

Compiled and edited by the LT Observatory project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 644583.

that would otherwise not be translated (user-generated content, user-generated engagement).

The table below shows a selection of organization categories for which MT could be particularly viable.

Sort of organization:	Why introduce/use MT?
(Global) companies of any size	Translating high volumes content (e.g. product documentation, online help, FAQs, knowledge bases, websites); reducing translation costs and turnaround time; personalized multilingual customer support (e.g. support emails, chat, forums); moderating and curating user-generated content (e.g. reviews, blogs, etc.), user-engagement content (e.g. social media); processing data for data analytics purposes (e.g. social media data, business intelligence).
Language service providers (LSPs)	(As part of) integrated translation solutions for global businesses or organizations; reducing costs and turnaround time; better matching customer expectations by setting different levels of quality (gisting, publishing, light post-editing); ensuring consistency and brand fidelity per customer.
Companies undergoing a localization process	Effectively localizing content into multiple languages on an ongoing basis; reducing translation costs and turnaround time.
Companies managing customer support via online channels	Offering personalized multilingual customer experience, also via social media; moderating and curating user-generated content; processing data for data analytics purposes (e.g. social media data).
Public sector	Translating high volumes content; reducing translation costs and turnaround time; ensuring eCitizenship and eInclusion of all citizens; instant interaction with citizens, also via social media (e.g. in case of natural disasters); processing data for data analytics purposes (e.g. cross-language informational retrieval).
Social media	Offering personalized customer experience; user-engagement content; processing data for data analytics purposes.

TAB. 1: Reasons to introduce MT by type of organization

Compiled and edited by the LT Observatory project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 644583.

2.2. RECOMMENDATIONS CONCERNING MACHINE TRANSLATION AND LANGUAGE RESOURCES FOR MACHINE TRANSLATION

This part of the Guide provides guidance and recommendations concerning machine translation (MT), and language resources (LRs) for MT. For each topic there is a question, an answer, and practical considerations, organized according to the needs and interests of different target groups.

This part is recommended for:

- End users
- LTs developers/providers
- LRs developers/providers
- Policy/decision makers

Q1: How to use translation technologies?

A: Translation technologies have been primarily developed to translate written data, i.e. turning a text in one language into a text in another or several other languages. Translation technologies broadly comprise two different types of technologies:

- Machine translation (MT), often called automated translation, aiming at automatically or semi-automatically translating text or speech, with several different approaches:
  - (depending on the degree of 'automated'): Fully automated machine translation, human-assisted machine translation (e.g. post-editing)
  - (depending on the technological approach): Statistical MT, rule-based MT, neural MT, adaptive MT, hybrid MT systems;
- Computer-assisted translation (CAT) with various CAT tools: CAT tools often comprise translation memory modules, which are a sort of LRs, and terminology database modules.

Increasingly, translation technologies are also applied for:

- Interpreting spoken data into different languages
- 'Transcreation', i.e. the translation not only into one or more different languages, but also into different sorts of texts (i.e. for instance reformulating a text addressing a specialist audience into a text addressing a non-specialist audience)

MT systems can be combined and integrated with:

- Corpus technologies using mono-, bi- or multilingual corpora for supporting MT or training MT systems,

Compiled and edited by the LT Observatory project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 644583.

- Tools for mono, bi- or multilingual LRs containing structured content needed for MT or directly to support MT,
- CAT tools,
- Authoring tools, technical documentation systems, etc.

Integration of these diverse tools remains a challenge even for mature adopters of MT. Text generated with these tools may include formatting codes and tags, which are reportedly difficult to deal with in an MT application and need to be often engineered around.

Adding to the complexity, a wide range of tools may be needed, such as aligners, taggers, tokenizers, lemmatizers, chunkers, parsers, disambiguation tools, concordancers, annotation tools, extraction tools, etc. for pre-processing and cleaning the data (e.g. when acquiring in-domain parallel data from the web, but also customer's data) for various MT-related processes. However, the results of applying this technology may be particularly useful not only for MT, but also for information gathering and business intelligence.

What to consider:

Fitting translation technologies well into the workflows and communication flows of an organization is decisive for the efficiency and effectiveness of the technology applied.

Consider the reuse and repurposing of the data generated with translation technologies, but also other technologies (e.g. word processors, authoring tools, desktop publishing systems, technical documentation systems, etc.) as an element in the operation, viz. value chains, of an organization. This is of particular relevance for the public sector: the reuse of public sector data would add an enormous potential of LRs in different field. Opening up these sources would help MT at large.

The efficient use of one or a combination of the translation technology tools/systems depends on several conditions: purpose of the translation, expected quality, language pairs, domains, text types, time and cost factor, volumes to be translated, re-use of existing LRs in the organization, need for brand fidelity, privacy and data security issues, role of translation technologies in the information flows of the organization, etc.

Thoroughly check integratability and interoperability aspects, especially 'content interoperability', if content is planned to be used or reused in more than one translation technology application, in other technologies, or in value-chains. Standards can greatly support implementing interoperability. See Technical regulations – especially standards on standards concerning LTs.

Q2: Do tools/systems need to be interoperable?

A: State of the art language technologies (LTs) should be technically interoperable, as they are increasingly needed in some or the other combined form. Besides, as quite some kinds of language resources (LRs) can be used across some or many of the above LTs, interoperability is a serious issue.

Compiled and edited by the LT Observatory project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 644583.

What to consider:

Be sceptical and scrutinize statements, such as "this LTs tool/system is fully multilingual", "this LR can be easily converted into other languages", "this LTs tool/system is interoperable with other tools/systems", this LR is interoperable with other LRs". "99% interoperability" often means in fact non-interoperability!

Interoperability should be achieved along three layers of interoperability: technical, organizational, and semantic interoperability. Standards can greatly support implementing interoperability along these three layers. See Technical regulations – especially standards on standards concerning LTs.

Organisations listed in European Directory of Language Technology Vendors and LT-Observe directory may assist in putting the right questions.

Q3: Does MT provide a good enough translation quality?

A: A better question would be "a translation good enough for what?"

Depending on the intended use of the output, the translation quality may be satisfactory for:

- somehow understanding a text? – gist translation, for instance real-time communication, such as online chats
- certain sorts of texts only? – for instance technical manuals already written in highly controlled language
- producing 'publishable quality' texts? – for instance for product catalogues produced by extracting data from product master data management systems, or texts that must meet special requirements (e.g. legal requirements concerning liability)
- translating texts without post editing? – depends on which kinds of text and for which purpose as well as for whom

Not all content, not all file formats and not all language pairs are suited to the same engine: The expected purpose of translation in relation to the required quality, content types and language pairs tend to play a determining role in engine performance. Satisfactory results can be achieved in certain domains (e.g. technical, administrative), and text types (e.g. patents, technical manuals, software documentation, vehicle assembly build instructions). Rule-based MT systems reportedly perform better in so-called "narrow domains" and certain language pairs (e.g. Japanese-German), while statistical MT systems are better suited for "broad domains" (e.g. user generated content) and languages of lesser distribution. Regardless the "tool wars", often an adequately performing MT system that fits into the workflow is preferable in terms of costs, time factor and benefits to a better performing MT that is not integrated into the workflow.

What to consider:

Determine the volumes of text, content types and language pairs to be machine translated for which purposes in relation to the expected or required quality of outcome – and the expected benefits under the consideration of time and cost factor.

Organisations listed in LT-Observe directory may assist in finding the right solution for your needs.

Compiled and edited by the LT Observatory project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 644583.

Q4: How to apply machine translation (MT) optimally in the organization?

A: The best machine translation solutions raise output quality, lower the human factor across the spectrum of applications, and therefore improve cost-effectiveness and availability of translation services. MT may be also capable of delivering brand fidelity through implementation of organization-specific language, while at the same time reducing the need for human post-editing, by aligning translations with organization's corporate language.

Many people use MT freely offered through the Internet for something like 'informative raw translation' (gist translation). This is different, if MT is used as one of the central systems in the organization: security and confidentiality, and reliable use of client's data and corporate terminology are the main benefits of customised MT service in professional environment.

What to consider:

To implement MT effectively in professional environments, it is important to see MT as a process. This requires a thorough analysis of how to integrate MT optimally into your organization's system environment, workflows, etc.

Determining what engine and what process will give the best results, under the consideration of time frame and cost factor and based on the following:

- The expected purpose of translation in relation to the required quality, content types and language pairs, etc.
- Would it need a combination and integration of MT systems with other technologies?
- What are the costs for:
  - the preparation/pre-editing of texts (and which kinds of text?) to undergo MT?
  - the human post-editing process (towards which quality level)?

Language resources (LRs) needed at what costs:

- What kind language resources are necessary to deploy the MT system you would need? Are the LRs available in your organization?
- What are the costs for the maintenance of
  - the MT system (possibly in combination with other language technology tools)?
  - the systems for processing and maintaining LRs?
  - different kinds of LRs for different roles in the MT process?

The integration or outsourcing of MT:

Do you want an internal engine that is customised to your needs, or do you plan to outsource all or parts of MT activities, from customising and processing to post-editing and maintaining?

Often, translation is outsourced (with or without MT). This means that valuable translation data does not become part of an organization's asset. Therefore, integrated solutions may be at the beginning more costly but could add value for future translation towards additional languages.

Legend: End users | LTs developers/providers | LRs developers/providers | Policy/decision makers

Compiled and edited by the LT Observatory project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 644583.

This issue is to be considered under strategic aspects, such as cost-effectiveness, quality management, legal implications, data security and confidentiality, knowledge of the organization.

Organisations listed in LT-Observe directory may assist in finding the right solution for your needs.

■ ■ The usefulness or even need of a organisational language strategy with respect to:

- formulating (or adapting) a organisation-specific language strategy
- taking into account aspects of legal or technical regulations, customer relations, etc.

Q5: What are the most important language resources (LRs) for use in machine translation (MT)?

A: The term language resource (LR) refers to a set of language or speech data and their descriptions in digital form. They are used for building, improving or evaluating natural language (human language) and speech algorithms or technologies. They are increasingly used for machine learning. Furthermore, they are also widely used in the language industry, in language and translation studies, in electronic publishing, for international transactions, etc.

Whichever type of machine translation model is used, the key to creating a good system is currently lots of (quality) LRs. LRs relevant for MT comprise a broad range of different kinds of structured content and corpora. For text-based data (as opposed to multimodal, e.g. audio or video data) they can be broadly differentiated into:

- Text corpora, such as
  - Parallel corpora, i.e. corpora where the same text appears in more than one language (particularly useful for machine translation)
  - Comparable corpora, i.e. corpora with texts of the same or very similar topic in the same or similar sort of text in different languages
  - Monolingual corpora
- Terminologies and similar data (such as domain-specific thesauri, classifications, nomenclatures, taxonomies, etc.)
- Lexicographical data and similar (such as dictionary data, treebanks, grammars, etc.)

Depending on the purpose of the use, other kinds of structured content (pictorial data, directories of proper names of all sorts, dialogue data, etc.) may also be or become necessary.

What to consider:

■ ■ Analyse carefully your needs for which kinds of LRs and the languages needed. Organizations listed in LT-Observe directory may assist in finding the right solution for your needs.

■ ■ ■ Set up the criteria for the usability of the LR for the specific project (language(s), domain, file format, time frame, budget, size, type of resource, license, documentation, etc). Consult the LT-Observe Catalogue to obtain operationally usable LRs for commercial purposes.

Legend: ■ End users ■ LTs developers/providers ■ LRs developers/providers ■ Policy/decision makers

Compiled and edited by the LT Observatory project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 644583.

■ ■ Consider the reuse and repurposing of LRs within the organization as an essential element in the operation, viz. value chains, of an organization. This is of particular relevance for the public sector: the reuse of public sector information would add an enormous potential of LRs in different fields. Opening up these sources would help MT at large.

Q6: What corpora are useful for machine translation (MT)?

A: Most corporate buyers believe parallel data LRs (two or more languages side by side) are what is needed to train a statistical machine translation (SMT) system. However, monolingual data in the target language are becoming increasingly useful. In statistical MT a crucial step is to develop a language model for the target language that selects the best translation. Therefore an MT engine needs to be trained on monolingual data as well as parallel data to produce a fluent output.

What to consider:

■ ■ Analyse carefully your needs for LRs. Organizations listed in LT-Observe directory may assist in finding the right solution for your needs.

■ ■ ■ Set up the criteria for the usability of the LR for the specific project (e.g. language(s), domain, file format, time frame, budget, size, type of resource, license, documentation, etc). Consult the LT-Observe Catalogue to obtain operationally usable LRs for commercial purposes.

■ MT developers need a considerable amount of data to build an MT system. As a rule of thumb, the closer the data that is used to train the system to the type of data that should be translated, the better the results will be. The universally preferred format is plain text or XML; TMX and XLIFF data formats may be preferable for parallel resources. Reportedly an MT developer may need at least 5 million tokens/500,000 segments of language data to build an MT system for a particular domain. In terms of size and quality, balance is needed.
The widely-used statistical MT, Moses, needs sentence-aligned data for its training process. If data is aligned at the document level, it is recommended to convert it to sentence-aligned data using a sentence aligner. See Best Practice Guide to LRs for Automated MT for a report on two sentence aligner tools, Hunalign and BSA.

■ ■ ■ More parallel data is needed in all the verticals: legal, tourism, etc., as well as for lesser covered languages (see gaps in LRs for MT). New MT models (e.g. neural MT) may have an impact on the need for data, e.g. for monolingual data.

