European Commission

Final study report on CEF Automated Translation value proposition in the context of the European LT market/ecosystem

FINAL REPORT

A study prepared for the European Commission

DG Communications Networks, Content & Technology by:

CROSSLANG TRANSLATION AUTOMATION

TILDE

ELDA

IDC ANALYZE THE FUTURE

Digital Single Market

This study was carried out for the European Commission by

Luc MEERTENS

Khalid CHOUKRI

Stefania AGUZZI

Andrejs VASILJEVS

Internal identification

Contract number: 2017/S 108-216374

SMART number: 2016/0103

DISCLAIMER

By the European Commission, Directorate-General of Communications Networks, Content & Technology.

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 Commission. The Commission does not guarantee the accuracy of the data included in this study. Neither the Commission nor any person acting on the Commission's behalf may be held responsible for the use which may be made of the information contained therein.

ISBN 978-92-76-00783-8

doi: 10.2759/142151

CONTENTS

Table of figures

List of tables

Definitions

In the present document, the following definitions apply:

- **CEF Automated Translation (CEF AT):** this is a building block referring to "machine-translation engines and specialised language resources including the necessary tools and programming interfaces needed to operate pan-European digital services in a multilingual environment. The Automated Translation building block is designed to serve any current or future CEF DSI requiring cross-lingual functionality." (SMART 2016/0103 tender specifications and CEF Telecom Regulation (EU) No 283/2014).[1]

- **eTranslation:** the machine translation service of the EC developed by DGT, used also by CEF AT.

[1] It should be noted that CEF Stakeholder Management Office (e.g. on CEF Digital website) refers to this building block as "CEF eTranslation".

Abbreviations

AI	Artificial Intelligence
ACL	Association for Computational Linguistics
ACM	Association for Computing Machinery
AMB	Architectural Management Board
AMTA	Association for Machine Translation in the Americas
API	Application Program Interface
ASR	Automated Speech Recognition
BDVA	Big Data Value Association
CAGR	Compound Annual Growth Rate
CAT	Computer-Aided Translation / Computer-Assisted Translation
CC	Creative Commons (license)
CEF	Connecting Europe Facility
CEF AT	CEF Automated Translation
CLEF	Conference and Labs of the Evaluation Forum
CLIR	Cross-language Information Retrieval
CNECT	EC's Directorate-General for Communications Networks, Content and Technology
CRM	Customer Relationship Management
CSA	Common Sense Advisory
DARIAH	Digital Research Infrastructure for the Arts and Humanities
DG DIGIT	EC's Directorate-General for Informatics
DGT	EC's Directorate-General Translation
DSI	Digital Service Infrastructure
DSM	Digital Single Market
EACL	European Chapter of the ACL
EAMT	European Association for Machine Translation
EC	European Commission
ECAS	EC's main authentication service
EEA	European Economic Area
EFTA	European Free Trade Association
ELDA	Evaluations and Language resources Distribution Agency
ELRA	European Language Resources Association
ELRC	European Language Resource Coordination
ELSNET	European Network in Language and Speech
ERP	Enterprise Resource Planning
EU28	Member States of the European Union
FP6	Sixth Framework Programme (2002-2006)
FP7	Seventh Framework Programme (2007-2013)
FTE	Full-time Equivalent
GALA	Globalization and Localization Association
GDP	Gross Domestic Product

GPL	GNU General Public License
GPU	Graphics Processing Unit
HLT	Human Language Technologies
HPC	High Performance Computing
HQ	Headquarters
HR	Human Resources
ICT	Information and Communication Technology
IEEE	Institute of Electrical and Electronics Engineers
IP	Internet Protocol
IPR	Intellectual Property Right
IR	Information Retrieval
ISCA	International Speech Communication Association
IT	Information Technology
IVR	Interactive Voice Response
JRC	EC’s Joint Research Centre
LDC	Linguistic Data Consortium
LISA	Localization Industry Standards Organization
LR	Language Resource
LSP	Language Service Provider
LSTM	Long Short-term Memory (neural network terminology)
LT	Language Technology(ies)
LT Innovate	Language Technology Industry Association
META-NET	Multilingual Europe Technology Alliance
MFF	Multiannual Financial Framework
ML	Machine Learning
MS	Member State
MT	Machine Translation
MT@EC	DGT’s MT service until 2013
NAACL	North American Chapter of the ACL
NAP	National Anchor Point (of ELRC)
NER	Named Entity Recognition
NIST	National Institute of Standards and Technology
NLP	Natural Language Processing
NLU	Natural Language Understanding
NMT	Neural Machine Translation
NRI	Network Readiness Index
OCR	Optical Character Recognition
OECD	Organisation for Economic Co-operation and Development
PCS	Public Cloud Software
PSI	Public Sector Information
QA	Question Answering
QE	Quality Estimation
RNN	Recurrent Neural Network
SaaS	Software as a Service

SCIC	Service for Conference and Interpretation
SME	Small or Medium-sized Enterprise
SMO	Stakeholder Management Office
SMT	Statistical Machine Translation
SWOT	Strengths, Weaknesses, Opportunities, and Threats (analysis)
TB	Terabyte
TM	Translation Memory
TMS	Translation Management System
TMX	Translation Memory eXchange
TTS	Text-to-Speech
UI	User Interface
VC	Venture Capital
VOIP	Voice over Internet Protocol
WER	Word Error Rate
XML	Extensible Markup Language

Abstract

This study positions CEF Automated Translation, a building block of the Connecting Europe Facility, in the European market for language technologies (LT). The study provides an analysis of the LT market in the EU (supply and demand), of LT adoption by public services in the EU, and of the EU's competitiveness with respect to the US and Asia in three LT areas, i.e. machine translation (MT), speech technology and cross-lingual search. Based on the results of the analyses, the study develops a business model for CEF AT by defining the latter's value proposition in the context of the market.

The analyses show that suppliers are often SMEs with local solutions, that public services have a strong interest for translation technology, and that the worldwide LT market, dominated by large players, has deficiencies regarding under-resourced languages, customisation needs, and security and privacy requirements. While CEF AT's current value proposition consists of providing a secured MT service to public administrations and of a language resource collection effort (ELRC-SHARE), the proposition may be extended in two ways, given the market deficiencies and CEF AT's mission as a multilingual enabler: a more elaborate MT offer or a broad LT offer, focusing on under-resourced languages and customisation while avoiding market distortion.

Résumé

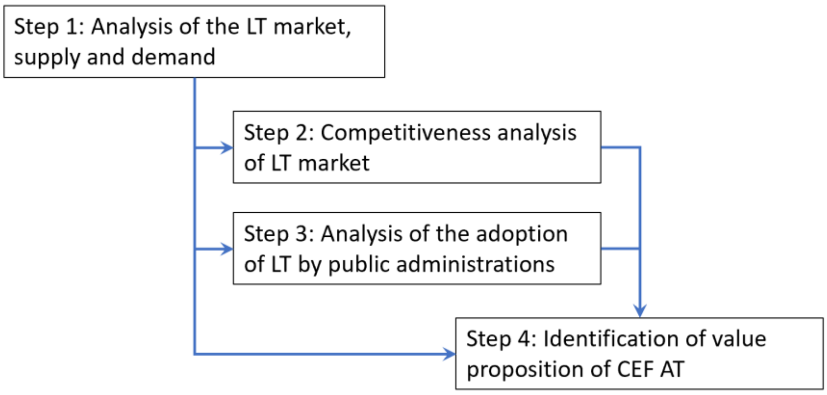
La présente étude positionne la plate-forme de traduction automatique du MIE (CEF Automated Translation), un élément constitutif du Mécanisme pour l'Interconnexion en Europe (Connecting Europe Facility), sur le marché européen des technologies du langage (TL). Elle présente une analyse du marché (offre et demande), de l'adoption des TL par les services publics européens, et de la compétitivité de l'UE par rapport aux États Unis et à l'Asie pour trois types de TL : la traduction automatique (TA), la technologie de la parole et la recherche multilingue. Basée sur les résultats des analyses, l'étude développe un modèle économique pour la CEF AT en définissant sa proposition de valeur dans le contexte du marché.

Les analyses montrent que la plupart des fournisseurs sont des PME offrant des solutions locales, que les services publics s'intéressent fortement aux technologies de traduction, et que le marché mondial des TL, dominé par de grandes entreprises, montre des défaillances quant aux langues pauvres en ressources, et aux besoins de customisation, de sécurité et de protection de données. En considérant ces défaillances et la mission de facilitateur multilingue de la CEF AT, sa proposition de valeur, qui consiste primordialement à fournir un service de TA sécurisé aux administrations publiques et d'une collecte de ressources linguistiques (ELRC-SHARE), pourrait être élargie par une offre de TA ou des TL plus élaborée et concentrée sur les langues pauvres en ressources et sur la customisation, mais en évitant une distorsion de marché.

Executive summary

This study positions CEF Automated Translation (CEF AT), a building block of the Connecting Europe Facility, in the European market for language technologies (LT) and develops a business model for CEF AT by defining the latter's value proposition in the context of this market. Furthermore, it suggests potential extensions of the current business model.

The methodology of the study consists of four steps. The first step is an analysis of the LT market of the EU (including Norway and Iceland) in terms of supply and demand. The second step is a competitiveness analysis of the LT market, leading to the identification of market deficiencies. In the third step, the adoption of LT by public administrations, both EU-level and national, is analysed. Finally, a value proposition is defined, the current business model of CEF AT is developed, and potential extensions of the latter are suggested. As shown in the below figure, the second and third step make use of the findings of the first step, while the fourth step builds upon all other steps. Each step corresponds to a task performed by one of the consortium members (Task 1 was performed by IDC, Task 2 by Tilde, Task 3 by ELDA and Task 4 by CrossLang). The consortium members were supported by two contractors (ILSP and DFKI).


Step 1: Analysis of the LT market, supply and demand
Step 2: Competitiveness analysis of LT market
Step 3: Analysis of the adoption of LT by public administrations
Step 4: Identification of value proposition of CEF AT

Step 1: Analysis of the LT market

The analysis has the objective of providing a complete overview of the European LT market together with a description of the emerging trends and an estimate of the growth in the revenues. A two-fold approach was undertaken, i.e. a combination of preliminary desk research with primary research (through questionnaires and interviews).

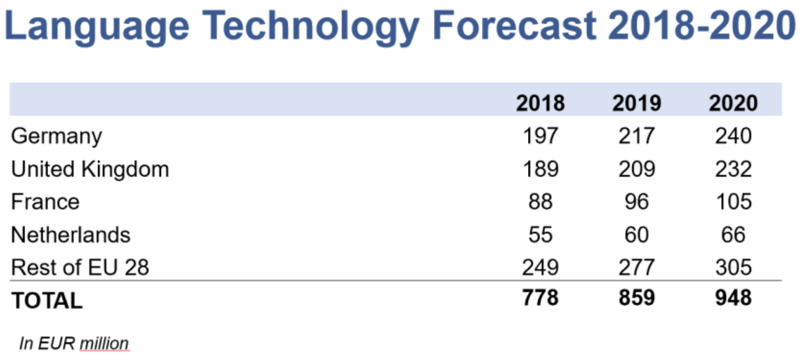
During the preliminary desk research, an exhaustive list of companies active in EU member states in the domain of LT was created; 473 of those fully qualify as LT vendors. Based on further desk research using public sources and in-house databases of the consortium member IDC, the total size of the LT industry within the EU26 plus Iceland and Norway in 2017 was estimated at approximately 800 million euro, which is a relatively small market in IT terms. Germany holds the largest share of the LT market followed by the UK. Forecasts predict this market to grow at an average rate of 10% between now and 2021.

The primary research consisted of analysing the responses of an online questionnaire and information provided through telephone interviews. The questions related to company profile, offering, revenue, state of play of the market etc. Invitations for filling out the questionnaire were sent out to top executives from companies that were identified during desk research. Based on the 51 responses and 8 subsequent telephone interviews, the consortium was able to get a picture of the market size, language offering, types of LT offered, customer segments, and perception of the future.

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.

In terms of language offering, it comes as no surprise that English, German, French, Spanish and Italian are of most importance to the LT vendors. As LT markets for most European languages are small, business opportunities are limited for vendors that focus on particular languages.

In terms of the types of LT offered, translation technology is considered as the biggest revenue contributor 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. The below table shows the countries with the largest market share.



Language Technology Forecast 2018-2020

	2018	2019	2020
Germany	197	217	240
United Kingdom	189	209	232
France	88	96	105
Netherlands	55	60	66
Rest of EU 28	249	277	305
TOTAL	**778**	**859**	**948**

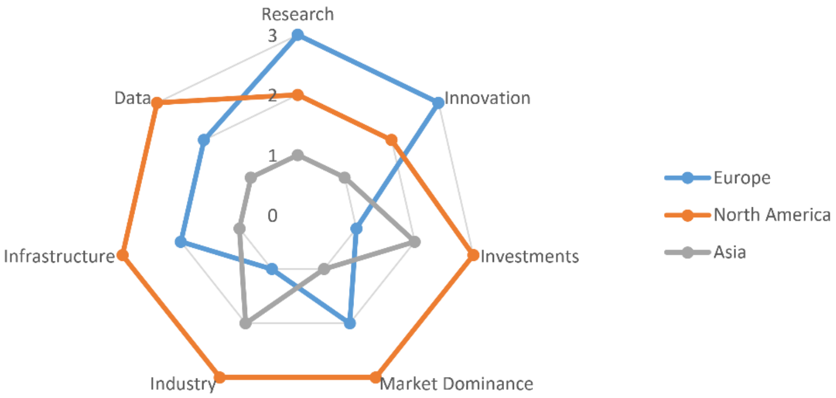
In EUR million

As for customer segments, vendors consider the public sector as the most important segment, though this sector accounts for only 20% of their revenues and lags behind the private sector in terms of profitability.

Most LT suppliers are quite positive when looking towards the future and expect the LT market to grow, as Artificial Intelligence will be increasingly part of LT. In that respect, Natural Language Understanding (NLU) in general and chatbot applications in particular were often mentioned as the emerging technology to look out for and are expected to become increasingly widespread.

Step 2: Competitiveness analysis

In order to investigate competitiveness in the LT market, three LT areas were selected, i.e. machine translation (MT), speech technology and cross-lingual search, and three regions, i.e. the EU, the US and Asia. This analysis made use of sources such as the findings of the market analysis performed in Step 1. For each LT type, the regions were compared across seven dimensions, i.e. Research, Innovation, Investments, Market Dominance, Industry, Infrastructure and Data. The below figure illustrates the comparison for the area of MT; a region's position for a dimension becomes stronger at it moves towards the outer lines. The results from the comparison using dimensions resulted in a SWOT analysis for the EU, involving strengths, weaknesses, opportunities and threats.



Comparative position of European machine translation market versus North America and Asia regions

As for weaknesses and threats, the EU industry is fragmented with many small players struggling to find a place in the market in order to compete with the global players, which dominate the market and upon which European businesses and public sector have become dependent. While the research position of the EU in the three areas is weakening, the global US players have a large competitive advantage in terms of research capacities, computing resources and available data. They distort the market, for instance by providing a free MT service, though MT is not their core business. They also have larger amounts of data at their disposal, because of copyright disparities between the EU (requirement of explicit permission by European entities) and the US (fair use copyright exception), and because the intensive use of their systems allows them to collect many user data. While EU industry is experienced with small and complex languages, they involve a limited market and restricted business opportunities, and the amount of accessible data for these languages is low.

As for strengths and opportunities, European MT developers have been successful in deploying services for the public sector through the support of EU-funded programmes. In the area of speech, the EU has demonstrated successful experience in multilingual infrastructure building projects which aim at reducing digital linguistic fragmentation across the EU. In the market for MT, speech technology and cross-lingual search as a whole, three deficiencies can be observed, providing opportunities for the EU:

1. There are gaps in the offering for small and complex languages. EU developers have built strong experience for such languages thanks to the multilingual market. While they provide limited business opportunities, as stated earlier, and the quality gap with the larger languages focused upon by global players increases, support for small and complex languages is an essential means to preserve cultural identity, to foster inclusiveness, and to guarantee equal digital opportunities

across languages, thereby supporting a key principle of the EU, language equality. Support for small and complex languages is also important in shaping a Digital Single Market.

2. There is lack of domain-specific and application-specific MT. Due to the need to search for niche markets, many European developers have accumulated strong experience in customised and domain-specific solution development in the areas of MT and speech. This experience could be helpful to meet the lack of customised systems.

