D2.1 Report on User Stories

Author(s): Thomas Thurner, SWC
Eduardo Mena, UNIZAR
Jorge Gracia. UNIZAR
John P. McCrae. NUIG
Eva Theodoridou, OUP

Date: 30/09/2019 (V1)
10/06/2020 (V2)

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825182. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

H2020-ICT-29b
Grant Agreement No. 825182
Prêt-à-LLOD - Ready-to-use Multilingual Linked Language Data for Knowledge Services across Sectors

D2.1
Report on User Stories

Deliverable Number:	D2.1
Dissemination Level:	Restricted
Delivery Date:	30/09/2019 (V1) 10/06/2020 (V2)
Version:	2
Author(s):	Thomas Thurner

Document History

Version	Date	Changes	Authors
0.1	10.09.2019	Initial	Thomas Thurner
	20.09.2019	Review	Eduardo Mena Jorge Gracia John P. McCrae
1.1	30.09.2019	Submit	Thomas Thurner
1.2	20.04.2020	Amendments after inputs demanded by reviewers (Chapter 4.2, 4.3.6, 4.3.7)	Thomas Thurner Eva Theodoridou
2	10.06.2020	Submit	Thomas Thurner

Table of Contents

Figures

Abbreviations

AI	Artificial Intelligence
ACL	Association for Computational Linguistics
API	Application Program Interface
ASR	Automated Speech Recognition
BMC	Business Model Canvas
CEF	Connecting Europe Facility
EC	European Commission
HR	Human Resources

IPR	Intellectual Property Right
LOD	Linked Open Data
LLOD	Linked Language Open Data
LT	Language Technology(ies)
ML	Machine Learning
NLP	Natural Language Processing
SaaS	Software as a Service
SME	Small or Medium-sized Enterprise
TTS	Text-to-Speech
XML	Extensible Markup Language

1. Executive Summary

Based on a survey, interviews and desk research, this Report on User Stories discusses the opportunities which are opened with the Prêt-à-LLOD project for the Language Technology (LT) market. As a general thesis, we target **medium and small** market participants with this analysis, as it is evident that the better part of the LT market is Europe is structured with such companies mainly active in the broader field of **text analytics and dictionaries**. The results of our survey shows that speech technologies as such are not so relevant for the European LT market.

This idea of a divided market leads to the potential of **cross-company, cross-platform and cross-resources solutions** reusable for various and independent acting LT providers. To come from a scattered market of competing players to a **vital and complementary exchange** between those players, some basic cornerstones are described in the Business Model Canvas of this Report:

- **Exploitation** of the Prêt-à-LLOD technological ecosystem **by organisations outside the consortium** via out-licensing and subscription models for accessing to multilingual language technology services and Linked Open Data (LOD).
- New methodologies for a **faster development** of domain-specific language resources.
- Contributions to ongoing standardisation work around **exchangeable** and interoperable **language technology components**, and vocabularies and interfaces for Linked Language Open Data (LLOD),
- New models and mechanisms for ensuring the **validity, maintainability and licensing** of language resources

The discussion of connected Business Model Canvas from other recent studies on the LT market is followed by a section which sketches exemplified **Business User Stories** as a discussion ground for further elaboration in the project:

- LT Middleware
- Lexica and Dictionaries
- Language Models and Algorithms
- Integrators

Together with the key figures on market, technologies and usage, these four Business User Stories, will guide the project in the **next phase of technical requirement elicitation**.

2. Introduction

2.1. Task description

The goal of this work package is to elicit (analyse and understand) business cases, (regulatory, technical, societal) needs and requirements for a community-driven ecosystem to support the lifecycle of LLOD. The goal of WP2 is to collect requirements for the pilots of WP4 and collect requirements for the research in WP3 and WP5. The work package will further investigate cross-domain synergies that could improve the universality of use of the standards and services provided by the project.

The aim of this task is to identify a wide range of innovative user scenarios, which are going to be enabled in this project. These scenarios will drive the process of user-level requirements gathering. By presenting scenarios for the use of Language Technology we further enable the developers to bridge the gap between the user's demands and the design of new technological approaches. The main goal of this first task is thus to specify the user needs and elaborate user scenarios that will guide the design and development of the functionalities.

3. Methodology

3.1. Development of business stories

To obtain resilient results for the further technical and commercial development of the project, the business stories development phase of the Prêt-à-LLOD project is founded on several pillars. In this very early stage of the project, we have chosen methods of elicitation in which results and findings can be combined into a bigger picture, describing the status and needs regarding the current or future implementations of the Language Technologies envisaged by the Prêt-à-LLOD project.

3.2. Question Matrix

To ensure appropriate results in the processing of all collected data for the business stories elicitation, we have built an interconnected question matrix, called the Prêt-à-LLOD-Core Question Matrix (PQ matrix). The initial design is based on the common concept of personas, stories and use cases amended by technology-related questions aligned with the basic Prêt-à-LLOD value chain model market model.

Basic Data

This questions block defines to which statistical pattern we are applying your answers of the following blocks.

You

Name	email	role	department	work experience	I hereby declare, that I agree the computation and use of this data for the use of the Pret-a-LLOD project. Any other use or processing is not permitted. Read more on the Pret-a-LLOD code of Conduct
					FALSE

Your company

Name	No. of Employees	Annual turnover	Sector	Regional coverage	Are you new to the Language Technology Market?	Branche

survey and I am answering it in a professional capacity

FALSE

Business Scenarios

The aim of this questions block is to identify and prioritze those building blocks of the Language Data Value Chain, which are of use for your innovative Language Technology Application

Your business case

Please explain to us your business use cases where language technologies are applied.

add text

Are there competitors to your Language Technology Product?	Ist this Language Technology Product new to your company (radical innovation)?	Is this Language Technology Product based on previous work of your company (increment innovation?)	Which kind of Service or Dataset stands in the centre of your Language Technology Product?	Who are your end-users?
			add text	add text

What is the business model of your service or resource provision?	Are there any constraints or rules to which your product must conform?

Your possible interact with Pret-a-LLOD value chain

	Discover	Prepare	Organize	Integrate	Analyze & Act
	In this project we will extend the working to analyzing and monitoring the resources directly in order to deduce metadata about the availability, technical quality and content of language resources.	Dataset transformation currently depends significantly on manual transformation. We move beyond this with the use of semantic ontologies which many formats (including XML, CSV, JSON) can be matched to (RDF).	We will investigate (i) the representation of rights information, (ii) the methodology to manipulate policies and provenance information; (iii) and new license composition algorithms.	Look at linking across linguistic data modalities, in particular corpora, lexicons, thesauri and ontologies.	We propose the Prêt-à-LLOD Workflows component will allow the deployment of language technology pipelines on the cloud, increasing the interoperability by using containerization technology
I can provide a solution here	☐	☐	☐	☐	☐
I need a solution here	☐	☐	☐	☐	☐
This is crucial for my innovation	☐	☐	☐	☐	☐

Language of Language Ressource included in my innovation	Format of Language Ressource included in my innovation	How do you represent your rights information of your data sets (term of contract, formal agreement)?	Do you use linked datasets in your innovation product?	Would you be interested in adjusting your current workflows to incorporate/modify based upon new langauge technology pipelines?
add text	add text	add text	add text	add text

Type of Language Resource included in my innovation	Preferred internal format for lexical resources (XML, JSON, TSV, ...)	What types of resources do you have linked?	Are you already using workflows for processing your (linked) data?
add text	add text	add text	add text

License of Language Ressource included in my innovation	Preferred internal format for linguistic annotations (JSON, TSV/CoNLL, ...)	What unit levels are currently linked in your innovation?	What services would you be intrested in receiving as part of standardized workflows?

Market Expectations

We want to ask on your assesment of the future (next five years) of Language Technologies

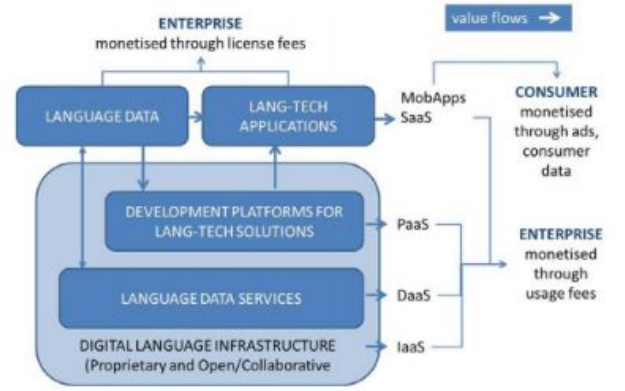
Potential of branches

	I see a general high market potential for this	I want to invest and innovate in this
Speech recognition	☐	☐
Speech synthesis	☐	☐
Text categorization	☐	☐
Text Summarization	☐	☐
Text Indexing	☐	☐
Text Retrieval	☐	☐
Information Extraction	☐	☐
Data Fusion and Text Data Mining	☐	☐
Question Answering	☐	☐
Report Generation	☐	☐
Spoken Dialogue Systems	☐	☐
Translation Technologies	☐	☐

Market focus

	When it comes to a distributed market model (see beside); where do you will focus your efforts
Language Data	☐
Language Apps	☐
Platforms	☐
Language Services	☐
Infrasructure	☐

	Do you know the European top players in Language Technologies? If possible, please name them for each of the branches.
Speech recognition	
Speech synthesis	
Text categorization	
Text Summarization	
Text Indexing	
Text Retrieval	
Information Extraction	
Data Fusion and Text Data Mining	
Question Answering	
Report Generation	
Spoken Dialogue Systems	
Translation Technologies	



from: LT-Innovate, Status and Potential of the European Language Technology Markets, January 2013

Figure 16: Collaborative Innovation

Adopters

	Characterize industries along their adaption of speech technologies	Characterize industries along their adaption of translation technologies	Characterize industries along their adaption of content technologies
Pioneers	^		
Adopters			
Followers			
Outdistanced			

Figure 1: Question Matrix

3.3. Public Survey

Based on the PQ matrix we developed a public survey to get empirical data about the business needs. In close cooperation of the consortium, we collected a list of stakeholders in the commercial LT sector, where the partners of Prêt-à-LLOD have close relations. GDPR limitations cause a more difficult identification and addressing of stakeholders. We had to go with an indirect addressing of stakeholders to not violate the spam and ethical regulation. It has to be mentioned, that this reduced the possible response rate substantially. Nevertheless, we reached out for industry in the sector with the survey with a sum of 300 contacts and a usual response rate of 10%.

The public survey is divided into the following chapters:

- About the respondent
- About the company
- About the companies business case

- About possible interact with Prêt-à-LLOD value chain
- General view on the potential for various LT branches
- The focus branches chosen by the respondent(‘s company)
- Adopters known in various LT branches

The public survey is online at https://swc4.typeform.com/to/nrm9Xt.

