Plan for the Advancement of Language Technology

October 2015

TABLE OF CONTENTS

1 Rationale

Both the Digital Agenda for Spain and the Spanish Strategy for Science, Technology and Innovation establish development of the digital economy and society as one of the global challenges that require the greatest efforts in Scientific and Technological Research, Development and Innovation (R&D&I); **they also underscore the leading scientific, technological and business position of the Information and Communications Technology (ICT) sector as one of Spain's strengths.** This industry is identified as a strategic innovative area with considerable potential to increase the competitiveness of the productive fabric, drive growth and boost job creation. Furthermore, in their Research and Innovation Strategies for Smart Specialisation (RIS3), the Autonomous Regions of Spain also highlight the potential for ICT to be a motor for the economy.

In this regard, the conclusions and recommendations of the ERAC Peer Review of Spanish Research and Innovation System, carried out by a group of European experts in R&D&I and presented on 24 July 2014, also encourage development of initiatives in **Strategic Innovation Areas**, led by businesses and the public administrations. **A set of priorities are set out therein, reflecting market opportunities, existing strengths and future potential in the competitive international arena.** The report lists three key principles for action:

1) **Identification of large-scale innovation projects** from a specialisation perspective, with a substantial R&D&I component and a significant contribution from information technology.

2) The need for **medium-term approaches and instruments** (5-10 years).

3) **Concentration of resources to improve efficiency and effectiveness of aid** and increase the impact, with effective changes in the economic and social development of the targeted sectors or regions.

The State Secretariat for Telecommunications and the Information Society (hereinafter SETSI, its Spanish acronym) has the following **duties**, among others:

- Studying, proposing and implementing **general telecommunications and information society policy**.
- **Promoting and developing** advanced telecommunications and information society **infrastructure and services.**
- Fostering and coordinating **plans, technology projects and programmes** to develop the information society.
- Preparing, managing and monitoring **programmes to promote the range of new technologies, services, applications and content** offered with regard to

telecommunications and the information society; also, drafting and coordinating this policy with corresponding European Union programmes and initiatives and other international programmes in this area.

SETSI is also responsible for **implementing and coordinating the Digital Agenda for Spain** and is the managing body for actions that form part of the 2013-2016 Strategic Action Plan for the Digital Economy and Society.

The ongoing development of the internet and ICT in general **provides access to enormous—and growing—volumes of textual information**. However, IT systems, which are able to easily process data, cannot directly process human language.

Language is one of the tools that set human beings apart. As for mathematics, it forms the foundation of our science and the basis of our information and communications technology. **Natural language processing technology brings together these two implements, applying scientific methods and information technology to the understanding of human language** and its diverse and plentiful tongues, dialects and means of communication. Tools such as search engines, intelligent personal assistants, text classifiers and machine translation have become essential to our day-to-day work, whatever our field. However, there are many other situations in which application of natural language processing and machine translation may be key to offering citizens new advanced services and optimising processes and productive resources in both business and in public administrations. Any step towards improved understanding, synthesis, classification or machine translation of unstructured textual information **generates value for society and may be applied to all business sectors**.

The language technology market is growing rapidly and reports by a number of consultancy firms[1] estimate **substantial growth in the global market over the coming years**, based on the explosion in applications seen in the last two years and the exponential growth in digital textual data.

In addition to this technological progress, other key factors support the timeliness of this plan.

Notably, at the **European Union** level, the European Commission has prioritised the establishment of a Digital Single Market (DSM). This brings with it challenges such as overcoming the barriers caused by the existence of many different languages within the EU. However, this linguistic diversity is also one of its greatest cultural assets. Language technology therefore plays a crucial role within the European Union.

Spanish is the world's second most widely spoken language after Mandarin Chinese, based on the number of native speakers, and **third in terms of total number of speakers, behind English.** Spanish is forecast to be the second most widely used language in business transactions globally, primarily due to growth in the Latin American market. The Spanish language's capacity

[1] *Gartner 2014, LT-Innovate, AltaPlana, etc.*

for internationalisation is huge, as nine in ten speakers are located outside Spain[2]. The fact that this language is shared with other nations represents, above all else, an opportunity to **strengthen ties with the Ibero-American community in terms of cooperation**.

Spain also has **internationally renowned organisations that specialise in the Spanish language**, including the Royal Spanish Academy, other royal academies and Instituto Cervantes.

Moreover, the country is home to **numerous globally respected research groups**, which work on the processing of the Spanish language and the co-official languages of Spain, as well as groups specialising in semantics. The researchers and many of the businesses in the sector are grouped in the Spanish Society for Natural Language Processing (hereinafter SEPLN, its Spanish acronym), which celebrated its 31st anniversary in 2015, as well as in a number of business associations.

However, development of applications for a specific language and, in many cases, for a particular area of knowledge, is dependent on the availability of technology and resources for the language in a designated field of expertise. In the case of Spain, the availability of such resources for the Spanish language—but to a lesser extent and with some considerable gaps—is similar to that for German or French, despite the number of Spanish speakers being much higher. For the co-official languages of Spain the level is lower. The cost of these resources is substantial and cannot be borne by small- and medium-sized enterprises (SMEs). To ensure that applications are available in Spanish and the co-official languages of Spain, the **number, quality, variety and availability of the supporting resources and tools must be improved**.

Furthermore, in anticipation of growth in the natural language processing and machine translation sectors, it would be advisable to expand **training** in these technologies to ICT professionals in the private and public sectors.

The **public administrations** should integrate natural language processing and machine translation technology into their processes, to **improve quality and increase the capacity of public services**, as well as **driving** demand. In some areas, such as healthcare and the judiciary, the administration plays a key role in development of new services based on better understanding or machine translation of the content managed.

What is more, the extraordinary potential value as a language resource of much of the information generated by the public sector represents an outstanding opportunity to develop the natural language processing industry[3]. The policy on Re-use of Public Sector Information (**RPSI**) is a means of developing this **linguistic linked open data**, as its aim is to make the data generated by the public sector in the course of its duties available to society as an open resource that may be used for financial gain.

1.1. The Plan for the Advancement of Language Technology

[2] *Information obtained from the 2015 report by Instituto Cervantes, "El español: una lengua viva". Chapter 1. Spanish in numbers. http://eldiae.es/wp-content/uploads/2015/06/espanol_lengua-viva_20151.pdf*

[3] *For example, this is reflected by the fact that the most downloaded item from the EU Open Data Portal is the EURPARL Parallel Corpus (https://open-data.europa.eu/es/data/).*

This situation has led SETSI to **foster the natural language processing and machine translation sectors through this targeted plan**, with a time horizon of five years and a geographical and institutional scope covering the various Autonomous Regions and co-official languages of Spain. This Plan expands on the content and scope of the Digital Agenda for Spain, as has been done since the Agenda was approved.

Given the **multidisciplinary** nature of language technology, the Plan is **interministerial** and is based on promoting language technology by coordinating all the actions of Spain's Central Administration, in conjunction with the authorities of Spain's Autonomous Regions.

The initiative to draft the Plan began with the following **SWOT analysis**:

STRENGTHS	WEAKNESSES
• **A high standard of research** in natural language processing, coordinated by the Spanish Society for Natural Language Processing. • **Good governance** of the Spanish language (Royal Spanish Academy and the Association of Spanish Language Academies in Ibero-America). • **Extensive experience in multilingualism**, due to the presence of co-official languages in Spain.	• The sector comprises **small- and medium-sized enterprises** that do not have the industrial capacity to compete on international markets or complete the value chain within Spain. • The **difficulty of transferring knowledge** from the research sector to the business sector, primarily due to the cross-sector and multidisciplinary nature of natural language processing.
OPPORTUNITIES	THREATS
• The Spanish language's high potential for **internationalisation** and cooperation with **Ibero-America**. • New **public services** for citizens and companies in strategic sectors (e.g. healthcare, tourism and education). • A **fast-expanding market** tied to innovation and development. • The potential of **RPSI** as a source of extremely valuable language resources for business and research.	• The **loss of economic and industrial competitiveness** of Spain and of Ibero-America. • The **digital underdevelopment of the Spanish language.** • The **digital extinction of co-official languages of Spain.** • The **flight of researchers and professionals** and the deterioration in the Spanish research sector.