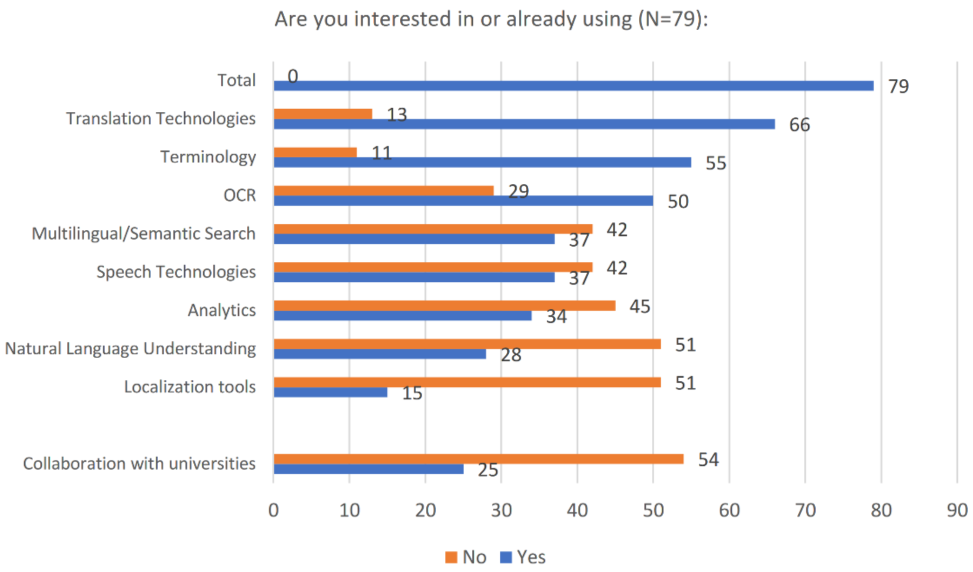
3. US global players pay little attention to security and privacy. As for privacy, the EU has well-established practices for the creation of open data and policies fostering public data sharing.

Step 3: Analysis of LT adoption by public administrations

In order to investigate the current adoption of LT services and solutions by public EU-level or national administrations and to analyse plans for LT adoption in the next few years, an online questionnaire was set up. Questions related to the population(s) served by an administration, the latter's level of interest in LT or use of specific types of LT, etc. The LT types mentioned in the questionnaire included those used in the market analysis questionnaire of Step 1.

Higher management staff in a total of 606 organisations in a wide range of domains, from social security to domestic affairs and utility services, was invited to fill out the questionnaire; the list of contacts was compiled from the lists of participants of workshops organised by the European Language Resource Coordination (ELRC); the consortium member ELDA is responsible for a group of countries within this organisation, which manages, maintains and coordinates resources in a large range of languages.

From the 79 complete responses to the online questionnaire, the following conclusions can be drawn. MT is clearly the type of LT most frequently used by public administrations. There is also a strong interest in related tools, like translation memories and terminology management systems. In that respect, it does not come as a surprise that many respondents are interested in knowing more about eTranslation and the ELRC. The figure below shows the number of administrations that use specific types of LT or have an interest in doing so.



Although many of the respondents were optimistic about their future needs for 2020 and beyond to deploy other LT types than MT, it seems that these technologies are not considered mature enough today to be used in their working environment. In that respect, technologies like optical character recognition (OCR) and speech technology (primarily speech recognition) appear to be on many administrations' radar for implementation, but adoption of these LT types remains rather low today.

In terms of vendors, EU-based players are often cited when referring to specific applications like translation management systems or translation memories. In most domains, however, major players – when cited – are predominantly US-based.

Collaboration between public administrations and academia appears to be strong, involving a third of the respondents, and mostly local or national universities. This high level of collaboration points towards a need for customisation and tuning of technologies.

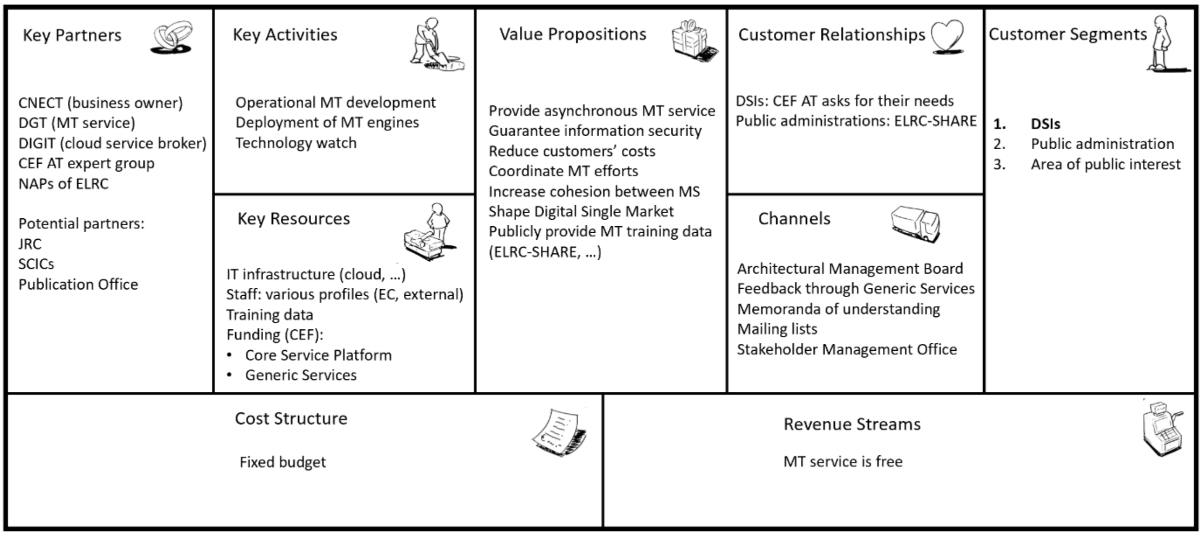
Step 4: Value proposition of CEF AT

As a final step in the study, a business model for CEF AT was developed, by defining the value proposition of the latter in the context of the LT market, and potential extensions to the model were suggested. A business model is the rationale of how an organisation creates, delivers and captures value. While such a model typically applies to companies, it can also be used for other organisation types, like public administrations, by taking care of specific constraints (policy-related, financial and operational). CEF AT's business model was developed using the business canvas approach, which describes a model through nine basic blocks, such as Customer segments, Key resources, and Value Proposition. These blocks cover the main areas of a business and are described by answering a set of questions.

The questions related to basic blocks in the business canvas approach were answered based on a number of information sources: EU policies related to multilingualism and LT, the findings of Steps 1 to 3, information on the current CEF AT implementation, and meetings of the consortium member

CrossLang with CEF AT staff. An additional source of information originated from the meetings of CrossLang with Digital Service Infrastructures (DSIs) in the framework of Smart 2016/0103 Lot 2.

The business model developed from the above information sources is shown in the below figure. It clarifies CEF AT's current value proposition, which is focused, on the one hand, on eTranslation, an asynchronous secured MT service which is offered to the DSIs for the multilingual deployment of their services and guarantees information security, and, on the other hand, on ELRC-SHARE, a language resource collection effort through which CEF AT publicly provides MT training data. CEF AT's main partners are the MT team at DG Translation, which deploys the eTranslation service, and DG DIGIT, which acts as a cloud service broker. CEF AT's activities are geared towards operational development and deployment. Its prime customers are DSIs, but it also serves public administrations and the area of public interest. It operates on a fixed budget.



Key Partners

CNECT (business owner)
DGT (MT service)
DIGIT (cloud service broker)
CEF AT expert group
NAPs of ELRC

Potential partners:
JRC
SCICs
Publication Office

Key Activities

Operational MT development
Deployment of MT engines
Technology watch

Key Resources

IT infrastructure (cloud, …)
Staff: various profiles (EC, external)
Training data
Funding (CEF):
- Core Service Platform
- Generic Services

Value Propositions

Provide asynchronous MT service
Guarantee information security
Reduce customers' costs
Coordinate MT efforts
Increase cohesion between MS
Shape Digital Single Market
Publicly provide MT training data (ELRC-SHARE, …)

Customer Relationships

DSIs: CEF AT asks for their needs
Public administrations: ELRC-SHARE

Channels

Architectural Management Board
Feedback through Generic Services
Memoranda of understanding
Mailing lists
Stakeholder Management Office

Customer Segments

1. **DSIs**
2. Public administration
3. Area of public interest

Cost Structure

Fixed budget

Revenue Streams

MT service is free

The above information sources also led the consortium to suggest two potential future business models to CEF AT: the *MT Business Model* extends the scale of the service and makes CEF AT an instrument facilitating the customisation of MT, the *LT Business Model* goes beyond MT and also involves customisation of LT in a broader sense and the provision of LT resources beyond MT training data. It should be stressed that these business models merely express an opinion of the consortium.

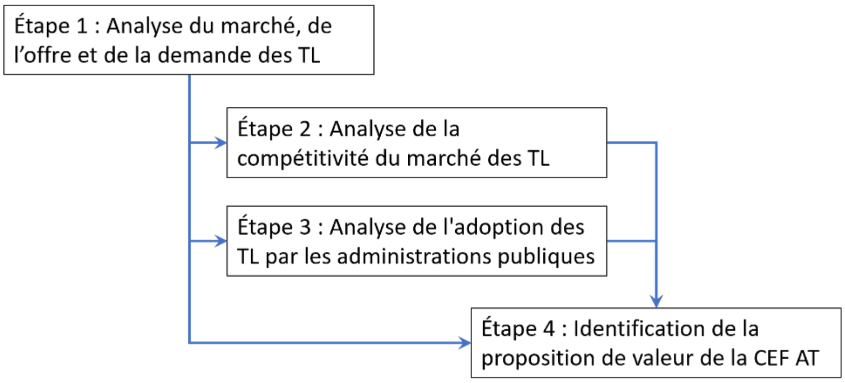
The MT Business Model is motivated by the fact that an increase in eTranslation demand is likely given the rising interest from DSIs and by the interest from public administrations in MT (see Step 3). In this model, real-time translation is provided, based on DSIs' interest in chat translation and on the speed expectancies shaped by the online service of global MT players. Customisation of MT allows eTranslation to distinguish itself from the service of global players on the level of domain adaptation, security and under-resourced languages, thus taking into account market deficiencies (see Step 2). Customisation takes place through projects involving specialised companies in order to avoid market distortion and focuses on valorisation (business aspects) and reusability of results. The MT Business Model has clear implications on the level of physical and financial resources, cost structure and revenue streams.

The LT Business Model is motivated by the fact that DSIs not only show a vivid interest in MT, but also in LT in general. The interest is not at the same level in public administrations, as they do not consider LT as mature enough, thus not ready to be invested in. However, the possibility to customise LT components in collaboration with specialised companies instead of using off-the-shelf tools to create such components could be a strong motivation for public administrations to start using LT tools. In that respect, the supplier database containing 473 LT companies can be a valuable asset for public services.

Sommaire exécutif

Cette étude positionne la CEF Automated Translation (CEF AT), un élément constitutif de la Connecting Europe Facility, sur le marché européen des technologies langagières (TL) et développe un modèle économique pour la CEF AT en définissant la proposition de valeur de cette dernière dans le cadre de ce marché. En outre, elle suggère des extensions potentielles du modèle économique actuel.

La méthodologie de l'étude comporte quatre étapes. La première étape consiste en une analyse du marché des TL de l'UE (y compris la Norvège et l'Islande) en termes d'offre et de demande. La deuxième étape consiste en une analyse de la compétitivité du marché des TL, conduisant à l'identification des déficiences du marché. Dans la troisième étape est analysée, l'adoption des TL par les administrations publiques, tant au niveau européen que national. Enfin, une proposition de valeur est définie, le modèle économique actuel de la CEF AT est développé et des extensions possibles de ce dernier sont proposées. Comme le montre le schéma ci-dessous, les deuxième et troisième étapes utilisent les résultats de la première, tandis que la quatrième s'appuie sur toutes les autres. Chaque étape correspond à une tâche effectuée par l'un des membres du consortium (la tâche 1 a été effectuée par IDC, la deuxième par Tilde, la troisième par ELDA et la quatrième par CrossLang). Les membres du consortium étaient assistés par deux prestataires (ILSP et DFKI).


Étape 1 : Analyse du marché, de l'offre et de la demande des TL

Étape 2 : Analyse de la compétitivité du marché des TL

Étape 3 : Analyse de l'adoption des TL par les administrations publiques

Étape 4 : Identification de la proposition de valeur de la CEF AT

Étape 1 : Analyse du marché des TL

L'analyse a pour objectif de fournir une vue d'ensemble complète du marché européen des TL, ainsi qu'une description des tendances émergentes et une estimation de la croissance des revenus. Une double approche a été adoptée, c'est-à-dire une combinaison de recherche documentaire préliminaire et de recherche primaire (au moyen de questionnaires et d'entrevues).

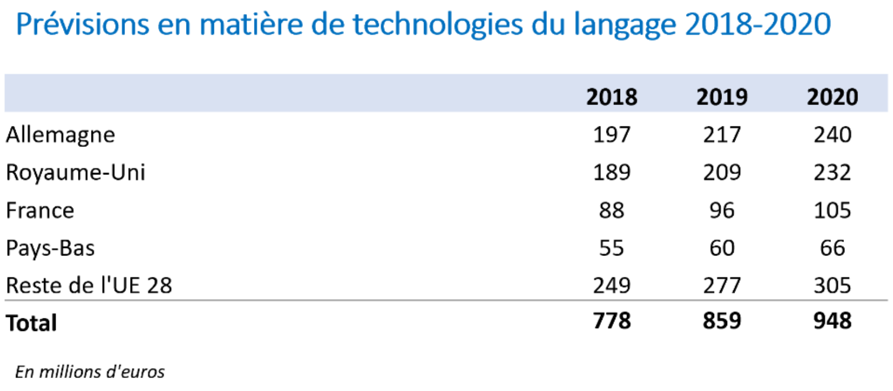
Au cours de la recherche documentaire préliminaire, une liste exhaustive d'entreprises actives dans le domaine des TL, dans les États membres de l'UE a été dressée ; 473 d'entre elles sont pleinement qualifiées en tant que fournisseurs de TL. Sur la base d'autres recherches documentaires utilisant des sources publiques et des bases de données internes d'IDC, membre du consortium, la taille totale de l'industrie des TL dans l'UE26 plus l'Islande et la Norvège en 2017, a été estimée à environ 800 millions d'euros, soit un marché relativement restreint en termes informatiques. L'Allemagne détient la plus grande part du marché des TL, suivie du Royaume-Uni. Selon les prévisions, ce marché devrait croître à un taux moyen de 10 % d'ici 2021.

La recherche primaire consistait à analyser les réponses d'un questionnaire en ligne et les informations fournies par le biais d'entretiens téléphoniques. Les questions étaient relatives au profil de l'entreprise, à l'offre, au chiffre d'affaires, à l'état du marché, etc. Des invitations à remplir le questionnaire ont été envoyées à des cadres supérieurs d'entreprises identifiées au cours d'une recherche documentaire. En se basant sur les 51 réponses et les 8 entrevues téléphoniques qui ont suivi, le consortium a été en mesure de se faire une idée de la taille du marché, de l'offre linguistique, des types de TL offerts, des segments de clientèle et de la perception de l'avenir.

Le marché des TL en Europe est très fragmenté et composé de petites et moyennes entreprises (PME), qui sont généralement des acteurs locaux apportant des solutions locales. La rentabilité est assez faible, la concurrence intense et les marges comprimées. L'UE ne bénéficie pas d'un acteur mondial unique et de premier plan. L'une des principales raisons de cette faible rentabilité globale des fournisseurs est la nécessité de continuer à innover, et les coûts liés à ce besoin.

En termes d'offre linguistique, il n'est pas surprenant que l'anglais, l'allemand, le français, l'espagnol et l'italien soient les langues les plus importantes pour les fournisseurs des TL. Étant donné que les marchés de TL sont restreints pour la plupart des langues européennes, les opportunités commerciales sont limitées pour les fournisseurs qui se concentrent sur des langues spécifiques.

En ce qui concerne les types de TL offerts, la technologie de traduction est considérée comme la plus importante source de revenus, suivie de la technologie vocale. Les technologies de recherche multilingue et sémantique sont les moins importantes en termes de revenus. Les répondants au sondage étaient très satisfaits de l'amélioration récente de la qualité de la traduction automatique. Le tableau ci-dessous indique les pays qui détiennent la plus grande part de marché.