3.4. Desk Research

As showed in 3.3, the collected statistical data is not stable enough to draw a clear, meaningful and relevant picture of the business stories, so we decided to do an additional comparative research on other resources, which recently carried out basic business and market research in the LT sector.

Source 1: **LT2013, Status and Potential of the European Language Technology Markets, January 2013** by LT-Innovate, the Forum for Europe’s Language Technology Industry, a not-for-profit organisation representing mostly SMEs involved in developing products using intelligent content, speech and translation technologies. LT-Innovate was founded in January 2012. As of 1 November, it has gathered 115 LT suppliers in 22 countries, as well as several dozens of other LT stakeholders. The European Language Technology industry generated an aggregate turnover of 19.3 billion € in 2011. LT is a very dynamic industry, with a yearly growth rate in excess of 10%.

Source 2: Preliminary: **Final study report on CEF Automated Translation value proposition in the context of the European LT market/ecosystem**, 2019, A study prepared for the European Commission DG Communications Networks, Content & Technology by: Crosslang, Tilde, Elda, IDC

Source 3: **Slator 2020 Language Industry Market Report,** 2020 by Slator AG, Swizerland
The Slator 2020 Language Industry Market Report provides a comprehensive view of the global language services and technology industry, lists the top growing verticals, it analyzes the market from a services perspective, listing more than 200 core and adjacent services provided by leading language service providers (LSPs), and presents a market outlook.

3.5. Workshop on Language Technology Market and Components Taxonomy

In a joint effort of the projects European Language Grid (www.european-language-grid.eu) and Prêt-à-LLOD (www.pret-a-llod.eu) a taxonomic description of fields, subdomains, techniques, solutions, and components are planned to be developed to foster exchange and interaction in the European Language Technology Sector - for both - research and industry. For that, the mentioned projects are driving a process, where such a taxonomy is built up, maintained and provided openly to the sector. As a goal, the Language Technology Market

and Components Taxonomy will be published in its first version in late 2019. A series of consultations and workshops cornerstone the efforts of the group:

- **8th Language Technology Industry Summit**, 24-25 June 2019, Brussels: Initial Workshop on the Language Technology Market and Components Taxonomy
- **European Language Services Industry Forum**, September 9, 2019, Karlsruhe: Workshop on the Language Technology Market and Components Taxonomy
- **Vienna Semantic Web Meetup** (in conjunction with the Prêt-à-LLOD plenary) , November 2019, Vienna: Launch of the Language Technology Market and Components Taxonomy

4. Findings

4.1. Public Survey

Based on the PQ matrix we developed a public survey to get empirical data about the business needs. We collected a list of stakeholders in the commercial LT sector, where the partners of Prêt-à-LLOD have close relationships.

4.1.1. Company profiles

The final results mentioned in this chapter represent the feedback of a total of 21 respondents, where 80% categorize themselves as Commercial Enterprise, 13.3% as Research and 6.7% as NPO/NGO (Figure 2). The half of the organizations surveyed declare their turnover between 1 and 5 M€, per year, and the other half is less than that, so we face a sample of respondents of small and medium enterprises in our survey. As Prêt-à-LLOD in general, and this survey in specific, looks for an analysis focusing on the disperse European Market of medium sized LT companies, we assume that the results discussed here (at least) indicate a valid business story development for the project.

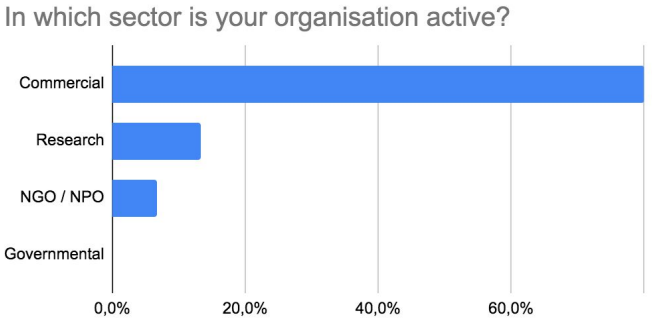


Figure 2. Sectors

Responses to the question on how long the respondents are present at the market shows that 69% is longer than 5 years on the market, from which can be assumed, that our business stories have to focus more on the servicing of existing businesses than on startups.

4.1.2. Technology ranking

Asking for the technologies the respective companies are focusing on, the various forms of text processing are listed on top followed with translation and speech technologies in the weaker ranks.

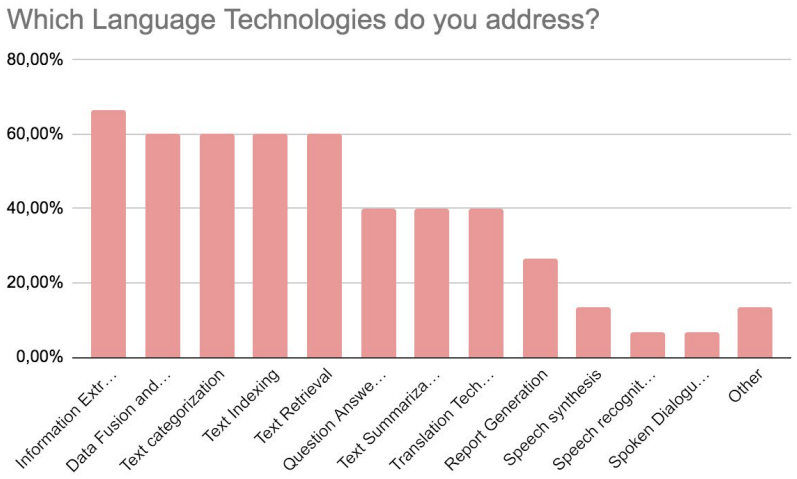


Figure 3. Technology Ranking

For the development of Business Stories, the fact that text processing technologies are focus technologies. This fact can be assumed as a typical characteristic of the European Language Technology Market. Such market bias has further its implications on how a resource and asset framework for the sector (as Prêt-à-LLOD) may be focused.

4.1.3. Target users

A strong focus lies in B2B relation, when it comes to the user groups and targeted customers. Specifically, Banking & Insurance, Media & Publishing, Life Science & Pharma, Oil & Gas, Publishers, IT companies, Tourism is meant in this respect.

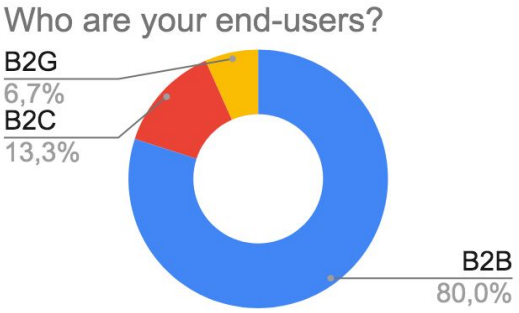


Figure 4. Target users

It is no wonder that listed branches are from sectors which we may call data-intensive branches and left out primary industries like mining, production of goods or agriculture.

4.1.4. Services or Datasets

Asking for the datasets and services which are in the core of existing products of respondents, follow the previous technology ranking (see Figure 3). Thus, those resources which are needed for text processing are in the focus of the respondents.

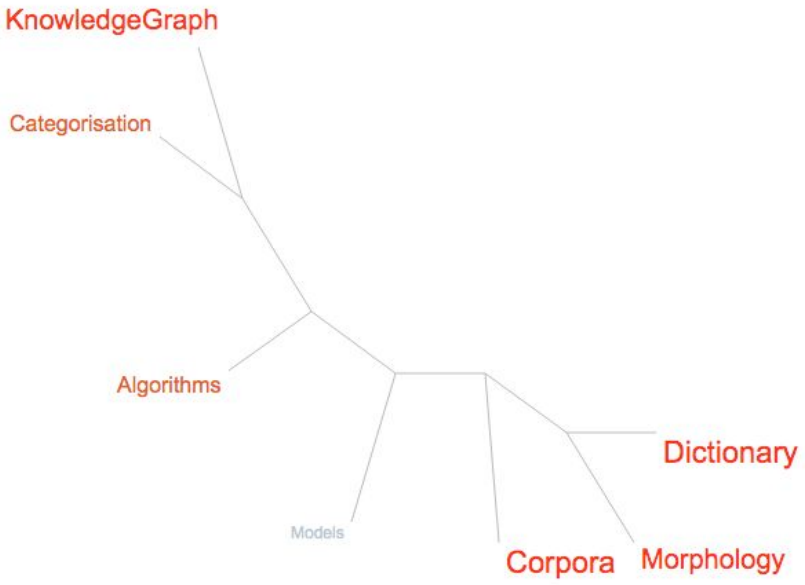


Figure 5: Resources demanded

4.1.5. Offers and Demands for Prêt-à-LLOD

Prêt-à-LLOD provides open solutions in language data and language services along the language technology value chain.

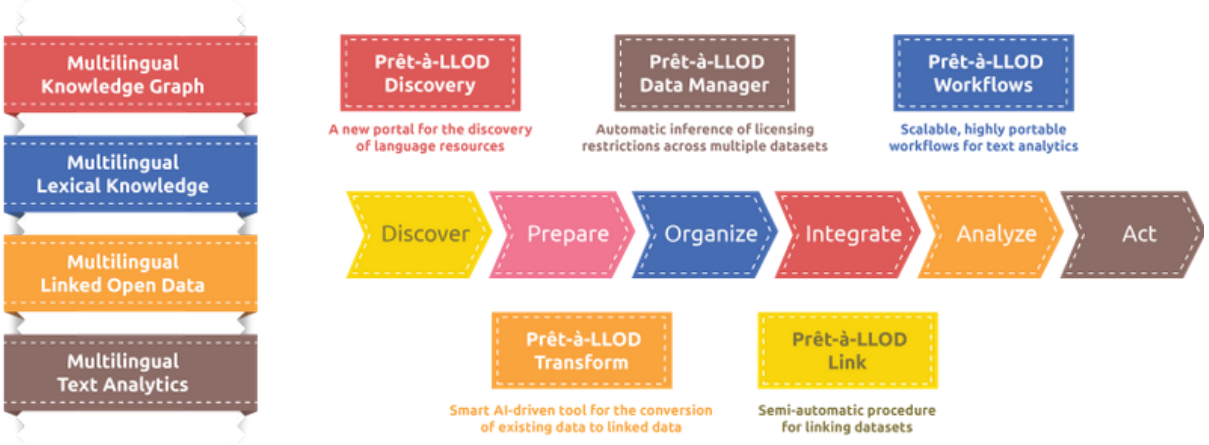


Figure 6: PaL Value Chain

Whilst for the discovery, the preparation as well as the integration and acting with language resources the market (respondents) is kind of balanced, between the demand and provision of resources. There seems to be a clear over-supply of resources when it comes to the organisation of language resources.