This analysis can be summarised in the following **key ideas**:

- The language technology sector is a **cross-cutting emerging sector**, linked to innovation, with the capacity to drive growth, competitiveness and high quality employment.

- Its **development is unstoppable**, but if we do not seize upon this opportunity, others will.

- Spain has the means, but the country's Central Administration must **promote and coordinate initiatives** in conjunction with the authorities of Spain's Autonomous Regions and with the countries of Ibero-America, to make the most of this opportunity.

The general aim of the Plan for the Advancement of Language Technology is therefore to **promote development of natural language processing and machine translation** in Spanish and Spain's co-official languages, by means of the following specific goals:

1. **Increasing the amount, quality and availability of linguistic infrastructure** in Spanish and in Spain's co-official languages.

2. Fostering the language industry by promoting **knowledge transfer from the research field to that industry**. Bolstering **internationalisation** of companies and institutions in the sector. Improving the **reach** of current projects.

3. **Improving the quality and capacity of public services**, integrating natural language processing and machine translation technologies, simultaneously **driving demand**. Supporting **creation, standardisation and distribution** of language resources created by the management activities performed by the public administrations.

The aim of the present Plan is for the advancement of language technology to be coordinated, **seeking synergies and eliminating overlapping**, in accordance with the recommendations of the Commission for the Reform of the Public Administrations (known as **CORA**, its Spanish acronym).

2 Background

2.1 An introduction to language technology

Natural language processing is the path towards **better automated understanding of human beings' greatest creation: language.** Language is the most commonly used and versatile form of communication. IT systems can easily process data (i.e. structured information with a unique, explicit meaning), such as tables containing millions of numbers. However, human language is much more complex. Meanings can change according to context, and can refer to implicit information. Nevertheless, the **volume of unstructured digital textual data is astounding** and is growing at a dizzying rate, making assistance necessary if it to be processed automatically.

Despite this difficulty, **today, the state of development of natural language processing technology allows us to use a multitude of useful applications**. What is more, the fast-paced development of this technology over the past decade heralds increasingly surprising results.

Natural language processing and machine translation technologies **make it possible to analyse texts and enable commonly used IT programs to work with them** in a number of different sectors, such as healthcare, education and tourism. Some of the tasks that may be performed include: named-entity recognition (peoples' or companies' names, brands or place names), filtering and classification of documents, automatic summarisation, data extraction, sentiment analysis, opinion mining, social media reputation management and monitoring, correction of spelling and grammatical errors, smart searches, optimised searches, automated responses to questions and intelligent personal assistants, and machine translation of texts. All of these applications can be summarised as the **leveraging of unstructured information to improve understanding of texts in documentary corpora**.

Properly applied, such textual analysis tools can help companies and organisations to **optimise many of their production processes and obtain highly valuable insights** from their own information and that available from our increasingly digitalised and global world. Language technology is a facilitating industry that forms part of the horizontal structure behind many applications and devices.

2.2 Analysis of the current situation

The language technology sector has been described in detail in a number of recent reports[4]. It has been estimated that the global language technology market generated an aggregate turnover of €19.3 billion in 2011, **and in 2015 it is forecast to reach €30 billion[5].** The same report outlines the current status of the European market, affirming that there are more than 500 companies linked to this industry. They tend to be SMEs concentrated in northern Europe that **cannot cover the entire European or international market**. The report recommends that these companies be supported through specific infrastructure, providing access to resources and technology.

[4] *Prepared by organisations such as LT-Innovate, Meta-Net, Gartner, CRACKER and LT_Observatory.*
[5] *"LT2013: Status and Potential of the European Language Technology Markets", LT-Innovate.*

To date, the machine translation market has been primarily European. It is a **strategic sector for the European Union**, which has 24 official languages, some of which are in danger of digital extinction. At present, the business sector comprises approximately 2,500 companies and 800 research centres related to natural language processing[6]. According to an EU report[7], in terms of sales and services the European machine translation sector reached a level of €8.6 billion in 2011 and **will reach €14.9 billion in 2015**. Growth rates in less mature foreign markets will be much higher.

The calls for the **H2020 Framework Programme** and the definition of the **Connecting Europe Facility (CEF)** both aim to roll out digital services at a pan-European level, including language services, to ensure the establishment of a mechanism to coordinate information and provide access to language resources for European languages.

Spanish is the second most widely spoken language in the world, after Mandarin Chinese, in terms of native speakers, and also the second most spoken language in terms of the total number of speakers. **Spanish speakers represent 6.7% of the global population**, amounting to 470 million people. Forecasts indicate that by 2030, Spanish speakers will constitute 7.5% of the global population[8].

The use of a common language triples Spain's export share to Spanish-speaking countries. Nine out of ten users of the technology promoted by this Plan reside outside Spain. Moreover, there are Spanish-speaking communities in North America with enormous growth potential. Our geographical and linguistic proximity and close historical ties also bring us closer to the other major language of the Ibero-American Community, Portuguese.

Furthermore, Spain has prestigious international organisations that specialise in the Spanish language, including the Royal Spanish Academy and Instituto Cervantes, with extensive reach in Ibero-America. There are also respected academies in specific areas of expertise such as medicine and engineering.

In addition, researchers and many of the businesses in the industry are grouped in the **Spanish Society for Natural Language Processing**, which celebrated its 31st anniversary in 2015. There are also **many internationally renowned research teams in Spain**, working on processing of the Spanish language and the co-official languages of Spain, as well as groups specialising in semantics. Several scientific and business groups have also been formed in Catalonia, the Basque Country and Madrid in relation to the language industry.

Although there is clearly the potential for this industry to thrive in Spain, the study by the Multilingual Technology Alliance (META) and the related META-NET Network of Excellence *Europe's Languages in the Digital Age*[9], highlights that the **lack of natural language processing resources for specific languages are the most significant barrier** that must be overcome for language industry applications to be developed in Europe, such as:

[6] *"Strategic Agenda for the Multilingual Digital Single Market", CRACKER and LT_Observatory.*
[7] *"LT2013: Status and Potential of the European Language Technology Markets", LT-Innovate.*
[8] *"El español: una lengua viva" 2015 Report, Instituto Cervantes.*
[9] *http://www.meta-net.eu/*

- **Optimisation of industrial processes** for language-based document management: translation of documents and tools for authors (e.g., spellcheckers, document generation).
- **Communication and intelligent personal assistants** (virtual assistants, human-machine communication for cars, customer service and interaction with robots; smart searches and automated responses to questions).
- **Smart processing of information and knowledge** (e.g., extraction and mining of data from text and content, classification of documents, automatic summarisation).
- **Assisted language learning.**

In European programmes, however, **there is no effective supply of resources to constitute a supporting infrastructure for each language**. The supply of resources for Spanish and Spain's co-official languages must be guaranteed, as although there is a language technology industry, its nature (most companies are SMEs) means that it cannot be expected to generate resources quickly enough to become self-sufficient.

2.3 Preparation of the Plan for the Advancement of Language Technology

The initiative to prepare a Plan for the Advancement of Language Technology came from the State Secretariat for Telecommunications and the Information Society (SETSI). For this purpose, a **Steering Committee was formed to prepare the Plan**, with the participation of the following bodies:

- Secretariat-General for Industry and Small and Medium-sized Enterprises
- State Secretariat for Research, Development and Innovation
- State Secretariat for Culture
- Under-Secretariat for Education, Culture and Sport
- State Secretariat for International Development Cooperation and Ibero-America
- Under-Secretariat of the Ministry of the Presidency
- Directorate for Information and Communications Technology of the General State Administration
- State Secretariat for Tourism
- Deputy Directorate-General for Information and Communications Technology of the General State Administration, of the Ministry of Health, Social Services and Equality

Other bodies that have shown interest in participating are also expected to join the Committee shortly.