Prévisions en matière de technologies du langage 2018-2020

	2018	2019	2020
Allemagne	197	217	240
Royaume-Uni	189	209	232
France	88	96	105
Pays-Bas	55	60	66
Reste de l'UE 28	249	277	305
Total	**778**	**859**	**948**

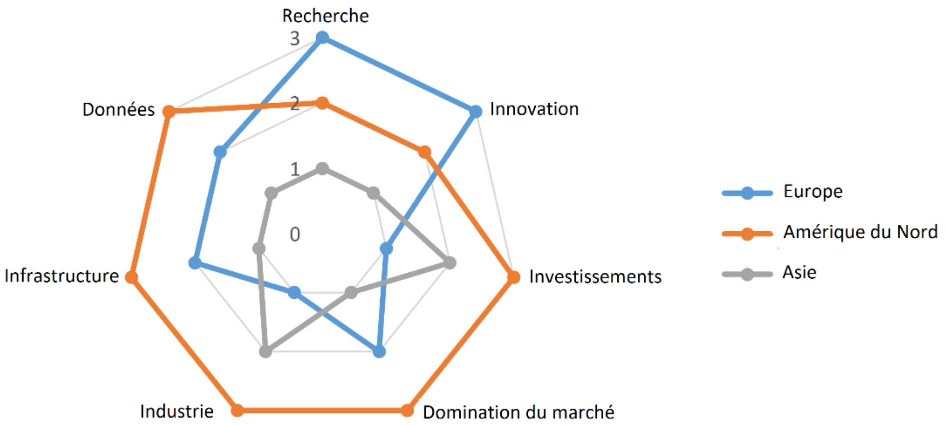
En millions d'euros

Concernant les segments de clientèle, les fournisseurs considèrent le secteur public comme le plus important, bien que ce secteur ne représente que 20 % de leurs revenus et accuse un retard par rapport au secteur privé en termes de rentabilité.

La plupart des fournisseurs des TL sont très positifs face à l'avenir et s'attendent à ce que le marché des TL se développe, car l'intelligence artificielle fera de plus en plus partie des TL. À cet égard, la compréhension du langage naturel (NLU) en général et les applications de chatbot en particulier ont souvent été mentionnées comme les technologies émergentes à surveiller et devraient devenir de plus en plus répandues.

Étape 2 : Analyse de la compétitivité

Afin d'étudier la compétitivité sur le marché des TL, trois domaines de TL ont été sélectionnés, à savoir la traduction automatique (TA), la technologie vocale et la recherche multilingue, et trois régions, à savoir l'UE, les États-Unis et l'Asie. Cette analyse a fait appel à des sources telles que les résultats de l'analyse de marché réalisée à l'étape 1. Pour chaque type de TL, les régions ont été comparées selon sept dimensions : recherche, innovation, investissements, domination du marché, industrie, infrastructure et données. La figure ci-dessous illustre la comparaison pour l'aire de la TA ; la position d'une région pour une dimension devient plus forte lorsqu'elle se déplace vers les lignes extérieures. Les résultats de la comparaison à l'aide des dimensions ont donné lieu à une analyse SWOT pour l'UE, comprenant les forces, les faiblesses, les opportunités et les menaces.



Position du marché de traduction automatique européen comparée à l'Amérique du Nord et l'Asie

En ce qui concerne les faiblesses et les menaces, l'industrie de l'UE est morcelée en ce sens que de nombreux petits acteurs luttent pour trouver une place sur le marché afin de concurrencer les acteurs mondiaux, qui dominent le marché et dont les entreprises et le secteur public européens dépendent désormais. Alors que la position de l'UE en matière de recherche dans ces trois domaines s'affaiblit, les acteurs mondiaux américains disposent d'un avantage concurrentiel important en termes de capacités de recherche, de ressources informatiques et de données disponibles. Elles faussent le marché, par exemple en fournissant un service de TA gratuit, bien que celle-ci ne soit pas leur activité principale. Ils disposent également de plus grandes quantités de données, en raison des disparités en matière de droits d'auteur entre l'UE (exigence d'une autorisation explicite des entités européennes) et les États-Unis (exception du droit d'auteur pour utilisation équitable), et parce que l'utilisation intensive de leurs systèmes leur permet de collecter de nombreuses données sur les utilisateurs. Bien que l'industrie de l'UE ait de l'expérience avec les langues peu répandues et complexes, elles impliquent un marché restreint et des opportunités commerciales limitées, et la quantité de données accessibles pour ces langues est faible.

En ce qui concerne les points forts et les opportunités, les développeurs européens de TA ont réussi à déployer des services pour le secteur public grâce au soutien de programmes financés par l'UE. Dans le domaine de la parole, l'UE a fait la preuve de son expérience réussie en matière de projets de construction d'infrastructures multilingues visant à réduire la fragmentation linguistique numérique dans l'UE. Sur l'ensemble du marché de la traduction automatique, des technologies vocales et de la recherche multilingue, trois lacunes peuvent être observées, ce qui offre des possibilités à l'UE :

1. Des carences existent dans l'offre pour les langues peu répandues et complexes. Les développeurs de l'UE ont acquis une solide expérience pour ces langues grâce au marché multilingue. Bien qu'elles offrent des opportunités commerciales limitées, comme indiqué plus haut, et que l'écart de qualité avec les langues les plus répandues sur lesquelles se concentrent les acteurs mondiaux se creuse, le soutien aux langues peu répandues et complexes est un moyen essentiel de préserver l'identité culturelle, de favoriser l'intégration et de garantir l'égalité des chances numériques entre langues, soutenant ainsi un principe essentiel de l'UE, l'égalité linguistique. La prise en charge des langues peu répandues et complexes est également importante pour la mise en place d'un marché numérique unique.

2. Une lacune existe dans la traduction automatique spécifique par domaine et par application. En raison de la nécessité de rechercher des marchés de niche, de nombreux développeurs européens ont accumulé une solide expérience dans le développement de solutions customisées et spécifiques par domaine, dans les secteurs de la TA et de la parole. Cette expérience pourrait être utile pour combler le manque de systèmes customisés.

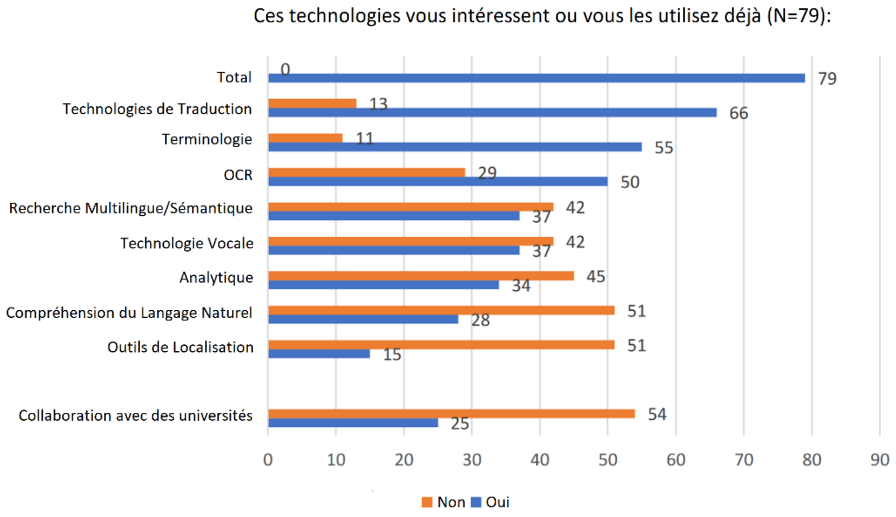
3. Les acteurs mondiaux américains accordent peu d'attention à la sécurité et à la protection de la vie privée. En ce qui concerne la protection de la vie privée, l'UE a des pratiques bien établies en matière de création de données ouvertes et de politiques favorisant le partage des données publiques.

Étape 3 : Analyse de l'adoption des TL par les administrations publiques

Un questionnaire en ligne a été mis en place afin d'étudier l'adoption actuelle des services et solutions de TL par les administrations publiques européennes ou nationales et d'analyser les projets d'adoption des TL dans les années à venir. Les questions se rapportaient à la ou aux populations desservies par une administration, au niveau d'intérêt de cette dernière pour les TL ou à l'utilisation de types spécifiques de TL, etc. Les types de TL mentionnés dans le questionnaire comprenaient ceux utilisés dans le questionnaire d'analyse de marché de l'étape 1.

Les cadres supérieurs de 606 organisations au total dans un large éventail de domaines, allant de la sécurité sociale aux affaires intérieures et aux services publics, ont été invités à remplir le questionnaire ; la liste des contacts a été établie à partir des listes des participants aux ateliers organisés par l'European Language Resource Coordination (ELRC) ; ELDA, membre du consortium, est responsable d'un groupe de pays dans cette organisation, qui gère, maintient et coordonne des ressources dans une grande variété de langues.

Les 79 réponses complètes au questionnaire en ligne permettent de tirer les conclusions suivantes. La TA est clairement le type de TL le plus fréquemment utilisé par les administrations publiques. Les outils connexes, comme les mémoires de traduction et les systèmes de gestion terminologique, suscitent également un vif intérêt. À cet égard, il n'est pas surprenant que de nombreux répondants soient curieux d'en savoir plus sur eTranslation et l'ELRC. La figure ci-dessous montre le nombre d'administrations qui utilisent des types spécifiques de TL ou qui ont un intérêt à le faire.



Bien que de nombreux répondants soient optimistes quant à leurs besoins futurs pour 2020 et au-delà pour déployer d'autres types de TL que la TA, il semble que ces technologies ne soient pas considérées comme suffisamment matures aujourd'hui pour être utilisées dans leur environnement professionnel. À cet égard, des technologies telles que la reconnaissance optique de caractères (OCR) et la technologie vocale (principalement la reconnaissance de la parole) semblent être la préoccupation de nombreuses administrations pour leur mise en œuvre, mais l'adoption de ces types de TL reste assez faible aujourd'hui.

En ce qui concerne les fournisseurs, les acteurs basés dans l'UE sont souvent cités lorsqu'il s'agit d'applications spécifiques telles que les systèmes de gestion de traduction ou les mémoires de traduction. Dans la plupart des domaines, cependant, les principaux acteurs - lorsqu'ils sont cités - sont principalement basés aux États-Unis.

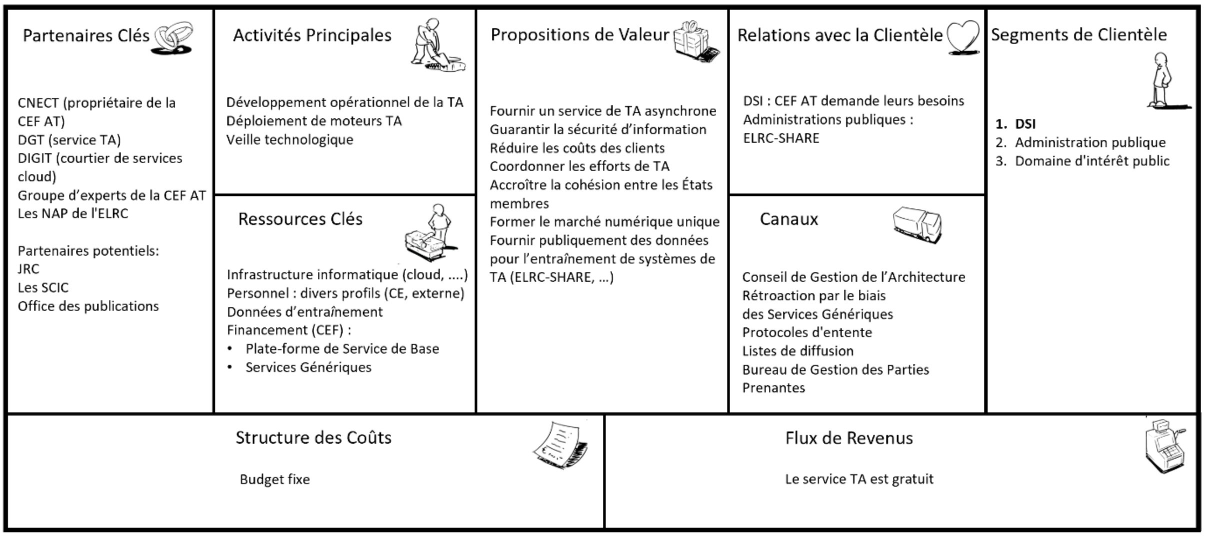
La collaboration entre les administrations publiques et le monde universitaire semble être forte, avec la participation d'un tiers des répondants, principalement des universités locales ou nationales. Ce haut niveau de collaboration indique un besoin de customisation et de mise au point des technologies.

Étape 4 : Proposition de valeur de la CEF AT

Comme dernière étape de l'étude, un modèle d'affaires pour la CEF AT a été développé, en définissant la proposition de valeur de ce dernier dans le contexte du marché des TL, et des extensions potentielles au modèle ont été suggérées. Un modèle économique est la justification de la façon dont une organisation crée, fournit et capture de la valeur. Si un tel modèle s'applique généralement aux entreprises, il peut également être utilisé pour d'autres types d'organisations, comme les administrations publiques, en tenant compte de contraintes spécifiques (politiques, financières et opérationnelles). Le modèle économique de la CEF AT a été développé en utilisant l'approche de canevas économique, qui décrit un modèle à travers neuf blocs de base, tels que les segments clients, les ressources clés et la proposition de valeur. Ces blocs couvrent les principaux domaines d'une entreprise et sont décrits en répondant à une série de questions.

Les réponses aux questions relatives aux blocs de base de l'approche de canevas économique ont été fournies sur la base d'un certain nombre de sources d'information : les politiques de l'UE en matière de multilinguisme et des TL, les conclusions des étapes 1 à 3, des informations sur la mise en œuvre actuelle de la CEF AT, et les réunions de CrossLang, membre du consortium, avec le personnel de la CEF AT. Une source d'information supplémentaire provient des réunions de CrossLang avec les infrastructures de service numérique (DSI, Digital Service Infrastructure) dans le cadre de Smart 2016/0103 Lot 2.

Le modèle économique élaboré à partir des sources d'information ci-dessus est illustré dans la figure ci-dessous. Il clarifie la proposition de valeur actuelle de la CEF AT, qui se concentre d'une part, sur l'eTranslation, un service de traduction automatique asynchrone sécurisé qui est offert aux DSI pour le déploiement multilingue de leurs services et guarantit la sécurité d'information, et d'autre part, sur ELRC-SHARE, une collecte de ressources linguistiques à travers laquelle la CEF AT offre des données publiques pour l'entraînement de systèmes de TA. Les principaux partenaires de la CEF AT sont l'équipe TA de la DG Traduction, qui déploie le service eTranslation, et la DG DIGIT, qui agit en tant que courtier cloud. Les activités de la CEF AT sont orientées vers le développement et le déploiement opérationnel. Ses principaux clients sont les DSI, mais elle sert également les administrations publiques et le domaine d'intérêt public. Il fonctionne avec un budget fixe.



Partenaires Clés

CNECT (propriétaire de la CEF AT)
DGT (service TA)
DIGIT (courtier de services cloud)
Groupe d'experts de la CEF AT
Les NAP de l'ELRC

Partenaires potentiels:
JRC
Les SCIC
Office des publications

Activités Principales

Développement opérationnel de la TA
Déploiement de moteurs TA
Veille technologique

Ressources Clés

Infrastructure informatique (cloud, ...)
Personnel : divers profils (CE, externe)
Données d'entraînement
Financement (CEF) :
- Plate-forme de Service de Base
- Services Génériques

Propositions de Valeur

Fournir un service de TA asynchrone
Guarantir la sécurité d'information
Réduire les coûts des clients
Coordonner les efforts de TA
Accroître la cohésion entre les États membres
Former le marché numérique unique
Fournir publiquement des données pour l'entraînement de systèmes de TA (ELRC-SHARE, ...)

Relations avec la Clientèle

DSI : CEF AT demande leurs besoins
Administrations publiques :
ELRC-SHARE

Canaux

Conseil de Gestion de l'Architecture
Rétroaction par le biais
des Services Génériques
Protocoles d'entente
Listes de diffusion
Bureau de Gestion des Parties Prenantes

Segments de Clientèle

1. **DSI**
2. Administration publique
3. Domaine d'intérêt public

Structure des Coûts

Budget fixe

Flux de Revenus

Le service TA est gratuit

Les sources d'information susmentionnées ont également conduit le consortium à proposer deux futurs modèles économiques potentiels à la CEF AT : le *Modèle Économique de la TA* élargie l'échelle du service et fait de la CEF AT un instrument facilitant la customisation de la TA, le *Modèle Économique des TL* va au-delà de la TA et implique également une customisation des TL dans un sens plus large et la mise à disposition de ressources de TL au-delà de données pour l’entraînement de systèmes de TA. Il convient de souligner que ces modèles économiques n'expriment qu'une opinion du consortium.