Discover	Prepare	Organize	Integrate	Analyze & Act
In this project we will extend the work to analyzing and monitoring the resources directly in order to deduce metadata about the availability, technical quality and content of language resources.	Dataset transformation currently depends significantly on manual transformation. We move beyond this with the use of semantic ontologies which many formats (including XML, CSV, JSON) can be matched to (RDF).	We will investigate (i) the representation of rights information, (ii) the methodology to manipulate policies and provenance information; (iii) and new license composition algorithms.	Look at linking across linguistic data modalities, in particular corpora, lexicons, thesauri and ontologies.	We propose the Prêt-à-LLOD Workflows component will allow the deployment of language technology pipelines on the cloud, increasing the interoperability by using containerization technology

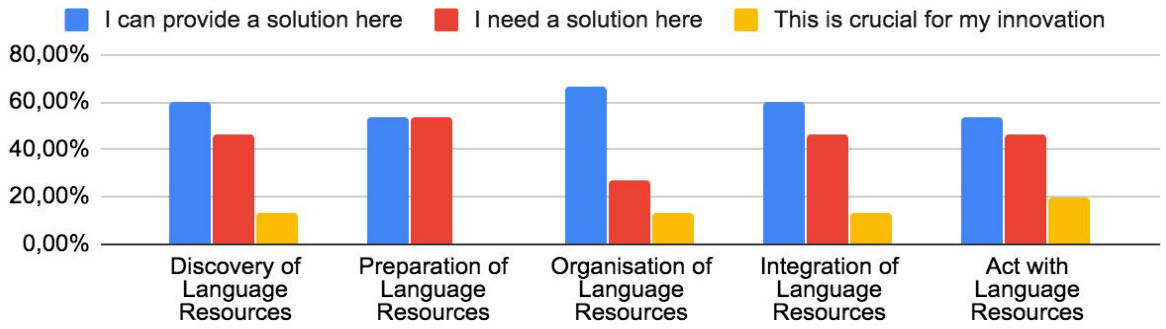


Figure 7: Provision and demand of solutions

On the other hand, respondents say that solutions for discovery, linking and workflow management of language resources may play a role in the innovation of new products and services.

Also, this feedback underpins, the thesis of a more incremental innovation tradition in the sector, build on an existing product (as seen also in the ratio regarding startups in the sector).

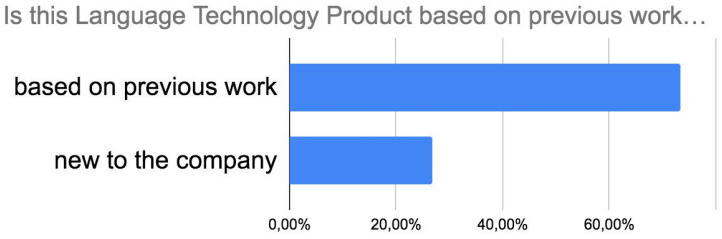


Figure 8: use of new technologies

4.1.6. Business Model

When trading language resources, respondents rely mostly on license fees (71.4%) followed by usage fees (35.7%) and equally subscription and open source (28.6%). This can be identified as a challenge for Prêt-à-LLOD, as the closed business regime of licensing will make the free trading difficult due to license incompatibilities, complicated contracting and fulfilment.

4.1.7. Market Potential

Analysing the respondents' feedback on evolving branches of the sector, the comparison to current focus branches shows markable gaps for some branches, where current activity is way behind the seen potential. Therefore, while question answering is seen as future potential in the field by 60%, only 40% are seeing themselves active in this branch.

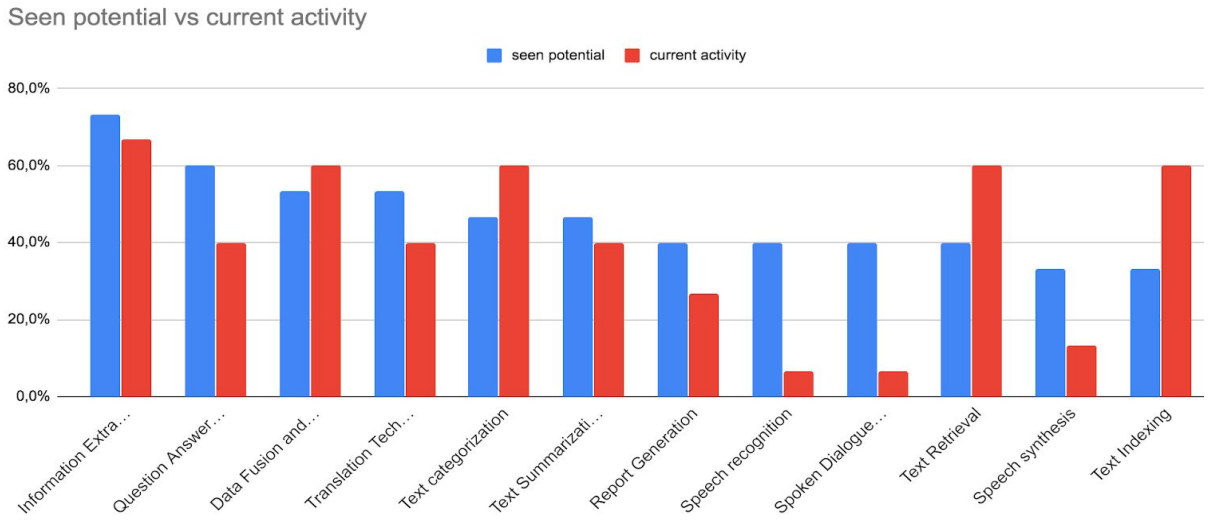


Figure 9: Seen potential vs. current activity

The gap gets even bigger in those branches which are anyway not in the focus of the respondents, like speech recognition, speech synthesis and spoken dialogue. The ratio of approx. 40% vs 5% shows the unexploited potential on the one hand and is an indicator of the potential which special resources in these fields (namely provided by Prêt-à-LLOD) may play.

So respondents are aware of a market situation, which is already saturated in several fields where they are active. So the question of how the competition on existing solutions is estimated brings a clear result of 86.7% competitive products in relation of only 13.3% products with a unique selling proposition.

In this given eco-system respondents see defence, communication, education and the financial Industry as the early adopters for LT. Which - not surprisingly - fits into the distinction between primary industries and the data-driven sector.

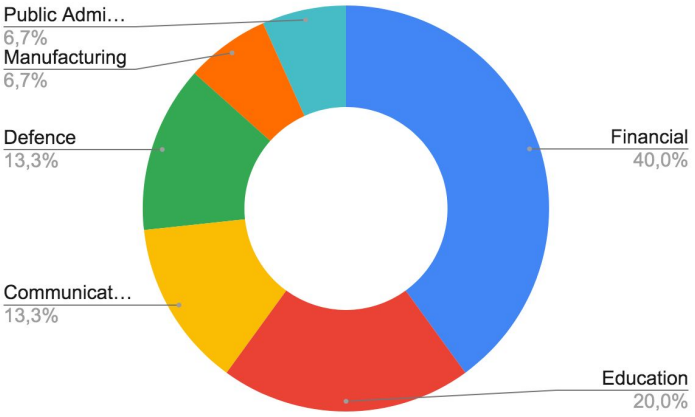
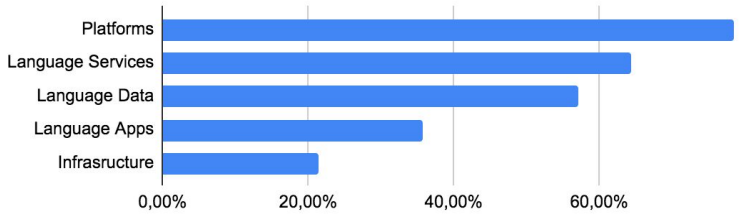


Figure 10: Market shares

4.1.8. Product Positioning in the Market

Asking respondents where they see themself positioned in terms of the product characteristics, you get the following picture:



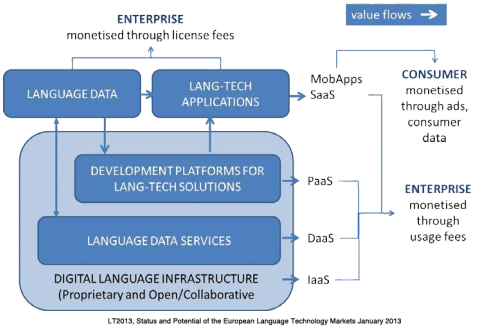


Figure 11: Positioning

4.2. Interviews

Based on the PQ matrix we developed an interview guide to underpin empirical data with selected qualitative data. This was carried out especially to strengthen findings from our public survey, which representativity is endangered by the low participation rate.

4.2.1. Interviewers Guide

1. How do you use language data in your pipeline? Is language data end-product or means to build product?
2. What types of language data are you using? And what would be most valuable for you?
3. What are the sources of the above data?
4. What problems are you experiencing with this data?
5. What languages are relevant for you now, and what languages would you see relevant in future?
6. How would you like to see the data quality improving: e.g. what format? What type of annotations? What cleaning process?
7. Would domain (genre) specific data improve the outcomes/product? 8. What volume is relevant?
8. What evaluation parameters are you using in order to assess the data quality of the source?

4.2.2. Interpretation

Interviewed companies (see Annex 8.1) are already using language resources, which are (mainly) free available on the internet. Tools which are searched and used are for Lemmatization, PoS Tagging, Transforming, Cleaning and Extraction of entities. Content language resources are for the training of ML, NLP, Chatbots etc. We find here the need for monolingual corpora with semantic annotations, semantic relations, synonyms, related terms, morphological lexicons and user-generated content e.g. social media.

There is no single point of accessibility to those resources, companies interviewed therefor use a broad variety of sources like: government portals, dictionaries, wordnet, list of synonyms, real user corpora, twitter bag of words and often have to create a resource from scratch in hire people to develop the sources or pay 3rd parties for getting the resources (datasift.com, scrapinghub.com, proxycrawl.com, socialgist.com)

Work with this language resources comes with different hurdles such as: missing domain classification, data which is not cleaned nor annotated, not well formed (acronyms, abbreviation), limited morphology applied at the tokens.

One of the scenario for the use of language resources from Prêt-à-LLOD is the extension of companies present products and services. E.g. introduce additional languguage coverage demanded by customers. Companies find there a potential for significant benefits in development time and costs.