Q7: How to find language resources (LRs) needed for machine translation (MT) or other purposes?

A: LRs are indispensable for the development of tools for machine translation (MT), but they are also expensive and labour-intensive to create or adapt e.g. for MT usability. It should be noted that the extent of the availability and coverage of LRs differs considerably from language to language (see gaps in LRs for MT).

The larger the organization, the higher the probability that at least some of the LRs needed exist already. However, if existing data are fragmented, partially outdated, mixed with heterogeneous data, or come

Compiled and edited by the LT Observatory project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 644583.

with legal and confidentiality problems, etc. obtaining them from external sources may be more cost effective and less time-consuming.

What to consider:

■ ■ ■ Set up the criteria for the usability of the LR for the specific project (e.g. language(s), domain, file format, time frame, budget, size, type of resource, license, documentation, etc).

■ ■ ■ Consult the LT-Observe Catalogue to obtain operationally usable LRs for commercial purposes.

Consult the repositories and catalogues, such as those provided by ELRA/ELDA, CLARIN, OPUS, META-SHARE.

■ ■ Consider the reuse and repurposing of LRs as an essential element in the operation, viz. value chains, of an organization. This is of particular relevance for the public sector: the reuse of public sector information would add an enormous potential of LRs in different fields. Opening up these sources would help MT at large.

Q8: How to evaluate the usability of language resources (LRs) needed for machine translation (MT)?

A: There is currently no clear answer from industry about any shared method for evaluating the ***usability*** of LRs. There are many aspects of the usability of language resources, and it only makes sense to talk about usability for a specific task. In general, the basic usability of the resource is evaluated by users based on **language (pair)**, **domain** and some **basic file format metadata**.

What to consider:

■ ■ ■ Set up the criteria for the usability of the LR for the specific project (e.g. language(s), domain, file format, time frame, budget, size, type of resource, license, documentation, file format, etc).

■ ■ ■ Check the LT-Observe Catalogue to obtain operationally usable LRs for commercial purposes.

Consult the repositories and catalogues, such as those provided by ELRA/ELDA, CLARIN, OPUS, META-SHARE.

■ ■ Analyse carefully your (immediate and future) needs for LRs. Check to which degree a LR must be interoperable with other LRs or LTs tools/systems and whether it can be scaled-up. This applies to both commercial or to open source. In spite of existing standards (mostly focusing on technical interoperability, see Technical regulations – especially standards), the interchange, re-use and re-purposing of LRs is not trivial and may require huge efforts, if not carefully planned. Organisations listed in LT-Observe directory may assist in finding the right solution for your needs.

Q9: How to add value to existing language resources (LRs) for machine translation (MT) purposes?

A: Many LRs in existing repositories present some challenges in an operational context. Valorisation of

Compiled and edited by the LT Observatory project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 644583.

LRs through optimization of existing metadata, and sometimes addition of new metadata, will be an invaluable aid to MT developers - as this means that the developers will actually be able to identify and select the resources they need.

What to consider:

Consider adding the following recommended metadata to LRs to be used for MT purposes:

Title, Resource type, Creator, Language(s), Availability, Modality, URL, Domain, Format, Size, Production date, Comment, Description, Tags, Contact person, Format description. For detailed description and examples, see Best Practice Guide to LRs for Automated MT.

Consider following a metadata standard, for example the Dublin Core metadata set for resources; even though some adjustments are recommended, such as specifying the production date of the resource (*Production date*), inclusion of categories that are not part of the metadata set (*Size, Comment, Modality, Availability, Tags*), and omission of certain categories. For details, see Best Practice Guide to LRs for Automated MT.

Collect best practice information: send LRs with initial valorisation to potential LR users, vendors and buyers and let them examine and test resources in relation to their own work context and requirements. A user-feedback system in a collection can help to continuously improve the resources. A mere rating system may not be useful, as such rating may differ depending on purpose and use of the LRs.

Q10: How to create language resources (LRs) needed for machine translation (MT), especially aligned corpora?

A: Performance of statistical machine translation (SMT) systems is depended on how well the training data correlates with the documents that are translated regarding genre, style and in particularly domain-specific data. As these types of data are often not available, a recommended technique to create new LRs is to collect the data, for example domain-relevant training data, by exploiting web-crawling approaches.

What to consider:

Set up the criteria for the usability of the LR for the specific project (e.g. language(s), domain, file format, time frame, budget, size, type of resource, license, documentation, file format, etc).

Check the LT-Observe Catalogue to obtain operationally usable LRs for commercial purposes. Consult the repositories and catalogues, such as those provided by ELRA/ELDA, CLARIN, OPUS, META-SHARE.

Be advised that the issues of Intellectual Property Rights (IPR) hamper the free use of materials from the web.

See Q11: What are the best practices to lawfully acquire data through web-crawling? for details.

Compiled and edited by the LT Observatory project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 644583.

Acquisition of in-domain parallel data can be divided into three phases:

- A focused search for and subsequently ranking of domain relevant websites. The links found at these websites are then regarded as candidate URL seeds with respect to identifying bilingual documents, and the detected candidate documents are evaluated.
- Cleaning up and preparing documents: Removal of duplicates and exclusion of boilerplate elements.
- The next step consists of sentence splitting and tokenization. The final step in the pipeline of making parallel data qualified as training data for an SMT system, is to secure that the sentences extracted are aligned with the highest quality possible.

See the Best Practice Guide to LRs for Automated MT for detailed instructions for each phase.

Useful open resource tools do exist that can help you through the pipeline steps (see the list of recommended tools), but no single place exists where open source software for generating high quality in-domain and sentence aligned corpora, was available at one single website or portal, requiring that the users themselves are left to hard code the software that integrates the applications into one workflow.

Be advised that the internet increasingly contains content that has already been translated automatically, which is polluting the linguistic quality of internet-based content in general. Currently, there seems to be no quick method for deciding whether a given content found on the internet is "human translated" rather than produced by a machine.

Q11: What are the best practices to lawfully acquire data through web-crawling?

A: The issues of Intellectual Property Rights (IPR) hamper the free use of materials from the web. The legal framework within which agile corpus acquisition would operate is governed by two sets of legislative provisions: copyright and database rights (intellectual property rights, IPR), and data protection (privacy and autonomy, i.e. confidentiality, anonymity and access arrangements). In some countries there is copyright exception for specific purposes. Certain materials exist which do not fall under the copyright protection law – this mainly concerns texts from public administration, which can therefore be very useful in this context.

What to consider:

It is recommended using data where the rights are cleared – e.g. LRs from ELRA, and other providers (CLARIN, META-SHARE, etc.) where license is regulated. These materials for the most part come with a price tag and/or licensing conditions governing their use. The existence of license conditions attached to the use of a specific resource means that an agreement between the IPR owner and the provider, e.g. ELRA (see http://wizard.elda.org/principal.php), has been entered into about the conditions regarding distribution and use of the resource.

The resources listed in LT-Observe Catalogue have been included based on the above considerations for commercial purposes.

Compiled and edited by the LT Observatory project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 644583.

The process of assessing whether to harvest Web data and obtaining permission to use these can be split up into three steps:

- Locating the data relevant for your MT systems in terms of number sources and especially what means and tools that are needed to handle these data (see above).
- Determining execution costs, i.e. will it, seen from a cost-benefit point of view, be worthwhile to conduct the time-consuming analysis of: conditions and terms for using the data at a Web site, and possibility of identifying responsible content providers
- Evaluate the collected information. Be advised that the negotiations about data usage rights can be cumbersome and time consuming to conduct.

See the Best Practice Guide to LRs for Automated MT for detailed instructions for each step.

Q12: What are terminological data and similar data good for in machine translation (MT) and how to generate them?

A: Terminologies are usually following a language-independent approach, which allows managing data in two or more, sometimes many languages. Terminological data are not only semantically structured, but also contain additional information such as definitions (or explanations, contexts, etc.), references etc. Some large terminology collections are monolingual, but most of them are bi- or multilingual. Others look monolingual (such as some harmonizing quantities and units), but they can easily be turned multilingual. Many terminology collections are of a highly-harmonized nature and therefore contain particularly reliable data.

Similarly, structured LRs are domain-specific thesauri, classifications, nomenclatures, taxonomies, translation memories and the like. They too are mostly multilingual and usually contain highly harmonized data.

What to consider:

Consider the reuse and re-purposing of terminologies within the organization as an essential element in the operation, viz. value chains, of an organization. Organisations listed in LT-Observe directory may assist in finding the right solution for your needs.

Terminology collections may contain equivalents that are rated as unlikely by the statistical machine translation (SMT) system models. If such an SMT system is integrated in translation service workflow, it is not possible to ensure high quality terminology (consistency, correctness) in the SMT suggestions. Training data can contain contradicting terminology, corporate specific synonyms or brand terminology. For this reason, effective adaptation of SMT systems can profit from customized client's terminology collections.

Clients require correct and accurate use of specific terminology, often corporate terminology. Clients may provide their own terminology collections, but in projects where the client-supplied terminology collections are not readily available and the use of specific in-house terminology is still required, the terminology firstly needs to be extracted from documents provided by the client.

Compiled and edited by the LT Observatory project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 644583.

Term extraction generally involves four steps:

- compilation of a specialized corpus,
- extraction of term candidates (useful tools do exist that can help you to extract term candidates, see the list of recommended tools),
- validation of the term candidates and
- automatic or semi-automatic creation of terminological records.

See the Best Practice Guide to LRs for Automated MT for detailed instructions for each step.

Q13: What other structured LRs could be used?

A: The usefulness of other kinds of bi- or multilingual structured LRs, such as bi- or multilingual directories of all sorts of proper names (many names differ, have aliases, or are differently spelled or pronounced in different languages), certain types of master data (e.g. product properties in enterprise resource management systems or more specifically in product master data management systems), ontologies, etc. is generally not yet fully recognized in the field of the LTs. In particular, master data for instance in trade often need to be multilingual – some of them are even standardized and maintained by maintenance agencies or registration authorities. So far, this source of LRs has not been tapped – probably due to difficulties in accessing the respective LRs.

What to consider:

■ ■ Consider the reuse and repurposing of LRs as an essential element in the operation, viz. value chains, of an organization. Organizations listed in LT-Observe directory that may assist in finding the right solution for your needs.

Q14: Are LRs used only for translation purposes?

A: Some kinds of LRs also have or can have other purposes, as resources in authoring tools, for technical documentation, procurement purposes, organizational knowledge management, in archives, etc.

What to consider:

■ ■ Consider the reuse and repurposing of LRs as an essential element in the operation, viz. value chains, of an organization. Organizations listed in LT-Observe directory may assist in finding the right solution for your needs.

■ LRs may be created for specific purposes and within particular frameworks, and this is the information that is crucial to pass on to other users of the resources. Here metadata come into the picture; when an LR is provided with sufficient metadata that thoroughly describe the resource, the users can decide for themselves whether it is likely that this resource can be used in relation to the particular task. See Q9: How to add value to existing language resources (LRs) for machine translation (MT) purposes?

Compiled and edited by the LT Observatory project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 644583.

Q15: What organizations would be interested to use or repurpose language resources (LRs) for what purpose?

A: This depends on the purpose of the different kinds of LRs. Language service providers (LSPs) may need to use any of the LRs mentioned e.g. for translation purposes (and the respective language technologies). MT developers or MT service providers may primarily be interested in bi- or multilingual text corpora (or parallel texts or comparable texts), terminological data and lexicographical data. Enterprises and public services usually have several or many (not to mention different kinds of) databases with structured content for different purposes, as well as data generated with text processors, word processors, authoring tools, desktop publishing systems, technical documentation systems, etc.

What to consider:

■ ■ Consider the reuse and repurposing of LRs as an essential element in the operation, viz. value chains, of an organization. This is of particular relevance for the public sector: the reuse of public sector information would add an enormous potential of LRs in different field. Opening up these sources would help MT at large.

Compiled and edited by the LT Observatory project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 644583.

3. FUNDING OPPORTUNITIES

3.1. INTRODUCTION

European research & innovation programmes offer a rich opportunity for innovative projects, but they are becoming more and more competitive. In addition, dedicated funding for LT/LR/MT is currently scarce. Therefore, the LT Observatory project investigated the national/regional support given to languages and language technology in the EU Member States, in particular national/regional funding opportunities including EUREKA and Eurostars schemes.

The national/regional funding landscape is heterogeneous in all respects:

- Institutions/agencies that manage funding programmes can be independent or associated with Ministries/government institutions at national or regional level;
- topics vary depending on national/regional priorities or are bottom up which allows for all kinds of innovative projects;
- beneficiaries differ depending on the programme type (e.g. research institutions/academia for scientific research, SMEs for market-near programmes etc.);
- conditions to participate differ (continuous submission schemes, calls, cut-off dates);
- eligibility of costs differs;
- the percentage of funding is based on state aid rules;
- the availability of funds is often not very transparent.

The "National funding opportunities" (http://www.lt-innovate.org/lt-observe/public-policy-observatory/national-funding-opportunities) provide information about the availability of funding programmes and give contact details and links to useful websites, usually in EN and the local languages. Programmes investigated were: National and/or regional funding programmes; ESIF – European Structural and Investment Funds (that may overlap with or finance national/regional programmes); EUREKA and Eurostars. Although only very few programmes have LT as a focus (Spain, Estonia, Ireland, Latvia, Lithuania, Sweden), national and regional programmes did already fund LT projects in the past within the framework of their R&D&I programmes (e.g. Austria, Germany, Belgium).