Le Modèle Économique de la TA est motivé par le fait qu'une augmentation de la demande d'eTranslation est probable étant donné l'intérêt croissant des DSI et par l'intérêt des administrations publiques pour la TA (voir étape 3). Dans ce modèle, la traduction en temps réel est fournie, en fonction de l'intérêt des DSI pour la traduction en instantané et de la vitesse anticipée façonnée par le service en ligne des acteurs mondiaux de la traduction automatique. La customisation de la traduction automatique permet à eTranslation de se distinguer du service des acteurs mondiaux au niveau de l'adaptation du domaine, de la sécurité et des langues manquant de ressources, en tenant compte des insuffisances du marché (voir étape 2). La customisation s'effectue par le biais de projets impliquant des entreprises spécialisées afin d'éviter les distorsions du marché et se concentre sur la valorisation (aspects commerciaux) et la réutilisation des résultats. Le Modèle Économique de la TA a des implications claires sur le niveau des ressources physiques et financières, la structure des coûts et les flux de revenus.

Le Modèle Économique des TL est motivé par le fait que les DSI montrent non seulement un vif intérêt pour la TA, mais aussi pour les TL en général. L'intérêt n'est pas au même niveau dans les administrations publiques, car elles ne considèrent pas que les TL soient suffisamment matures, donc pas prêtes à être investies. Cependant, la possibilité de customiser les TL en collaboration avec des entreprises spécialisées au lieu d'utiliser des outils standard pour créer des composants de TL pourrait être une forte motivation pour les administrations publiques à commencer à utiliser les outils de TL. À cet égard, la base de données des fournisseurs qui contient 473 sociétés de TL, peut constituer un atout précieux pour les services publics.

1. Introduction

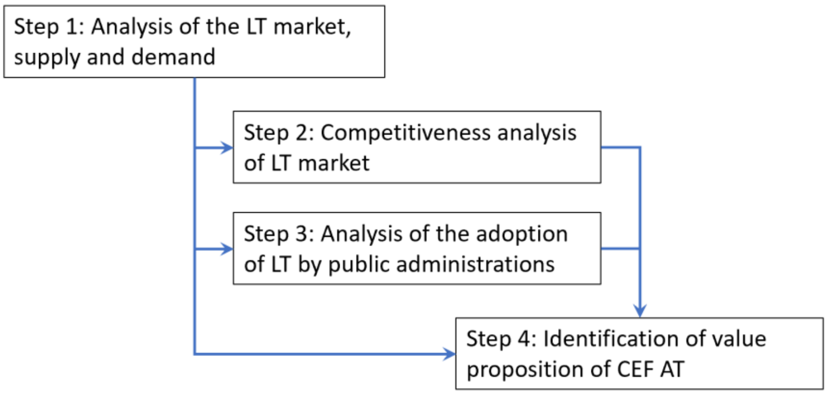
The Connecting Europe Facility (CEF) is a key EU funding instrument to promote growth, jobs and competitiveness through targeted infrastructure investment at European level. CEF Telecom is a key instrument that facilitates cross-border interaction between public administrations, businesses and citizens by deploying Digital Service Infrastructures (DSIs) and broadband networks. Some of these DSIs are building blocks, i.e. they belong to the set of generic and reusable DSIs and provide basic functionality, such as e.g. secure communication between IT infrastructures. Among these building blocks is CEF Automated Translation (CEF AT). Its mission is to provide multilingual support to DSIs so that individuals, administrations and companies in all countries of the European Economic Area (EU Member States, as well as Iceland, Liechtenstein and Norway) that participate in the CEF Telecom Work Programme can access public services in their own language.

The present study is the outcome of Lot 1 of the SMART 2016/0103 project, the objective of which is to position CEF AT in the European market for language technologies (LT) and to describe the building block's value proposition. The consortium of the project consists of the following organisations:

- CrossLang (full partner, project leader), a consulting and systems integration company (SME) specialised in translation automation technology;
- Tilde (full partner), an LT company specialising in the development of multilingual data technologies, such as machine translation (MT);
- ELDA (full partner, Evaluations and Language Resources Distribution Agency), a language resources broker;
- IDC (full partner, International Data Corporation), a global market intelligence, events, and advisory firm in the domain of ICT;
- ILSP (supporting subcontractor, Institute for Language and Speech Processing);
- DFKI (supporting subcontractor), the German Research Center for Artificial Intelligence.

In order to position CEF AT in the European LT market and determine its value proposition, the consortium applied a methodology consisting of four steps. The first step is an analysis of the LT market of the EU (including Norway and Iceland) in terms of supply and demand. The second step is a competitiveness analysis of the LT market, leading to the identification of market deficiencies. In the third step, the adoption of LT by public administrations, both EU-level and national, is analysed. Finally, the consortium established CEF AT's value proposition by defining the business model of CEF AT and suggests possible future extensions of the model to CEF AT. As shown in Figure 1, the second and third step make use of the findings of the first step, while the fourth step builds upon all other steps.

Figure 1 Steps of methodology


Step 1: Analysis of the LT market, supply and demand
Step 2: Competitiveness analysis of LT market
Step 3: Analysis of the adoption of LT by public administrations
Step 4: Identification of value proposition of CEF AT

Each step corresponds to a task performed by one of the consortium members (Task 1 was performed by IDC, Task 2 by Tilde, Task 3 by ELDA and Task 4 by CrossLang). Each step makes use of its specific methodology, although there are some commonalities across the steps. Details on these specific methodologies are provided in subsequent sections of this document and in the annexes. In the present section, we give a general overview of these methodologies and relations between the steps.

The first step has the objective of providing a complete overview of the European LT market together with a description of the emerging trends and an estimate of the growth in the revenues. A two-fold approach was undertaken, i.e. a combination of preliminary desk research with primary research. The first approach consists of the creation of an extensive LT supplier database and the use of public sources as well as an in-house database of IDC, while the second approach makes use of an online questionnaire and information provided through telephone interviews. Invitations for filling out the questionnaire or participating in an interview were sent out to companies and top executives identified during desk research. The LT supplier database created in this first step is not only useful in order to select potential respondents for an online questionnaire or potential interviewees, but also in the context of the fourth step, as it can serve as a source of information for public administrations in need of system customisation. Such customisation is part of the suggested future business models, as explained below.

The second step investigates the competitiveness in the LT market. Three LT areas were selected, i.e. MT, speech technology and cross-lingual search, and three regions, i.e. the EU, the US and Asia. The analysis made use of sources such as the findings of the market analysis performed in step 1. For each LT type, the regions were compared across seven dimensions, such as Research, Data and Innovation. The results from the comparison using dimensions resulted in a SWOT analysis for the EU, involving strengths, weaknesses, opportunities and threats. A SWOT analysis allows detecting market deficiencies, and hence opportunities for the EU.

In the third step, the current or potential future adoption of LT services and solutions by public EU-level or national administrations is investigated using an online questionnaire. The taxonomy of LT types used in the questionnaire includes the LT types used in the market analysis questionnaire of step 1. Invitations to fill out the questionnaire were sent out to higher management in a wide range of organisations, selected through participant lists of workshops in which ELDA is involved.

In the final step of the study, a business model for CEF AT was developed, by defining the value proposition of the latter in the context of the LT market. A business model is the rationale of how an organisation creates, delivers and captures value. While such a model typically applies to companies, it can also be used for other organisation types, like public administrations, by taking care of specific constraints (policy-related, financial, operational). CEF AT's model was developed using the business canvas approach, which describes a model through nine basic blocks, such as Customer Segments, Key Resources and Value Proposition, that cover the main areas of a business. The blocks were described by answering a set of questions based on several information sources, such as the findings of steps 1 to 3 and a meeting of CrossLang with CEF AT staff. In addition, the consortium used the information sources to suggest potential future extensions of the business model to CEF AT. These extensions include aspects like scaling up of the MT service and customisation of systems. It should be stressed that these extensions merely express an opinion of the consortium.

The remainder of this study is structured as follows. In Sections 2 to 5, the four steps sketched above are elaborated in detail. Each section starts with a summary or preface, provides information on the specific methodology used, describes the results, and ends with conclusions or recommendations. At the end of the study, final conclusions are provided, pertaining to the whole study, and annexes for the different tasks are included. The last annex reports on the presentation of the study at the 1st CEF eTranslation Conference in Brussels and on the subsequent panel discussion.

2. Task 1: LT market analysis

2.1. Summary

Task 1 of Smart 2016/0103 (“Analysis of the Language Technologies (LT) market at EU and Member State level, including Norway and Iceland”) led to the “Report on the analysis of the European Language Technologies market and possible shortcomings of the European LT market”, included below. The task was performed by IDC (the task leader) in collaboration with the consortium partners. It has the overarching objective to provide a complete overview of the European market of language technologies, a description of the emerging trends and a forecast estimate of the growth in the revenues of the key European players.

The bulk of the graphics and analysis presented in the report of Task 1 results from 51 online survey responses from top executives (typically CEO, President, Chairman) of larger language technology vendors operating in Europe, at the date of June 28, 2018. The potential targets of this survey are represented by 179 language technology vendors. The sample was qualified from an initial list of 1052 market players that was reduced to a group of 473 'interesting vendors'. The methodology adopted by the study team to select the appropriate sample of vendors representing the target of this survey exercise is presented in Annex A.

To gather insight into the LT market, IDC conducted 8 telephone interviews with key players providing LT services in Europe. These interviews involved discussions around the state of the current LT market and its recent developments. Some of the key areas covered during the telephone interviews also included competition, future strategic developments, industry-specific views, and accuracy of LT services. Telephone interviews were a crucial part of the market research study as they helped not only to promote a quality discussion with key leaders in the LT market, but also to check validity of data by analysing a research question from an extra perspective. This allowed the study to have thorough, accurate, and validated information on the LT market.

Overview of the Language Technologies Market in Europe

- The research results show that the market is dominated by US multi-national players who play a major role in Europe. Indigenous vendors are predominantly niche players serving local markets, among those, the largest vendor is SDL with annual European LT revenues of €13M.
- In terms of offering, the analytics is the most common product/service sold by the vendors in our sample, followed by natural language understanding technologies.
- However, the most relevant revenue source is represented by natural language understanding and translation technologies.
- In terms of delivery model, the vendors mainly rely on a mixed model of on-premise and cloud-based solutions.
- When we look at the revenue growth as well as at the profitability of these companies, data shows that a quarter of the vendor marketplace is not making notable revenues in their business.
- From a vertical market perspective, Government, Banking, Telecommunications and Professional Services are the primary industry markets targeted by LT vendors.

- Half of the considered sample is represented by small enterprises (between 10 and 99 employees).

Sizing and forecasting the Language Technologies Market in Europe

IDC predicts that the language technology market in the EU28 (plus Norway and Iceland) will grow from €706 million in 2017 to €1,040 million in 2021 at a compound annual growth rate (CAGR) of 9.8%. The data shows that the language technology market is growing significantly and will continue to expand over the next three to five years. Half of this market is represented by search technologies, followed by natural language understanding technologies, showing the highest growth rate (11% CAGR) to 2021 among the categories considered. From a country perspective, Germany holds the largest share of the LT market with a value of €179M in 2017 growing to nearly €270M in 2021, followed by the UK, that will grow to €255M in 2021. Government, Banking, Telecommunications and Professional Services represent the largest markets for LT technologies. However, although the public sector is seen as the most important market for LT vendors, only 20% of their revenues are sourced from this sector.

When we consider the languages for which LT services are provided, English, German, French, Spanish and Italian are of greatest importance to the vendors. In terms of market trends and based on the inputs collected by the vendors through the online survey, natural language processing (NLP) represents the key emerging trend in terms of adoption of LT, followed by text analytics and speech recognition.

In terms of marketplace innovation and new entrants, our survey data shows that a limited part of the vendors are interested in becoming part of a specialised language technology innovation lab or digital hub. Only 25% have done so. In addition, only 38% of the EU companies in the sample has external venture capital funding.

Our analysis also considered the market from a demand side perspective, with a particular focus on some vertical industries in which language technologies play a key role. The estimate took into account the spending for these products and services by the companies active in those markets.

IDC data shows that language technologies spending in the Healthcare industry accounts for €26M in 2018 and is expected to grow to €34M by 2021, showing a 2016-2021 CAGR of 9.8%. Machine translation and speech recognition are the most common applications in this market. The second market considered is Manufacturing. Language technologies spending in the Manufacturing industry accounts for €186M in 2018 and is expected to grow to €252M by 2021 showing a 2016-2021 CAGR of 10.4%. The most common and promising use cases are for the analysis of operational data, factory automation, and the analysis of online customer behaviour. Language technologies spending in the Telecom industry accounts for €39M in 2018 and is expected to grow to €53M by 2021 showing a 2016-2021 CAGR of 10.0%. The use of chatbots and conversational intelligent assistance for customer handling is one of the many use cases for the incorporation of cognitive computing and artificial intelligence technologies within organisations. Public sector is one of the biggest markets of adoption. IDC predicts that language technologies spending in the Government sector counts for €97M in 2018 and is expected to grow to € 126M by 2021 showing a 2016-2021 CAGR of 9.3%.

Finally, the analysis considered the Media sector. Language technologies spending in the Media industry accounts for €25M in 2018 and is expected to grow to €33M by 2021 showing a 2016-2021 CAGR of 9.4%. In this market space language technologies can play a role for example in automatic subtitling and speech recognition can be heavily deployed to convert spoken interviews into written form.

Key findings of the analysis

- **The LT market is very fragmented and composed by SMEs.** The LT market in the EU is very fragmented and there is a lack of large indigenous players. European players are all SMEs, where SDL is the largest. 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".** One of the positive effects of large players such as Google, Microsoft and Apple from the local vendors' point of view 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.

2.2. Introduction

Task 1 of Smart 2016/0103 ("Analysis of the Language Technologies (LT) market at EU and Member State level, including Norway and Iceland") led to the "Report on the analysis of the European Language Technologies market and possible shortcomings of the European LT market", included below. The task was performed by IDC (the task leader) in collaboration with the consortium partners. It has the overarching objective to provide a complete overview of the European market of language technologies, a description of the emerging trends and a forecast estimate of the growth in the revenues of the key European players.

The report of Task 1 is structured as follows:

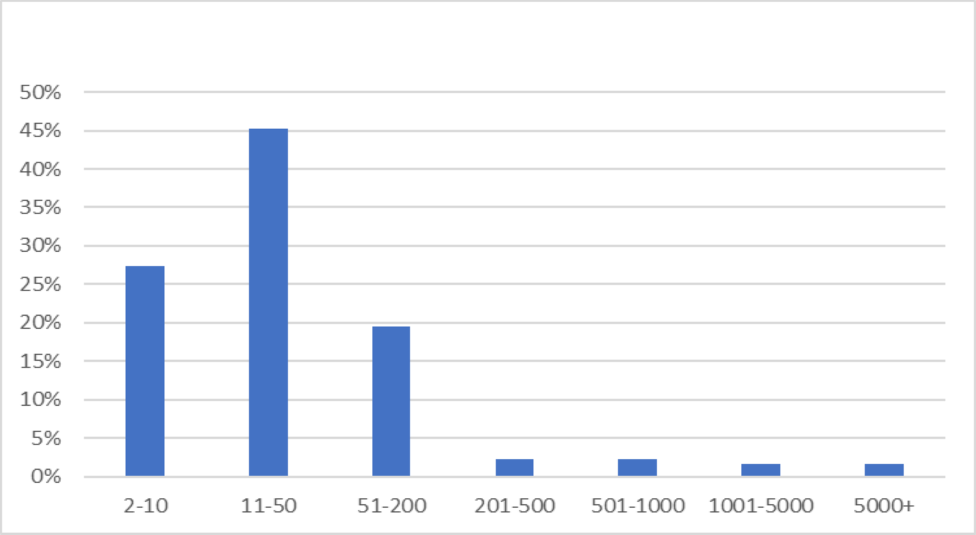
- Section 2.2 is the introduction to the task;
- Section 2.3 provides the key definitions adopted in this task and offers a qualitative overview of the language technology market in Europe providing an analysis of the main players;
- Section 2.4 is dedicated to the presentation of the language technology market sizing and forecast by providing data about the dynamics of the supply side and a qualitative analysis of the demand side. In addition, this section offers a description of the key vertical markets where the language technologies are mostly used, by presenting a number of key emerging use cases;
- Section 2.5 draws the conclusions of the analysis;
- Annex A is included to present the methodology designed and implemented in Task 1, including the research tools.

The bulk of the graphics and analysis presented in the report of Task 1 results from 51 online survey responses from top executives (typically CEO, President, Chairman) with larger language technology vendors operating in Europe. This top executive involvement provides a high degree of the reliability and accuracy of the data collected, making this a highly valuable piece of research.

The 51 respondents come from a group of 179 vendors that were selected from a long list of potential market players and targeted by the online survey.

This list of 179 vendors (our sample) reveals some interesting characteristics about the language technology vendor landscape as shown below. More details about the selection approach are presented in Annex A.