4.3. Desk Research

The following sections are dedicated to presenting the market data resulting from the online survey targeting a group of 179 vendors and collecting responses from 51 companies. [1]

4.3.1. Company Profiles

Crosslang, Tilde, Elda, IDC [1] see the market as follows: *Only 14% of vendors had revenues over €10M. Nearly half (48%) had revenues below €1M. 52% of our sample had between 10 and 99 employees, and 26% had less than 10 employees, representing nearly 80% of the market. This means there is a long tail of very small vendors, a few leading large vendors and very few mid-market vendors.*

4.3.2. Technology Ranking

Other's than in Prêt-à-LLOD's study Crosslang, Tilde, Elda, IDC [1] follows another classification scheme in technologies of the LT sector:

- Translation technologies including machine translation (MT), translation memory (TM) and translation management systems (TMS);
- Speech technologies including automated speech recognition (ASR) and speech synthesis (text-to-speech or TTS), interactive voice recognition (IVR);
- Natural language understanding (NLU) technologies (e.g. virtual assistants, chatbots, and question answering systems using AI technologies and others);
- Analytics including information retrieval (IR) text analytics, sentiment/opinion analysis, topic modeling, decision support systems);

[1] **Final study report on CEF Automated Translation value proposition in the context of the European LT market/ecosystem**, 2019, A study prepared for the European Commission DG Communications Networks, Content & Technology by: Crosslang, Tilde, Elda, IDC

- Search systems (enterprise search, multi-lingual and semantic search).

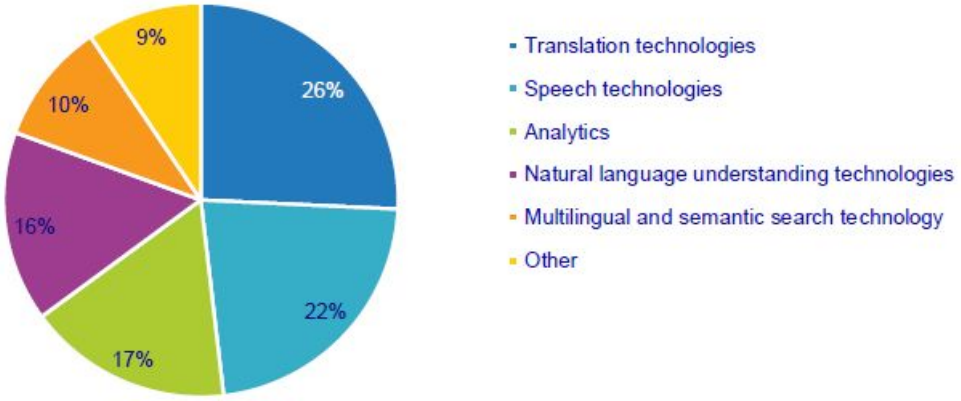


Figure 12: Segmentation - Source 2

In Crosslang, Tilde, Elda, IDC [1], the types of LT offered leaded by translation technology followed by speech technology. Multilingual and semantic search technology are the least important in terms of revenue. Respondents in the survey were quite pleased with the quality increase they experienced recently in automatic translation accuracy.

Here with Speech Technologies (22%) and Analytics (17%) as top technologies, CEF AT - Study *(*Crosslang, Tilde, Elda, IDC [1]*)* shows another picture than our Prêt-à-LLOD survey.

4.3.3. Companies Business Case

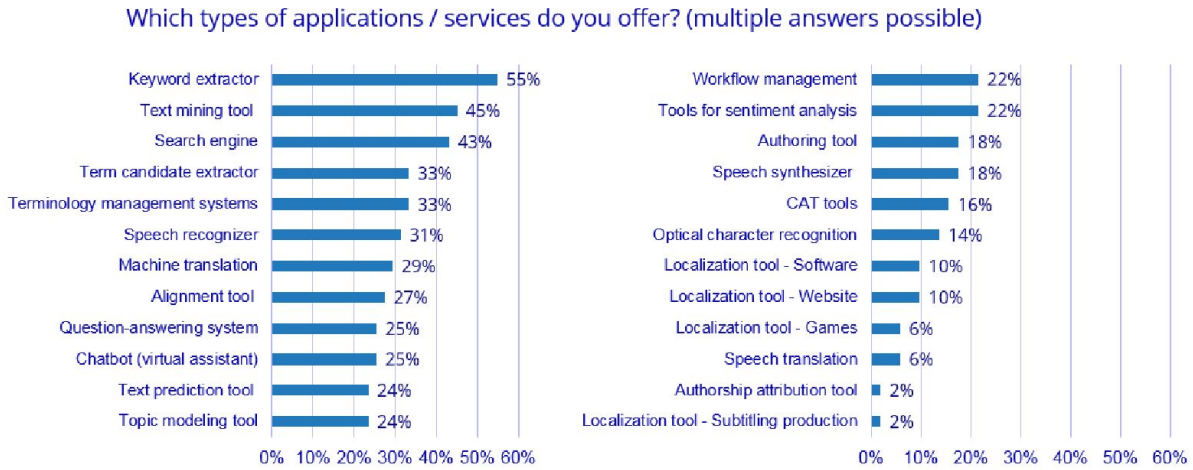


Figure 13: Business Case

4.3.4. Market Niches

According to Crosslang, Tilde, Elda, IDC [1], the LT market in Europe is very fragmented and composed of Small and Medium-sized Enterprises (SMEs), which are typically local players providing local solutions. Profitability is quite low, competition intense and margins are

compressed. The EU does not benefit from one global and leading player. One of the main reasons for this low overall vendor profitability is the need to keep innovating and the cost related to this need.

*When searching for possible market opportunities for European Language Technology, the analysis of the market presence of Google is a must. According to Google, as of May 2017 their multilingual machine translation service offers over 100 languages and counts over 500 million daily users (in May 2017). In August 2017, German technology company DeepL launched DeepL Translator, that uses neural machine translation to rival the capabilities of Google Translate (*Crosslang, Tilde, Elda, IDC [1]*).*

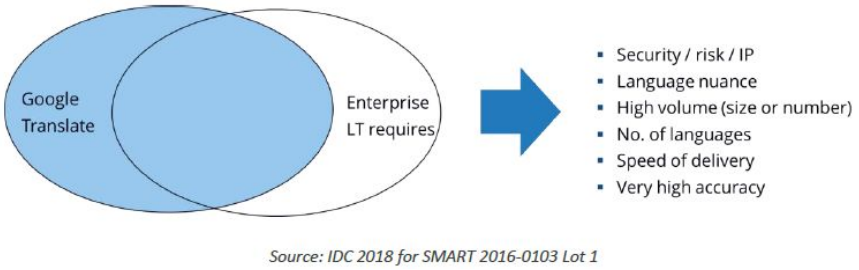


Figure 14: Google Translate

*However, market share and brand visibility remain for the most part with Google. Nevertheless, for many large enterprises, Google Translate is not sufficient due to the size and complexity of the LT task and the level of security and degree of accuracy required. This is the market opportunity that is currently being exploited by local players (*Crosslang, Tilde, Elda, IDC [1]*).*

4.3.5. Market Characteristics

Key Findings by CEF AT - Study *(*Crosslang, Tilde, Elda, IDC [1]*).*

- *The LT market is very fragmented and composed by SMEs and lacks of large indigenous players. Their go-to-market is often to tackle niche markets where competition is less intense.*
- *Profitability is on average quite low. Market players need to fight to reach and to maintain profitability, as margins are compressed.*
- *The LT market is relatively small. As of today, the relative size of the LT market is not huge especially if compared to the overall IT market.*
- *LT is a growing market. Language technologies are growing markets, where customers today have more awareness of benefits also due to marketing of large players.*
- *Competition is intense. Despite LT being a growing market, it is also a market where competition is fierce, and players need to keep innovating, as well as to go to market with the right solution at the right time and often through the right channel and deploy the appropriate partnerships.*

- *"Large non-European players are a blessing and a curse". From the local vendors' point of view, one of the positive effects of large players such as Google, Microsoft and Apple is that they strongly contribute to create or increase market awareness. On the other hand, they are tough competitors who offer mass market free software which is difficult to compete with, especially for SMEs.*
- *Automatic translation accuracy has increased strongly over the past 2-3 years. Even if 100% accuracy is most likely a utopia, accuracy is on the increase and players are keeping working on it to offer better services to their customers.*
- *Speech generation and natural language understanding will improve. Language generation and natural language understanding will improve contributing strongly to higher acceptance of LT technologies.*
- *Chatbots will be increasingly widespread. The chatbot market is maturing quickly and they are becoming a natural part of language translation technologies.*
- *The Artificial Intelligence (AI) market is growing strongly. The AI market will grow at more than 40% compound annual growth rate to 2021. AI will be increasingly part of LT technologies and will boost LT market.*

4.3.6. Language Service Providers (Slator 2020 Language Industry Market Report)

The Slator 2019 Language Service Provider Index (LSPI)[2] were selected based on their revenues and market activities for 2017 and 2018, as they represent a meaningful composite of leading vendors.

The 2019 Slator LSPI begins with a Leaders group that represents the top 30 or so leading language service providers and where assigning a rank is meaningful. The index feature a Challenger group composed of companies with significant revenues but where assigning a rank is no longer meaningful given the overall fragmentation of the industry.

So the LSI is an Index of leading language service providers, whose sources of revenues are derived from services such as translation, localization, language technology, interpretation, subtitling and dubbing and other related services.