This guide is geared at any person/entity that has a project idea in the area of LT/MT/LR and is looking for financing. Primarily, it targets companies (start-up or SME or large company), but also research institutes/universities that engage in market-near research and development. Funding opportunities exist for all these entities, but not all are for all types of beneficiaries or projects.

3.2. FAQS CONCERNING FUNDING OPPORTUNITIES

FAQ1: What funding options are available?

A: Funding can be grants (not repayable) or loans (repayable) or a mix. The funding landscape is very heterogeneous and conditions vary largely. Examples are: in France, there is a scheme where a successful project has to repay the grant; in Malta, you can receive a grant up to a maximum limit, with the rest as tax credit.

[4] For explanation on "smart specialisation" see: http://www.lt-innovate.org/lt-observe/how-does-esif-work

Compiled and edited by the LT Observatory project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 644583.

In general, national/regional funding programmes apply a bottom-up approach, i.e. there are no specific topics into which a proposal must fit. However, programmes differ considerably and it is recommended to check out programmes, calls etc. regularly.

Most national and regional programmes are limited to legal entities that have their business seat in their territory. Some have additional conditions, e.g. that the results must benefit the region or fall under the smart specialization[4] of the region. Therefore, all information of funding sources are listed by country, and, where applicable, per region inside this country.

In countries where there are only regional agencies (e.g. Belgium), the location is even more crucial. But there are schemes to cooperate with partners outside the region e.g. again Belgium, but also other countries where there are bilateral programmes (e.g. SK, SI), and of course, EUREKA and Eurostars (minimum 2 partners from 2 different (EUREKA/Eurostars) Member States).

FAQ2: What is the difference between funding types: National/Regional/ ESIF, Eurostars and Eureka?

Conditions/requirements	National/regional	ESIF	EUREKA/Eurostars
Territorial benefit	Sometimes yes	Yes	No
European benefit	No	No	Yes
Single applicant	Yes	Yes	No
Partnership applicant	Possible	Possible	Yes
Predefined topics	No	Yes (S3 priorities)	No

National Funding reflects the provision of funds from the respective nation's government.

Regional Funding represents the provision of funds dispatched by the region's government.

ESIF represents European Structural and Investment Funds. They comprise the European Social Fund, Cohesion Fund, European Agricultural Fund for Rural Development (EAFRD), European Regional and Development Fund (ERDF) and European Maritime & Fisheries Fund (EMFF). Funding follows pre-defined priority axes (Smart specialization strategies – S3 - for R&I projects). For more details on ESIF, see: http://www.lt-innovate.org/lt-observe/how-does-esif-work

Eureka is an intergovernmental scheme implemented for market oriented projects. The initiative is supported by public national and private funding, but the consortiums determine the project objective.

Eurostars is a European funding programme respectively devoted to SME's specializing in R&D. However other participants such as universities, large enterprises, research institutions can act as collaborators of the project.

For available resources per country see: http://www.lt-innovate.org/lt-observe/public-policy-observatory/national-funding-opportunities

Compiled and edited by the LT Observatory project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 644583.

FAQ3: How can I receive funding?

A: A concise guide through the funding maze, with direct links to the country web page, is the "Path through the Funding Maze". It is recommended to read it before engaging in any project writing.

The below table gives and overview and provides links for quick checks.

Questions	National/regional	ESIF	EUREKA/Eurostars
Where can I apply for funding?	National/regional funding agencies; they figure with contact data in the respective country pages at: http://www.lt-innovate.org/lt-observe/public-policy-observatory/national-funding-opportunities	Managing authorities and related agencies can be found in the ESIF document: http://www.lt-innovate.org/sites/default/files/InfoGuide_ESIF_Funding_Opportunities_2016_web.pdf	You find contact data of the National Project Coordinators (NPC) at the country pages (see national/regional). Or you can go directly to EUREKA: http://www.eurekanetwork.org/eureka-countries OR Eurostars: https://www.eurostars-eureka.eu/eurostars-countries/europe
What are the conditions?	Differ depending on country and programme.	Differ depending on country and programme.	At least to legal entities from at least two different (EUREKA/ Eurostars[5]) Member States
Do I get a loan or a grant?	Depends on country and programme.	Depends on country and programme.	Eurostars: grants; EUREKA mostly grants or mix; only NL gives only loans. **Attention**: Some EU Member States **do not** provide funding for EUREKA/ Eurostars[6]!
Do I get 100% funding?	As a private company, never (due to state-aid rules); Universities/ research institutions depend on programme and country.	As a private company, never (due to state-aid rules); Universities/ research institutions depend on programme and country.	As a private company, never (due to state-aid rules); Universities/ research institutions depend on programme and country.
Do I need a partner?	In most programmes not.	In most programmes not.	Yes. At least one partner from another EUREKA/Eurostars country (see footnote 5).

Compiled and edited by the LT Observatory project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 644583.

Can I have a partner?	In most cases yes.	In most cases yes.	You need at least one partner, see above.
Can the partner be from another country?	Most programmes are territorial, i.e. beneficiaries must be from the country/ region. But there are exceptions (e.g. Flanders).	Most programmes are territorial, i.e. beneficiaries must be from the country/ region. Sometimes a partner outside the country/region is possible if it benefits the region.	At least one partner MUST be from another country.
Can the partner be from a non- EU country?	In most cases not (except some multilateral programmes).	No.	Yes, if from a EUREKA/ Eurostar country (see footnote 2)
Can the partner be a research institute?	In most cases yes.	In most cases yes.	In most cases yes.
Can the partner be a large enterprise?	In most cases yes but sometimes without funding.	In most cases yes but sometimes without funding.	In most cases yes but sometimes without funding.
Can I apply for national/regional funding although I am a partner in a H2020 project?	Yes.	Yes. The only condition is that the ESIF funding covers other activities than the H2020 funding (principle of no double funding).	Yes. For Eurostars, the only condition is that the ESIF funding covers other activities than the H2020 funding (principle of no double funding).
Are there deadlines to observe?	Often, there is a continuous submission scheme with cut-off dates. Sometimes there are calls with deadlines.	There may be calls; sometimes also cut-off dates.	Several cut-off dates per year.
Can I receive funding as a start-up?	In most cases yes. Depends on the respective financial viability rules.	In most cases yes. Depends on the respective financial viability rules.	In most cases yes. Depends on the respective financial viability rules.
Can I receive funding as a large enterprise?	In most cases yes. Funding for LEs is always lower than for SMEs (around 20% for grants) due to state-aid rules.	Often yes. Funding for LEs is always lower than for SMEs (around 20% for grants) due to state-aid rules.	Depends on the country. Sometimes LE can participate but without funding.

Compiled and edited by the LT Observatory project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 644583.

<table>
<tr><td>What type of project idea receives funding?</td><td>It must be innovative – this is valid for almost all programmes. Often, projects are split between “industrial research” and “experimental development”. Only rarely there are specific programmes for LT (e.g. in Spain or Estonia).</td><td>A project idea that falls into the priority axes of the country/region. For LT, this can be ICT (horizontal) or verticals like eHealth, tourism, cultural developments etc.</td><td>Innovative ideas (bottom-up) with a potential impact beyond just the partners.</td></tr>
<tr><td>Can any topic/idea receive funding?</td><td>Most programmes are bottom-up, i.e. any topic is possible if all other conditions are fulfilled.</td><td>See above.</td><td>Yes, EUREKA and Eurostars are also bottom-up programmes.</td></tr>
<tr><td>What can I do if my topic/idea is currently not covered by the funding schemes?</td><td colspan="3">Make sure you get in contact with the respective funding agency for potential future calls. If your idea is not considered for funding, scrutinize the idea: Is it really innovative? Could it fit under another type of programme?</td></tr>
<tr><td>Is my company considered to be a SME?</td><td colspan="3">Check here: ec.europa.eu/growth/sme
<table><tr><th>Company category</th><th>Staff headcount</th><th>Turnover</th><th>or</th><th>Balance sheet total</th></tr><tr><td>Medium-sized</td><td>< 250</td><td>≤ € 50 m</td><td></td><td>≤ € 43 m</td></tr><tr><td>Small</td><td>< 50</td><td>≤ € 10 m</td><td></td><td>≤ € 10 m</td></tr><tr><td>Micro</td><td>< 10</td><td>≤ € 2 m</td><td></td><td>≤ € 2 m</td></tr></table></td></tr>
<tr><td>How can I get information?</td><td colspan="3">http://www.lt-innovate.org/lt-observe/public-policy-observatory/national-funding-opportunities</td></tr>
<tr><td>How can I get advice on my specific situation?</td><td colspan="3">Many countries offer services to check out project ideas – take these opportunities! Sometimes contact points (e.g. EUREKA/Eurostars) can also help for other (national/regional) programmes as they work for the same funding agency. A good relationship with funding agencies helps creating trust and eventually will lead to success.</td></tr>
</table>

[5] 36 Eurostars participating countries: EU, NO, Iceland, Switzerland, Canada, South Africa, Israel, Turkey, South Korea; 41 EUREKA Member States: EU plus European Commission, Russia, Ukraine, Turkey, Switzerland, Montenegro, Makedonia, Serbia, San Marino, Monaco, Israel, Iceland, Norway plus Associated countries: Canada, South Africa, South Korea.

[6] EUREKA: EE, UK, SK, CY; Eurostars: EE, IT, GR.

Compiled and edited by the LT Observatory project that has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No. 644583.

3.3. STEPS TOWARDS NATIONAL/REGIONAL FUNDING – VADEMECUM

This Vademecum will help you taking advantage of the funding schemes available, no matter if you are an experienced funding person or a newcomer. The eight steps show the main paths towards success. First, three simple questions (that are not as simple in the end) have to be answered: WHAT, WHERE, WHO. These are followed by the HOW that is subdivided into 5 steps: INFORMATION, COMMUNICATION, IDENTIFICATION, ACTION, MARKETING.

Here we go:

1 You must have a clear idea **WHAT** you want to do, in short: for what activity do you want money? Depending on this answer, different types of programmes are available. Here some examples:

LT tools or applications development → any market-near development programme (e.g. experimental development) might fit, or, if research is involved, any R&I programme.

LR creation → any vertical programme (eHealth, tourism etc.) might be useful, depending on the LR theme.

Multilingual solutions → these could fall into programmes for the improvement of SMEs (eCommerce, internationalisation), or any market-near programme (see above).

2 For national and regional funding programmes, the location plays a crucial role. **WHERE** are you located? Most programmes are limited to legal entities that have their business seat in their territory. Some have additional conditions, e.g. that the results must benefit the region or fall under the smart specialisation of the region. Therefore, all information of funding sources are listed by country, and, where applicable, per region inside this country.

In countries where there are only regional agencies (e.g. Belgium), the location is even more crucial. But there are schemes to cooperate with partners outside the region e.g. again Belgium, but also other countries where there are bilateral programmes (e.g. SK, SI), and of course, EUREKA and Eurostars.

3 **WHO** will perform the project's work? First, you need to know your **status**, or better said, the status of your company or organisation:

Company: Are you a SME – Small or Medium sized company? Are you a micro-enterprise? Are you a large enterprise?

STATUS:

These definitions follow the EU rules (Source: ec.europa.eu/growth/sme), see above 3.2 FAQs.

But there are always exceptions, like funding for SMEs by ZIM in Germany that limits the head count with 499. Or France that makes a difference between large enterprises with less or more than 2000 employees. Some countries also have different funding percentages, depending if you are a micro, small or medium-sized enterprise.

If you established that you are, let's say, a SME, you might want to establish if you are R&D performing, i.e. if some of your turnover is used for research and/or development activities. Some countries make funding depending on that (UK), or provide higher funding for R&D performing SMEs, like Eurostars in CZ, DK, or FI.

Compiled and edited by the LT Observatory project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 644583.

ALONE OR PARTNERSHIP?

The next question to ask oneself is: Do I plan to perform the work alone, or in **partnership**? If the latter, will the partner(s) be in the same region, or outside, or even outside the country?

National and regional programmes have different rules: In most of them you can act alone, with partner(s) from the region or with a subcontractor (that can be a company or a research institute or a university). Some allow a partner to be outside the region (e.g. Flanders). For EUREKA and Eurostars, you need at least one partner in another Member State (of the EUREKA programme, not necessarily the EU).

HOW TO…

The **HOW** is in every process the most tricky element. It is important to get it right to have a realistic chance to grasp the opportunities offered by national and regional funding schemes. Therefore, the "HOW" is sub-divided into 5 interrelated steps:

Access to **Information** – establish a relationship with your national/regional funding agency through **Communication – Identification** of best possible funding scheme – **Action** in submitting a proposal – **Marketing** of the result as a Success Story for your Company, your partners but also the Funding Agency.