Figure 2 LT vendor size / employee numbers



Source: IDC 2018 for SMART 2016-0103 Lot 1 – N = 179

72% of our sample have less than 50 employees and 92% of vendors have less than 200 employees. This reflects the embryonic small size of both the language technology industry and the vendor community. At a global level, there are only 14 vendors with more than 200 employees and this list is shown below in descending order of size.

Table 1 List of large language technology vendors operating in Europe

COMPANY	BANDED EMPLOYEES
NUANCE	5000+
LexisNexis	5000+
Lionbridge Technologies Inc.	5000+
TomTom	1001-5000
AMPLEXOR	1001-5000
SDL	1001-5000
Bertin Technologies	501-1000
Intersystems	501-1000
Televic	501-1000
SDI Media Latvia	501-1000
Lesson Nine GmbH (babbel.com)	201-500
Appen	201-500
Burning Glass Technologies	201-500
Collibra	201-500

Source: IDC 2018 for SMART 2016-0103 Lot 1

It must be noted that the analysis presented in Task 1 put a special focus on the key domestic language technology vendors that play a key role as local and European players. Global vendors, like Google, Amazon and Microsoft have been considered in the market sizing exercise.

It is interesting to note that few if any of these companies are dedicated to the business of language technology. SDL is probably closest to this description. The analysis shows that language technology is usually part of a portfolio of technology product and service interests in the larger players. For example, Nuance, the largest company in the market, has interests in AI, speech recognition, biometrics, and analytics across many different industry segments and product categories which include language technology, but is not confined to language technology. To date, the relatively small size of language technology market has barely been able to support a dedicated focus, resulting in small number of larger entities with diverse interests and a long tail of small dedicated niche players often serving only local language technology markets. There is virtually no 'middle market' to speak of.

73% of the vendors in our potential survey respondents (of 179) were based in 7 larger EU countries: The United Kingdom, France, Germany, Italy, Spain, The Netherlands and Belgium. The remaining 27% were scattered across 15 other countries: Austria, Czech Republic, Estonia, Finland, Greece, Hungary, Ireland, Latvia, Luxembourg, Norway, Portugal, Romania, Slovakia, Slovenia and Sweden. Hence, virtually every country in the EU has one or more vendors that offer language technology. This reflects the local market focus of the language technology industry.

2.3. Overview of the LT Market

The analysis presented in the report of Task 1 is based on the following key segments of the language technology software market:

- 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 questions answering systems using AI technologies and others);
- Analytics including information retrieval (IR) text analytics, sentiment/opinion analysis, topic modeling, decision support systems);
- Search systems (enterprise search, multi-lingual and semantic search).

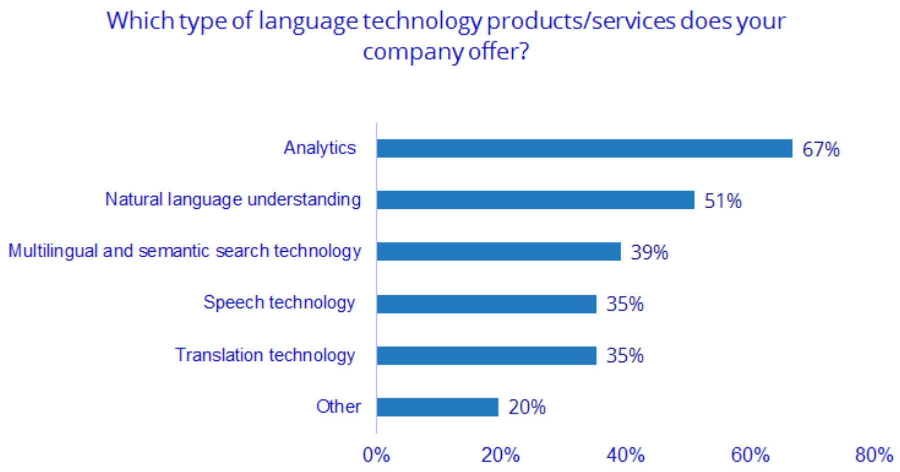
The following sections are dedicated to presenting the market data resulting from the online survey targeting a group of 179 vendors and collecting responses from 51 companies (sample).

2.3.1. Main players and innovators

Nowadays, the European market is dominated by US multi-national players (including Microsoft, Nuance, Amazon, IBM, Google, Apple and Facebook) who have a pan-European presence. Indigenous vendors are predominantly niche players serving local markets. The presence of these large players dissuades local entrepreneurs and innovators from market entry. Few new software companies have entered the European LT market in the past decade, although innovators and start-ups are now starting to appear across Europe. Currently, however, local players are mostly long-standing and well-established.

2.3.2. Current LT offering

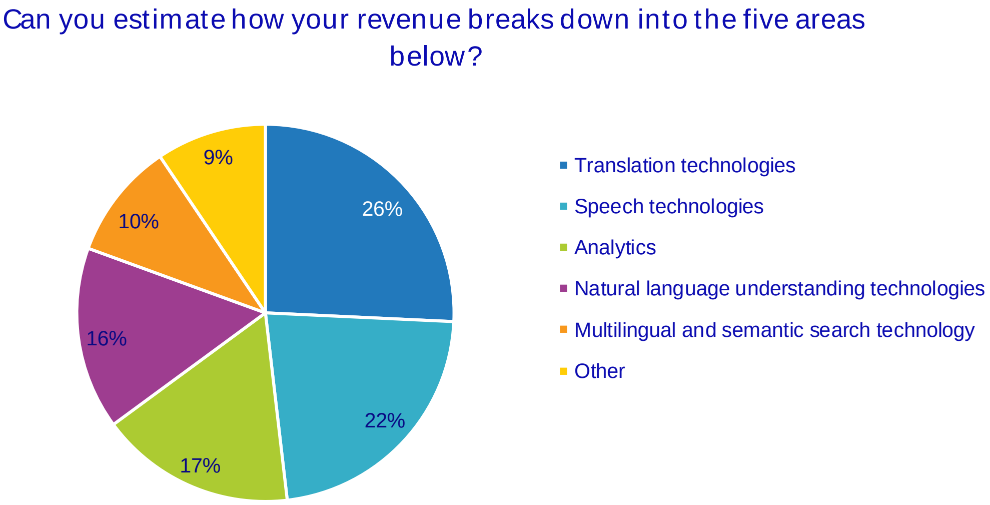
Figure 3 Product / services offered by the survey respondents



Source: IDC 2018, Online Survey, N=51, for SMART 2016-0103 Lot 1

The primary research carried out in this study revealed that analytics was the most popular of the 5 product areas, chosen by 2/3rds (67%) of respondents, followed by natural language understanding. Search, speech and translation are closely clustered together. 40 of the 51 respondents offered products or services in 1 (14 respondents), 2 (12 respondents), or 3 (14 respondents) of these product categories.

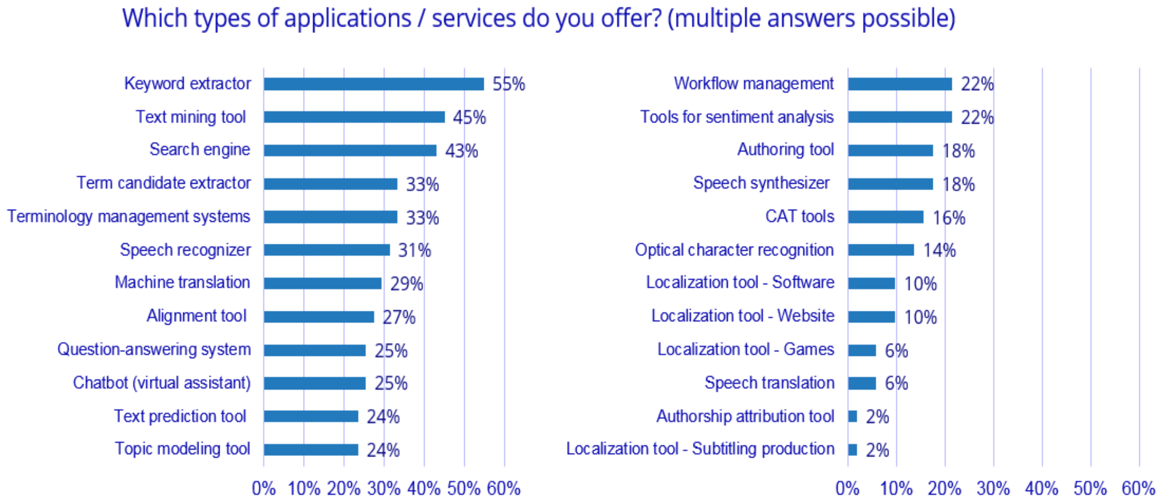
Figure 4 Overview of the main revenue sources of the survey respondents



Source: IDC 2018, Online Survey, N=39, for SMART 2016-0103 Lot 1

Although analytics is the most popular language technology offered by vendors (see Figure 3), the biggest revenue contributor is translation technologies which represents 26% of vendor revenues, followed by speech technologies at 22% of revenues. Analytics follows in 3rd place with only 17% of revenues, so is therefore not relatively a big revenue contributor, but is widely offered by the vendors. Natural language understanding, and multilingual and semantic search technology are the least important in revenue terms.

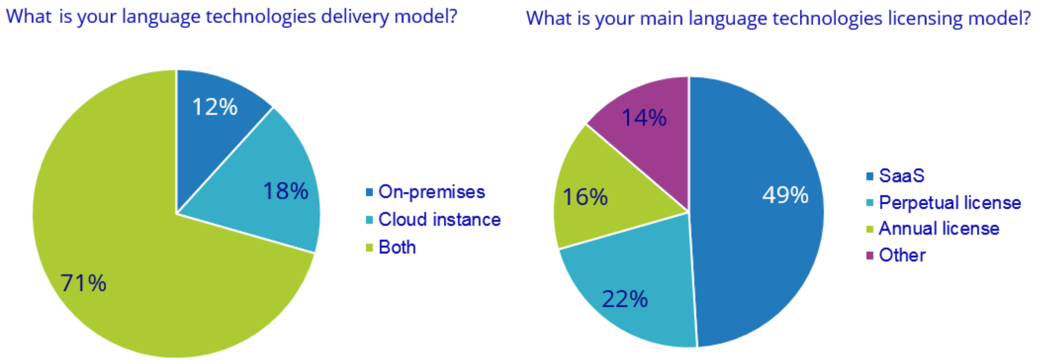
Figure 5 Overview of the applications/services offered by the survey respondents



Source: IDC 2018, Online Survey, N=51, for SMART 2016-0103 Lot 1

When asked about the product components sold by vendors, keyword extractors came out on top with 55% of responses, followed by text mining (45%) and search engines (43%). 20 product technologies garnered over 10% of responses which indicates that vendors generally have a wide and varied LT toolset.

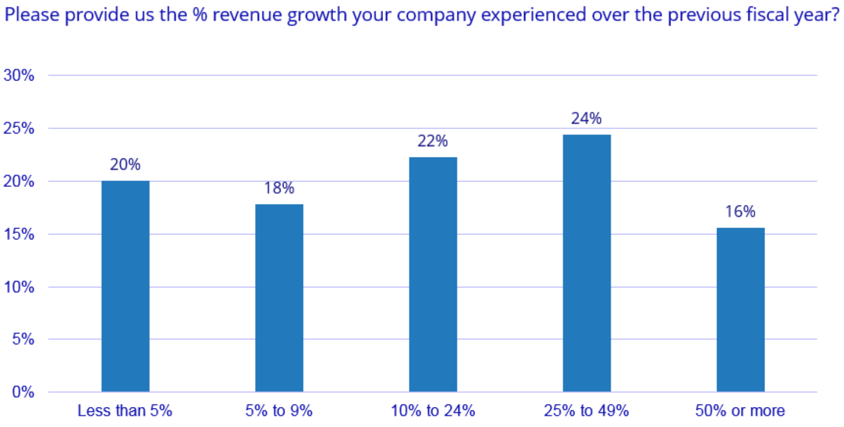
Figure 6 Overview of the delivery and licencing models offered by the survey respondents



Source: IDC 2018, Online Survey, N=51, for SMART 2016-0103 Lot 1

The preferred delivery model for these products and services is 'mixed' – a hybrid mix of on-premise and Public Cloud, rather than Cloud only' or 'on-premise only'. SaaS subscription pricing is used by 49% of respondents, the other 51% offer various types of licence agreements.

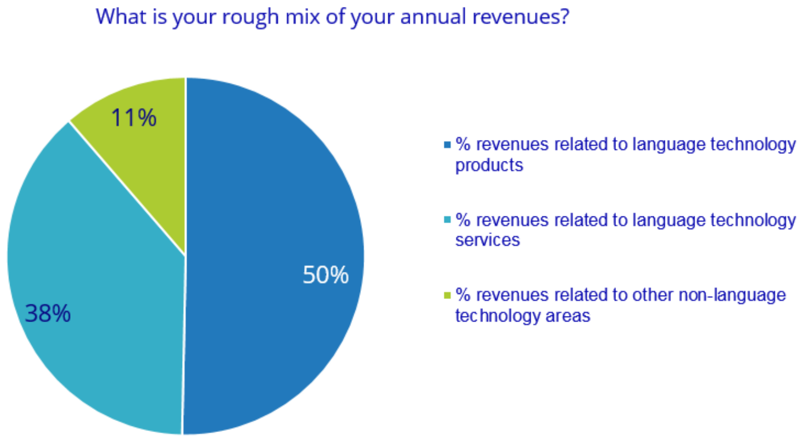
Figure 7 Overview of the revenue growth in 2017 of the survey respondents



Source: IDC 2018, Online Survey, N=45, for SMART 2016-0103 Lot 1

62% of our sample had revenue growth more than 10% which indicates a buoyant market for LT for many existing players, however 1/5 of the market players had little revenue growth in 2017. Data indicates that the market is changing rapidly. As confirmed by the survey results, market participants generally feel optimistic about the demand and revenue growth, supported by strong trends like the introduction of chat bots and the opportunities introduced by AI. This may not be realistic in case of smaller players depending on limited key clients, and consequently without a stable revenue base.

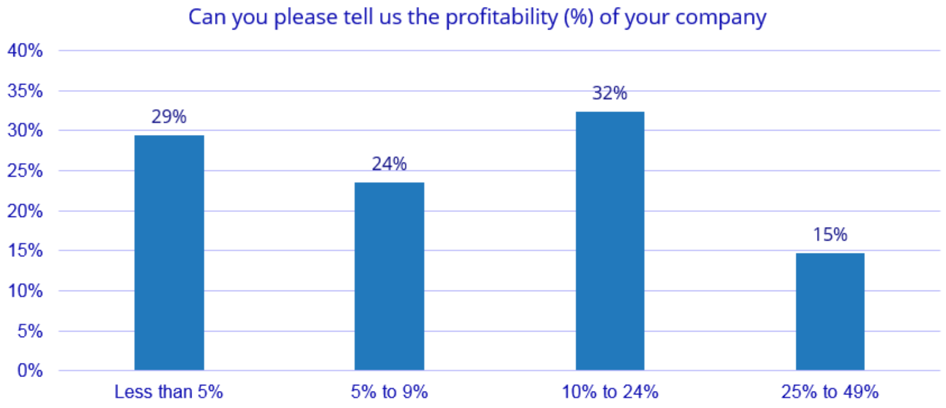
Figure 8 Overview of the software and services revenue mix of the survey respondents



Source: IDC 2018, Online Survey, N=43, for SMART 2016-0103 Lot 1

Fifty percent of revenues were attributed to LT product sales and 38% to LT services, which illustrates the combined software and services nature of the vendor offers. Only 11% had other sources of revenue, indicating that LT vendors are very dependent upon LT product and services' sales.

Figure 9 Overview of vendor profitability in 2017 of the survey respondents



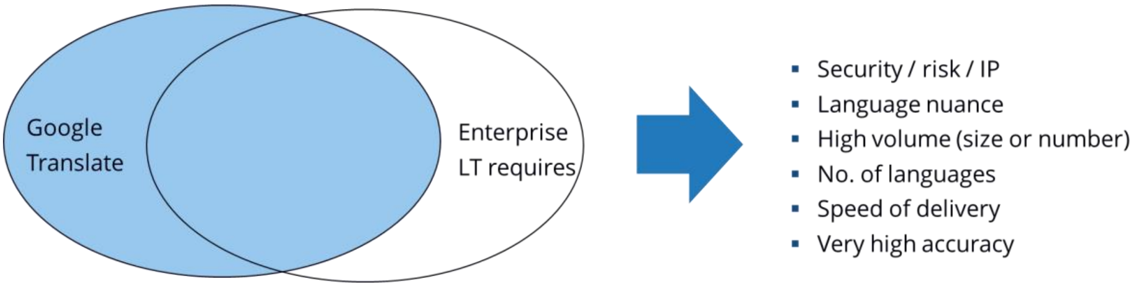
Source: IDC 2018, Online Survey, N=34, for SMART 2016-0103 Lot 1

Vendor profitability is variable. 29% of vendors are barely profitable with less than 5% profits, yet 15% of the respondents make over 25% profit. In general, these are lower margins than would be expected. Based on an ongoing monitoring of the global software market, at IDC we see that global enterprise software product margins are typically 90%, and services margins of 30-40% are not uncommon which enables forward high investments in R&D and sales and marketing.