	Company Name	Head-Quarter	Mil $ 2017	Mil $ 2018	Growth	Ownership
1	TransPerfect	USA	705.0	614.8	14.7%	Private (Phil Shawe)
2	Lionbridge*	USA	650.0	590.0	10.2%	PE (HIG Capital)
3	LanguageLine Solutions	USA	480.0	451.0	6.4%	Teleperformance
4	SDL	United Kingdom	412.4	388.3	12.4%	Listed UK

[2] https://slator.com/language-service-provider-index/the-slator-2019-language-service-provider-index/

5	RWS	United Kingdom	398.4	219.8	86.6%	Listed UK (Exec Chairman A. Brode, 32.97%)
6	Keywords Studios*	Ireland	286.4	181.8	65.1%	Listed UK
7	Welocalize	USA	227.0	200.0	13.5%	PE (majority - NEP)
8	SDI Media	USA	225.0	221.0	1.8%	100% owned by Imagica Robot Holdings Inc.
9	STAR Group	Switzerland	177.0	166.2	7.4%	Private
10	Amplexor International	Luxembourg	174.8	171.1	7.1%	Saarbrücker Zeitung media group
11	CyraCom International	USA	147.0	139.0	5.8%	Private
12	Acolad Group (Technicis)*	France	134.0	52.8	165.9%	PE (Majority-owned by Naxicap)
13	BTI Studios*	Sweden	114.5	115.2	4.2%	PE (Altor and Shamrock Capital)
14	Semantix	Sweden	111.4	106.9	13.0%	Majority-owned by PE (Segulah V L.P.)
15	thebigword	United Kingdom	101.5	77.0	27.8%	Private
16	Pactera Technology International	China	100.0	85.0	17.6%	Private (HNA Group)
17	Honyaku Center	Japan	99.9	91.7	3.9%	Listed Japan
18	Ubiqus	France	84.3	84.0	5.1%	PE (Euromezzanine, Indigo Capital), other private
19	Voice & Script International	United Kingdom	77.9	66.1	38.8%	Private
20	LanguageWire*	Denmark	69.0	33.8	114.2%	PE (CataCap)
21	IYUNO Media Group	Singapore	58.3	34.8	67.7%	Private, VC (Softbank)
22	Stratus Video	USA	56.0	47.4	18.1%	PE (Kinderhook)
23	KERN Global Language Services	Germany	54.4	54.4	4.9%	Private
24	Morningside Translations	USA	48.8	42.8	14.0%	Private
25	SeproTec Multilingual Solutions	Spain	43.3	37.3	16.3%	Private

26	Certified Languages International	USA	41.4	36.8	12.5%	Private
27	Livewords	Netherlands	39.7	35.0	19.0%	PE (Bencis Capital Partners)
28	Akorbi	USA	39.3	36.3	8.4%	Private
29	CSOFT International	China	38.9	41.3	-5.8%	Private
30	ZOO Digital	United Kingdom	28.6	16.5	73.3%	Listed UK
31	Apostroph Group	Germany	26.5	25.0	11.1%	PE (ECM)

Language Service Providers (LSP) with significant revenue which form the midfield of the highly fragmented language industry. The Challenger Companies section contains a list of smaller companies whose revenues we ascertained during the course of our research into the top LSPs globally. The language service industry is highly fragmented and there are lots of companies in the mid-field — meaning it is extremely difficult to continue to rank LSPs as company revenues decrease to below the USD 25m mark. Although LSPs included in the list of Challenger Companies have been ordered by 2018 revenues, the list should by no means be taken to be a complete one of companies of this size (ca. USD 10–25m). Of course, there is still value in making this data available. Companies may choose to use it as a benchmark for their own performance and growth, as an indication of growth in the language services industry, and as a starting point for evaluating strategic options including M&A.

	Company Name	Head-Quarter	Mil $ 2017	Mil $ 2018	Growth	Ownership
32	Janus Worldwide	Austria	23.6	19.9	18.4%	Private
33	Translated	Italy	23.4	20.8	17.9%	Private
34	Lan-bridge Communication	China	21.7	17.7	29.6%	Private
35	Argos Multilingual	Poland	21.0	15.4	36.6%	Private
36	EC Innovations	Singapore	20.8	17.9	16.2%	Private
37	CBG Konsult & Information AB	Sweden	20.7	18.3	11.5%	Private
38	Awatera	Russia	19.4	21.6	8.1%	Private investors and management shareholders
39	Rozetta Corp.	Japan	18.8	17.0	5.2%	Listed Japan
40	Nordisk Undertext	Sweden	18.6	5.5	266.7%	Private

41	Versacom	Canada	18.3	19.1	4.2%	Private
42	NLG GmbH	Germany	18.1	17.9	6.1%	Private
43	Transline Gruppe Gmbh	Germany	17.8	15.6	19.2%	PE (LEAD Equities Group) and Dr.-Ing. Wolfgang Sturz
44	TRSB Inc.	Canada	17.6	15.9	20.0%	Private (Serge Belair)
45	LanguageLoop	Australia	16.8	16.9	3.4%	Government agency
46	Interpreters Unlimited	United States	15.5	12.1	28.1%	Private (Sayed Ali)
47	Easytranslate	Denmark	15.3	10.6	51.5%	Private
48	EVS Translations	Germany	15.0	12.0	25.0%	Private
49	Lingsoft	Finland	14.8	11.2	16.0%	Private (majority owner Juhani Reiman)
50	Lylo	France	14.2	11.2	26.8%	Private
51	Language Connect	United Kingdom	14.0	11.9	24.7%	The Hut Group
52	itl Institut für technische Literatur AG	Germany	13.7	14.2	1.7%	Private (Christine Wallin-Felkner)
53	MasterWord Services	United States	13.7	14.6	-6.2%	Private
54	Lingo24	United Kingdom	13.7	12.0	21.2%	Private
55	Propio Language Services	United States	12.5	8.2	52.4%	Private
56	Straker	New Zealand	12.3	8.3	44.3%	Listed Australia
57	Diction	Switzerland	12.2	11.3	9.1%	Private
58	Geneva Worldwide	United States	12.0	11.6	3.7%	Private
59	HansemEUG	Korea	11.0	12.0	-8.3%	Private
60	EGO TRANSLATING COMPANY	Russia	10.9	12.6	4.0%	Private

61	FastTranslator.com	Netherlands	10.3	11.5	-5.9%	Private
62	e2f	United States	10.3	8.8	17.0%	Private
63	Dynamic Language	United States	10.2	11.3	-10.1%	Private
64	Sandberg Translation Partners	United Kingdom	9.9	9.9	5.9%	Private
65	Iota Localisation Services	Ireland	9.4	8.9	11.4%	Private
66	Tolingo	Germany	9.4	9.6	2.5%	Private and VC (Acton and others)
67	Translate Media	United Kingdom	9.0	10.1	-5.6%	Private
68	Kaleidoscope GmbH	Austria	8.6	9.0	0.0%	Private
69	Linguaserve	Spain	8.0	7.0	19.5%	Private

2018 was a positive year overall, with strong double-digit growth for many of the leading 40 or so players. Growth among Leaders was marginally stronger than among Challengers. While one driver of the Leaders' outperformance was M&A, the data suggests larger LSPs are indeed growing faster than smaller ones on average. Only one company on the Leaders list reported negative revenue performance in 2018, while a total of five on the Challenger list did so.

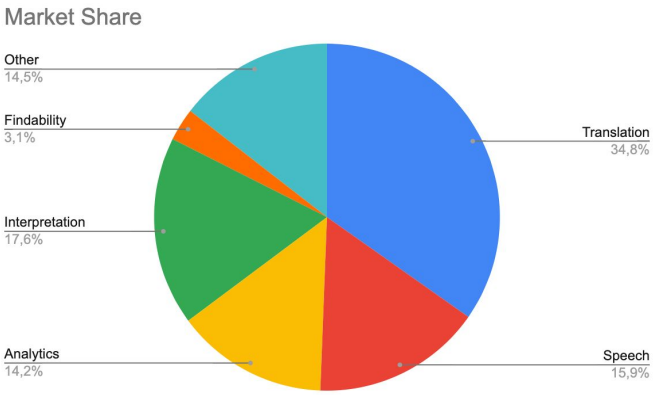


Figure 15: Market acc. SLATOR 2019

4.3.7. Content / Lexical Resources Providers

Company Name	Languages offered	Main products	NLP services (products)
Taus	600 language pairs 17 content types	• Parallel language data • Customized corpus (domain specific) • Matching data • Dashboard: quality evaluation, reporting, benchmarks • Interface for viewing/downloading samples	• Lemmatization • PoS Tagging • Phrase extraction • not explicitly stated
Bitext	80 languages and variants	• Morph analysers • Embeddable services (e.g. chatbox)	• Lemmatization • PoS Tagging • Phrase extraction • Sentiment Categorization • Language Identification
Lexical Computing Ltd	90+	• Corpora creation • Text analysis software • word lists • n-gram lists • word databases • lexicons • Text analytics API	• Morphological analysis • Tagging • Stemming/Lemmatization • Parsing
Linguistic Data Consortium	80 languages and variants	• Resource provider • Corpus creation software	• Alignment tools • Manual annotation tools • Conversion tools
Ravenpack	English (others?)	• Big data analytics for financial services (risk management, investment, competitive intelligence) • Vizualization tool • Custom dataset creation tool	• Not explicit, but it must include: • Lemmatization • PoS tagging
Lingea	5 - 44 (product dependant)	• Translation and localization (MT) • Audio processing (corpora compilation) • Text processing (not sure if NLP analysed) • Typesetting solutions (multiformat) • Language identification • Spoofing tools: Word hyphenation, Thesaurus, spellchecker, Diacritics • Term translator	• Tagging • Stemming/Lemmatization • Morphological analyser
European Language Resources	Mostly European languages but	• Language resource provider • Parallel Laxicon	• NA

Association	also a range of other (mostly) majority languages in the world		
Corpus data	English, Spanish, Portuguese	• Full-text corpus data in 3 formats: db, vertical, linear • Word frequency data • Ngrams • Collocates	• Lemmatization • PoS tagging
Sketch engine	90+	• Tagged/annotated corpora • Interface, corpus building, corpus querying and text analysis tool	• depending on languages (Lemmatization, PoS tagging)
Event Registry	30+	• news analytics platform: • media intelligence: news feed, historical data, informative visualizations • media monitoring: • news api	• NO
Kantantmt	90 language pairs	• Automated Translation Platform • API	• MT
Verilogue	English (others?)	• searchable document (transcripts, real dialogue) repository. Word embeddings can be created from audio, chat, transcripts, etc.	• text extraction • (medical) metaphor analysis • manual annotation
welocalize	250+	• Platform	• Text Extraction • Sentiment Analysis • Image and Video Annotation • Categorization • Classification
Figure 8	English (others?)	• NLP services • Platform that annotates, adds labels, on userdata	• bespoke
Appen	180+	• high-quality, human annotated datasets • AI-assisted data annotation platform	
Lionbridge	Several	• Machine translation and computer-aided translation services and products (translation memories) • Translation and localization services • Data Creation • Annotation • Linguistic Services	• Tokenization • Lemmatization • Tagging

5. Business Model Canvas (BMC)

In this section we develop the user stories in the common description framework of the business canvas[3]. Business Model Canvas is a strategic management and lean startup template for developing new or documenting existing business models. It is a visual chart with elements describing a firm's or product's value proposition, infrastructure, customers, and finances. It describes in brief (Definition by Wikipedia[4]):

- ***Key Activities***: *The most important activities in executing a company's value proposition. An example for Bic, the pen manufacturer, would be creating an efficient supply chain to drive down costs.*
- ***Key Resources***: *The resources that are necessary to create value for the customer. They are considered assets to a company that are needed to sustain and support the business. These resources could be human, financial, physical and intellectual.*
- ***Partner Network***: *In order to optimize operations and reduce risks of a business model, organizations usually cultivate buyer-supplier relationships so they can focus on their core activity. Complementary business alliances also can be considered through joint ventures or strategic alliances between competitors or non-competitors.*
- ***Value Propositions***: *The collection of products and services a business offers to meet the needs of its customers. According to Osterwalder (2004), a company's value proposition is what distinguishes it from its competitors.*
- ***Customer Segments***: *To build an effective business model, a company must identify which customers it tries to serve. Various sets of customers can be segmented based on their different needs and attributes to ensure appropriate implementation of corporate strategy to meet the characteristics of selected groups of clients.*
- ***Channels***: *A company can deliver its value proposition to its targeted customers through different channels. Effective channels will distribute a company's value proposition in ways that are fast, efficient and cost-effective. An organization can reach its clients through its own channels (store front), partner channels (major distributors), or a combination of both.*
- ***Customer Relationships***: *To ensure the survival and success of any businesses, companies must identify the type of relationship they want to create with their customer segments.*
- ***Cost Structure***: *This describes the most important monetary consequences while operating under different business models.*
- ***Revenue Streams***: *The way a company makes income from each customer segment. Several ways to generate a revenue stream*

This BMC are meant as proposals of either the generic business mechanic we have in mind (see 5.1) or of more specific business ideas, describing a niche product (like 5.2).