In turn, the Steering Committee created a **Committee of Experts to prepare a preliminary Report** to support the creation of the Plan for the Advancement of Language Technology[10]. The Committee of Experts comprises:

- Royal Spanish Academy
- Instituto Cervantes
- Institute of Spain (comprising nine specialised Royal Academies)
- National Library of Spain
- Spanish Royal Academy of Engineering
- Research groups specialised in natural language processing and machine translation (SEPLN)
- Key companies in the Spanish ICT sector
- Business groups specialised in natural language processing and machine translation
- Policy-makers involved in the re-use of public sector information (RPSI policy)
- Directorate for Information and Communications Technology of the Administration (known by its Spanish acronym, DTIC)
- Commission for the Reform of the Public Administrations (CORA)
- State Secretariat for International Cooperation and Ibero-America.

The analysis performed by the Committee of Experts is summarised in the following **SWOT analysis**:

STRENGTHS

- Development of lines of research in natural language processing in Spain, **covering almost all areas currently being studied internationally**. Spain has its **own, established, robust resources and tools** for basic natural language processing and machine translation of Spanish, Catalan, Basque and Galician and also for English, as well as extensive **public sector information** that may be converted into language resources.
- **A wealth of Spanish researchers**, who participate in projects, associations and standardisation groups at European and global levels. There are **more than 30 consolidated research groups** organised into associations and networks, giving rise to nine spin-offs, covering almost all of the areas currently being studied internationally.
- The possibility of quickly establishing **models for collaboration with research centres** thanks to national transfer programmes (Strategic National Technical Research Consortiums, National Business Research Consortiums, industrial doctorates).
- Spanish companies' and public administrations' **extensive experience in management of**

[10] *Report on the current state of language technology in Spain, in the context of the Digital Agenda for Spain.*

STRENGTHS

multilingualism, which may be exportable**.**

- Spain's leading position in the **potential Spanish-language market**, with some 470 million speakers, 54 million of whom are in the USA, a leading economy. A number of Spanish companies have opened **offices in the USA and Ibero-America**, demonstrating the viability of internationalisation of the technology developed in Spain and reflecting Spain's flourishing ICT sector, which is proving its ability to compete at a global level.
- Spain forms part of the prestigious Ibero-American networks of the **institutions related to the languages of Spain** (e.g., Royal Spanish Academy, Institute of Catalan Studies), which collaborate to regulate language at the international level.
- The low **cost of re-using and raising the profile** of all of the available materials for specialised language, due to its existing structure and organisation and proven **track record in strategic initiatives** under national plans.
- The existence of **databases on terminology**, synonyms, thesauruses and place names, which can be universally exported. The existence of an experienced **industrial machine translation framework** with key clients in the public administrations.
- The existence of operational machine translation initiatives, such as the Plata platform[11].
- Spain's accession to the European Directive on the re-use of public sector information.
- Past **involvement in European initiatives**, such as Meta-Share, Clarin-Eric, and ELRA, and European projects such as OPENER, NEWSREADER, and QT-Leap.

WEAKNESSES

- **Insufficient inter-company collaboration and collaboration between companies and research groups**, which prevents re-use of data and tools, multiplies investment by companies, and reduces the effectiveness of other areas, e.g. marketing. **Lack of knowledge and coordinated investment and limited sharing** by companies and academia of tools with extensive coverage, making it difficult to apply methods that ensure re-use, resulting in duplication and dispersion of work when building corpora, tools, etc.
- **Lack of interdisciplinary collaboration** due to the limited interaction between linguists and IT specialists when creating and sharing resources and applications, due to the rigid structure in universities and the lack of research centres focusing on natural language processing and machine translation.
- **Cuts to financing of research** into natural language processing and machine translation, hindering progress in research and preventing retention of specialised teams.
- Insufficient **investment in training** of highly qualified specialists. A lack of educational opportunities related to postgraduate qualifications, masters and targeted grants, leading to a **limited number of spin-offs** in Spanish universities.
- Insufficient **financing facilities to support internationalisation** of companies and R&D&I projects. However, the **amount of investment required is substantial** in terms of both time and money, to develop new technological tools.
- Inadequate **capacity for investment** in research and innovation and a **limited capacity for**

[11] *http://administracionelectronica.gob.es/ctt/plata*

WEAKNESSES

internationalisation of products, as a result of a lack of capitalisation, as most companies are SMEs or microenterprises.

- **Insufficient basic research and technological development** in relation to natural language processing and machine translation by public research agencies and a **decline in the number of research groups working on natural language processing and machine translation** in major companies in Spain (e.g., IBM).
- Limited availability of resources and tools for **Latin American Spanish**.
- Weakness with regard to processing of specialised texts, as Spanish is **not a major language in scientific literature and patents**.
- **Limited knowledge among Spanish companies of the standards, licences and business models** already agreed at the European level.
- **Lack of recognition of this discipline in the national academic community**, despite being an interdisciplinary research area. Training is scattered amongst different specialities, reducing the field's visibility.
- **Lack of greater insight into the characteristics of this technology**, with limited program licence costs, but high investment requirements in terms of localisation and adaptation (language and areas of activity), which would allow for much more efficient processes in many applications within the public administrations.
- Lack of a **specific interoperability standard**.
- Largely undefined marketing strategy, making it very difficult to access the market.
- Lack of an established RPSI culture in the various stakeholder groups.

OPPORTUNITIES

- Now is an **excellent time in Europe to develop natural language processing and machine translation**, following the recent accession of new states. Infrastructure, associations, foundations and distributors have already worked together on formal matters: standards, licences and business models. There are now **European and international initiatives**, offering linguistic data through open data portals, which could be joined.
- **Urgent demand for applications** related to social media, big data and open data, establishing short-term goals that can already be met through mixed development consortiums, which would considerably raise the profile of this field.
- **Existence of multiple areas of use**: tourism, healthcare, justice, education, etc., in which horizontal processes can be optimised and systemised, providing proof of concept and insight for future projects, with regard to their possibility of generating **re-usable resources**.
- Value may be generated by defining **linguistic linked open data sets** within the RPSI strategy.
- The presence of more than **300 researchers in Ibero-America**, primarily in Mexico, with whom to collaborate on development of infrastructure in the region.
- The expansion of social networks and big data processing, putting language industries in an excellent position to explore new areas of action and to access basic resources to continue to improve systems.
- **Many players** in the sector. Specifically, **horizontal and global requirements** have been

OPPORTUNITIES

identified in relation to the industry in all the public administrations.

- The existence of **doctorate programmes in Spain** dealing with natural language processing.
- The existence of **EU R&D&I programmes** to help finance new projects in this field.
- The possibility of **improving accessibility** for groups with functional limitations.
- The existence of the Semantic Interoperability Centre and the availability from the Technology Transfer Centre of material to find **shared solutions within** the public **administrations**.

THREATS

- **A decline in competitiveness** with regard to other countries, such as the USA, in development of resources and tools for natural language processing and machine translation of Peninsular and Latin American Spanish.
- A lack of agreement on standards and licence models, among the various **European associations and organisations**, which may erode Spain's industry if it does not participate more actively.
- The possibility of Spanish and the co-official languages of Spain disappearing as languages used in **specialised fields** if publications in Spanish are not promoted, along with their digital distribution and use.
- A need for **investment and planning**, to offer **high-quality linguistic data** through open data portals, making this activity unsustainable without targeted financing.
- **Competition from major corporations**, in the area of research and development, with research groups dealing with natural language processing and machine translation, in Spanish and the co-official languages of Spain.

3 Specific Aims of the Plan

The general aim of the Plan for the Advancement of Language Technology is to **develop the natural language processing and machine translation industries** in Spain, specifically for Spanish and the co-official languages of Spain.

This general aim can be broken down into the following **specific aims**:

1. **Development of linguistic infrastructure**

    - Advancing the natural language processing industry for Spanish and the co-official languages of Spain, providing general **linguistic infrastructure** (resources and processors).

    - **Reducing the gap** between linguistic infrastructure in Spanish and the co-official languages of Spain and those in **English**, in terms of the quantity, quality and availability of general natural language processing and machine translation infrastructure.

    - Ensuring **public availability** of high-quality linguistic infrastructure in Spanish and the co-official languages of Spain, **free of charge or at a low cost** (at least for innovative SMEs, the research sector and the public administrations).