4 **INFORMATION** on national and regional funding programmes are compiled at the LT-Observe web pages:

http://www.lt-innovate.org/lt-observe/public-policy-observatory/national-funding-opportunities

The map on the page links to the respective country:

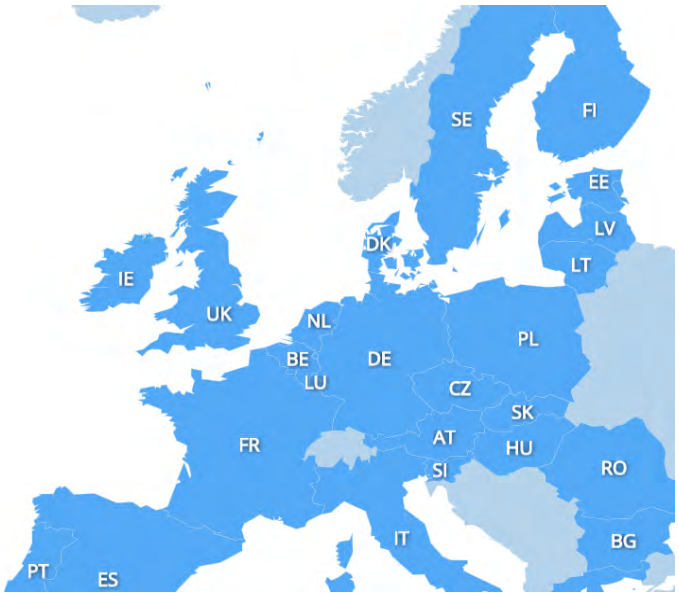

FIG. 1: Screen shot funding opportunities per country

By clicking on the country, you get the country information and, where appropriate and available, regional information. Each country page looks like that:

Compiled and edited by the LT Observatory project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 644583.

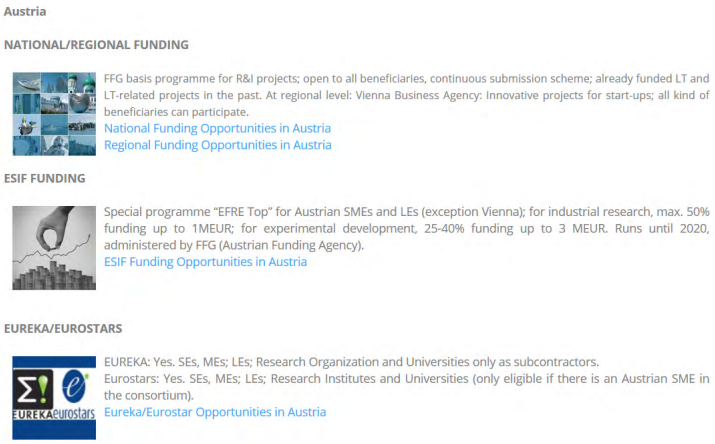

FIG. 2: Sample country page

The three funding segments "National/Regional funding", "ESIF" and "EUREKA/Eurostars" refer to *interactive pdf pages* that contain all useful information, including contact names, addresses, e-mails and telephone number.

The information provided allows for an ex ante assessment if and which funding resources are available. The compilation of this information in one single space saves time as this information is usually dispersed over many websites.

Information is provided in English, but links to information in local languages are included.

Contact data like names, e-mails and telephone numbers are given. This allows getting quickly in contact with the person(s) in charge, to obtain accurate and updated information.

5 **COMMUNICATION** with funding agencies: A good relationship to local, regional or national funding agencies is a great help: for you, but also for them. They like to know their constituencies; they like to have success stories. Experience shows that good relationships sooner or later yield results: You get regular information, you may be invited to info-days or (more and more often) webinars, and you get to know what makes them tick. And eventually, it will be your project idea that makes them tick.

Public funding can of course be tried without these preliminaries, but the risk is that you submit a project that corresponds to *your* understanding what "they" want to finance – but this is not necessarily *their* understanding.

To avoid disappointment and waste of time, it is therefore recommended to contact funding agencies always BEFORE investing time in writing a proposal.

6 As a result of the communication exercise, you may be able to IDENTIFYING the right programme for your project idea. Or, ideally, you identify it together with the agency in question. They may also give hints as to important elements that must be included, or to upcoming opportunities.

7 Once the funding programme is identified, it is ACTION time or better said: Proposal writing time. Usually, this is less time-consuming than EU projects (e.g. Horizon 2020) but still it is an effort and a time investment. Therefore, all guidelines for the programme should be read and all information provided so that a fair judgement of the project is possible. In many cases, there are continuous submission

Compiled and edited by the LT Observatory project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 644583.

schemes, and decisions are taken at several cut-off dates per year. Hence, the process is most often quicker than at EU level.

8 The last step: **MARKETING** your success. If you were successful and obtained some funding, you may want to promote it. Why? Because it shows that your idea is worthwhile funding; because it shows that the way you approached it created trust that you can perform; because obtaining financing helps you enlarge your product or service portfolio. The (business or scientific) world should know about it! Put it on your website, and when first results are there, promote them. LT-Observe would be happy to promote such success stories on their LangPolNews and LangTechNews!

To summarize the eight steps through the funding maze:

1. Be clear about WHAT you want to do
2. Establish WHERE the seat of your legal entity is located
3. Make up your mind about WHO shall perform the project
4. Get INFORMATION on funding programmes
5. Establish good COMMUNICATION with local/regional/national funding agencies
6. IDENTIFICATION the best suited funding programme
7. ACTION time: Write the proposal
8. MARKETING your success - in terms of receiving funding and promoting results!

Compiled and edited by the LT Observatory project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 644583.

4. RECOMMENDATIONS FOR POLICY MAKERS

In order to derive to valid recommendations, a SWOT analysis was carried out to define the status quo of LT/MT in Europe. This analysis is based on discussions carried out during the course of the LTO project and the events organised by it and where partners participated. It therefore gives a broad picture that was confirmed by a majority of stakeholders in the LT community.

Strengths	Weaknesses
▸ Europe is still a leader in translation technology and has strong positions in speech technology and analytics; ▸ Many research groups in the field of LT; ▸ Organisations and initiatives in the field of LT and LRs; ▸ Several (federated) LR repositories.	▸ Inconsistent support for LT; ▸ Lack of an overall EU LT strategy, in particular with regard to the Digital Single Market; ▸ Unfavourable environment for LT start-ups ▸ Dispersed and uncoordinated actions in the field of LT and LRs that leads to a fragmented market at all levels (languages, LRs); ▸ Lack of awareness of benefits of LT; at policy level (EU an national), as well as demand and supply side.
Opportunities	**Threats**
▸ LT as key enabler of eCommerce, eGovernment, Big Data, Artificial Intelligence; ▸ Key technology that could differentiate Europe: machine translation, dialogue interaction, robotics, sentiment text analysis; ▸ Global growth of the LT sector; ▸ CEF.AT for public services: can become an infrastructure for the European language Infrastructure (basic layers e.g. LRs) and a driver for the European LT industry ▸ Greater provision of multilingual public services; ▸ Multilingualism is engrained in Europe's culture, hence an integral part of European thinking; ▸ Bringing the technology and experiences from the multilingual EU context to the global level; ▸ Cross-integration and interoperability towards a real LT-ecosystem.	▸ Quick-acting competition (US, China, etc.); ▸ Large foreign companies have been systematically acquiring promising EU companies; ▸ European public institutions and companies rely on language technologies developed outside Europe; ▸ "Free" solutions from foreign providers that harvest data; ▸ Unfavourable regulatory environment: IPR, data protection and privacy regimes do not allow data harvesting as, e.g. in the US.

TAB. 2: SWOT analysis for LT in Europe, with emphasis on MT in the framework of the CEF

Legend: End users | LTs developers/providers | LRs developers/providers | Policy/decision makers

Compiled and edited by the LT Observatory project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 644583.

Given the strategic importance of LT, ignoring the challenges (and therefore opportunities) would fundamentally set back the economic and political development of Europe for generations to come. Therefore, we recommend the following:

4.1. CREATE A EUROPEAN LT STRATEGY

A borderless infrastructure is needed to guarantee access to all (official) languages and tools for all European citizens, businesses, researchers, and public administrations. Currently, many single modules and solution are available, but a full LT-ecosystem with interoperable plug-ins is not yet a reality. Similar missing links can be observed in other vertical value chains, such as media and publishing, teaching and learning, pharma/chemicals, health care, and more.

The Action Plan given below gives an indication as to how the most pressing developments can be realized with the help of public stakeholders:

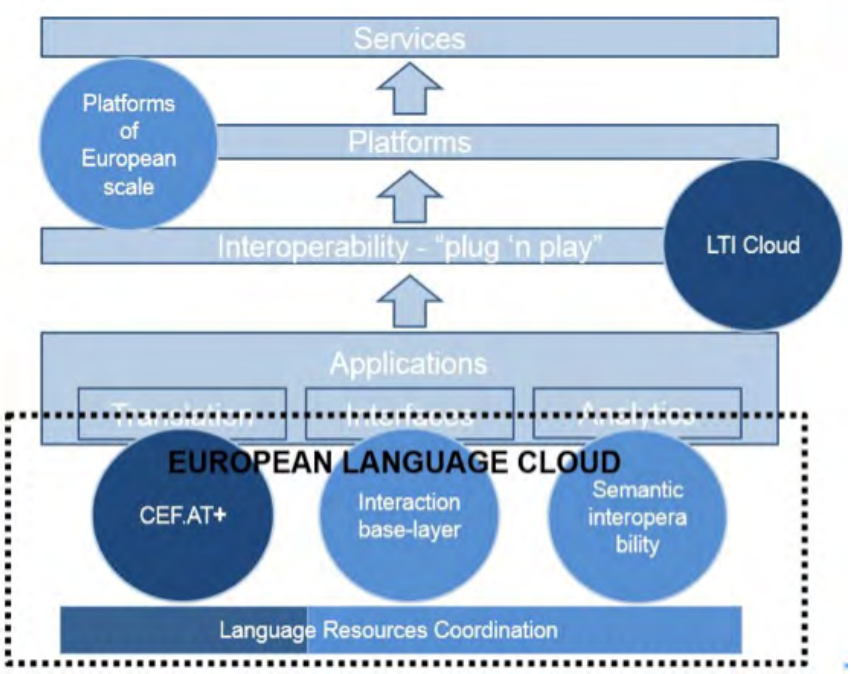

FIG. 3: Action plan at European level

The dark blue elements are currently being implemented. These include the Automated Translation (AT) for the Connecting Europe Facility intended to make all Digital Service Infrastructure multilingual, and the LTI Cloud that federates tools from the LT businesses across the EU in a common cloud marketplace to provide easy access, especially for SMEs. At the infrastructure level, there are two urgent action areas - speech technology and semantic interoperability. The creation of a Europe-wide infrastructure needs initial financial support at EU level to avoid immediate market fragmentation into language or sector.

At the platform level, much can be done by the private sector or through research and innovation projects. However, truly European scale platforms are needed to implement the Digital Single Market.

Legend: End users LTs developers/providers LRs developers/providers Policy/decision makers

Compiled and edited by the LT Observatory project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 644583.

This is not only a question of strategy, but of concerted action and financing.

At the Luxembourg Round Table (13 December 2016), the “European Language Infrastructure” was suggested by the LTI President (J. Hummel) and found approval by all present. The idea of having a basic, open infrastructure (mainly open LRs for all purposes) on which (public and private) services can be deployed found large approval by the entire audience.

It should be noted that some representatives of media called for an opening of CEF.AT for “industries of public interest” (like media) but this suggestion was met with criticism as it might distort competition.

The strategy pursued by the Spanish “Plan for the Advancement of LT” is based on 4 pillars of which the 1st is a “support for the development of a linguistic infrastructure” (the 3 other pillars are: 2: Promoting the language industry; 3: The public administrations as drivers of the language industry and 4: Flagship projects). Ideally, a European Language Infrastructure should be aligned to national infrastructures and developed in cooperation with them.

4.2. SUPPORT FOR LT/LR/MT AT NATIONAL/REGIONAL LEVEL

There are two crucial ways to support this drive to a multilingual digital singe market at national/ regional level so that Europe remains multilingual and the digital economy is open for all:

-Reuse of public sector content for language resources. Less-used languages in the EU suffer from machine translation efforts due to the lack of substantial digital language resources. New resources must be found or created, and public sector information is the most obvious place to search for relevant content to build such multilingual resources.

- Use of national/regional funding for innovative language projects for national or regional languages would provide an excellent source of support for this strategic endeavour. Results from any projects involving local/regional language data collection/creation could plug smoothly into the different layers of the ecosystem as described above.

4.3. CREATE AN IPR REGIME THAT SUPPORTS LT

One way to accelerate the accessibility of large sources of language content to feed the MT would be make an exemption under European copyright law. This should be adapted to accommodate the possibility of “decompiling” documents into their linguistic components (sets of phrases that do not allow the original document to be recreated) for the purposes of developing new language resources for machine translation. This action could be modelled on the existing provision from the "reverse engineering/decompilation" exception inscribed in the EC Software Directive - art. 6).

At least, Europe should opt for a clause like the “fair use clause” in the US that would allow web crawling for LR collection for MT training.

Compiled and edited by the LT Observatory project that has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No. 644583.

5. CONCLUSIONS

The common aim of all the stakeholder in the European machine translation ecosystem should be ubiquitous MT deployment across Europe, drawing on language resources, technology expertise, and awareness-raising and marketing through LSPs, public services, online services, commercial platforms, mobile MT services, etc.

In terms of language resources for machine translation, the current situation can be summed up as follows:

- Many resources spread over different repositories, but very few for commercial use.
- Not all MT practitioners know major repositories.
- About 150 LRs (corpora, terminology, lexica) identified that correspond to minimum criteria.
- There are huge gaps in terms of languages and domains:
  - Even large languages (FR, DE, ES) appear more often in multilingual corpora than bilingual ones; the situation is worse for less used languages;
  - 19 domains identified in the corpora but with very different coverage (amount/languages);
  - Most domains are public sector LRs.
- The creation and sharing of new is LRs difficult due to European copyright, privacy and data protection rules.