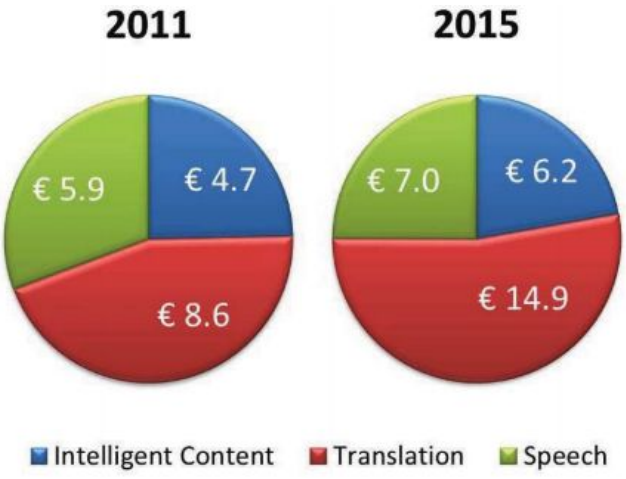
[3] The Business Model Canvas was initially proposed by Alexander Osterwalder (Osterwalder, Alexander (2005-11-05). "What is a business model?". *businessmodelalchemist.com).*
[4] https://en.wikipedia.org/wiki/Business_Model_Canvas on 29.09.2019

5.1. The LTi Market Model

The LT-Innovate-Study 2013[5] do not provide a full blown BMC, but discusses trends and growth in the sector accordingly.

To measure the scale of the LT market involves modelling its components, and hypothesising about size, segmentation and growth rates. At present no analysts follow the LT market as a whole, though many track its components and sub-components in different ways; larger pure-play LT companies are tracked, as are the LT-related developments and products in the larger software companies. As will become evident in the discussion of trends, the borders between the technological segments (speech, translation, content) are fuzzy at best, and much of the real innovation in LT is happening at the edges, where different types of intelligent services are combined in Unified LT applications (speech and translation, intelligent content and translation, etc.)

LT2013 by LT-Innovate[4]



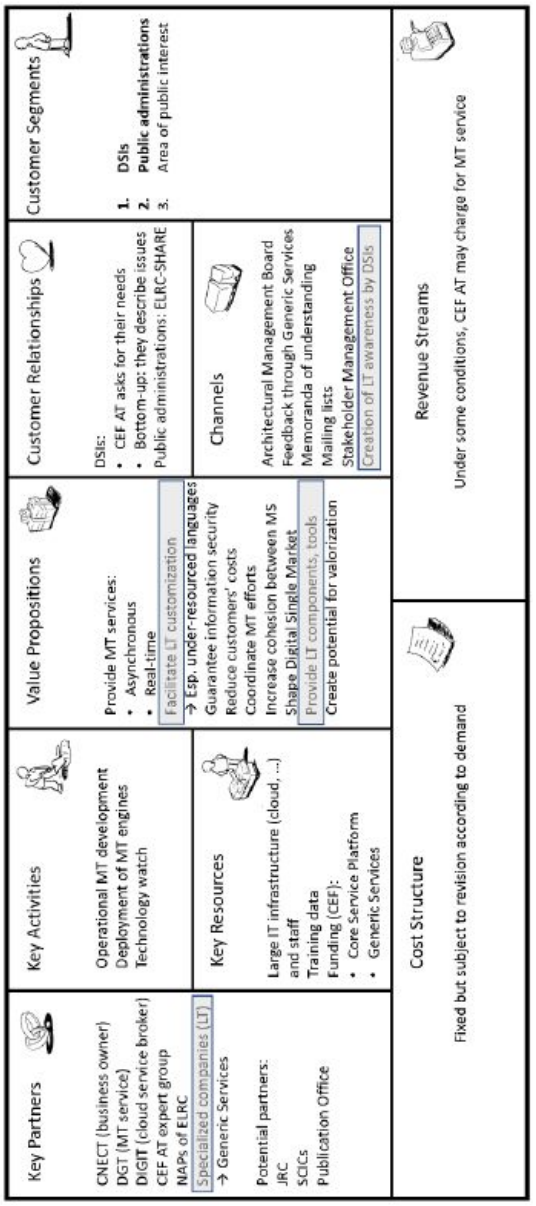


Figure 16: Worldwide Language Technology Software & Services Market

[5] **LT2013, Status and Potential of the European Language Technology Markets, January 2013** by LT-Innovate

5.1.1. Trends and Growth – Speech Technology

The market is heavily dominated by speech recognition, with a long history of commercial use, positioned as cost-saving technology. Speech transcription services (e.g. in the healthcare domain) are increasingly offered through cloud services.

Improvements in the quality of TTS, combined with platforms requiring interactivity (such as mobile, gaming) are driving new opportunities for speaking applications. Notable features are naturalistic voices in many more languages, used in education and gaming environments, as well as interactive access to the web (Voice Portals).

Major markets for Speech:

- *Call Centre is a core global market*
- *Medical reporting and transcription is growing (especially in the USA for compliance with new Electronic Health Records regulations)*
- *Large and stable government customer base (including specialised defence applications)*
- *Speedy growth in consumer markets on devices and social platforms.*

LT2013 by LT-Innovate[4]

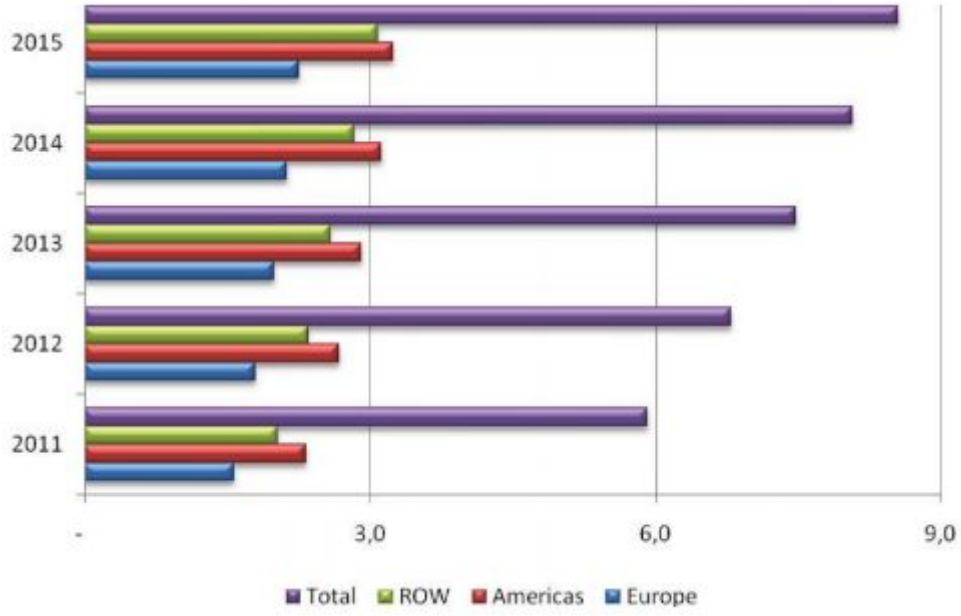


Figure 17: Worldwide Speech Technology Software & Services Market 2011-2015 (€B)

5.1.2.Trends and Growth – Translation Technology

According to the LT-Innovate Study 2013 (Source 1), the estimate of the size of the 2011 translation technology market, including software and services, is €8.6B, the vast majority spent on technology-based services; direct software revenue is only 7% of the market.4 The five-year CAGR for translation is 14.6%. Translation is the LT application least susceptible to full-scale commoditization, at least for the foreseeable future.

By 2015, the value of the translation technology market is forecast to grow to €14.9B; while services continue to constitute most of the spending, the share attributable to software grows to 12%. Growth rates in the software share of the market compound as the new translation platforms mature. Services continue to grow, but at a slower pace, averaging 13% per year. The expansion of the size of the market is driven primarily by technology.

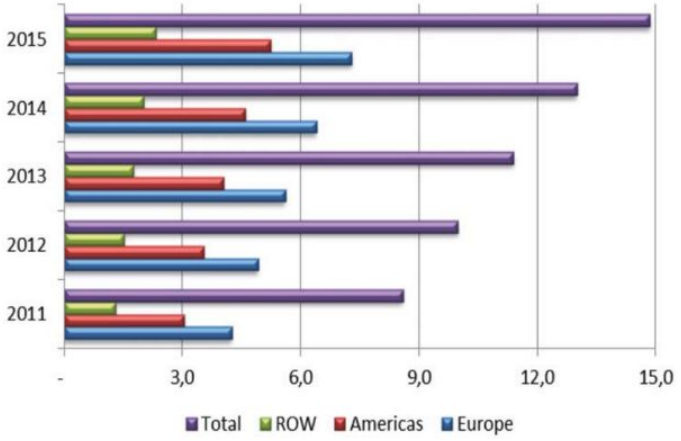


Figure 18: Worldwide Translation Technology Software & Services Market

LT2013 by LT-Innovate[4]

5.1.3.Trends and Growth – Intelligent Content Technology

Customers are looking for faster (and therefore more actionable) insight, to handle a diverse range of data and content resources, to solve core business problems. New search-based applications supporting a specific task or workflow (e-discovery, fraud detection, voice of the customer, sales prospecting, research, customer support) integrate domain knowledge to support the particular task, including industry taxonomies and vocabularies. Search is embedded within the process.

On the content side, we see intelligent automatic authoring of regulated documents in industries such as chemicals and pharmaceuticals. Intelligent creation of content overlaps with translation, as both promote the use and management of standard terminology, and use linguistic analysis of text to achieve clearer and more translatable content.