2.3.3. Key market trends

As the most well-known and popular free translation product on the market, Google Translate currently represents the first player in the EU machine translation market. 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. 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.

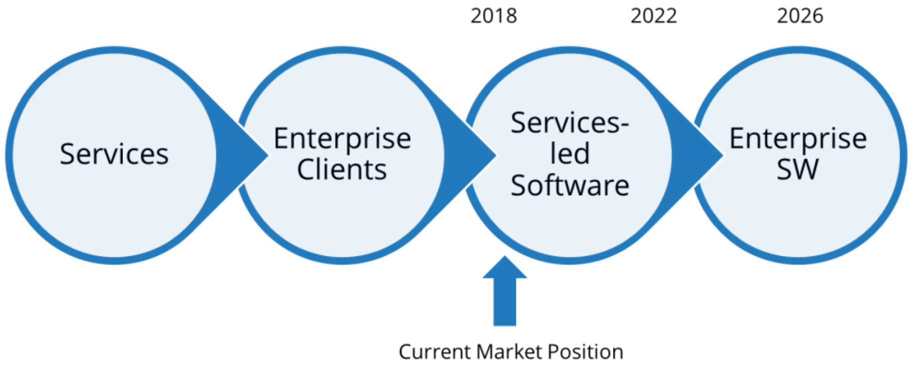
Figure 10 Where Google Translate is not sufficient



Source: IDC 2018 for SMART 2016-0103 Lot 1

Google Translate is accurate enough for the ad hoc use of many SMEs. However, as the need for translation accuracy increases, the greater is the need for more sophisticated language technology solutions. This is particularly evident for industries like Investment banking, Telecommunications, IT, and Pharmaceuticals that may require multiple industry-specific language translations of a highly technical nature. Many of the local language technology solutions vendors seek to service this need, especially when in-house language technology solutions are required for ongoing operational requirements.

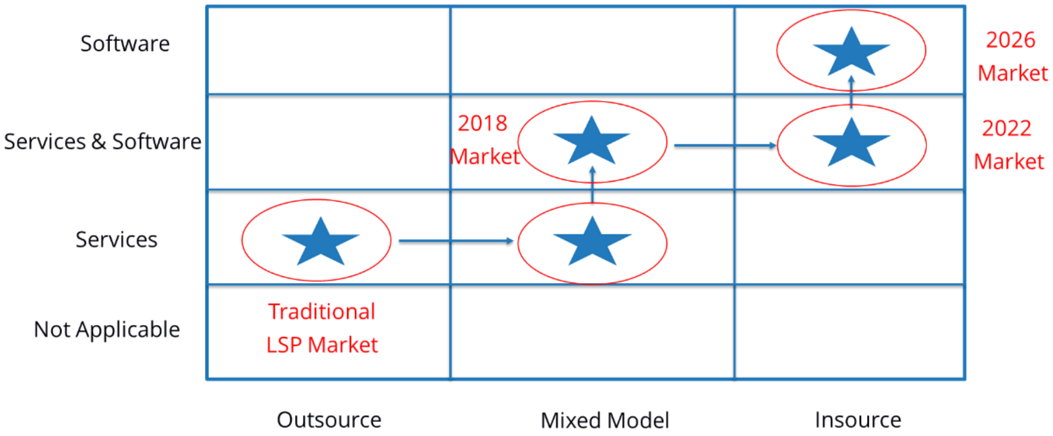
Figure 11 Vendor route map to the LT enterprise software market



Source: IDC 2018 for SMART 2016-0103 Lot 1

Outsourced language services were the genesis of the language industry in the EU. Services still form an important part of enterprise language requirements, as generally it is perceived that although good, the outputs produced by current technology available on the market still needs manual revision to ensure the appropriate adequacy. As the market and the product technology matures, LT will progressively become part of the enterprise IT architecture stack much as relational database management systems are today.

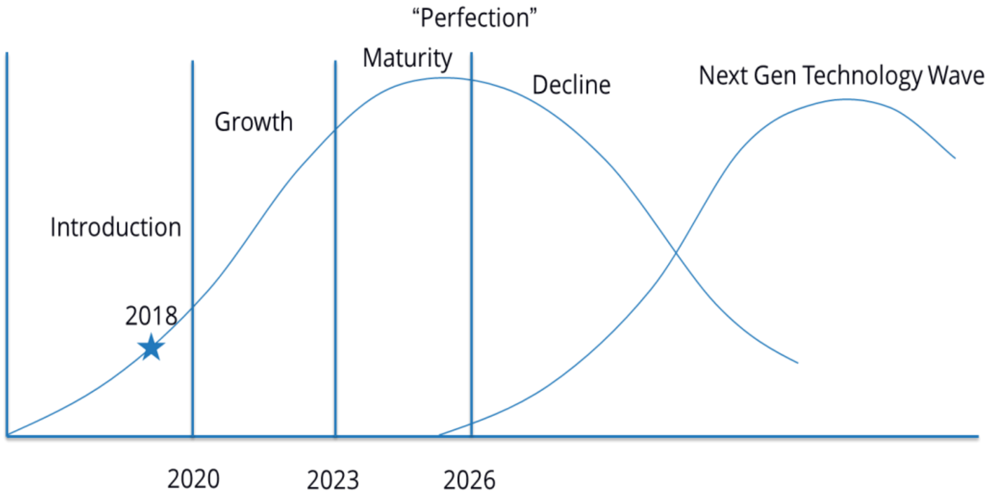
Figure 12 Enterprise LT software market potential



Source: IDC 2018 for SMART 2016-0103 Lot 1

The development of language technology to become embedded in the enterprise IT architecture stack will occur as forecast above. IDC believes that language technology, like many other categories of enterprise software, will rapidly move from the mixed hybrid model of today to an insourced model of predominantly Public Cloud Software (PCS) staffed and maintained by in-house staff and supported remotely by external vendors. The shift from statistical and rule-based approaches to neural systems are creating a huge jump in performance of language technologies in comparison to earlier incarnations. Key technologies driving this trend include mass data (input-output pairs), faster GPU computer clusters, and standardised algorithms.

Figure 13 LT software market life cycle



Source: IDC 2018 for SMART 2016-0103 Lot 1

According to the analysis carried out for this study, the current generation of LT products and services will mature in the 2023 – 2026 timeframe to be replaced by the next generation of fully automated self-managing machines which will start to emerge around 2026. IDC believes that by 2026 Machine Translation technology will deliver translations that will reflect the subtle nuances in most European languages including sarcasm, innuendo, and sentiment.

2.3.4. Overview by vertical markets

According to our research, Government, Banking, Telecommunications and Professional Services are the primary industry markets for LT vendors. However, revenues are spread across 18 vertical markets and many more sub-markets. This will be discussed in detail later on and a graphic of research results will be provided.

2.3.5. Overview by company size

In our sample, 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. This will be discussed in detail later on and a graphic of research results will be provided.

2.3.6. Overview by country

In our survey the respondents' headquarters are located mainly in larger central and northern European countries – France, Germany, the UK, the Netherlands and Belgium. However, the larger sample of 179 companies considered is located across 22 EU countries indicates a wide distribution of LT vendors across the EU. This will be discussed in detail later on and a graphic of research results will be provided.

2.3.7. Key findings

This section presented an overview of the language technology software market in Europe, based on the results of two research exercises. First, an in-depth desk research was carried out on publicly available sources as well as on IDC research pieces on this domain. In addition, an extensive field research was conducted through an online survey on 179 selected language technology players in Europe. 51 responses to the online survey were elaborated and analysed.

- The research results show that the market is dominated by US multi-national companies, which play a major role in Europe. Indigenous vendors are predominantly niche players serving local markets, among these the largest vendor is SDL with annual European LT revenues of €13M.
- In terms of offering, analytics is the most common product/service sold by the vendors in our sample followed by Natural Language Understanding technologies.
- However, the most relevant revenue sources are Natural Language Understanding and Translation technologies.
- In terms of delivery model, the vendors mainly rely on a mixed model of on-premise and cloud-based solutions.
- When we look at the revenue growth as well as to the profitability of these companies, our data shows that a quarter of the marketplace is not making notable revenues in their business.
- From a vertical market perspective, Government, Banking, Telecommunications and Professional Services are the primary industry markets for LT vendors.
- Half of the considered sample is represented by small enterprises (between 10 and 99 employees).

2.4. The LT market in Europe

2.4.1. Market size and forecast

The present section is devoted to the presentation of the market sizing exercise carried out within this study to estimate the size of the European language technologies market and the forecast to 2021.

The initial sizing of the market is based on a model that is reliant on different sources, with inputs from:

- An extensive preliminary desk research carried out on publicly available sources in the preliminary phase of the study;
- IDC's Worldwide Semiannual Software Tracker that monitors the software industry with frequent releases of semiannual software revenue estimates. This tracker provides total market size and vendor shares for 80 software markets. Measurement for this tracker is total software revenue, which includes license, maintenance, and subscription revenue (including public cloud services);
- A primary research effort (Computer Aided Web Interviews survey) addressing a solid and qualified group of LT players across Europe, represented by 179 companies of which 51 completed the online survey in the period between mid-May and early June.

The IDC model developed for this study builds on a robust forecasting and sizing expertise and unique knowledge of the worldwide software market. The general software market growth rates have been adjusted to the specific context of this study, also based on the inputs from the online survey.

This IDC market sizing and forecasting of the language technology market in the EU28 (plus Norway and Iceland) provides spending from the historical year 2016 through to the five forecast years of 2017–2021. The total size EU28 language technology market is provided for 3 different variables, by tech, by country and by industry. To measure the overall language technology market, the following segments have been considered:

- Speech technologies
- Translation technologies
- Natural Language Understanding technologies
- Analytics
- Multilingual and semantic search technology

A description of the LT technologies included in these segments is provided in Section 2.3.

IDC predicts that the language technology market size in the EU28 (plus Norway and Iceland) will grow from €706M in 2017 to €1,040M in 2021 at a compound annual growth rate (CAGR) of 9.8%.

One of the main drivers underpinning the growth this market is represented by Artificial Intelligence being used to develop applications ranging from chatbots and conversational interfaces to predictive and prescriptive applications that offer advice and recommendations. This market is focused on tools and API frameworks for applications and technologies based on Artificial Intelligence (AI), Machine Learning, and cognitive computing and is mostly using unstructured data and information as the fuel to drive these applications. A key component of this market is the use of embedded tools focused on extracting, processing, and understanding a wide range of unstructured content such as text, images, speech, and video for use in these AI-based applications.

The strong overall growth rate represents both the maturation and broad adoption of the current generation of information access technologies and applications such as Deep Learning and other forms of Machine Learning, Natural Language processing, generation, and understanding as well as semantically enabled knowledge extraction technologies including knowledge graphs and reasoning systems.

The language technology market is growing significantly and will continue to expand over the next three to five years. Vendors that are participating in this market should actively consider adding a full range of capabilities such as conversational technologies, Natural Language Processing, image and video analytics, Deep Learning, Machine Learning, hypothesis generation, and predictive analytics and adding more to their offerings to provide a complete suite of functionality for enterprise and commercial developers. Since these applications are highly reliant on unstructured information analysis and manipulation, vendors that offer strong capabilities in these areas should be able to provide tools that allow cognitive/AI applications to exploit these assets.

IDC predicts that the cognitive/Artificial Intelligence market will grow from around €1.5B in 2018 to around €5.5B in 2021 at a compound annual growth rate (CAGR) of more than 40%. Growth in this market continues to be driven by increases in AI software platforms, including conversational AI platforms being used to develop applications ranging from chatbots and conversational interfaces to predictive and prescriptive applications that offer advice and recommendations. The AI software platforms market continues to grow at rapid pace and involves billion-dollar software firms as well as a wealth of start-ups around the globe.

Many organisations are continually looking for ways to make the jobs of knowledge workers more efficient and productive, given the increasing amount of information that these workers must deal with daily. Other organisations are looking for new ways to increase sales, reduce costs, or understand their customers better by using various types of automation coupled with big data. To that end, some of these organisations have begun to evaluate a range of technologies including speech recognition, content analytics or automated translation tools.

The proliferation of data created by individuals through their devices creates opportunities to better understand consumer preferences and develop strategies to address them on a personalised basis. Enterprises use customer data assets to gain a competitive edge, and to offer differentiated and personalised products and services.

Table 2 Total market EU 28 including Norway and Iceland (in EUR million), LT by technology type

		2016	**2017**	**2018**	**2019**	**2020**	**2021**	**CAGR**
Language technologies	Speech technologies	86	90	97	106	116	129	8.4%
	Translation technologies	56	61	67	74	82	90	10.0%
	Natural language understanding technologies	114	122	138	155	174	193	11.0%
	Analytics	50	54	59	65	71	77	8.9%
	Search Systems	346	380	417	459	505	552	9.8%
Total Market		**652**	**706**	**778**	**859**	**948**	**1,040**	**9.8%**

Source: IDC 2018 for SMART 2016-0103 Lot 1

Table 3 Total market EU 28 including Norway and Iceland (in EUR million), LT by country

		2016	**2017**	**2018**	**2019**	**2020**	**2021**	**CAGR**
Language technologies	Austria	14	14	15	16	17	18	5.3%
	Belgium	22	23	25	27	29	32	7.9%
	Czech Republic	4	5	6	7	7	8	14.7%
	Denmark	19	21	23	25	27	30	8.7%
	Finland	17	18	20	22	24	26	9.2%
	France	76	81	88	96	105	114	8.5%
	Germany	164	179	197	217	240	268	10.3%
	Ireland	4	5	5	6	6	7	9.7%
	Italy	43	45	48	52	56	59	6.6%
	Netherlands	47	51	55	60	66	72	9.1%
	Norway	13	14	15	17	19	21	10.4%
	Poland	5	6	8	9	11	14	22.0%
	Portugal	4	4	4	4	5	5	6.4%
	Rest of EU28 plus Iceland	17	21	26	32	38	43	20.4%
	Spain	20	20	22	23	25	26	5.3%
	Sweden	27	30	33	36	40	44	10.1%
	United Kingdom	157	170	189	209	232	255	10.1%
Total EU		**652**	**706**	**778**	**859**	**948**	**1,040**	**9.8%**

Source: IDC 2018 for SMART 2016-0103 Lot 1

Table 4 Total market EU 28 including Norway and Iceland (in EUR million), LT by industry

		2016	2017	2018	2019	2020	2021	CAGR
Language technologies	Banking	75	80	88	97	107	117	9.5%
	Construction	6	6	7	7	8	9	9.1%
	Discrete Manufacturing	85	94	105	116	129	143	10.8%
	Education	16	17	18	20	22	24	8.8%
	Government	81	88	97	106	116	126	9.3%
	Healthcare Provider	22	23	26	28	31	34	9.8%
	Insurance	19	20	22	25	27	30	9.7%
	Media	21	23	25	28	31	33	9.4%
	Personal and Consumer Se	9	10	11	12	13	14	9.5%
	Process Manufacturing	68	74	81	89	99	109	9.8%
	Professional Services	57	62	69	77	86	96	11.0%
	Resource Industries	11	12	13	14	16	17	9.7%
	Retail	55	59	65	72	79	87	9.6%
	Securities and Investment S	19	21	22	24	27	29	8.5%
	Telecommunications	33	35	39	43	48	53	10.0%
	Transportation	26	28	31	34	37	41	9.3%
	Utilities	19	21	24	26	29	32	10.7%
	Wholesale	31	33	36	39	43	47	8.5%
		652	**706**	**778**	**859**	**948**	**1,040**	**9.8%**

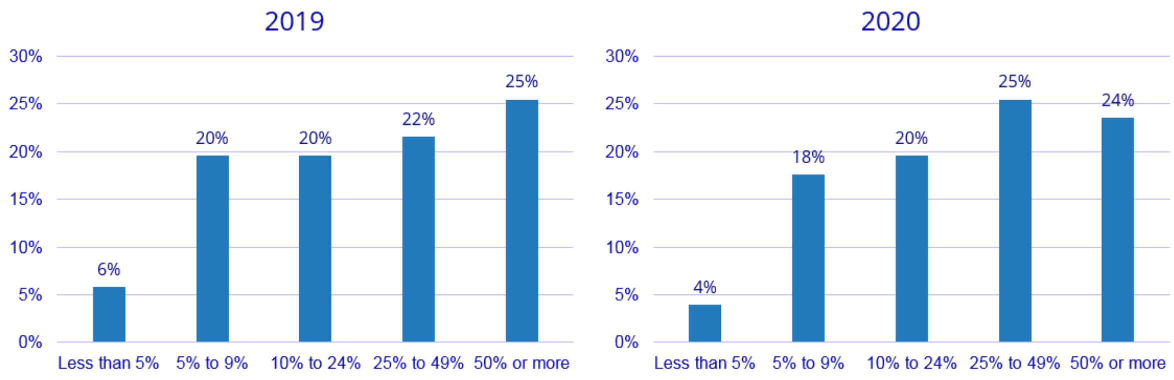
Source: IDC 2018 for SMART 2016-0103 Lot 1

2.4.2. Supply-side analysis

2.4.2.1. Key supplier trends

Vendors are optimistic about the future of the language technology market, with around 1 in 4 vendors expecting revenue growth levels projected to reach 50% or more over the next 3 years.