    - Ensuring **coordination in the development of linguistic infrastructure,** avoiding overlapping and seeking synergies. Making common linguistic resources generation tools available, as well as linguistic infrastructure assessment campaigns.

    - Adopting **technical interoperability standards**, an appropriate licence policy and mechanisms for protecting personal data when generating language resources.

    - Promoting **methods of automatic generation** of language resources.

2. Promoting the **language technology** industry

    2.1. Improving the **visibility** of knowledge from the sector, and its **exchange**

    - Raising the **profile** of the industry for natural language processing and machine translation.

    - Transferring **Spain's excellence in research** to this industry. Attracting doctorate-level specialists back to Spain, and training new researchers and developers in these areas.

    2.2. Supporting **internationalisation and commercialisation** of the industry

    - Improving the **internationalisation** of companies in the sector, especially in the Ibero-American and North American markets.

- Using the tools provided through the Digital Agenda for Spain's Internationalisation Plan, which contains a framework for direct financing of companies' internationalisation (e.g., ICEX, AECID).

- Supporting **major Spanish corporations** that are well established in Ibero-America and other emerging markets, in their role as **drivers of the industry**.

- Strengthening **development cooperation with the Ibero-American community** in order to lead the implementation of the technologies of natural language processing and machine translation in Spanish. Institutional support for the project from Ibero-American countries and involvement therein.

3. The **public administrations as promoter of the language industry**

   3.1. Creating **shared platforms** for natural language processing and machine translation within the public administrations

   - **Improving the quality and capacity** of public services, by including natural language processing and machine translation technology.

   - Providing the public administrations with **shared tools** for natural language processing and machine translation.

   - Pursuant to the **CORA** recommendations, simplifying the roll-out of new services based on natural language processing and machine translation, seeking synergies and applying economies of scale. The aim is to save costs by sharing resources between public administrations and eliminating duplication of efforts, making procedures more efficient and improving insight. The platform will be based on re-usable and interoperable components, preferably with unrestrictive, open source licences.

   - Making shared tools available, to **generate, assess and leverage** language resources.

   - Sharing infrastructure and systems between **innovative projects**, supported by public institutions. They could also be used following completion of the projects as online demonstration centres.

   - Applying shared infrastructure and systems to the **research sector**, facilitating the development of new components and analysis of large corpora of documents.

   - Sharing services with other **institutions in the sector, including**: Royal Spanish Academy, Institute of Spain, Ibero-American organisations.

   3.2. Linguistic resources of the public administrations and policy for **re-use of public sector information**

- Promoting the natural language processing industry for Spanish and the co-official languages of Spain, providing **language resources generated from public sector information**.
- Ensuring that the language resources generated from public sector information are **publicly available, free of charge or at a low cost**.
- Ensuring **coordinated development of language resources** generated from public sector information, preventing overlapping work and seeking synergies.
- Developing and using **shared tools** to generate and assess language resources generated from public sector information.
- Adopting **technical interoperability standards**, an appropriate licence policy and mechanisms for protecting personal data.

The aim of the Plan is for the advancement of the language industry to be coordinated, **seeking synergies and eliminating duplication of efforts**, in accordance with the CORA recommendations.

4 Structure of the Plan

An interdepartmental **Steering Committee** has been formed to prepare the Plan for the Advancement of Language Technology, bringing together the bodies within the General State Administration that play key roles in this area. In turn, the Steering Committee requested a preliminary report from a **Committee of Experts** comprising representatives from the sectors involved (research, the business sector, the specialised Academies, the public administrations). The report concerns the state of language technology in Spain, also including recommendations on developing the industry in our country.

The measures proposed in the Plan aim to contribute to achieving the strategic objectives of the **Digital Agenda for Spain** and the specific objectives set out in this Plan. The measures are based on the following pillars.

Pillar 0: Governance

The first pillar aims to **define a mechanism for the various stakeholders in the industry to coordinate, collaborate on and exchange ideas, documents and information**, in order to assess the current state of the sector, periodically evaluate progress with the Plan, and determine subsequent actions. This tool for coordination and control will comprise the **Steering Committee** and a **Committee of Experts**.

Pillar I: Support for development of linguistic infrastructure

Natural language processing and machine translation are based on the use of **linguistic processors** that allow, for example, for extraction of headwords, named-entity recognition (e.g., people, organisations, places), word-sense disambiguation, computing of semantic similarity of documents, automatic classification thereof, simplification of textual context, and automatic summarisation.

To perform these tasks, specific **language resources** are required, depending on the language being processed or translated, the field of the information being processed and other specific factors. These language resources include parallel corpora (collections of documents in various languages), lists of proper nouns (e.g. people, organisations, brands, place names), terminology lists, and dictionaries, all in the correct formats to be used in language processors.

The aim of this pillar is to **develop general language processors and resources.**

Pillar II: Promoting the language industry

The aim is to support **knowledge transfer between the research field and the business sector**. This transfer must be strongly supported by attracting and generating talent in the language technology sector, in order to retain the knowledge accumulated by Spain's universities and transfer knowledge with existing or new companies in the industry, in order to foster innovation.

Specialised industry-focused training is also required for the IT sector and the various vertical sectors that are potential users of this technology. Current projects and technical capabilities must be marketed more effectively.

Consideration is also being given to **support for internationalisation of companies and institutions** in the industry. Maintenance and creation of language resources with high potential for internationalisation must be strengthened, and their adaption to target export markets must also be bolstered. The reach of this action may be extended by the prestige and dissemination capacity of institutions such as the Royal Spanish Academy. It is essential to **raise the international profile of Spanish companies and researchers** in coordination with Spain Export and Investment (known as ICEX, its Spanish acronym), through trade missions, participation in international congresses, showcase projects, etc. Particular attention shall be paid to cooperation with the Ibero-American Community.

Pillar III: The public administrations as promoters of the language industry

The services provided by the public administrations are potential beneficiaries of language technology. Due to their multifaceted nature, there are numerous applications. Potential benefits include:

- Improving the capacity, quality and automation of **citizens services,** which are increasingly mobile, whether by phone or over the internet.
- Improving the capacity and quality of **translation between Spanish and the co-official languages of Spain**, as well as reducing its cost.
- Increasing the **capacity, quality and speed of administrative processing**, incorporating steps featuring automatic natural language processing.
- Improving **accessibility** for persons with disabilities.
- Expanding the in-depth, up-to-date knowledge about the sector as a whole, for evidence-based public policy.

In the spirit of the CORA proposals, the Spanish public administrations are considering the need to eliminate overlapping administrative work and provide high-quality shared public services. As a result, in relation to this pillar, it is proposed that **shared platforms be created for language processing and machine translation for the public administrations**.

Furthermore, the RPSI policy presents an excellent opportunity to develop the natural language processing industry, due to the considerable potential value of much of the information generated by the public sector as a language resource. These resources comprise, for example, legal and contractual texts, sworn translations, specialised dictionaries, entity names (e.g. people, place names, organisations, companies, publications), terminology lists, and parallel translation corpora. It is proposed that support be provided for creation and leveraging of **linguistic linked open data**.

Pillar IV: Flagship projects

Flagship projects are those undertaken by the public administrations or the private sector **applying language technology to strategic sectors** which aim to prove capabilities and benefits, generate business and create resources that can be re-used in other projects. Such projects also serve as learning experience for future projects in other sectors.

Flagship projects aim to find synergies with other measures in the Plan for the Advancement of Language Technology, of a horizontal, cross-cutting nature (general language infrastructure, linguistic linked RPSI, natural language processing and machine translation platforms of the public administrations), through actions in specific sectors, **covering the entire value chain, resulting in market-ready products and services**.

The requirements applied to select the initial flagship projects in the public sector were as follows:

- Synergies with strategic sectors of the economy or Spanish public services.
- High economic and social impact.
- Development of Spain's language industry.
- Specification and commitments of the competent bodies.
- Generation of re-usable resources.
- Synergies with other measures in the Plan for the Advancement of Language Technology and particularly with the generation of language resources.
- Accumulation of experience for future projects. Show-casing the capacities and benefits of language technology.