Although with its horizontal/Enterprise focus, IC Technology is used in all industries, markets leading thetake-up of advanced search and analytics include:

- *Banking and Financial Services*
- *Communications, Media and Services*
- *Government*
- *Manufacturing*
- *Natural Resources*

LT2013 by LT-Innovate[4]

5.1.4. LTi's 2013 conclusions for the European Market

The forecasts in the model predict that the Translation segment will continue to dominate the European LT market, and will grow to be a larger overall share (65%) by 2015. Intelligent Content remains the smallest segment in Europe, and speech is only slightly larger. The assumptions of the model are based on recent trends in the respective segments, notably the dilution of the European industry in both speech and content through acquisition by off-shore companies. By contrast, consolidation in the translation industry has historically been Euro-centric; acquiring European translation company signals, by definition, a desire to continue to operate in the local market and develop local linguistic talent and resources. Moreover, translation technology development has historically been a European strength.

Europe's share of the worldwide market will increase slightly to 38% over the five year period, compared to 42% in the Americas. However due to the imbalance between LT segments, that share is significantly lower (24% in 2015) for the software portion of the market; as we have noted, sales of technology-supported human translation services far outweigh sales of translation software, and will continue to do so during the forecast period. The strength of the European market for translation reflects both the depth and excellence of the industry in Europe, and the need for translation into the many languages of Europe on a large scale.

Large-scale multilinguality, in turn, is an inhibitor for growth in both the speech and content markets, where products and applications must be deployable in local languages.

LT2013 by LT-Innovate[4]

Factors that could change the assumptions behind the market model:

- *Faster and more extensive deployment of content applications in more European languages, in a coherent framework for all languages.*
- *Development – and integration – of speech components (for recognition, generation, and identification/verification) in more European languages, affordably available for European app and solution developers.*
- *Large-scale deployment of open source machine translation in open environments using shared resources.*
- *Large-scale sharing of resources (paid and free) throughout the European industry.*
- *Development of vertical and industry-specific platforms for LT development and deployment, engaging whole industries in cooperative initiatives (analogous to SWIFT in banking).*

LT2013 by LT-Innovate[4]

Technological Barriers

LT is a highly complex technological domain that represents the intersection of several disciplines, including the many sub-domains of linguistics, mathematics, and information science. LT functionality remains (and may always remain) a work in progress, with few genuine technological breakthroughs. The most significant, for current technology, was the combination of NLP and computational linguistics with statistical modelling, which began more than thirty years ago, and is now a feature of many LT implementations including both speech and text.

Most improvements using today's technologies are incremental and rely particularly on the ability to access and maintain ever larger and more finely tuned linguistic data. Lack of access to that data will constrain the technological development of LT. Acquiring and using it may rely on cooperation between the LT industry and the different constituencies that own, need and use it. Collaboration between the industry and data owners will be needed. Also regulation of the use of such data should be made much more open, and core data (such as terms, concepts, and ontologies) should be standardised and shared in an open environment.

LT2013 by LT-Innovate[4]

5.2. The LT-BMC developed by CEF AT - Study

According to Crosslang, Tilde, Elda, IDC [1] the LT Business Model provided in the "Final study report on CEF Automated Translation value proposition in the context of the European LT market/ecosystem" goes beyond CEF AT and also involves customisation of LT components in a broader sense. The fact that this assumption by the proposed Business Model is closer or near the market and characteristics we look for in the Prêt-à-LLOD business stories.

5.2.1. Customer segments

- Digital Service Infrastructures: Providing customisation services to them.
- Public administrations
- The area of public interest. For instance, museums that are involved in cross-border collaboration

5.2.2. Value proposition

- Making customers service/content multilingual
- Reduce costs, by automating the translation
- Translating documents or text snippets, within a short delivery time

Figure 15: LT-BMC developed by CEF AT

5.2.3. Channels

- The AMB (Architectural Management Board) coordinates architectural activities of building block
- End users provide technical feedback to DGT.
- There are also memoranda of understanding between entities working for CEF
- Information is provided via channels like mailing lists

5.2.4. Customer relationships

- Communication on needs in one direction - between CEF AT and DSIs.

5.2.5. Key resources

- Physical resources in the form of infrastructure and data for training the MT system.
- As for intellectual (human) resources, CEF AT's activities are performed by a variety of profiles. These profiles include machine translation experts, project managers, software developers for integration, UI (user interface) developers, testers, cloud expertise,
- As for financial resources, the budget for the Core Service Platform is provided by CEF.

5.2.6. Key activities

- CEF AT focuses on operational development and deployment of engines.

5.2.7. Key partnerships

- CNECT is the business owner, while DGT, DIGIT are business providers, providing the eTranslation and cloud service.
- JRC, SCICs (Service for Conference and Interpretation), Publication Office are potential business partners. CEF AT expert group, NAPs (National Anchor Points) of ELRC are partners.
- Some of the above partners perform key activities. The MT team at DGT provides the eTranslation service. DIGIT is a cloud service broker.

5.2.8. Cost structure

- The budget is fixed. The cloud consumption is proportional to the translation needs.

5.3. The Prêt-à-LLOD-BMC

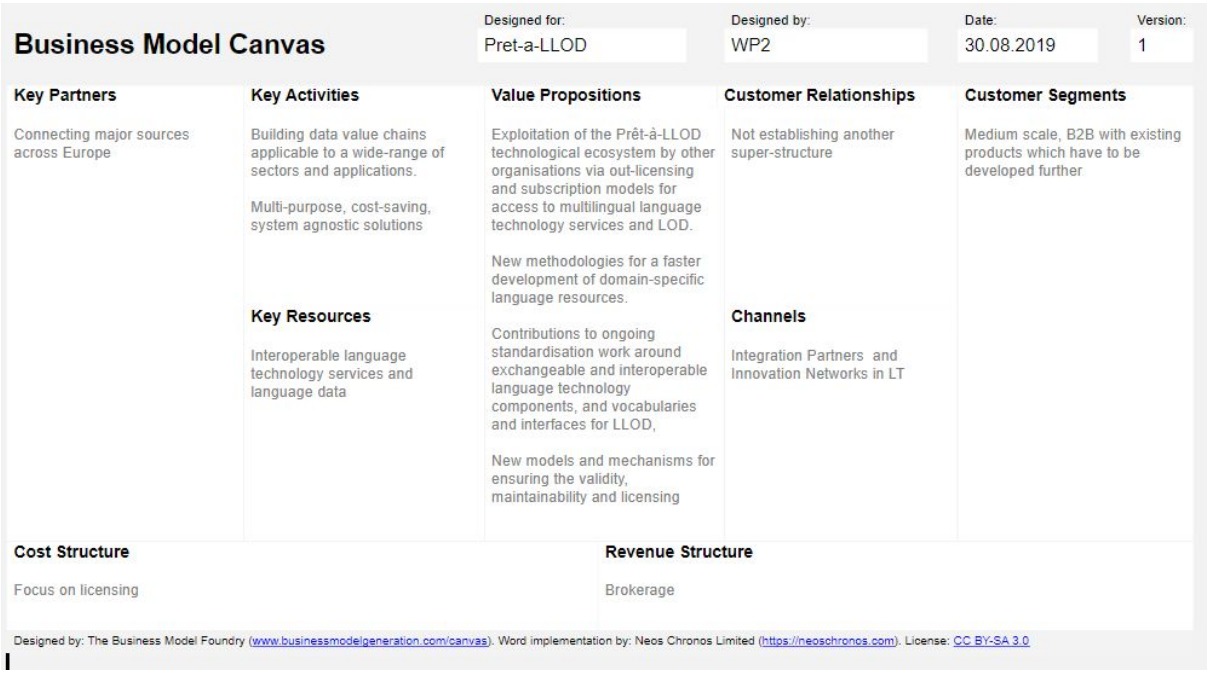


Figure 19: Prêt-à-LLOD-BMC

5.3.1. Key Activities

Building data value chains applicable to a wide-range of sectors and applications

This project's principal objective is to utilize linked open data and language technologies in order to create groundbreaking cross-sectoral applications. Prêt-à-LLOD targets multi-purpose, cost-saving, system agnostic solution creating a new methodology building data value chains applicable to a wide-range of sectors and applications and based around language resources and language technologies that can be integrated by means of semantic technologies, in particular the usage of Linguistic Linked Open Data (LLOD)

5.3.2. Key Resources

Interoperable language technology services and language data

Prêt-à-LLOD provides an ecosystem to support the development of novel linked data-aware language technologies. We will provide data discovery tools based on metadata aggregated from multiple sources, methodologies for describing the licenses of data and services, and tools to deduce the possible licenses of a resource produced after a complex pipeline.

5.3.3. Partner Network

Connecting major sources across Europe

The Prêt-à-LLOD services built on the existing Linghub portal, cover major dataset sources across Europe and the world including EUDAT and Datahub and in particular language resource repositories including ELRA, LDC, Metashare, CLARIN, and the European Language Grid.

5.3.4. Value Propositions

Utilize LT to facilitate customers projects and products

The project envisions the improvement of the data value chain by providing concrete tools that utilize language technologies and linked data in order to facilitate customers LT projects and products.

- **Exploitation** of the Prêt-à-LLOD technological ecosystem **by other organisations** via out-licensing and subscription models for access to multilingual language technology services and LOD.
- New methodologies for a **faster development** of domain-specific language resources.
- Contributions to ongoing standardisation work around **exchangeable** and interoperable. **language technology components**, and vocabularies and interfaces for LLOD,
- New models and mechanisms for ensuring the **validity, maintainability and licensing** of language resources.

5.3.5. Customer Segments

Medium scale, B2B with existing products which have to be developed further

Following the characteristics of our market survey, customers are found in smaller and medium sized European companies, which are targeting all kinds of text-processing Language Technologies. Products of these customers are highly specialized and often limited on a specific language (home-)market.

The target customer is B2B oriented and has already passed through it's start-up phase. So developed solutions of the customers may benefit from Prêt-à-LLOD by incremental development and increase of cost efficiency.

5.3.6. Channels

Integration Partners and Innovation Networks in LT

The technological ecosystem developed by all the partners together will support each of the industrial partners in reducing costs and time-to-market for providing their products to their sectors. Stakeholders of these sectors - as there are partners in the consortium, steering group, research and startup-scene - may integrate the technologies developed in Prêt-à-LLOD into their products and thus tailor the solutions to particular markets/sectors in which they operate.

5.3.7. Customer Relationships

Not establishing another super-structure

In not establishing another super-structure for the LT sector, Prêt-à-LLOD builds a sustainable ecosystem, to ensure the sustainability of the Prêt-à-LLOD outcomes beyond the duration of the project. In close connection with other ongoing initiatives such as the future European Language Grid, the single players in the ecosystem carry on growth, innovation and commercial success.

The ongoing work on vocabulary and interface specifications and ongoing community building establishes refreshed contact from and to customers permanently.