All stakeholders in the European MT ecosystem should participate in and support the following shared actions:

- To strengthen EU initiatives (including funding) to:
  - coordinate national/regional initiatives
  - create a European Language Infrastructure
- To coordinate among existing (federated) LR repositories, also in terms of standardizing licensing models
- To raise awareness of the benefits of LT at all levels, including benefits of CEF.AT for public sector
- To open up LRs from public sector for all purposes, including commercial users
- To increase the number of LRs in terms of languages and domains
- To improve LRs in terms of:
  - Metadata
  - Accessibility
  - Usability
- To consider ways to make LRs creation and sharing easily & lawfully possible in Europe and discuss regulatory framework (IPR) with the authorities in charge.

Compiled and edited by the LT Observatory project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 644583.

6. ANNEX I: LANGUAGE TECHNOLOGY AND LANGUAGE RESOURCES FOR MT IN THE EU

We are living in a connected world. Digital technology allows for seamless, ubiquitous communication that brings the world closer together than ever. Language is often a *de facto* barrier to the full deployment of political or economic goals like the Digital Single Market - by nature a multilingual market. Like all digital technologies, **language technologies** (LT) are "vehicles" for reaching an ultimate goal.

By their nature, language technologies are closely related to content in various written and spoken languages and formats, and communication in a multitude of situations. Language technologies can provide solutions to multilingual needs in relation to economic and societal challenges. For more general information on language technology and frequently asked questions about language technology, see Chapter 2.

Language resources (LRs) are the fuel of most modern state-of-the-art language technologies: machine translation systems learn from translated text, named entity recognition systems from annotated language data, parsers from syntactically annotated data, speech recognition from transcribed speech and vast amounts of plain text etc. Insufficient amounts of training data limit the quality of the language technology system in question. This has serious consequences for applications: Data Value Chains may end up locked into language silos as translation is unreliable, the information extracted from Big Language Data (one of the largest and most significant Big Data sources) will be noisy (and unreliable), etc.

In 2012, the Language in the Digital Age White Papers showed that out of the 31 European languages covered, 21 have fragmentary, low or non-existent language resources support; only English is considered as having "very good support". The work done in the Language Technology Observatory project has revealed similar gaps in LR support for machine translation, in particular for the use by SMEs that represent the majority of European LT companies.

The European landscape currently exhibits the following shortcomings in terms of LRs for machine translation:

- Quantity gap

For LRs to be useful in an operational context they need to be available in large quantities. The quantities available today are largely insufficient to have a positive impact on the quality of MT in a commercial context.

- Usability gap

Very few LRs that are presently on offer in existing repositories correspond to minimum quality requirements on metadata. Moreover, it appears that only very few LRs are made available by repositories in a way that enables their straightforward commercial use. The latter is often restricted in the first place; licensing conditions are not clearly spelled out; contact persons to obtain additional information are not identified; where LRs are made available for a price, the usability-price relationship is often considered inadequate; etc. Hence, there is a need to improve the usability of existing LRs and to reflect seriously on the conditions at which they can/should be made available for commercial use (particularly when they have been compiled with the support of public money).

Compiled and edited by the LT Observatory project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 644583.

- Awareness gap

There is clear evidence that a large part of the demand side is unaware of the offer. The supply side is often unaware of the needs of the demand side in terms of languages, size, domain, text sorts, and metadata needed for an informed decision.

- Coverage gap

In terms of LRs needed for machine translation, in particular statistical machine translation, only English has a reasonable coverage in relation to volume as well as in relation to domains, and very limited number of languages has moderate support.

For corpora only English has a good coverage in terms of combinations with other languages. Spanish, German, French, Latvian, Romanian, Croatian, Polish and Lithuanian are moderately covered in relation to other languages. Maltese, Danish, Czech and Slovak are poorly covered. All languages, including English have gaps in relation to domains. Eurovoc top categories that are not represented in metadata of any language are: Trade, Finance, Transport, Agriculture Forestry and Fisheries, Production, Technology and Research, Geography, International Organizations. Besides, many subdomains within many other top categories are not represented in any or only in very few languages. For terminological resources, English, French and German have a good coverage of certain domains.

For full comparison of coverage gaps, see the Analysis of the coverage situation (see http://www.lt-observatory.eu/sites/default/files/docs/D1_4.pdf).

In order for all European languages to be on a roughly equal situation, these gaps need to be filled in the short term. Filling these gaps doesn't necessarily require developing new methods and techniques from scratch. Rather, the idea is to provide language technology and language resources that put **all European languages** on a roughly equal footing. This Guide provides some guidelines for doing that.

Compiled and edited by the LT Observatory project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 644583.

7. ANNEX II: STAKEHOLDERS AND ENABLERS

This chapter describes the stakeholders and enablers of LTs and LRs that may come in different types. The following sections are aimed at explaining what is the role of the stakeholders and enablers, and how they can play different roles. Many stakeholders may be enablers, but not all – and some enablers may also be stakeholders.

LT-OBSERVE

LT Observe (http://www.lt-innovate.org/lt-observe/directory/organisations) displays a list of LT developers and providers, LR providers, LSPs, research institutions, solution providers, or public authorities, which constitutes the directory of stakeholders. This directory is focused on the machine translation (MT).

7.1. WHO ARE THE MOST RELEVANT STAKEHOLDERS OF LTS AND LRS?

The most relevant stakeholders (stakeholder groups) of LTs and LRs are:

- LTs developers/providers
  - LTs developers/providers (i.e. organizations/experts/managers in the field of LTs – and the tools/systems they develop/provide)
  - LTs integrators (i.e. organizations/experts/managers engaged in the field of LTs integration)
- LRs developers/providers (i.e. organizations/experts/managers in the field of LRs – and the LRs they develop/provide)
- LTs- and LRs-related services developers/providers and users (i.e. organizations/experts/managers in the fields of services pertinent to LTs and LRs – and the services they develop/provide)
- Language service providers (LSPs) being users of LTs and LRs are a strong stakeholder group
- Researchers in the fields of LTs and LRs and related disciplines

Based on the refined above taxonomy, the following main stakeholder groups have been identified in connection with MT (as well as LTs and the respective LRs, if necessary) that are relevant for this MT EcoGuide:

- Researchers
- Developers
- Service providers
- Users
- Policy/decision makers

Each of these can further be subdivided into several categories. Most of these groups may have activities that belong to two or more of the categories above play.

Compiled and edited by the LT Observatory project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 644583.

7.1.1. DEVELOPERS AS STAKEHOLDER GROUP

Developers of MT systems may develop MT systems (as well as LTs and the respective LRs, if necessary) or adapt existing MT systems under contract with individual customers/users or market their own solution to several/many customers/users (see D1.3).

Developers often are focused on 'developing', and

- not considering possibly necessary services, such as maintenance services, consultancy services etc.
- not fully respecting aspects of integratability, interoperability, and extendibility (referring to scalability and upgradability).

On the one hand, their products (viz. MT systems as well as LTs and the respective LRs, if necessary) can be regarded as enablers for the economy at large. On the other hand, official language policies and organizations' language strategies (duly comprising aspects of MT systems as well as LTs and the respective LRs, if necessary) can be strong enabler for developers.

A sufficient availability of LRs of all sorts for developing, testing and training LTs, in particular MT systems would be a strong enabling factor for LTs developers and the LI at large.

7.1.2. SERVICE PROVIDERS AS STAKEHOLDER GROUP

The importance of service providers with respect to MT systems (as well as LTs and the respective LRs, if necessary) more often than not is underestimated. This particularly refers to vendor-independent service providers, especially LT consultants.

Besides, there is a highly developed and thriving part of the language industry in the form of the LSPs which are providing language services of all sorts to customers in industry, in the public sector as well as in the tertiary sector (i.e. NGOs, NPOs etc.).

European LSPs have been early adopters of LTs and are still the worldwide leaders in the production of post-edited MT (Lommel and DePalma, 2016). Open-source systems and hosted MT applications made MT accessible also to smaller LSPs that found a solution for the growing translation volumes in post-edited MT. The adoption has usually developed organically over time: LSPs have started to outsource human translation globally, this has been followed by experimenting with various (generic) MT applications and post-editing, and finally, by adopting (customized) MT into the core processes of LSPs. Adaptive MT may show an even bigger reception of MT at all levels of the translation service industry, and especially those LSPs that have observed MT sceptically until now.

The large – and some of the smaller – LSPs are sometimes also LTs developers (including MT systems and the respective LRs, if necessary), LRs developers, and/or providing pertinent services. LSPs are heavily invested in LTs (including MT systems and the respective LRs). They can be considered as strong enablers for enterprises or other organizations with a great need for automatically handling large volumes of language data.

A sufficient availability of LRs of all sorts for developing, testing and training LTs, in particular MT systems would be a strong enabling factor for LTs developers and the language industry at large.

Compiled and edited by the LT Observatory project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 644583.

7.1.3. USERS AS STAKEHOLDER GROUP

The stakeholder group 'users' is – as can be expected – highly fragmented. Besides, some kinds of users are also enablers and/or developers and/or service providers. Their combining characteristic is the need for LTs, LRs or LSPs for various purposes. They may be:

- Large organizations in the private sector or large non-profit organizations (NPOs) or large institutions in the public domain
- Small and medium enterprises (SMEs) or smaller non-profit organizations (NPOs) or smaller institutions in the public domain
- One-person enterprises (OPEs) or individual experts
- Any individual in need of LTs, LRs or LSPs.

7.1.4. POLICY AND DECISION MAKERS AS STAKEHOLDER GROUP

This stakeholder group can be broadly subdivided into:

- policy makers, comprising politicians and other leading figures in politics, law making and public administration at different levels
- decision makers in all kinds of organizations – depending on the size and organizational structure. This may comprise decision makers
  - at different levels, such as top management, sector management or division management
  - in different kinds of stakeholders' organizations

LT-OBSERVE

LT Observe (http://www.lt-innovate.org/lt-observe/directory/organisations) displays a directory of policy and decision makers. It comprises the name of the stakeholder, logo, city, country, website links.

7.2. WHAT IS AN ENABLER?

In the context of this Guide, enablers may be organizations (or parts thereof), individuals at different levels (such as decision makers, experts, etc.), products (such as LTs or LRs) or services, etc. In addition, there are a sort of meta-enablers, such as policies and strategies, standardization and certification, consultancy services, R&D and training. Needless to mention, investors can also be an important kind of enabler.

In connection with MT (as well as LTs and the respective LRs, if necessary) the following enablers or enabling factors have been identified under a broad perspective:

- Language policies and strategies by public institution
- Language policies and strategies of organizations
- Legal regulations, in particular legislation (such as the law on supporting the Catalan language, or laws defining the official language(s) of a country

Compiled and edited by the LT Observatory project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 644583.

- Technical regulations, especially standards and standardization
- Certification (in particular standards-based certification
- R&D and training with respect to the above
- Consultancy services with respect to the above
- Funding through public institutions or private investors (see chapter Funding opportunities)

There may be large-scale enablers and small-scale enablers as well as enablers of any size in between.

Possible roles and functions of enablers under the perspective of the MT EcoGuide are:

- Provide support and assistance to MT and LRs developers/providers as well as MT users
- Promote good MT solutions for given purposes in a generic way, i.e. under the perspective of integratability, interoperability, extendability (referring to scalability and upgradability)
- Promote and support good practice with respect to interoperability of operations, tools and LRs in a generic way, i.e. under the perspective of workflow integration and value chains
- Facilitating cooperation among stakeholder groups, in order to speed up the development and exchange of appropriate language resources

7.2.1. LANGUAGE POLICIES BY PUBLIC INSTITUTIONS AS ENABLERS

To develop a language and language industry related policy or strategy is a highly complex task involving many stakeholders. It needs thorough preparation and considerable systematic persuading efforts before and during the process if its implementation.

Large public governance institutions, such as the EU Institutions, national governments, regional governments/administrations can be strong enablers with respect to MT (as well as LTs and the respective LRs, if necessary), if they formulate respective policies and promotion/implementation strategies that involve language (see http://www.lt-innovate.org/lt-observe/public-policy-observatory).

Implementation and or promotion strategies for public language policies can make use of various instruments, such as:

- Legal regulations
- Promotion of and support for R&D activities (see examples at the EU level and national levels on LT-Observe)
- Support for technical regulations and standards-based certification
- Promotion of pertinent consultancy services
- Political support for innovation and investments in the development or deployment of language technologies that support the policies.

A sufficient availability of LRs of all sorts for developing, testing and training LT, in particular MT systems, would be a strong enabling factor for economic and societal development. The consideration of LRs in policies and strategies and especially opening up and making LRs available for commercial use is therefore a crucial element in language policies and strategies

Compiled and edited by the LT Observatory project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 644583.

7.2.2. LANGUAGE STRATEGIES IN LARGE ORGANIZATIONS AS ENABLERS

The larger the organization and the more the organization is globalized or globalizing, the more a full-fledged language strategy may become necessary and can be considered as an enabler. At enterprise level – especially in multinational enterprises – decisions on strategies, such as to invest in MT (as well as LTs and the respective LRs, if necessary) can be major enablers. The same is true for strategic decisions in large branches, subsidiaries or departments of the enterprise.

According to modern management models, decision makers of large organizations are concerned with quite some management issues, such as:

Stakeholders, comprising internal stakeholders (employees, customers, suppliers, investors) and external stakeholders (such as government and public administrations, public/media/NGOs, competitors etc.)

- Interaction, comprising resources, norms & values, concerns & interests
- Societal environment, comprising society, nature, technology, economy
- Processes, comprising management processes, business processes, support processes
- Structuring forces, comprising strategy, structure, culture
- Modes of development, comprising renewal, optimization

The implementation of a full-fledged whole-organization language strategy would concern more or less all of the above-mentioned issues. Under each of them LTs, LRs or the respective consultancy services should be considered.