The vertical sectors in which work will be performed over the coming months include **healthcare, tourism and education**. The aim of the Plan is to achieve the greatest possible dissemination and re-use of any developments. Further vertical projects are therefore planned in other strategic areas in the near term.

5 Measures

The section below sets forth all of the **measures that constitute the Plan, organised according to its pillars and specific lines of action**. Operational plans shall be designed for these measures, and indicators established to gauge their progress. The impact and state of this field shall be studied periodically, as well as the adoption of language technologies in the different public administrations.

Pillar 0: GOVERNANCE OF THE PLAN

Purpose:

Implementing the Plan demands **the involvement of many stakeholders** and encompasses a variety of measures. Therefore, it is necessary to establish a **mechanism for coordination, control and follow-up** of the Plan able to ensure concerted stakeholder action and synergy amongst the Plan's actions in order to achieve its goals.

Goals:

The goals of this Pillar are the following:

- **Identify the competent bodies and administrations**, as well as the operational units that are going to participate in implementing the Plan.
- Identify the **chains of command** and define the **decision-making procedures**. Ensure stakeholder commitment and channel their participation.
- Create a **coordination body for the competent bodies** with the necessary infrastructure. Ensure concerted stakeholder action and synergy amongst the Plan's actions in order to avoid overlapping, and to synchronise actions and share resources.
- Carry out the **operational planning of the Plan** and of the operational plans of each Pillar of the Plan. Conduct **evaluations** of the initial situation and of the outcomes of the Plan's successive phases.
- Provide the Plan's **implementation with flexibility** to enable proper adaptation to unfolding developments and to take advantage of lessons learned. Ensure transparency.
- Create **a liaison body** with the stakeholders involved in the technologies of natural language processing and machine translation.

Measures:

The Ministry of Industry, Tourism and Trade, through its **State Secretariat for Telecommunications and the Information Society (SETSI)**, is the central government body that will **coordinate implementation of the Plan** for the Advancement of Language Technology.

- **Measure 1:** Creation of a **Steering Committee**:
  A *Steering Committee of the Plan for the Advancement of Language Technology* shall be created, with the **following duties**:

  - Establishing a **mechanism for coordination, collaboration, exchange of experiences and mutual assistance** between the competent bodies of the national, regional and local administrations, seeking complementarity and avoidance of overlapping actions in order to jointly optimise the benefits expected from all the actions of the Plan for the Advancement of Language Technology.
  - **Evaluating** periodically **the situation of the Plan's different pillars of action,** in order to plan in detail (operational planning) and evaluate the Plan's progress.
  - **Appointing and dismissing** the members of the Committee of Experts to be defined below. Establishing a streamlined decision-making process.
  - The Steering Committee may create **working groups** for specific tasks.

  The Steering Committee of the Plan for the Advancement of Language Technology shall initially comprise:

  - State Secretariat for Tourism.
  - State Secretariat for Research, Development and Innovation.
  - State Secretariat for Development Cooperation and for Ibero-America.

- State Secretariat for Culture.
- State Secretariat for Telecommunications and the Information Society.
- Under-Secretariat for Education, Culture and Sport.
- Under-Secretariat of the Presidency.
- Under-Secretariat of Health, Social Services and Equality.
- Directorate for Information Technologies of the General State Administration.
- Secretariat-General of Industry and of Small and Medium-sized Enterprises.

The development of the different phases of the Plan for the Advancement of Language Technology, and its own development, especially as regards vertical projects, shall determine **which bodies of the General State Administration are competent at any given time.**

- **Measure 2:** Creation of the **Committee of Experts**:
  A Committee of Experts of the Plan for the Advancement of Language Technology shall be created, with the **following duties:**

  - **Provide technical advice** to the Steering Committee for the Plan.
  - Serve as a **liaison mechanism** for the stakeholders involved in achieving the goals of the Plan for the Advancement of Language Technology.
  - Facilitate **collaboration and exchange of experiences and best practices** in the process of carrying out the Plan. Collaborate on disseminating the Plan in their respective fields.

  The members of the Committee of Experts of the Plan for the Advancement of Language Technology shall combine **representativeness and effectiveness**. Membership may vary over the course of the Plan's implementation. The Committee of Experts may create working groups for specific tasks. The Committee of Experts should include **representatives from, at least, the following areas:**

  - Researchers in natural language processing.
  - Natural language processing industry.
  - Academic and institutional fields related to language.
  - Public sector involved in implementation of the Plan.

- **Measure 3:** Operational Planning
  Among the tasks of the Steering Committee is the **operational planning** for the Plan over the course of its lifetime. The Committee is responsible for the planning cycles and the phases of each cycle (evaluation of the initial situation, operational objectives, design and evaluation of alternatives, action proposals, coordination of actions, evaluation of outcomes). The competent bodies shall approve the operational plan. The plans cited in the different pillars and lines of action form part of this operational planning, which is the responsibility of the Steering Committee.

- **Measure 4:** Evaluation of the situation and implementation of the Plan
  A principal element of the operational planning cycles is an objective and sufficiently precise periodic evaluation both of the situation and the outcomes of each phase of the Plan's implementation. These evaluation tasks will be coordinated, as much as possible, with other plans of the Digital Agenda for Spain.

- **Measure 5:** Coordination with other public administrations
  It is essential to ensure close coordination and cooperation on several pillars with other public administrations: Spanish Regional Administrations, the European Union and Ibero-America.

Below is a **table indicating which lines of action require a special degree of coordination between them.**

		THE PLAN'S LINES OF ACTION					
		Line 1.1	Line 2.1	Line 2.2	Line 3.1	Line 3.2	Line 4.1
THE PLAN'S LINES OF ACTION	Line 1.1				X	X	X
	Line 2.1					X	X
	Line 2.2						X
	Line 3.1	X				X	X
	Line 3.2	X	X		X		X
	Line 4.1	X	X	X	X	X	

For each of the Pillars described below, the governance actions (already noted above in this section) specific to them are repeated.

Pillar I: DEVELOPMENT OF LINGUISTIC INFRASTRUCTURE

Line 1: Development of linguistic infrastructure

Purpose:

Linguistic infrastructure is understood to mean the entire set of **language processors, resources, and evaluation campaigns** taken together as a whole.

Natural language processing and machine translation require **language processors** (e.g., named-entity recognisers, disambiguators, semantic proximity calculators) and **language resources** (e.g. parallel corpora, dictionaries, taxonomies). These can be general (e.g. the function of a word in a sentence) or dependent on the field being analysed (e.g. dictionaries of medical terms). Moreover, it is necessary to evaluate the quality of these processors and resources through evaluation campaigns of specific tasks. Priority will be given to their development in those areas of interest in vertical projects.

The reason for this line of action of the Plan is that Spanish, despite being the world's third most important language by number of speakers, and the second by native speakers, **lags far behind English in terms of the linguistic infrastructure available.** The situation of Spain's co-official languages is even worse. Therefore, it is necessary to develop linguistic infrastructures in Spanish and in Spain's co-official languages to **fuel the development of Spain's natural language and machine translation industry.**

This linguistic infrastructure must be made available to **users with open licences** in order to reap the benefits of economies of scale from the re-use of components and cost-sharing, as well as the qualitative benefits of collaborative maintenance of open resources.

The **public administrations** are the ideal actor to develop linguistic infrastructures, because of their size, of the synergies with RPSI policy, and to guarantee open access to these resources.

Goals:

The priority goals of this pillar are the following:

- **Promote the natural language processing industry,** in Spanish and Spain's co-official languages, by making general linguistic infrastructures available.
- **Close the gap** between the linguistic infrastructures available in **English** and those in Spanish and Spain's co-official languages, regarding the quality, quantity and availability of general language resources for natural language processing and machine translation.
- Ensure the **availability to the public, free or at a low cost** (at least to innovative SMEs, research fields and the public administrations), of quality linguistic infrastructure in Spanish and Spain's co-official languages.
- Bring the industry to the **vanguard of innovation,** facilitating the open use of basic linguistic infrastructure.
- Ensure **coordination in the development of linguistic infrastructures,** avoiding overlapping and seeking synergies. Make available **common tools** for generating and evaluating linguistic infrastructures.
- Adopt **technical interoperability standards**, an appropriate licence policy, and mechanisms for protecting personal data in the generation of language resources.
- Promote **methods of automatic generation** of language resources.