Figure 14 Vendor expected revenue growth rate levels 2019 and 2020

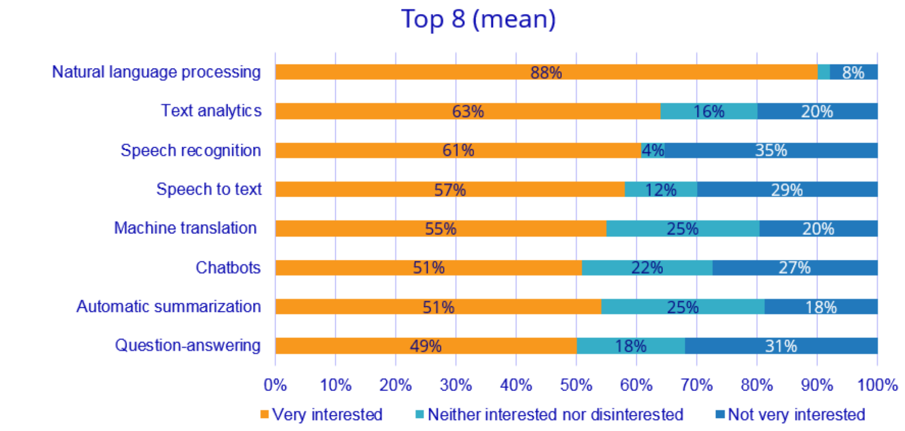


Source: IDC 2018, Online Survey, N=51, for SMART 2016-0103 Lot 1

2.4.2.2. *Emerging language technologies*

Natural Language Processing is of high interest to 88% of the online survey respondents (51) among the technologies considered.

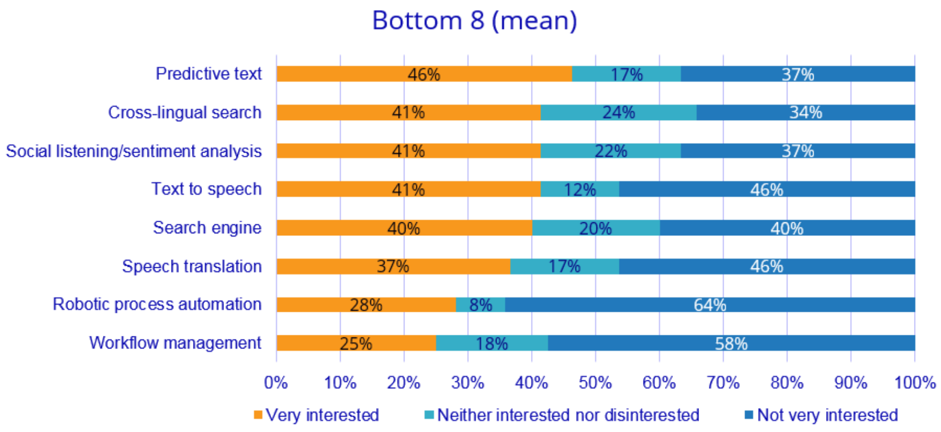
Figure 15 Higher interest in adopting new language technologies



Source: IDC 2018, Online Survey, N=51, for SMART 2016-0103 Lot 1

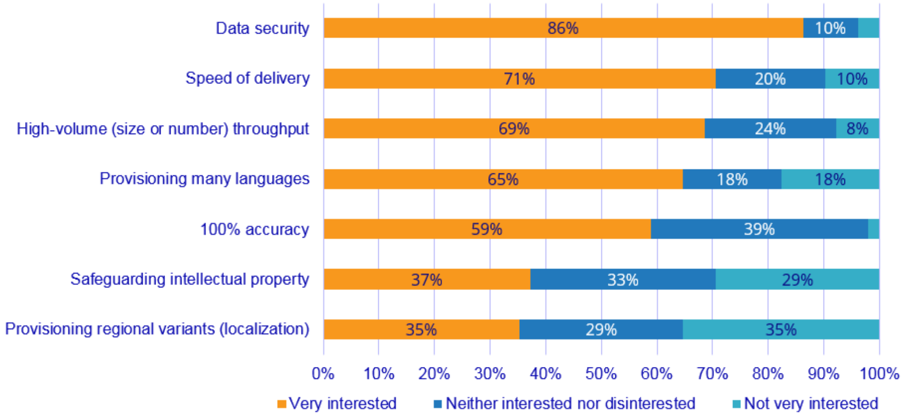
Text analytics is in 2nd place with 63% of our sample interested in adopting it. After these two, there is a clustering of technologies vendors are interested in. A further 12 technologies are of interest to over 1/3 of our interview sample. It is also noticeable that in 15 of the 16 technology areas, over 1/5 of respondents show little interest. While the results reveal the heterogenous nature of the market which may lead to subjective value judgements about customer demand for the new technologies flooding into the market, research in this area can offer much-needed clarity to vendor executive decision-making. The charts in Figure 15 and Figure 16 show how the interest and the perceived potential of emerging technologies is distributed across the sample of respondents.

Figure 16 Lower interest in adopting new language technologies



Source: IDC 2018, Online Survey, N=51, for SMART 2016-0103 Lot 1

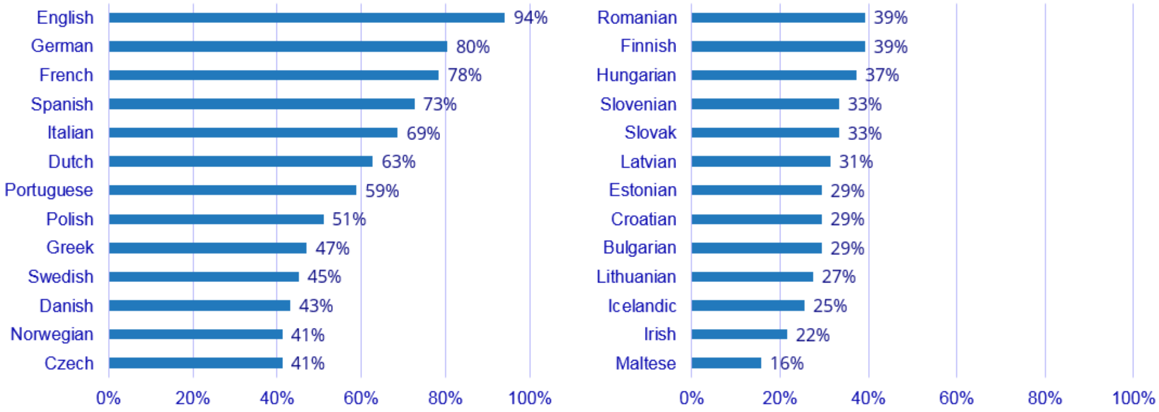
Figure 17 Needs of customers with regards to service delivery



Source: IDC 2018, Online Survey, N=51, for SMART 2016-0103 Lot 1

For enterprise service requirements, data security is by far the most important category, with 86% of responses. Speed, volumes, language availability and accuracy follow as primary drivers of customer demand, according to our sample. The requirement for solutions customisation another key customer consideration.

Figure 18 Languages for which services are provided



Source: IDC 2018, Online Survey, N=51, for SMART 2016-0103 Lot 1

The chart above shows the EU language focus of the vendors. English, German, French, Spanish and Italian, the languages with the largest populations of native speakers, are the most important to the vendors. Interestingly, 26 languages were mentioned in all, which shows that smaller country local languages services are being provided for, as local vendors offer services for the niche markets where the large US multi-national vendors are less influential as their main interests are in the higher volume markets such as French, German and Spanish.

2.4.2.3. New market entrants

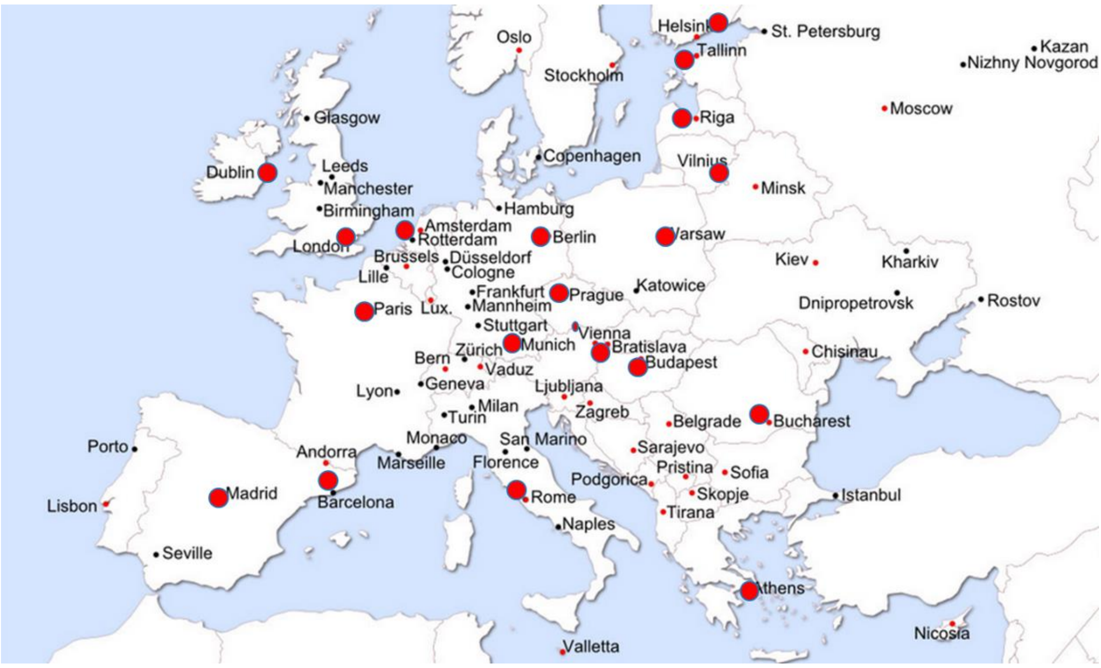
Only 44% of the EU companies in the sample has external / VC funding. In general, VC interest has been subdued in the LT market. Traditional language services have not been strongly on the VC radar, with the exception of those language technology firms, which can provide a compelling link to an AI/ML proposition.

For example, London-based start-up Memrise has over 35 million users of its 20-language AI-powered application which uses a predominantly a 'freemium' model to compete with Google Translate and the like. Founded in 2010, in June 2018 Memrise has just raised $15.5m in a Series B round led by Octopus Ventures and Korelya Capital, including existing investors Avalon Ventures and Balderton Capital. This brings Memrise's funding to date to $22m.

Other recent VC investments include an Irish start-up firm, Aylien, which received $2.35m from Atlantic Bridge University for its Text analytics tools in 2017. Another example is Transfluent from Finland, which received $825K from crowdfunding. For the VC world however, these are small amounts. The likelihood is VCs are dissuaded from investing by the freemium pricing strategies of the large US multi-national vendors.

As the collected data shows, vendors are not greatly interested in becoming part of specialised language technology innovation labs or digital hubs. Only 25% have done so.

Figure 19 Map of European innovators and accelerator hubs

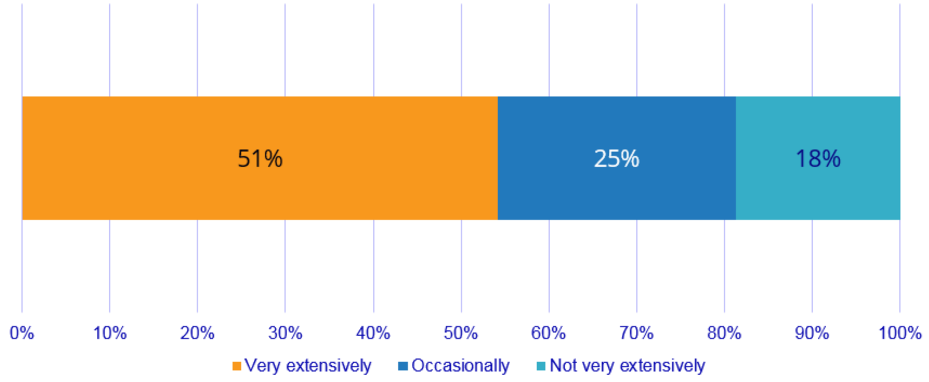


Source: IDC 2018 for SMART 2016-0103 Lot 1

The red dots in Figure 19 show where IDC has identified 19 European language technology Innovators and Accelerator Hubs. Innovators are typically sources of influential scientific and technical publications and patents. DFKI in Germany is an example, as are the University of Edinburgh (MT), and the University of Cambridge (dialog). Accelerators are the first 'movers' in the market, which are replicated by others. An example is Arria NLG PLC, which offers Artificial Intelligence technology for data analytics and information delivery. Arria is also an example of an accelerator entering the LT market encouraged by the application of new technologies to language processing. For the most part, the attention of start-ups is on the major commercial centres where they can expect to find the best sales opportunities for language technology.

Although few vendors are involved with VCs or Innovators and Accelerator Hubs, the opposite is true when it comes to collaboration with academic and research institutions. 51% of our vendor sample collaborate extensively and a further 25% work with academic and research institutions on an occasional basis.

Figure 20 Collaboration with academic and research institutions

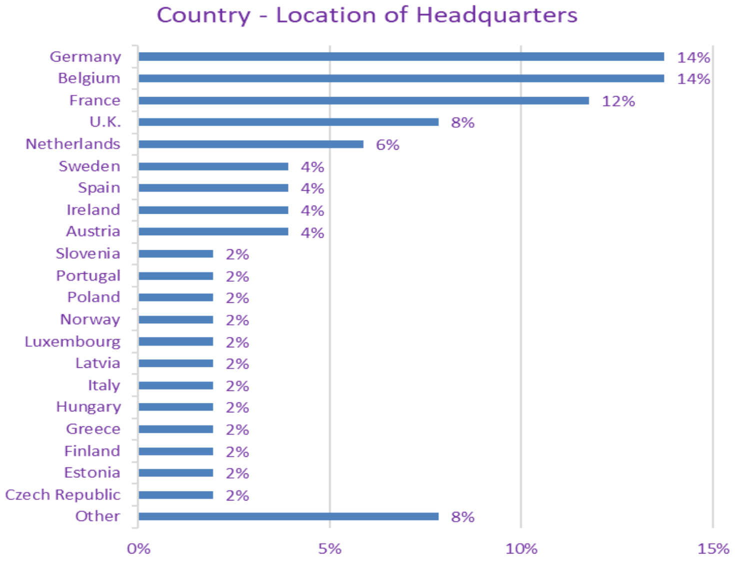


Source: IDC 2018, Online Survey, N=51, for SMART 2016-0103 Lot 1

2.4.2.4. Supplier demographics

The large countries, Germany, France, UK and the Netherlands attract the most head offices of LT vendors, although there is a long tail, which implies a quite balanced distribution of companies across the EU. The preliminary in-depth research carried out by the consortium to identify and qualify the sample of companies to be targeted through the online survey led in some cases to an over-representation of some countries, like Belgium that is a somewhat over-represented in this survey in IDC's opinion where the consortium could leverage a network of good contacts with the local players.