> *A sufficient availability of LRs of all sorts for developing, testing and training LT, in particular MT systems, would be a strong enabling factor for innovation and business development.*

7.2.3. LEGAL REGULATIONS AS ENABLERS

Legal regulations and similar authoritative regulations are important enablers as they can strongly impact new developments. Usually they are based on one or the other public policy. Usually they more or less directly impact R&D activities, standardization and certification, services of all sorts, innovation and investment activities.

There is a close relation and interdependence between legal regulations and technical regulations, which is often not recognized.

7.2.4. TECHNICAL REGULATIONS – ESPECIALLY STANDARDS – AS ENABLERS

Technical regulations especially standards – are governing our lives as much as law (or legal regulations) is doing. As such they are also important enablers. Standards are of different nature from law, though complementary to each other. In cases of conflict or damage, technical standards are second to law, but if a standard is referred to in a law, it becomes part of the law.

If a standard pertinent to the fields covered in the MT EcoGuide would be referred to by law, this standard would become a legal enabler because of the legal pressure on stakeholders to abide by the law and adapt existing language industry products or services under political, legal or economic requirements.

Compiled and edited by the LT Observatory project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 644583.

Such standards in combination with the respective laws would for instance become a big enabler pushing towards interoperability and sustainability of systems/tools, products and services which has become a big issue in the language industry today.

The development of technical standards is the main purpose of standardization. Official standards bodies develop (de jure) standards at international, regional and national level – sometimes even at provincial or state level. They develop – depending on their sphere of authority – international or regional or national standards. This does not hinder situations, where a national standard becomes so widely acknowledged world-wide that it is recognized as international standard. Besides, ISO and CEN, IEC and CENELEC, ITU and ETSI mutually acknowledge their standards as international.

In addition – especially in the fields of the ICTs – there are numerous industry standards ('de facto standards').

Another formal distinction of standards depends on their degree of obligingness of the respective standards document (or deliverable). In ISO today there are:

- IS – International Standards: highly normative
- TS – Technical Specifications: normative under certain conditions
- TR – Technical Reports: informative nature

Besides, there may be basic or fundamental standards, and more or less vertical or more or less horizontal standards.

From various sources (e.g. IN LIFE deliverable D9.7) it can be concluded that other kinds of technical standards mentioned in ISO/IEC Guide 2:2004, can be subdivided according to their content or nature of content as follows:

- terminology standards (referring to individual terms and definitions)
- product standards
- service standards
- testing standards
- process standards
- interface standards
- methodology standards
- ICT hardware and software standards
- content standards (or data standards)
- Most of the standards are a mixture of two or more of the above. A standard may cover
- a comparatively small thematic field or application
- a comparatively big thematic field or application

Sometimes a single standard (such as the GSM standard for telecommunications) becomes the enabler of a whole industry, in other cases it needs a set of standards (such as the Internet governance related standards).

Standardization is a broad and highly complex field. Definitely "standardization" can function as enabler under several aspects. According to ETSI on "Society needs standards", standards support:

Compiled and edited by the LT Observatory project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 644583.

- Safety and reliability
- Implementation of government policies and legislation
- Interoperability
- Business benefits, such as
  - Open up market access
  - Provide economies of scale
  - Encourage innovation
  - Increase awareness of technical developments and initiatives
- Consumer choice

> *The EU Rolling Plan for ICT Standardisation (see http://ec.europa.eu/growth/sectors/digital-economy/ict-standardisation/) is a short- to medium-term work programme in ICT standardisation that is arranged by topic, linking EU policies to standardisation activities.*

Standardization activities are – and should be at least in the area of official ('de jure') standardizing organizations – open to experts of all walks of life: industry, administration, academic research, consumers' interests. There are many misconceptions about standardization, such as in:

- SMEs: is it worthwhile to engage and spend time as well as money for standardization activities?
- Academia: are standards of relevance to research?
- NGOs and NPOs: what is the benefit for us?

In large enterprises and public administration, the benefits of standardization activities are not doubted.

SMEs can not only benefit from the know-how transfer in the course of standardization activities, but also influence the content of technical standards. Academia should be aware that tools and methods they use are heavily influenced by standards and that standards (formal or de facto) make it possible to share and to build on each other's results. NGOs and NPOs could make better use of standards for achieving their goals. To all parties in standardization the following is valid:

- The very fundamental principles of standardization are geared towards integration and interoperability as well as sustainability under various aspects and different levels.
- The gain of know-how by being active in standardization outnumbers by far the information that can be gathered from the text of the standard.
- Last but not least it may be mentioned that hardly any standardization organization does not have an explicit or implicit language policy (at least implicitly) and does not invest in ICTs – some in particular in LTs sometimes including MT.

A certain knowledge of standards and standardization should be the rule in organizations and institutions. Often, active involvement in standardization activities may prove to be highly beneficial to the organization/institution – if not necessary.

> *For recommended LT- in LRs-relevant standards, see recommendations by CLARIN, and business-relevant standards and guidelines in language industry.*

Compiled and edited by the LT Observatory project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 644583.

7.2.5. CERTIFICATION – IN PARTICULAR STANDARDS-BASED CERTIFICATION – AS ENABLER

As quality is an important cost, image and market success factor, certification is defined as a procedure by which a first, second or third party gives written assurance that a process, product, service, skill or competence conforms to specified requirements. If these requirements are specified in a standard, the certification process would assess the standards compliance of the respective product or service, individual's skill/competence (up to the respective training and training material). Even the successful implementation of a language policy could be certified. Successful implementations of pertinent standards-based certification schemes in the fields of concern in this MT EcoGuide are for instance LICS, the Language Industry Certification System (see: www.lics-certification.org) and the European Certification and Qualification Association (ECQA) (see: www.ecqa.org).

In the context of the MT EcoGuide, certification may refer to:

- Tools or systems of the language technologies (LTs)
- Language resources (LRs)
- Language services and their provision by language service providers (LSPs)
- eCertification of LTs experts
- Training of LTs experts and the respective training material
- Implementation of a language policy or strategy
- Skills and competences of pertinent consultants and experts (for example ISO 17100:2015 Translation services -- Requirements for translation services, see http://www.iso.org/iso/catalogue_detail.htm?csnumber=59149; and ECQA Terminology Manager, see http://www.ecqa.org/index.php?id=52, etc.)

7.2.6. R&D AND TRAINING AS ENABLER

Research & Development are one of the most important enabling factors in society in general and for the availability and use of language technologies, language resources for various purposes. There are a number of R&D and similar initiatives that have driven innovation in the LT and LR sectors, such as

- LT-Innovate (see http://www.lt-innovate.org) is the language technology industry association with a number of activities and sub-networks
- CLARIN (see https://www.clarin.eu/) is the European Research Infrastructure for Language Resources and Technology
- META (The Network of Excellence forging the Multilingual Europe Technology Alliance, see http://www.meta-net.eu/) with its sub-components and sub-networks
- The Federation "Cracking the Language Barrier" (see http://www.cracking-the-language-barrier.eu/) brings together a number of European research and innovation projects as well as related community organisations working on or with cross-lingual or multi-lingual technologies, in neighbouring areas or on closely related topics
- FLaReNet (Fostering Language Resources Network, see http://www.flarenet.eu/)

  ELRA/ELDA (European Language Resources Association/Evaluations and Language Resources Distribution Agency, see http://www.elra.info/)
- TAUS (Translation Automation User Society, see http://www.taus.net/)

Compiled and edited by the LT Observatory project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 644583.

The European Commission has been instrumental in setting incentives for these R&D initiatives to flourish by asking the related research and industry communities to carry out R&D projects such as EuroMatrix (see http://www.euromatrix.net/), Euromatrix Plus (see http://www.euromatrixplus.net/), Let's MT (see https://www.letsmt.eu), CASMACAT (see http://www.casmacat.eu), TCSTAR (see http://www.tcstar.org/), AbuMaTran (see http://www.abumatran.eu/), QT LaunchPad (see http://www.qt21.eu/launchpad/) and others.

In the 20th century machine translation R&D has been one of the most important driving factors in the development of information and communication technologies in general. Long before translation studies emerged as an academic discipline in the humanities, machine translation was already operational in the Cold War period, although its performance was extremely poor in the first decades.

Education and training in language technologies in general and in machine translation in particular have been another key enabler in progress in this industry sector. Numerous B.A. and M.A. programmes today are dedicated to translation studies, many of them including or even focusing on translation technologies, specialized translation, translation corpora, multilingual terminologies and related topics. The European Master of Translation (EMT, see https://ec.europa.eu/info/education/european-masters-translation-emt/european-masters-translation-emt-explained_en_en) is a quality label initiated by the DGT (Directorate General for Translation) of the EU-Commission in cooperation with universities all over Europe. It is imperative that translators are trained in using MT systems, computer assisted translation systems, terminology database management systems, as translators are not only the most important users of language resources such as corpora and termbases, but also important producers of such resources.

7.2.7. CONSULTANCY SERVICES AS ENABLER

Consultancy services – especially if based on formal recognition, such as certification – are assisting any kind of stakeholder and, therefore, can be considered as enabler. Consultancy is insofar closely intertwined with standardization, as:

- Consultants should know about the standards in their respective consultancy field
- Formal qualification of consultants should include the knowledge about standards
- Pertinent certification systems for consultants should be based – as much a possible – on pertinent standards.

As with standardization and certification activities, consultancy services can also be considered a service to industry & business and to society at large. Given their importance for many aspects in an organization, consultancy services should follow good practice. Given the speed of development in nearly all walks of life, this complies regular formal fresh up and upgrading of the competences of the consultants – e.g. in the form of skill & competence certification.

In the whole LI area covered it needs more and more consultants and their expertise on the whole area covered by LI to provide enterprises with the theoretical and practical knowhow needed to use or introduce language technologies and language resources. On the one hand, there is a need for consultants having a broad overview of existing developments, as well as of developments, which are likely to lead to major changes in the ICT- and content-world. This applies to consultants rendering services to

- governments or other major administrative public services with respect to language policies
- enterprises in need of upgrading their language- and LTs-related capacities

Compiled and edited by the LT Observatory project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 644583.

- LTs enterprises under pressure to cope with the speed of ICTs innovation
- LRs developers with respect to apply methodologies and LTs tools/systems recognized as good practice
- LSPs with respect to the need of innovation from within
- peer consultants or consultancy enterprises which need to upgrade themselves.

On the other hand, consultants specialized on individual aspects of LTs and LRs are also needed.

The above applies first of all to vendor-, system- and approach-independent consultants who advise organisations (incl. LT and LR developers as well as LSPs) in such a way that inappropriate decisions can be avoided, which might turn out to be costly in the end.

Potential customers of consultancy services, therefore, are advised to screen thoroughly the professional background and credentials of consultants they want to hire.

Consultants could tremendously upgrade their role and reputation if they can show a record of certification and achievements.

7.2.8. FUNDING THROUGH PUBLIC INSTITUTIONS OR PRIVATE INVESTORS AS ENABLERS

Funding innovation is always a potential enabler or at least, an incentive. However, in many cases money is only one factor, the other being recognition, collaboration forging or paving the way towards future networking opportunities.

While single projects (in particular for national regional languages) can find support in national/regional programmes, federating efforts, infrastructures or large-scale networking coordination needs support at a supra-regional level, such as the EU.

Unfortunately, seed money for initial ideas or VC at European level is scarce, therefore private investors usually stick to the home market. This effect results in a fragmented market which, again, leads to no critical mass when it comes to competition from outside Europe, e.g. the US.

With regard to “complementary funding opportunities” see Funding opportunities.

7.3. INTERRELATIONS AND INTERDEPENDENCIES BETWEEN STAKEHOLDERS, ENABLERS AND APPLICATION ASPECTS

Under a LI perspective, standardization, certification and language policy/strategy are not only closely related to each other, but also highly interdependent. The rapidly growing market of the LI has led to a differentiation of demands on the customer side, and LI products and services offered by the LI. This and the general demand for system integration and interoperability have triggered the need for language strategies (or policies), standardization, as well as for quality assessment systems (see: certification). This development has also had a great impact on the competences and skills taught at higher educational institutions (including the respective academic certification systems) as well as on LRs and pertinent teaching and training services.

Compiled and edited by the LT Observatory project that has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No. 644583.