Measures:

- **Measure 1:** Select technical **interoperability** standards, **licence** policies and mechanisms for protecting **personal data** in the generation of language resources.

- **Measure 2:** Purchase or develop **common tools** for generating and evaluating linguistic infrastructures**.**

- **Measure 3:** Draw up and implement a linguistic infrastructure development **plan.** Create an **inventory** of the currently available linguistic infrastructures. Evaluate the **evolution** of the quantity, quality and availability of linguistic infrastructures.

- **Measure 4:** Facilitate public access to the existing linguistic infrastructures.

Pillar II: PROMOTING THE LANGUAGE TECHNOLOGY INDUSTRY

Line 1: Raising the industry's profile and improving knowledge transfer

Purpose:

In Spain, the machine translation industry and—to an even greater extent—that of natural language processing, is **unfamiliar within the public administrations** and within the majority of the **productive sectors**. Furthermore, it is necessary to **increase the number of experts** in natural language processing and machine translation to guarantee the growth of this field. In academic circles the situation is similar, with an inadequate offering of training opportunities.

Goals:

The goals sought by this line of action are the following:

- **Raise the profile of the industry** for natural language processing and machine translation.
- Transfer **Spanish research excellence** to the industry. Ensure the availability of doctorate-level specialists and train new researchers and developers in these areas.

Measures:

- **Measure 1:** Design a **plan** to raise the industry's profile and improve transfer.

- **Measure 2:** Plan and coordinate actions aimed at raising the industry's **profile** and improving training in this field with other public administrations, especially and Ibero-America.

- **Measure 3:** Improve **training**: Include specific courses on language technology in university programmes, promote the creation of online training (MOOCs), hackathons, support for industrial doctorate and master's programmes, and specialised grants.

- **Measure 4:** Enhance **visibility:** Basic training seminars for SMEs and professionals; Conventions, forums and participation in national and international trade fairs. Promote cloud computing (SaaS). Coordination with the portal proposed in Pillar I.

Line 2: Support internationalisation and commercialisation of the industry

Purpose:

Spanish is a transnational language with nearly 470 million speakers, representing a **huge market**. Moreover, nine out of ten users of the technology promoted in this Plan are outside of Spain. This represents a golden **opportunity to expand the scope of internationalisation**, as well as to expand institutional and economic cooperation with the countries of Ibero-America.

Goals:

The main goals of this line of action are the following:

- **Enhance the internationalisation** of the companies in this industry, especially in the Ibero-American and North American markets.
- **Promote use of the tools provided by the Internationalisation Plan of the Digital Agenda for Spain**, which includes a programme for directly funding companies for internationalisation activities (e.g., ICEX, AECID).
- Support the **leadership role of major Spanish corporations** established in the Ibero-American market and other emerging markets.
- **Strengthen development cooperation** with the Ibero-American community to lead the implementation of the technologies of natural language processing and machine translation in Spanish. Institutional support and participation of the different Ibero-American countries in this project. Work jointly on the creation of open language resources and data.

Measures:

- **Measure 1:** Design an internationalisation plan.
- **Measure 2:** Cooperation with Ibero-America (e.g., Ibero-American Summit, IODC 2016 Madrid). Collaboration with the Secretariat-General for Ibero-America (SEGIB) and other Ibero-American institutions.
- **Measure 3:** Coordination with **AECID programmes** and use of the Network of Science Counsellors Abroad and of the existing associations of Spanish scientists abroad.
- **Measure 4:** Possibility of integrating natural language processing and machine translation into the areas that are currently being funded within the framework of **Strategic Action Plan for the Economy and Digital Society**.
- **Measure 5:** The ICT sector is one of the priorities for the **Invest in Spain** programme to expand foreign investment in Spain. Natural language processing and machine translation should be included in the subsectors for which Spain offers the most attractive investment opportunities.

- **Measure 6:** Possibility of including natural language processing and machine translation in the agreements (**MoU**) signed with Ibero-American countries (or from other regions) in the future.

- **Measure 7:** Promote the development of linguistic infrastructure and availability of open public information on **variants of Spanish**.

- **Measure 8:** Identify trade fairs, conventions or other events which, with ICEX collaboration, could be **marketing** vehicles for products or projects being carried out by Spanish companies in this industry.

- **Measure 9:** Conduct a study regarding the **state of the art** of the internationalisation of **companies from Spain** in the industry in those countries having the biggest markets for internationalisation.

- **Measure 10:** Study the possibility of providing grants to incubators or accelerators, or of presenting twinning projects between small and large companies.

III: THE PUBLIC ADMINISTRATIONS AS DRIVERS OF THE LANGUAGE INDUSTRY

Line 1: Natural language processing and machine translation platforms in the public administrations

Purpose:

The services provided by the public administrations are **clearly beneficiaries of language technology**. Due to its cross-cutting nature, there are myriad opportunities for applying this technology to improving citizen services. Some examples, among many, of these benefits are:

- Improve the capacity, quality and automation of **citizen services,** which are increasingly mobile, whether by phone or the internet. Provide advanced services of based on natural language processing and machine translation.

- Improve capacity and cut costs for **translation between Spanish, other EU languages, and the co-official languages of Spain.**

- Increase the **capacity, quality and speed of administrative processing**, incorporating steps featuring automatic natural language processing.

- **Improve accessibility** for persons with disabilities.

- **Improve in-depth, up-to-date knowledge of the data** generated by different industries in order to craft public policy. This knowledge can be extracted through computer analysis of large corpora of documents (e.g. clinical histories, applications for assistance, public procurement, patents).

Moreover, following the CORA recommendations, the Spanish government is considering the **need to avoid overlapping, to simplify, to apply economies of scale, and to provide quality public services**. Consequently, this line proposes the creation of common platforms for natural language processing and machine translation, based on reusable components from natural language processing and machine translation for the public administrations.

In addition, these platforms can strengthen the **leading role of demand** from the public administrations and provide demonstration services to showcase applications of natural language processing and machine translation technology. Furthermore, these platforms can be useful for innovative sectors and research.

Goals:

This line's goals are the following:

- **Improve the quality and capacity of public services,** incorporating natural language processing and machine translation technology.

- Provide **common tools** to facilitate natural language processing and machine translation in the public administrations.

- Pursuant to the **CORA** recommendations, simplify, achieve synergies and apply economies of scale to the roll-out of new services based on natural language processing and machine translation. Seek to cut costs through **resource sharing among the public administrations**, eliminating duplication of efforts; also, by improving the efficiency of procedures and improving knowledge. The platform will be based on re-usable and interoperable components, preferably with unrestrictive open-source licenses.

- Provide **common tools for generating**, evaluating and exploiting language resources.

- **Processing flows should be carried out** both on the common physical platform and through simply copying them onto local platforms in order to meet the need for a high level of confidentiality or in cases of very large volumes of data that are difficult to move.

- Share infrastructures and systems between **innovative projects** supported by public institutions. They can also be used after the end of these projects as online demonstration centres.

- Share common infrastructure and systems for their application in **research fields,** facilitating the development of new components and analysis of large corpora of documents.

- **Share services with other institutions in the field**: Royal Spanish Academy, Institute of Spain, Ibero-American bodies.

Measures:

- **Measure 1:** Design a development **plan** for natural language processing and machine translation platforms in the public administrations.

- **Measure 2:** Establish a clear organisational and financing structure to ensure its continuance beyond the lifetime of the Plan.

- **Measure 3:** Creation of a **common natural language processing and machine translation platform for the public administrations,** with the following essential requirements:

  - Facilitate the launching of **advanced services** based on natural language processing and machine translation in the national and regional administrations.

  - Develop a **scalable** infrastructure based on components for the parallel processing of large corpora of documents.

  - Maintain the guarantees of **confidentiality** appropriate for public services.

  - Add different **components** and language resources to the flow of language processing with different licensing models and processing methods.

  - **Common tools** for anonymisation, editing, post-editing of machine translations, etc.