5.3.8. Cost Structure

Focus on licensing

Cost of usage of resources of the Prêt-à-LLOD ecosystem have to be inline with the customers business models. As the focus lies on licensing, the cost structure for LT resources have to follow this paradigm.

So license clearing and license merging will become central for Prêt-à-LLODs exploitation. It influences also on how adaptable the resource-usage from and to Prêt-à-LLOD may be organized.

5.3.9. Revenue Streams

Brokerage

As common for the growing landscape of data markets and in compatibility with them, Prêt-à-LLOD should go for brokerage fees for the intermediate service between two parties. Bilateral trade between parties (not touching Prêt-à-LLOD's infrastructure) may agree on another value exchange independently.

6. Exemplified Business User Stories

A user story can be described as a high-level statement of a requirement that does not go into excessive detail. It describes the functionality or feature that a product is expected to deliver to the user. Stories encourage iterative development and can be refined as many times as possible to reach agreement and understanding among stakeholders. User stories may be expressed by presenting the role, the goal or the value first. BAs should however choose whichever format is best for expressing their requirements by considering the context.

The user story is placed as a short narrative and used as a reminder of the conversation between the customer and the developer. It is usually expressed as: Name + Brief Narrative + Success Criteria.

6.1. LT Middleware

6.1.1. Actors

Small and medium sized companies developing software packages

6.1.2. Narrative

The middleware developed is solving a specific problem within a bigger LT stacks. It focuses on one or a couple of function within they stack, like Cross Lingual Linking, RDF Representation, Monitor Corpora, Lexical covering of nouns, verbs, adjectives and adverbs, cognitive synonyms (synsets), linking dictionaries, Sense Level Linking, Linguistic annotation, Cloud computing integration, or General Text Analytics.

Those solutions **consume and provide solutions** of a bigger LT stack, but they never are capable of solving customers' problems isolated. They need the embedding in a bigger ecosystem like Prêt-à-LLOD.

6.1.3. Success Criteria

Integrators play a key role to recombine the LT stack to customer satisfying solutions. Comparability of the stack components, standardisation of resources, and a sufficient number of vendors are key to establish and successfully grow such an eco-system of interlinked providers of middleware solutions.

6.2. Lexica and Dictionaries

6.2.1. Actors

Providers of lexica, monolingual and bilingual dictionaries, sentence databases, audio databases, transcripts, morphologies, and wordlists.

6.2.2. Narrative

At the heart of the LT stack stands a couple of linguistic resources, which allow the LT to analyse, categorize, review, judge and interpret language in written or spoken form. Based on various linguistic and lexical traditions and methods this resource has to be aligned and harmonized before they can be used interlinked and related. Lexical methods have to be intertwined with AI driven analysis.

6.2.3. Success Criteria

Usage aware (or even agnostic) language resources play a key role in getting a resource pool which powers the whole LT stack and value chain equally

6.3. Language Models and Algorithms

6.3.1. Actors

Providers of Morphologies, Language Models, Classifiers and Embeddings, algorithms like SVM, Deep Learning, Naive Bayes, and others.

6.3.2. Narrative

Models for ASR (automatic speech recognition), Datasets to train acoustic models, language models, dictionaries, annotated corpora for sentiment analysis, topic detection and detection of named entities. Tools to allow customers/partners to customize models and extend them.

Provision of resources is despite direct data or services in the field of models, methods and algorithms to compute the resources.

6.3.3. Success Criteria

Comparable and combinable models which may fuel services and applications within the LT stack have to be well documented, accessible, adaptable and easy to process.

6.4. Integrators

6.4.1. Actors

Software and consultancy companies which include LT into an overall solution serving companies in BI, Value Chain Management, Customer Relations, Marketing or general Knowledge Management.

6.4.2. Narrative

LT focused solution providers (re)-combinate resources of the Prêt-à-LLOD stack to satisfy customer demands. Whereas those integrators have a good knowledge of the customers domain in specific but do not have in-depth knowledge of the inner life of the LT components they are compiling. Reliable and solid components are key for solutions which consist of several components chained together.

Typical solutions are Search, Recommendation, Chatbots, Knowledge Management and HR in fields like Pharma, Education, Media, Finance, Automotive Industry, etc.

6.4.3. Success Criteria

Reliable and solid components are key for solutions which consist of several components chained together.

7. Annex

7.1. Specific Business User Stories - Chatbot Improving Access to HSE Services

Within Prêt-à-LLOD specific Use Cases are defined (see section 6). The further user stories coming directly and specific from Prêt-à-LLOD partners, imply a more detailed description and on-the-point requirements. Even these use cases are not generally representative, so they provide another view on an specific angle of the business dtories pointed out in Chapter 6.

7.1.1. Actors

- General users wishing to access HSE (Irish Health Service) benefits, schemes and allowances.
- Social Worker or Benefits Officer wanting to explain to a member of public what is available from the HSE.

7.1.2. Narrative

The HSE (Irish Health Service) currently operates a manual 'HSELive' service, providing help to the Irish public in navigating the Irish public health system. They would like to enhance the service by means of a chatbot. This chatbot will in particular improve access to the Irish Health Service's schemes and allowances programme (https://www2.hse.ie/costs-schemes-allowances/). Prêt-à-LLOD tools will be used to support the chatbot in interpreting the user's questions into plain language. The chatbot will be an extension to Derilinx GovAssist https://chatbot.staging.derilinx.com/ and may incorporate Open Data from https://data.ehealthireland.ie/group/pcrs.

7.1.3. Success Criteria

- New iteration of user-suggested questions and answers which are used to improve the model.
- User's question has been answered to their satisfaction.
- Some measure of the consistency of responses collated.

7.2. Post review Interviews

	Company A	Company B	Company C	Company D

How do you use language data in your pipeline? Is language data end-product or means to build a product?	* getting sources from the internet (free) and using tools provided by python libraries (such as lemmatizer, PoS tagger, etc.) * the data comes from web and they transform it, clean it, extract entities (i.e., date, country, location, organization), relationships (rdf graph) * To be provided by derilinx: text relationship extraction logic data is then stored in elastic search and rdf database subgraph	* Different for each tool. Chatbox is the most complex. * Technologies they use: AI, NLP; but no ML * rule based approaches Languages: they really support 15 but they state many more.	* train word-embedding models * they have different components each own with different requirements: annotations (sentiment analysis vs. emotional analysis vs. entity type analysis, relational detection, semantic relations), size/volume	* Is corpus consumed by all your products? NO * How will those products be improved by our corpus? A lot * training: ML, Ai many pieces --> 23 models per language (several models for the same linguistic task)
What types of language data are you currently using?	generic and domain (government, citizen information) specific	* monolingual corpora with semantic annotations, semantic relations, synonyms, related terms. * domain desired but not high priority * Ngrams not very informative. However, there is another team working with autocomplete that might be interested.	* user generated content e.g. social media client specific content (proprietor data) * patient reports * medical expert reports	* corpora * morphological lexicons * PoS tagging very important
What are the sources of the above data?	* Crawling government portal to extract citizen information (json) open data Irish, Spanish portals: download resources in csvs, json	* dictionaries, * wordnet, * list of synonyms, * real user corpora (since people communicate with chat boxes while making typos, grammar errors) * twitter bag of words * group based on actions: buy+synonyms, return+synonyms, cancel order + synonyms, etc. * For new languages they hire people to develop the sources for them	* all resources they are using are paid resources that crawl social media, blogs, forums (https://datasift.com/ https://scrapinghub.com/ , www.proxycrawl.com/crawling-api https://socialgist.com/ * few more that I did not get their name * delivery method: API	open source, crawling, acquired resources from public institutions

What problems are you experiencing with currently consumed data?	* Lack of conversational training data *Not enough data for domain classification	* scarce resources in many languages. * data not cleaned or annotated	* Rarely training data fit for purpose (domain, annotations) * expensively produce their own training data * stay away from supervised learning that is data demanding * Rarely training data in languages they are interested in * Data (user content) is not cleaned, not well formed (acronyms, abbreviation) * Crawling limitations: 100 sources per month for not more than 2 years.	* limited morphology applied at the tokens. * not rich morphology as nordic languages: difficulties to distinct (run) good lexicon * Poor quality of annotations, esp. in EN, which leads to POS ambiguity. They suffer from not having enough morphological information that helps them disambiguate across POS tags. e.g., 'run' infinitive vs. present tense vs. noun
What languages are relevant for you now, and what languages would you see relevant in future?	* 1st phase - Irish * 2nd phase - English, Spanish * 3rd phase - German, French * 4rth phase - rest of Europe	* languages: developing their tools for new languages can take up to 1 year!!. Developing a simple terminology database in a new language can take 1 month, 3-4 months to build language resources and then fine-tuning tools on that language. They would be very interested in acquiring monolingual corpora and if they like our resources they would be willing to collaborate whenever a customer asks for a new language.	* Future: expand by including the linking perspective; what is scientifically reported Vs, what is observed in clinical trials vs. what matters most by patients.	* english, nordic (5 languages), french, german * general language kid corpora could be interested
How would you like to see the data quality improving: e.g. what format? What type of annotations? What cleaning process?	#	e.g. what format? What type of annotations? What cleaning process? How would you like to see the data quality improving: Script Next Steps format: currently sql --> in future json annotations: PoS tagging, lemmatization, tokenization (only for some languages), stemming, related terms, semantic annotations, entity detection, off-topic conversation detection (e.g. Hello, Hate robots, etc.), detect location, concept tagging. This highly depends on use cases	most are answered already. need for minimum 10 years of data, annotated, semantically enriched, format: plain text, json for metadata; no interest in derived products, they can generate those on Next Steps their own cleaning: anonymization, remove markup from html,	Updated corpora (every six months to update models since updating) 10 B PoS, lemmatization as much distinction as possible (e.g. pronouns) rich morphology regional metadata on document level high quality: data in itself and annotations xml is preferred They showed interest in the possibility of identifying subcorpora by region General text (news, blogs) is already ok.

		cleaning process: detect structure of language: post/preposition, special characters		
Would domain (genre) specific data improve the outcomes/product?	Significantly.	low resource languages depending on customer request - no internal portfolio	Medical, pharmaceutical.	genres: news, blogs, domain: not interested
What volume is relevant?	once they start training they will get back to us	few thousands sentences work for them	* 2000 data points per class for sentiment model training * 100000 data points for emotion classifier * datapoint definition: depends on the task's granularity; in most cases it is synonym to sentence, but can be aspect inside sentence	Billions
What evaluation parameters are you using in order to assess data quality of the source?	they have an automated validator through which they run the input data in order to verify quality. did not understand much.	linguistic parameters.	* How easy can their application interact with the delivery method (API) What time ranges are covered, more better (10years+) * What linguistic parameters are they offering? * What data representation format is being used? (json)	(feedback on sample): tokens,