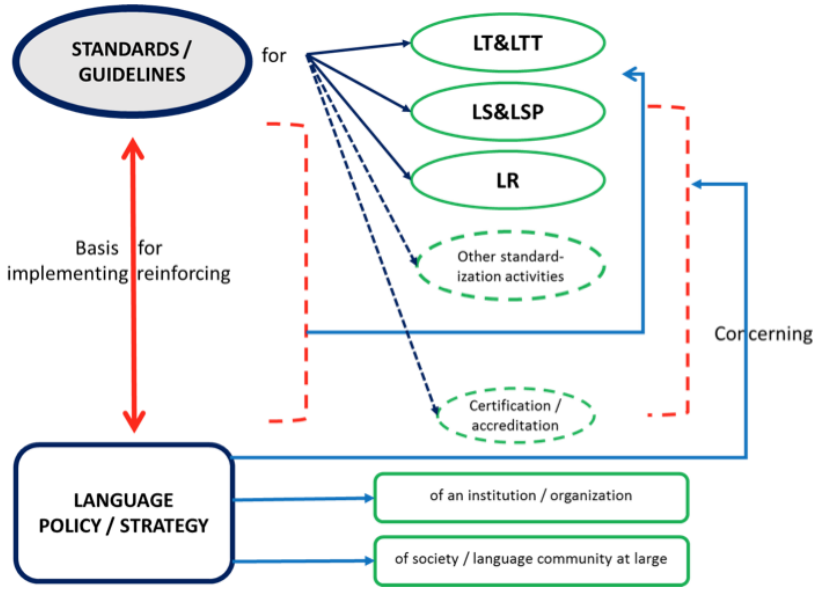

FIG. 4: Relations between the various aspects of the language industry

Standardization, certification and language policy/strategy development can also be considered as a service:

- standardization activities are a service to industry & business and to society at large;
- (particularly standards-based) certification schemes support the pursuit of a high degree of quality, reducing costs and potential for conflict thus also helping to establish a high level of trust;
- an explicit language policy, as it helps to establish a better understanding for national development strategies (if there is an official national language policy) or enterprise globalization/ localization strategies (if there is an enterprise-specific general language policy/strategy). “eCertification” (or ICT certification) refers to certification activities (mostly based on provider-specific training) by which an individual gains a credential in a particular ICT skill or more generally a range of skills. While it can be recognized that certification provides value in both the labor and product segments of the ICT market, a report [CEPIS 2007] describes over 600 often overlapping qualifications from over 60 providers as a “certification jungle”, causing confusion to prospective users. The above-mentioned market certifications contrast and co-exist with the historic university based education system, leading to the phrase “parallel universe”. They are seen in academic quarters as developing skills not education, and product ability not underlying theory, little more than marketing aids to the commercial interests of the vendors. On the other hand, their global application contrasts with the national or even self-accreditation of most university degrees.

Compiled and edited by the LT Observatory project that has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No. 644583.

8. ANNEX III: SUPPORT FOR LANGUAGES

From the chapter Stakeholders and enablers it becomes clear that the support for languages, in particular through a national language policy or strategy of some kind or other, can be a strong development and innovation factor not only for language technologies, language resources and language service providers, but for whole language communities. The same applies – however, with different means of implementation – for any larger organizations or institutions. In reality, language policies and strategies largely overlap.

Language technologies help maintaining the linguistic diversity, and at the same time foster cross-lingual communication, be it for e-commerce, e-government, culture or education.

8.1. LANGUAGE POLICIES AND STRATEGIES

8.1.1. PUBLIC LANGUAGE POLICIES AND STRATEGIES

Language policies, especial 'national' language policies, can be a strong cultural-linguistic support for a language community. Language policies are not only complex from a socio-culturally and socio-economically point of view, but also involve highly sensitive issues. Implemented at different levels they are or can be an enabler for MT- and other LTs- as well as LRs-related developments.

A 'national' (public) language policy may be conceived for

- one country/region,
- several countries where the same language is used (albeit with certain variations),
- parts of more than one country (or communities in cross-border regions) where the same language is used (albeit with certain variations),
- the whole or parts of a national or regional government.

Language policies are mainly apparent where a country/region wants to preserve its language and fosters it through language learning or preservation strategies. Instead of an explicit language policy, taking into account language in a set of other policies may well be as effective.

Public language policies and strategies hold a potential to effectively support other types of policies, such as innovation policy, trade policy, information policy, ICT policy etc. – in reality, these cross-fertilizing policies are rarely implemented. For details on national/regional strategies in Europe and examples of best practices, see Language policies by public institutions as enablers

8.1.2. ORGANIZATIONAL LANGUAGE POLICIES AND STRATEGIES

The more an organization – not only multinational enterprises, but in fact also SMEs – does business at a global level, the more it has to deal with language, or rather localization issues. Given the aggregated costs of dealing with many localization issues individually without coordination, a language strategy is commended.

Large administrative organizations or major parts of it can also decide to develop a language strategy to address the multilingual challenge in communication.

The following aspects may, inter alia, be considered when developing an organizational language policy or strategy:

Compiled and edited by the LT Observatory project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 644583.

- integration into innovation strategy, trade strategy, information strategy, ICT strategy etc.
- benefits in terms of positive development of an organization
- decision makers' aspects: purpose, benefits, options, human resources/content resources,
- commercial aspects (while not excluding non-commercial aspects)
- introducing/upgrading LT systems/tools (duly considering pertinent services)
- developing or making use of existing LRs (duly considering pertinent services)
- making use of existing or develop consultancy services
- using standards or engaging in standardization activities
- getting certified or engage in the development of certification schemes
- considerations of size of enterprise, customer, service provider,
- consideration (due to the speed of development of the ICT industry) of imminent developments: Internet of Things,Industry 4.0, Big Data … (duly analysed well beyond the "hype" aspects)

8.2. SUPPORT AT THE EU-LEVEL

Currently, the EU has 500 million citizens, 28 Member States, 3 alphabets and 24 official languages, some of them with a worldwide coverage. Some 60 other languages are also part of the EU's heritage and are spoken in specific regions or by specific groups. In addition, immigrants have brought a wide range of languages with them; it is estimated that at least 175 nationalities are now present within the EU's borders.

Linguistic diversity is enshrined in Article 22 of the European Charter of Fundamental Rights ("The Union respects cultural, religious and linguistic diversity"), and in Article 3 of the Treaty on European Union ("It shall respect its rich cultural and linguistic diversity, and shall ensure that Europe's cultural heritage is safeguarded and enhanced.").

Languages are traditionally associated with culture and education: Language as an expression of culture, and language learning as an essential part of education. Therefore, many language policies at EU level were initiated by DG Education & Culture. However, the European Commission has supported Human Language Technologies for more than 40 years. There was a considerable effort made throughout 1980-1990 which resulted in some pioneering Machine Translation and Translation Memory technologies. Financial support for language technologies reached a peak during the 7th Framework Programme (DG CNECT, then DG InfSo).

The current EU ambition to create a Digital Single Market revives the support for language technologies, e.g. for cross-border transactions: More and more commercial transactions are being done online and there are more consumers using the Web that do not speak English than those who do. Recent e-commerce statistics indicate that two out of three EU customers buy only in their own language. This suggests that language is a significant barrier to a truly Europe-wide Digital Single Market. Language barriers do not only impact e-commerce activities, but also have their repercussion on access to content and online services. This refers particularly to eGovernment services that will be taken care of by the Connecting Europe Facility (CEF). The multilingual element of this initiative is spearheaded by the multilingual building block CEF.AT based on DG Translation's MT@EC tool that is open to all public institutions of all Member States and disposes of a corpus of all official EU languages. CEF and Horizon 2020 hold a potential of working hand-in-hand for funding relevant projects that support CEF's multilingualism.

Compiled and edited by the LT Observatory project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 644583.

8.3. SUPPORT FOR LANGUAGES AND LTS/LRS/MT AT THE NATIONAL/REGIONAL LEVEL

National and regional language strategies are mainly apparent where a country/region wants to preserve its language and fosters it through language learning or preservation strategies. It is far rarer to find strategies that involve language technologies at national level.

One of the first attempts was made by France in the early 21th century with its Technolangue programme (2003-2006) that is recently taken up again for a potential "Technolangue II". Countries with official languages that are also languages of lesser distribution (for example Ireland) are keener to engage in technologies that can help their language to gain momentum. This can be also seen in the Baltic countries where language strategies promote their official national languages, often as contrast to formerly used languages like Russian. Some Member States promote the learning of several languages in order to enhance one's own plurilingual portfolio, in line with EU educational policies. In some instances, regions have developed their own language strategies (e.g. Wales or Catalonia).

The socio-economic element in assessing languages and language technologies is missing in all Member States policies. Although there are hardly any Member States that have a real strategy in terms of language technology, there are some promising strategies, even best practices.

LT-OBSERVE

LT Observe (http://www.lt-innovate.org/lt-observe/directory/organisations) displays a directory of national policies and strategies in Europe. It includes the information on the official and co-official languages, policies or strategies in the field of language, and policies or strategies in the field of language technologies. It also displays the country's stakeholders, and policy and decision makers, as well as other relevant organizations.

Currently, Spain can be named as a best practice example for an all-encompassing language technology strategy that was published in its "Plan de Impulso" in 2015. Through this strategy Spain dedicates more than 90 million EUR to language technologies for its national and regional languages. As such, it is currently the highest doted national initiative in the area of language technologies (see http://www.lt-innovate.org/lt-observe/spain).

Another example is Ireland with its "20 years strategy for the Irish language" 2010 to 2030 that explicitly includes language technologies (see http://www.lt-innovate.org/lt-observe/ireland).

Compiled and edited by the LT Observatory project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 644583.

9. ANNEX IV: LIST OF ABBREVIATIONS

CAT	computer-assisted translation
CEF	Connecting Europe Facility
CEN	European Committee for Standardization
CENELEC	European Committee for Electrotechnical Standardization
CMS	content management system
DTP	desktop publishing
ESO	European standards organization
ETSI	European Telecommunication Standards Institute
G11N	globalization (in the sense of the localization industry)
GSM	Global System for Mobile Communications
GUI	graphical user interface
HLT	Human Language Technology (domain)
HLTs	human language technologies (subject)
I18N	internationalization (in the sense of the localization industry)
ICT	Information and Communication Technology (domain)
ICTs	information and communication technologies (subject)
IEC	International Electrotechnical Commission
IS	international standard
ISO	International Organization for Standardization
ITU	International Telecommunication Union
LR	language resource
LRs	language resources
LSs	language services (subject)
LSP	language service provider
LT	language technology
LTs	language technologies
MT	machine translation
NGO	non-governmental organization
NPO	non-profit organization
R&D	research and development
SaaS	software as a service
SMT	statistical machine translation
TMS	terminology management system
TR	technical report (in standardization)
TS	technical specification (in standardization)

Compiled and edited by the LT Observatory project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 644583.

10. ANNEX V: GLOSSARY

adaptive machine translation	Machine translation that enables use of available data sources to create customized and personalized MT engine for each user, adapted to user domain, style, and feedback. SDL XMT is an example of this approach.
controlled language	A subset of natural languages that use restricted grammar and vocabulary in order to reduce or eliminate ambiguity and complexity Controlled languages enable authors to write texts that are easily comprehensible. . A controlled language is an interesting solution for authors who write texts for translation. (http://www.muegge.cc/controlled-language.htm, accessed 2016-12-21)
enabler	The original meaning of "enabler" is: one that enables another to achieve an end. (http://www.merriam-webster.com/dictionary/enabler, accessed 2016-08-03). Depending on the nature of the "one" it can mean many things. Today it is used in several contexts, e.g. in the field of business it is defined as capabilities, forces, and resources that contribute to the success of an entity, program, or project. (http://www.businessdictionary.com/definition/enablers.html, accessed 2016-08-03)
extendibility→extensibility	<software engineering> systems design principle where the implementation takes future growth into consideration NOTE: Extensibility is a systemic measure of the ability to extend a system and the level of effort required to implement the extension. Extensions can be through the addition of new functionality or through modification of existing functionality. The central theme is to provide for change – typically enhancements – while minimizing impact to existing system functions.
gist translation	Use of (human or machine) translation to create a rough translation of the text that allows the reader to understand the essence of the text.
globalization	A broad range of processes necessary to prepare and launch products and company activities internationally. All the business issues associated with this decision have to be addressed, such as integrating localization throughout a company after proper internationalization of the product design.
hybrid machine translation	Machine translation that takes advantage of statistical and rule-based machine translation approaches and combines them. Examples include SYSTRAN, Omniscien Technologies (formerly Asia Online), and PROMPT.
integratability	<information technology> capability of using processes for linking together different computing systems and software applications physically or functionally, to act as a coordinated whole NOTE: There are different kinds of system integratability, e.g. aiming at horizontal integration, vertical integration, star integration (also known as spaghetti integration), continuous integration etc. A common data format – viz. an application-independent (or common) data format – Is an integration method to avoid every adapter having to convert data to/from every other applications' formats,
interoperability	<information technology> degree or extent to which diverse environments (hardware and software systems, products or components) are able to exchange information without loss of content and in a manner transparent to the user (modified ISO 29362:2008) NOTE: It was early recognized that technical interoperability may not suffice without 'organizational interoperability' – and further 'semantic interoperability' (semIOp). Under the requirements of multilinguality and multimodality – e.g. in connection with eAccessibility, eInclusion mContent – interoperability needs an extension towards 'content interoperability'.
internationalization	An approach to facilitate localization into a multitude of target markets by designing a software application so that it can be adapted to various languages and regions without engineering changes. Deployment of internationalization is considered strategically at the beginning of content development, not after original content is already developed.