  - This platform will make available **general-purpose language resources** (Pillar I) and **field-specific resources** (mainly those specialised resources necessary for the development of the vertical projects promoted in Pillar IV).

  - This will be a venue for exploiting and standardising the language resources generated under the aegis of the **RPSI** policy.

  - It will enable different models of implementation and distribution (embedded, local cluster, remote, and implementation at supercomputing centres).

Line 2: Language resources of the public administrations and policy for re-use of public sector information

Purpose:

The considerable potential value as a language resource of much of the information generated by the public sector represents an **outstanding opportunity to develop the natural language processing industry.** An illustrative example is that the most downloaded element on the EU's open-data portal is the EURP multilingual parallel corpus of translations (https://open-data.europa.eu/es/data/).

Among other resources, the Administration has: place names, personal names, trademarks, names of organisations, names of companies, taxonomies, glossaries, high-quality translated texts (e.g. sworn translations), field-specific corpora (e.g. legal, medical) and classified texts. It is necessary to **adapt these valuable assets to formats that are reusable by language processors.**

Moreover, the **RPSI** policy presents a **channel for developing linguistic linked open data**, as its aim is to make the data generated by the public sector in the course of its duties available to society as an open resource to be used for financial gain.

Goals:

The goals of this line of action are the following:

- **Promote the natural language processing industry** in Spanish and in Spain's co-official languages, making available to them the language resources generated from public-sector information.

- Ensure the **availability** of the language resources generated from public-sector information, either **free of charge or at a low cost.**

- Ensure **coordination in development of language resources** generated from public sector information, preventing overlapping and seeking synergies.

- Develop and use **common tools** for generating and evaluating language resources generated from public sector information.

- **Adopt technical interoperability standards**, an appropriate licence policy and mechanisms for protecting personal data.

Measures:

- **Measure 1:** Carry out these actions within the framework of the **RPSI policy.**

    - Introduce in RPSI the concept of **linguistic linked open data**.

    - Promote this concept within the public administrations.

    - Introduce the concept of linguistic linked open data at IODC 2016, in collaboration with **Ibero-America**, to raise the profile of the linguistic linked open data policy.

- **Measure 2:** Select technical **interoperability** standards, open-licence policies and personal data protection mechanisms.

- **Measure 3:** Make available the **common tools** necessary for generating and exploiting these language resources (e.g. anonymisers, text alignment, processing flows) on the natural language processing platform of the public administrations foreseen in this Plan.

- **Measure 4: Identify** those corpora of public-sector information that could be converted into language resources.

- **Measure 5: Catalogue** these open language resources within the open data portal, with an advanced user experience.

- **Measure 6:** Facilitate **availability** of these resources on the natural language processing and machine translation platforms of the public administrations that are foreseen in this Plan.

- **Measure 7:** Draw up a **plan** for generating language resources from public sector information.

Pillar IV: FLAGSHIP PROJECTS ON NATURAL LANGUAGE PROCESSING TECHNOLOGY

Line 1: Flagship projects on natural language processing technology in the public administrations

Purpose:

Launching the other measures in the Plan, which are horizontal and cross-cutting, with actions in specific public services with high social impact **encompassing the entire value chain** and resulting in market-ready products and services, in order to enhance the capacities and benefits of natural language processing and machine translation technology.

Public administration projects in the field of natural language processing and machine translation require, in general, **horizontal and vertical scalability for processing large amounts of information and serving a great many users simultaneously; a component-based approach; and the use of standards that guarantee maximum re-use and interoperability.** Moreover, it is necessary to facilitate the possibility of transferring data processing for a number of reasons, such as the excessive size of the data or for confidentiality restrictions.

Goals:

The goals pursued by this line of action are the following:

- Identify **new public services** or **enhance the capacity and the quality of existing public services**, by implementing natural language processing technology.
- Serve to **demonstrate the capacities and benefits** of natural language processing technology.
- **Generate re-usable resources** for other projects in different fields.
- Serve as **lessons learned for future projects.**
- **Serve for immediate implementation** of the Plan's horizontal measures; use of common infrastructures and platforms.

Measures:

- **Measure 1:** Carry out a limited set of projects implementing natural language processing technology in strategic public services with great social impact. The projects will be selected according to the **following requirements:**

- **Commitment of the competent bodies.** Ensure leadership by those who know the problem well and are competent to solve it.

- **Detail.** Respond to identified problems that justify the suitability and timeliness of launching the project.

- High **economic and social impact.**

- Development of the **entire value chain**.

- Generation of **re-usable resources**. Avoid generating resources that are dependent on proprietary solutions and technology, preventing resource portability.

- **Synergies with the other measures** in the Plan for the Advancement of Language Technology and, in particular, with the generation of language resources and with the public administrations' natural language processing and machine translation platform.

- **Ability to demonstrate the capacities and benefits** of language technology.

- Particular attention given to **acquiring experience** for future projects.

The sectors in which work will start first will be health, tourism and education (first projects in 2016-2017). However, the possibility is envisaged of incorporating other flagship projects later on, applied to other areas (justice, citizen services, and sector monitoring are clear candidates). Also envisaged is the incorporation of Spain's Autonomous Regions.

6 Relation with the Goals of the Digital Agenda for Spain

The following table shows the relation between the lines of action of the Plan for the Advancement of Language Technology (PALT) and the goals of the Digital Agenda for Spain (ADpE, its Spanish acronym).

Goal	Sub-goal	Lines of action	Identifier (ID)	PALT Lines
2. Develop the digital economy	2.1. Encourage the transformative use of ICT in Spanish companies	2.1.3. Foster the development of specific ICT solutions adapted to the needs of productive sectors that are insufficiently served by current ICT.	ADpE- 2.1.3.	Line I.1.
	2.3. Promote the online production and distribution of digital content.	2.3.3. Simplify the conditions for re-using public sector information.	ADpE- 2.3.3.	Line III.2.
	2.7. Strengthen the ICT industry by developing technological projects in public services.	2.7.6. Develop services for professionals and the general public, based on availability of Digital Medical Records in the National Health System.	ADpE- 2.7.6.	Line IV.1.
	2.7. Strengthen the ICT industry by developing technological projects in public services.	2.7.10. Promote standards fostering interoperability among healthcare ICT, tele-assistance and tele-medicine, through mechanisms of collaboration with the industry.	ADpE- 2.7.10.	Line IV.1
	2.7. Strengthen the ICT industry by developing technological projects in public services.	2.7.15 Use e-learning environments for the implementation of specific educational plans and for expanding the classroom concept in time and in space.	ADpE- 2.7.15.	Line IV.1
	2.7. Strengthen the ICT industry by developing technological projects in public services.	2.7.16. Establish formats to be supported by the learning tools and systems in the field of public digital educational content.	ADpE- 2.7.16.	Line IV.1
	2.7. Strengthen the ICT industry by developing technological projects in public services.	2.7.17. Encourage use of digital and technological platforms, and platforms involving quality teaching resources, by the entire educational community.	ADpE- 2.7.17.	Line IV.1

Goal	Sub-goal	Lines of action	Identifier (ID)	PALT Lines
3. Improve e-administration and adopt digital solutions for the efficient provision of public services	3.1. Advance towards a public administration that is integrated into society, with quality public services focused on citizens and companies.	3.1.2. Shift current public services towards citizen-oriented services, making them customisable, proactive, accessible from different platforms, adapted to user needs and user-friendly, aimed at vital events, and with guaranteed quality and security.	ADpE- 3.1.2.	Line IV.1
	3.1. Advance towards a public administration that is integrated into society, with quality public services focused on citizens and companies.	3.1.5. Foster the re-use of public sector information to enable the development of high-value services contributing to the promotion of economic activity and the generation of services of value for citizens and companies.	ADpE- 3.1.5.	Line III.2.
	3.4. Promote cooperation and collaboration with organisations, companies and social partners regarding e-administration.	3.4.1. Promote the sharing of e-administration experiences, projects, services and applications developed by all public administrations, companies and organisations, and establish forums for such sharing.	ADpE- 3.4.1.	Line IV.1
	3.4. Promote cooperation and collaboration with organisations, companies and social partners regarding e-administration.	3.4.2. Establish a new framework for relations with organisations, companies and social partners, to contribute to invigorating the ICT market, especially through the study of mechanisms enabling public-private collaboration.	ADpE- 3.4.2.	Line IV.1.
5. Promote the R&D&I in ICT system	5.1. Increase the effectiveness of public investment in R&D&I in ICT.	5.1.1. Coordinate the supported strategic lines with all the public stakeholders involved in promoting R&D&I in ICT.	ADpE- 5.1.1.	Line III.2.
	5.1. Increase the effectiveness of public investment in R&D&I in ICT.	5.1.2. Facilitate collaboration between companies and public research institutions through initiatives that strengthen mutual knowledge of capacities and needs, such as knowledge maps, technological platforms, and open innovation.	ADpE- 5.1.2.	Line I.1. Line III.1. Line IV.1.
	5.1. Increase the effectiveness of public investment in R&D&I in ICT.	5.1.3. Draw up a Plan to adapt R&D&I in ICT management systems in order to increase transparency, foster the participation and collaboration of applicants, and facilitate access to public resources.	ADpE- 5.1.3.	Line III.2.