Compiled and edited by the LT Observatory project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 644583.

localization	The adaptation of a product (incl. computer software) or communication to a community of speakers with respect to cultural, linguistic, legal, political and other aspects
machine translation, MT	Also automated translation. The application of computers aiming at automatically or semiautomatically translating text or speech from one natural language to another. Increasingly, MT is also used for speech translation.
neural machine translation, NMT	Utilizes neural networks, which are trained by deep learning techniques. Training is computationally expensive. This and other obstacles have so far limited NMT to research environment, but Google Translate and Systran announced shift to NMT in 2016, so the development of this approach should be closely followed.
operational usability	Operational Usability of Language Resources means being able to easily access: ▸ Links from LT Observe to (parallel) language resources that match an end-user's needs ▸ Peer-reviewed by experts ▸ Bearing valid metadata, including production date, ownership and contact information ▸ Either free of charge or at a reasonable price for commercial purpose ▸ Provide description on domain/content of the LR ▸ All listed in the LT Observe at http://www.lt-innovate.org/lt-observe/resources-list)
parallel data	Also bitext. A collection of sentences in two different languages, which is sentence-aligned. This means that each sentence in one language is matched with its corresponding translated sentence in the other language.
post-editing	Process by which humans review, edit, and improve the quality of machine translation output.
rule-based machine translation, RBMT	Machine translation that utilizes linguistic rules covering morphological, semantic, and syntactic regularities and lexical items and maps them from source to target languages. Examples include Apertium, Lucy, and GramTrans.
scalability	<information technology> capability of a system, network, or process to handle a growing amount of work, or its potential to be enlarged in order to accommodate that growth NOTE: Scalability, as a property of systems, is generally difficult to define and in any particular case it is necessary to define the specific requirements for scalability on those dimensions that are deemed important. It is a highly significant issue in electronics systems, databases, routers, and networking. A system whose performance improves after adding hardware, proportionally to the capacity added, is said to be a scalable system.
stakeholder	A person, group or organization that has interest or concern in an organization or in an issue NOTE: Individual stakeholders can affect or be affected by the organization's actions, objectives and policies. ... Not all stakeholders are equally powerful. (http://www.businessdictionary.com/definition/stakeholder.html, accessed 2016-08-04).
statistical machine translation, SMT	Machine translation that utilizes statistical translation models generated from the analysis of bilingual and monolingual training data and relies on large quantities of these types of data. The output of SMT can be less accurate, but more natural sounding than rule-based machine translation output. Examples include the open-source Moses (see http://statmt.org/moses/; released under GNU Lesser General Public License) that is often customised in professional context, and online generic services such as Google Translate and Bing.
standardization	"Activity for establishing, with regard to actual or potential problems, provisions for common and repeated use, aimed at the achievement of the optimum degree of order in a given context" "NOTE 1 – In particular, the activity consists of the processes of formulating, issuing and implementing standards." "NOTE 2 – Important benefits of standardization are improvement of the suitability of products, processes and services for their intended purposes, prevention of barriers to trade and facilitation of technological cooperation." (ISO/IEC Guide 2:2004)
upgradeability	<information technology> capability of a ICT system to be replaced by a newer or better version, in order to bring the system up to date or to improve its characteristics NOTE: Although developers usually produce upgrades in order to improve a product, there are risks involved—including the possibility that the upgrade will worsen the product.

Compiled and edited by the LT Observatory project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 644583.

11. ANNEX VI: LIST OF TOOLS

In the table below, the tools for getting statistical machine translation relevant language data and tools on the web are listed. Detailed guidelines how to stepwise approach acquiring domain specific parallel data via web-crawling are described in the Best practice guide to LRs for automated MT. This list is not to be regarded as exhaustive.

Name	Functionality	Supported file formats	Availability	Available from:
STRAND	Identification of mutual translations on the Web. Only access to STRAND bilingual databases	The STRAND database format	GNU public license (GPL)	http://www.umiacs.umd.edu/~resnik/strand/
Bitextor	Mining of parallel documents on the Web	HTML,XHTML, XML	GNU public license (GPL)	https://sourceforge.net/projects/bitextor/
ILSP Focused Crawler	Research prototype for acquiring domain-specific monolingual and bilingual corpora	HTML, XML	GNU GPL, v. 3.0 license	http://nlp.ilsp.gr/redmine/projects/ilsp-fc/wiki/Introduction
Nutch framework	Web crawler	No information	Apache License, version 2.0	http://nutch.apache.org/
SpotSigs	Filtering near duplicates	HTML, XML	GNU public license (GPL)	https://sourceforge.net/projects/spotsigs/
Boilerpipe	Detection of Boilerplates etc	HTML	Apache License version 2.0	https://github.com/kohlschutter/boilerpipe
HunAlign	Sentence aligner	TXT	GNU LGPLv3	https://github.com/danielvarga/hunalign
Geometric Mapping and Alignment (GMA)	Sentence aligner	TXT	GNU public license (GPL)	http://nlp.cs.nyu.edu/GMA/
Bilingual Sentence Aligner (BSA)	Sentence aligner	TXT	Microsoft Research end user license agreement (MSR-EULA)	https://www.microsoft.com/en-us/download/details.aspx?id=52608

TAB. 3: Tools for getting statistical machine translation relevant language data and tools on the web (the list is not exhaustive)

Compiled and edited by the LT Observatory project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 644583.

In the table below, the tools for (semi)-automated term extraction are listed. Detailed guidelines how to stepwise extract (corporate) terminology are available in the Best practice guide to LRs for automated MT. This list is not to be regarded as exhaustive.

Name	Languages supported	Supported file formats	Availability	Available from:
AlchemyAPI	EN, FR, DE, ES, IT, PT, RU, SV	HTML; TXT, or url	Commercial	http://www.alchemyapi.com/api/keyword-extraction
AntConc	Any	TXT, XML, HTML	Free	http://www.antlab.sci.waseda.ac.jp/
Fivefilters	Any	Plain text via web interface or url	Free	http://fivefilters.org/term-extraction/
Lexterm	Any	TXT, *.csv	Free	https://github.com/LexTerm
SDL MultiTerm Extract	Any	TXT, RTF, *.doc, *.xsl, *.ppt, HTML, TMX[1]	Commercial	http://www.sdl.com/cxc/language/terminology-management/multiterm/extract.html
TaaS	EU official languages, RU, TR	PDF, *.doc, *.xsl, *.ppt, TXT, RTF, XLIFF, HTML, XML, MIF	Basic version free; Premium version commercial	https://term.tilde.com/technology
TerMine[2]	Any	Plain text via web interface, TXT, HTML, PDF	Free	http://www.nactem.ac.uk/software/termine/#form
Terminology Extraction by Translated	EN, IT, FR	Plain text via web interface	Free	http://labs.translated.net/terminology-extraction/
SynchroTerm	All EU official languages, except ET, GA; HR; LV; MT	*.doc, *.xsl, RTF, TXT, HTML, PDF, TMX	Commercial	http://www.terminotix.com

TAB. 4: Term extraction tools (in alphabetical order; the list is not exhaustive)

Compiled and edited by the LT Observatory project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 644583.

12. ANNEX VII: REFERENCES

12.1. INPUT DOCUMENTS

D1.1 Report on resources (http://www.lt-observatory.eu/sites/default/files/docs/D1_1.pdf)

D1.2 On-line catalogue on resources (http://www.lt-observatory.eu/sites/default/files/docs/D1_2.pdf)

D1.3 Best practice guide to LRs for automated MT (http://www.lt-observatory.eu/sites/default/files/docs/D1_3.pdf)

D1.4 Analysis of coverage situation (http://www.lt-observatory.eu/sites/default/files/docs/D1_4.pdf)

D2.1 Database of stakeholders (http://www.lt-observatory.eu/sites/default/files/docs/D2_1.pdf)

D3.2 ESIF Funding Opportunities (http://www.lt-innovate.org/sites/default/files/InfoGuide_ESIF_Funding_Opportunities_2016_web.pdf)

D3.1 National and regional funding opportunities (http://www.lt-observatory.eu/sites/default/files/docs/D3_1.pdf)

D3.4 Guidelines to Funding Opportunities (http://www.lt-innovate.org/sites/default/files/National_regional_funding_opportunities_online_rev.pdf)

D5.1 Position Paper and Preliminary joint Strategic Research and Innovation Agenda for the LT/MT field (http://www.lt-observatory.eu/sites/default/files/docs/D5_1.pdf)

D5.2 Strategic Research and Innovation Agenda for the LT/MT field (http://www.lt-observatory.eu/en/our-results)

Multilingual Europe: The Crowning Touch to the Digital Single Market - A Call for Action (http://www.lt-innovate.org/lt-observe/document/multilingual-europe-crowning-touch-digital-single-market-call-action)

European Platform for the Multilingual Digital Single Market (http://www.lt-innovate.org/sites/default/files/Multilingual%20Platform%20Concept.pdf)

How multilingual is Europe's Digital Single Market? (https://ec.europa.eu/commission/2014-2019/ansip/blog/how-multilingual-europes-digital-single-market_en)

Resolution of the Riga Summit 2015 on the Multilingual Digital Single Market (http://www.rigasummit2015.eu/summit-resolution)

12.2. OTHER REFERENCES

- Documents:

European Commission, Directorate-General for Internal Market, Industry, Entrepreneurship and SMEs (2016). 2016 Rolling Plan on ICT Standardisation. (http://ec.europa.eu/DocsRoom/documents/15783/attachments/1/translations/en/renditions/pdf)

European Commission, Directorate-General for Regional and Urban policy (2014). Enabling synergies between European Structural and Investment Funds, Horizon 2020 and other research, innovation and

Compiled and edited by the LT Observatory project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 644583.

competitiveness-related Union programmes. (http://ec.europa.eu/regional_policy/sources/docgener/guides/synergy/synergies_en.pdf)

Gazzola, M. (2016) European Strategy for Multilingualism: Benefits and Costs. (http://www.europarl.europa.eu/RegData/etudes/STUD/2016/573460/IPOL_STU(2016)573460_EN.pdf)

Infoterm et al. (2012). Overview on the language industry (LI) products and services. (http://www.celan-platform.eu/assets/files/CELAN_D2-1_Annex 1_Overview_fv1-2.pdf)

LT-Innovate (n.n). LT industry definition/taxonomy. (http://www.lt-innovate.org/lt-observe/document/lt-industry-definition-taxonomy)

Lommel, A. R. and DePalma, D. A. (2016). How Europe Is Driving the Shift to MT. Common Sense Advisory.

Perez Giraldo, S. B. et al. (2012). Investigation of business -relevant standards and guidelines in the fields of the language industry. (http://www.celan-platform.eu/assets/files/CELAN_D2-1_Annex%202_Standards_fv1-2.pdf)

Rehm, G. and Uszkoreit, H. (eds). (2012). Language in the Digital Age White Papers Series. (http://www.meta-net.eu/whitepapers/overview)

Thicke, Lori (2013). The industrial process for quality machine translation. JoSTrans, 19/2013. (http://www.jostrans.org/issue19/art_thicke.php)

Wacker, P. (2016). Response to European Parliament / STOA Experts' Questionnaire on Market and Economic Impact of the Human Language Technology Sector. (http://www.lt-innovate.org/sites/default/files/STOA HLT Interview - LT-Innovate.pdf)

- Websites:

AbuMaTran: http://www.abumatran.eu/

AlchemyAPI: http://www.alchemyapi.com/api/keyword-extraction

Bilingual Sentence Aligner: https://www.microsoft.com/en-us/download/details.aspx?id=52608

Bitextor: https://sourceforge.net/projects/bitextor/

Boilerpipe: https://github.com/kohlschutter/boilerpipe

Business Dictionary: http://www.businessdictionary.com/definition/ enablers.html, http://www.businessdictionary.com/definition/stakeholder.html

CASMACAT: http://www.casmacat.eu

CLARIN: https://www.clarin.eu/

Cracking the Language Barrier Federation: http://www.cracking-the-language-barrier. eu/

DGT Translation Memory: https://ec.europa.eu/jrc/en/language-technologies/dgt-translation-memory

ECQA: www.ecqa.org

Eureka: http://www. eurekanetwork.org/eureka-countries

Euromatrix Plus: http://www.euromatrixplus.net/

Euromatrix: http://www.euromatrix.net/

European Association for Machine Translation: http://www.eamt.org/mt.php

Compiled and edited by the LT Observatory project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 644583.

European Language Resource Coordination (ELRC): http://lr-coordination.eu/

European Master in Translation: https://ec.europa.eu/info/education/european-mashttps://ec.europa.eu/info/education/european-masters-translation-emt/european-masters-translation-emt-explained_en_enters-translation-emt/european-masters-translation-emt-explained_en_en

Eurostars: https://www. eurostars-eureka.eu/eurostars-countries/europe

Fivefilters: http://fivefilters.org/term-extraction/

FlaReNet: http://www.flarenet.eu/

Geometric Mapping and Alignment: http://nlp.cs.nyu.edu/GMA/

HunAlign: https://github.com/danielvarga/hunalign

ICT standardization: http://ec.europa.eu/growth/sectors/digital-economy/ict-standardisation/

Let's MT: https://www.letsmt.eu/Login.aspx

Lexterm: https://github.com/LexTerm

LICS: http://www.lics-certification.org

LT at DFKI: https://www.dfki.de/lt/lt-general.php

LT-Innovate: http://www.lt-innovate.org

Merriam Webster: http:// www.merriam-webster.com/dictionary/enabler

META-NET: http://www.meta-net.eu

MOSES: http://statmt.org/moses/

Nutch framework: http://nutch.apache.org/

OPUS: http://opus.lingfil.uu.se/

Panacea: http://www.panacea-lr.eu/

QT LaunchPad: http://www.qt21.eu/launchpad/

SDL MultiTerm Extract: http://www.sdl.com/cxc/language/terminology-management/multiterm/extract.html

SpotSigs: https://sourceforge.net/projects/spotsigs/

Standards and formats – recommendations by CLARIN: https://www.clarin.eu/content/standards-and-formats

STRAND: http://www.umiacs.umd.edu/~resnik/strand/

SynchroTerm: http://www.terminotix.com

SYSTRAN: http://www.systransoft.com/

Taas: https://term.tilde.com/technology

TAUS: http://www.taus.net

TCSTAR: http://www.tcstar.org/

TerMine: http://www.nactem.ac.uk/software/termine/#form

Terminology Extraction by Translated: http://labs.translated.net/terminology-extraction/

Compiled and edited by the LT Observatory project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 644583.