Goal	Sub-goal	Lines of action	Identifier (ID)	PALT Lines
	5.2. Foster private investment in R&D&I in ICT.	5.2.1. Encourage private investment in R&D&I in the electronics industry and in ICT through the strategic use of public procurement and public-private collaboration.	ADpE- 5.2.1.	Line IV.1.
	5.2. Foster private investment in R&D&I in ICT.	5.2.2. Strengthen co-investment funds with the private sector in R&D&I applied to ICT.	ADpE- 5.2.2.	Line IV.1.
	5.4. Broaden Spanish participation in R&D&I in ICT at the international level.	5.4.1. Develop schemes for co-financing and promoting Spanish participation in European and international R&D&I in ICT programmes.	ADpE- 5.4.1.	Line II.2.
	5.4. Broaden Spanish participation in R&D&I in ICT in the international sphere.	5.4.3. Increase Spanish representation in international programmes and initiatives.	ADpE- 5.4.3.	Line II.2.
6. Promote digital inclusion and the training of new ICT professionals	6.2. Digital capacity-building and training of new ICT professionals.	6.2.2. Maximise effectiveness in the management and allocation of training funds aimed at ongoing training in ICT, both for private sector and public sector staff.	ADpE- 6.2.2.	Line II.1.
	6.2. Digital capacity-building and training of new ICT professionals.	6.2.3. Allocate part of the resources available for ongoing training to ICT professionals' capacity-building and acquisition of digital skills.	ADpE- 6.2.3.	Line II.1.
	6.2. Digital capacity-building and training of new ICT professionals.	6.2.4. Re-orient ICT-related vocational training.	ADpE- 6.2.4.	Line II.1.
	6.2. Digital capacity-building and training of new ICT professionals.	6.2.5. Foster the improvement of university courses aimed at training ICT professionals.	ADpE- 6.2.5.	Line II.1.

7 Schedule

Measure - Phase (Quarter/Year)	Q4 2015	Q1 2016	Q2 2016	Q3 2016	Q4 2016	Q1 2017	Q2 2017	Q3 2017	Q4 2017	Q1 2018	Q2 2018	Q3 2018	Q4 2018	Q1 2019	Q2 2019	Q3 2019	Q4 2019	Q1 2020	Q2 2020	Q3 2020	Q4 2020
Pillar 0: Governance	▓	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░
Line I: Governance	▓	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░
Pillar I: Development of Linguistic Infrastructures	▓	▓	░	░	░	░	░	░	░	░	░	▓	░	░	░	░	░	░	░	░	░
Line 1: Development of linguistic infrastructures	▓	▓	░	░	░	░	░	░	░	░	░	▓	░	░	░	░	░	░	░	░	░
Pillar II: Promoting the Language Technology Industry		▓	▓	░	░	░	░	░	░	▓	▓	░	░	░	░	░	░	▓	▓	░	░
Line 1: Raising the industry's profile and improving knowledge transfer		▓	▓	░	░	░	░	░	░	▓	▓	░	░	░	░	░	░	▓	▓	░	░
Pillar III: The Public Administrations as Promoters of the Language Industry	▓	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░
Line 1: Natural language processing and machine translation platforms in the public administrations	▓	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░
Line 2: Language resources of the public administrations and policy for re-use of public sector information	▓	▓	░	░	░	░	░	░	░	▓	▓	░	░	░	░	░	░	░	░	░	░
Pillar IV: Flagship Projects on Natural Language Processing Technology	▓	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░
Line 1: Flagship projects	▓	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░	░

▓ Procurement ░ Implementati

8 Budget

Pillars and Measures	SETSI		OTHER BODIES		ALL	
	Total € (thousands)	2016 € (thousands)	Total € (thousands)	2016 € (thousands)	Total € (thousands)	2016 € (thousands)
Pillar 0: Governance	**1,260**	**420**	**0**	**0**	**1,260**	**420**
Line I: Governance	1,260	420	0	0	1,260	420
Pillar I: Development of Linguistic Infrastructures	**14,083**	**3,231**	**16,349**	**0**	**30,432**	**3,231**
Line 1: Development of linguistic infrastructures	14,083	3,231	16,349	0	30,432	3,231
Pillar II: Promoting the Language Technology Industry	**1,780**	**480**	**400**	**0**	**2,180**	**480**
Lines 1 and 2: Raising the industry's profile and improving knowledge transfer	1,780	480	400	0	2,180	480
Pillar III: The Public Administrations as Promoters of the Language Industry	**6,180**	**1,540**	**0**	**0**	**6,180**	**1,540**
Line 1: Natural language processing and machine translation platforms in the public administrations	4,140	1,100	0	0	4,140	1,100
Line 2: Language resources of the public administrations and policy for re-use of public sector information.	2,040	440	0	0	2,040	440
Pillar IV: Flagship Projects on Natural Language Processing Technology	**49,090**	**8,378**	**0**	**0**	**49,090**	**8,378**
Line 1: Flagship projects	49,090	8,378	0	0	49,090	8,378
Total	**72,393**	**14,048**	**16,749**	**0**	**89,142**	**14,048**

9 Indicators

Pillars and Measures	Indicator	Date
Pillar I: Development of Linguistic Infrastructures		
Line 1: Development of linguistic infrastructures	Standardised European indicators of general resources (e.g. META-NET[12]).	2020
Pillar II: Promoting the Language Technology Industry		
Lines 1 and 2: Raising the industry's profile and improving knowledge transfer	Indicators from the Internationalisation Plan. Catalogue of the language technology sector.	2020
Pillar III: The Public Administrations as Promoters of the Language Industry		
Line 1: Natural language processing and machine translation platforms in the public administrations	Measures of use of the platforms.	2020
Line 2: Language resources of the public administrations and policy for re-use of public sector information.	Standardised European indicators of re-usable resources of the administrations.	2020
Pillar IV: Flagship Projects on Natural Language Processing Technology		
Line 1: Flagship projects	Standardised European indicators of specific resources. Measures of use.	2020
Overall goal of the Plan	Growth of the language technology sector.	2020

[12] *Source: DG Translation, European Commission: Translation figures 2014 LT-Innovate, META-FORUM 2015, META-NET.*

Annexes

Annex I: Relation with the European strategy

EU Layer	EU Line	PALT Pillar	PALT Line
I. Innovative technology solutions for the multilingual digital single market	1.2. Technology solutions for public services	Pillar V: Flagship projects	Line 1: Flagship projects on natural language processing technology in the public administrations
II. Language technology services, platforms and infrastructures	2.1. Development of language technology services, platforms and infrastructures	Pillar I: Development of linguistic infrastructures	Line 1: Development of linguistic infrastructures
IV. Horizontal measures	4.2. Standards and interoperability	Pillar III: The public administrations as promoters of the language industry	Line 2: Language resources of the public administrations and policy for re-use of public sector information
	4.3. Open data		

Source: "Strategic Agenda for the Multilingual Digital Single Market. Technologies for Overcoming Language Barriers towards a truly integrated European Online Market" META-NET