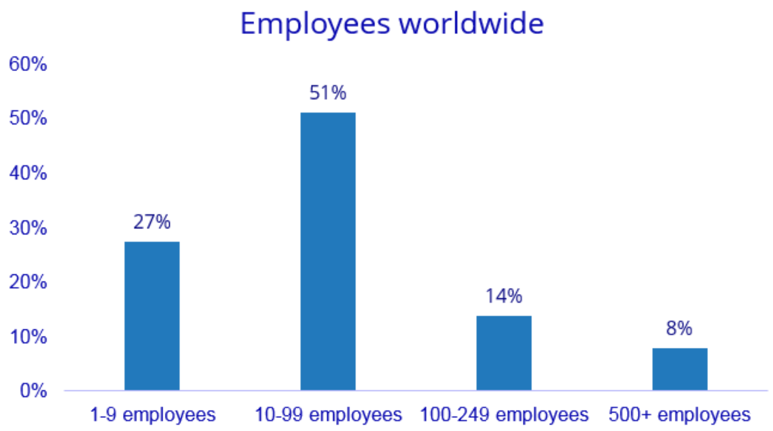
Figure 21 Country location of headquarters



Source: IDC 2018, Online Survey, N=51, for SMART 2016-0103 Lot 1

Nearly 4 out of 5 vendors (78%) have less than 99 employees globally.

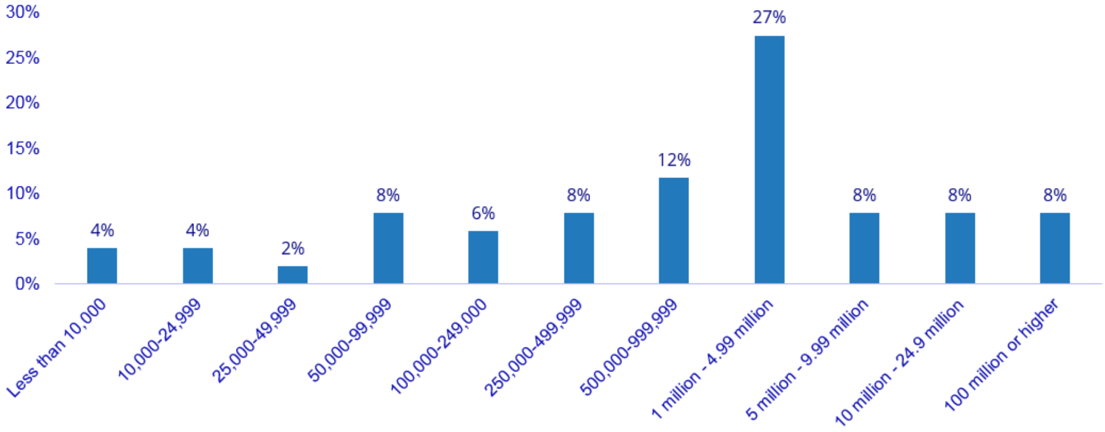
Figure 22 Number of employees by company



Source: IDC 2018, Online Survey, N=51, for SMART 2016-0103 Lot 1

49% of our sample had less than €1M revenues, which is very small in software industry terms.

Figure 23 2017 revenues of survey respondents

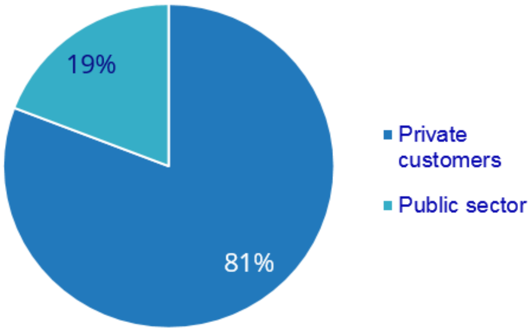


Source: IDC 2018, Online Survey, N=42, for SMART 2016-0103 Lot 1

2.4.2.5. Industry markets served

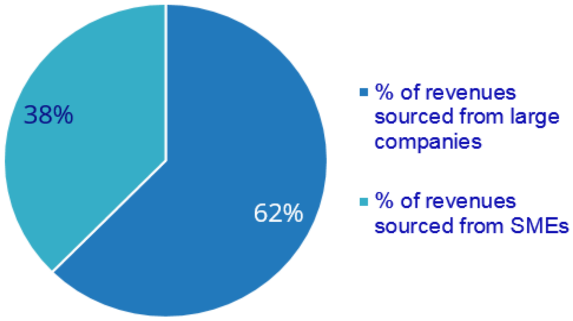
Although the Public sector is seen as the most important market for LT vendors, being the most served industry (as shown in Figure 26), only 19% of vendor revenues are sourced from this sector. This means that from the vendors' point of view, in terms of profitability, the public sector lags behind the private sectors.

Figure 24 Revenues from public sector bodies vs private customers (survey respondents)



Source: IDC 2018, Online Survey, N=40, for SMART 2016-0103 Lot 1

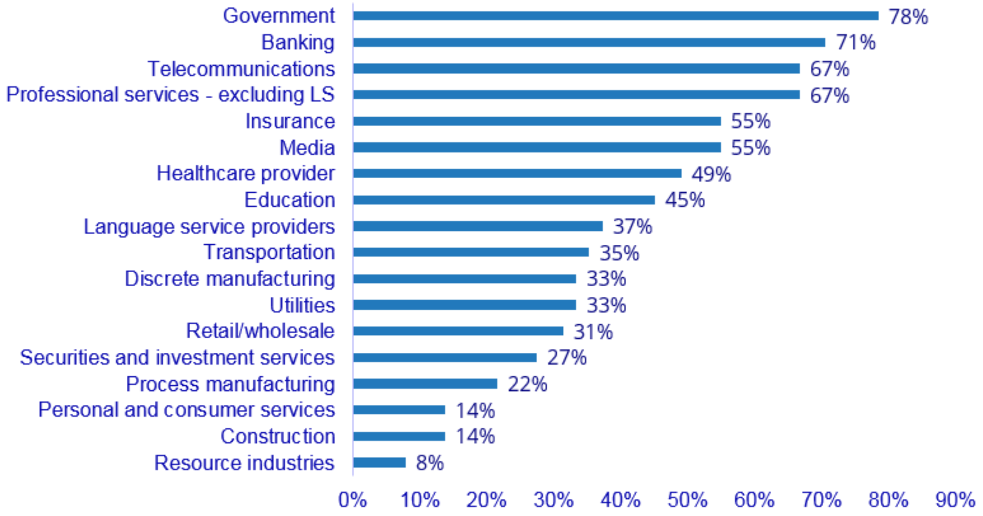
Figure 25 Revenues from SMEs vs large companies (survey respondents)



Source: IDC 2018, Online Survey, N=33, for SMART 2016-0103 Lot 1

Vendor revenues from larger companies outstrip those from small companies by approximately 2:1. Competing is tough for small companies and investing in and undertaking R&D is not necessarily easy. Positioning in niche markets and partnerships are often good ways to increase revenues and to consequently maintain (or sometimes increase) profitability levels.

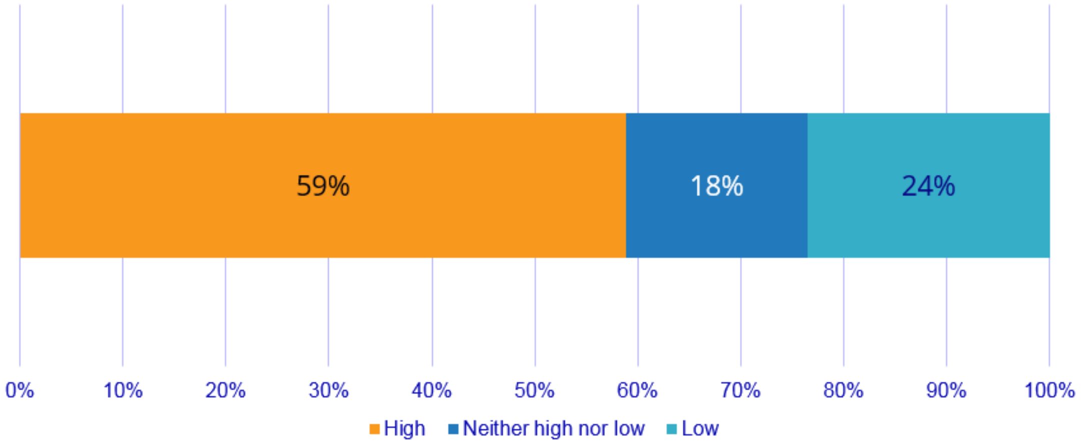
Figure 26 Industry sectors served



Source: IDC 2018, Online Survey, N=51, for SMART 2016-0103 Lot 1

Figure 26 represents a wide array of industries that are using LT, headed by Government, Banking, Telecommunications and Professional Services.

Figure 27 Customer requirement for industry sector-specific vendor expertise



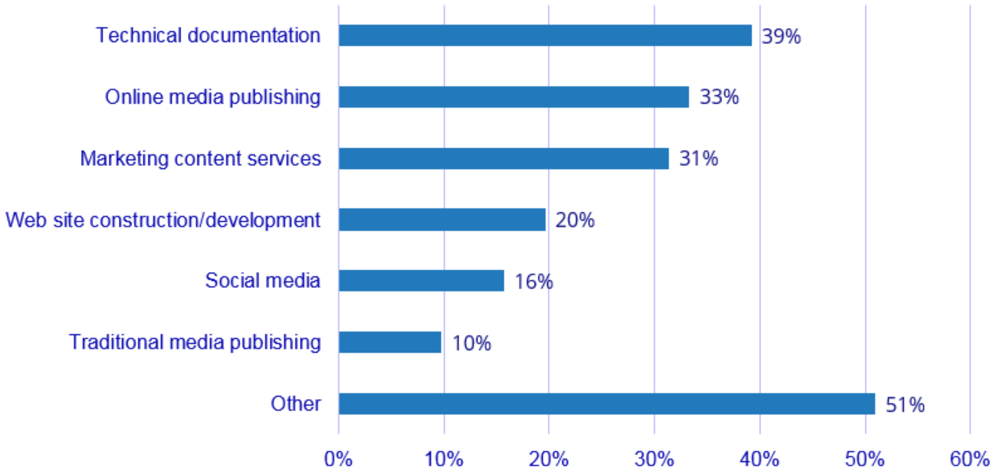
Source: IDC 2018, Online Survey, N=51, for SMART 2016-0103 Lot 1

59% of vendors believe that it is important to specialise in specific vertical markets. Understanding the terminology and semantics of specific industries is key to win their confidence and business.

2.4.2.6. End users' applications for LT

The survey results show four end user application areas that are the most important to customers, according to vendors: technical publishing (39%), online media publishing (33%), marketing content services (31%), and web site construction/development (20%).

Figure 28 Key application areas provided to end customers according to survey respondents



Source: IDC 2018, Online Survey, N=51, for SMART 2016-0103 Lot 1

'Others' includes a wide range of specific language technology application areas including human-machine-interaction, Customer Service automation, medical documentation, Voice of the Customer/Employee/Patient, Contract analytics, Robotic Process Automation, Automotive and mobile speech UI, Healthcare document management, Customer Experience, User Interface, HR, recruitment and labor market analytics; IVR, Speech analytics; Quality Assurance, conference interpretation, and Qualitative analysis.

2.4.3. Demand-side analysis

This section presents a qualitative analysis of the key vertical industries addressed by the language technology vendors.

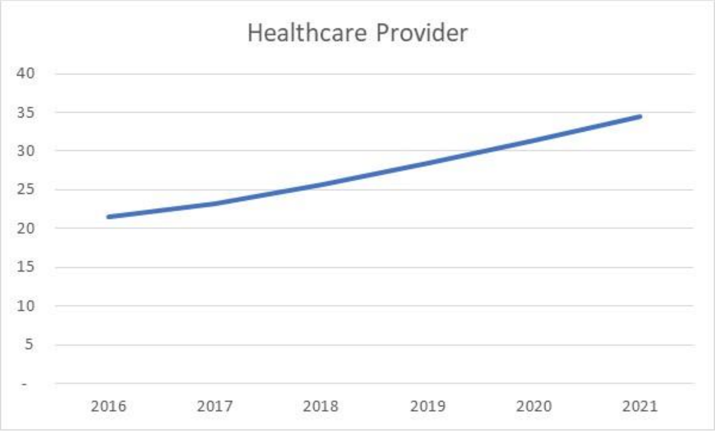
Verticalisation strategies are key as peculiarities in many sectors cannot be underestimated. The vocabulary used by a doctor, for example, is different from that of a lawyer. At the same time, from a purely software perspective, doctors' and lawyers' needs in speech recognition are similar. Verticalisation works on a case-by-case basis and it is clear that if vendors' specialisation for an industry is proving to be beneficial, this can become a differentiator factor in decisions to invest in further vertical markets.

In the following sections, we provide an analysis of the key industry sectors our research focussed on, the underling use cases, trends and challenges. This analysis relies also on the market size by industry, provided in Section 2.4.1.

2.4.3.1. Healthcare

Language technologies spending in the Healthcare industry accounts for €26M in 2018 and is expected to grow to €34M by 2021, showing a 2016-2021 CAGR of 9.8%.

Figure 29 EU 28 including Norway and Iceland (in EUR million), LT in healthcare industry



Source: IDC 2018 for SMART 2016-0103 Lot 1

Language technologies are crucial for the healthcare sector. Essentially, they make it possible to recognise and extract data from free text or speech and transform information locked in textual formats (publications, pathology reports, electronic health records, clinical notes or web content, etc.) into high quality structured data that can be used by computing processes. The use of large amounts of high-quality clinical data make it possible to optimise quality of care, improve overall patient experience, reduce costs and drive innovation and research in the health sector. The widespread adoption of language technologies in Healthcare is primarily driven by the increasing use of intelligence and analytics tools to obtain smarter clinical information.

The most common applications are:

- Machine translation
- Speech recognition
- Question answering
- Knowledge extraction
- Classification

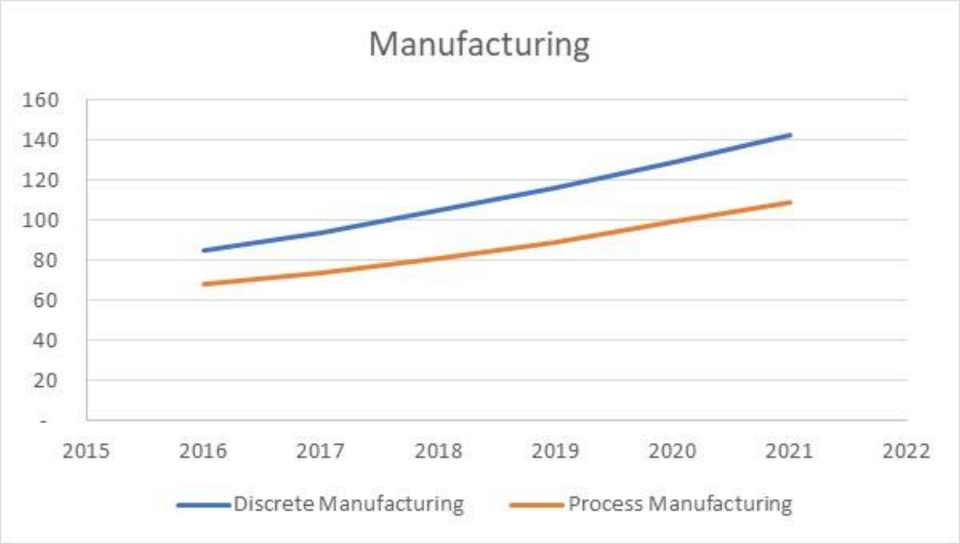
These applications feed into healthcare information systems and analytics solutions. This allows to provide additional and higher quality evidence to support clinical decisions, research processes, compliance, revenue cycle management and healthcare services planning. Language technologies solutions can capture information contained in health corpora by automatically identifying, extracting and structuring the information. Extracted information can be mapped to ontologies, terminologies and other formal representations of health information to feed decision support and research.

Also, extracting relevant data elements from clinical narratives constitutes a basic enabling technology to unlock knowledge and support more advanced reasoning applications such as diagnosis explanation, disease progression modelling and intelligent analysis of the effectiveness of treatment. For example, a clinician can ask a computer to extract a patient's diagnosis from a large data set or unstructured visit notes within an electronic health records (EHR) system: NLP is able to process all the available data and identify and extract the relevant information.

2.4.3.2. Manufacturing

Language technologies spending in the Manufacturing industry accounts for €186M in 2018 and is expected to grow to €252M by 2021 showing a 2016-2021 CAGR of 10.4%.

Figure 30 EU 28 including Norway and Iceland (in EUR million), LT in manufacturing industry



Source: IDC 2018 for SMART 2016-0103 Lot 1

The European Manufacturing industry is very varied and fragmented with a huge number of SMEs playing a relevant role. Along the manufacturing supply chain, raw materials may originate from different countries and, similarly, manufacturing plants can be geographically dispersed. Communication may therefore be also affected by language and cultural barriers.

Automation of documentation translation is not homogeneous across countries, manufacturing sub-industries and company size. It depends also on the overall IT readiness of manufacturers.

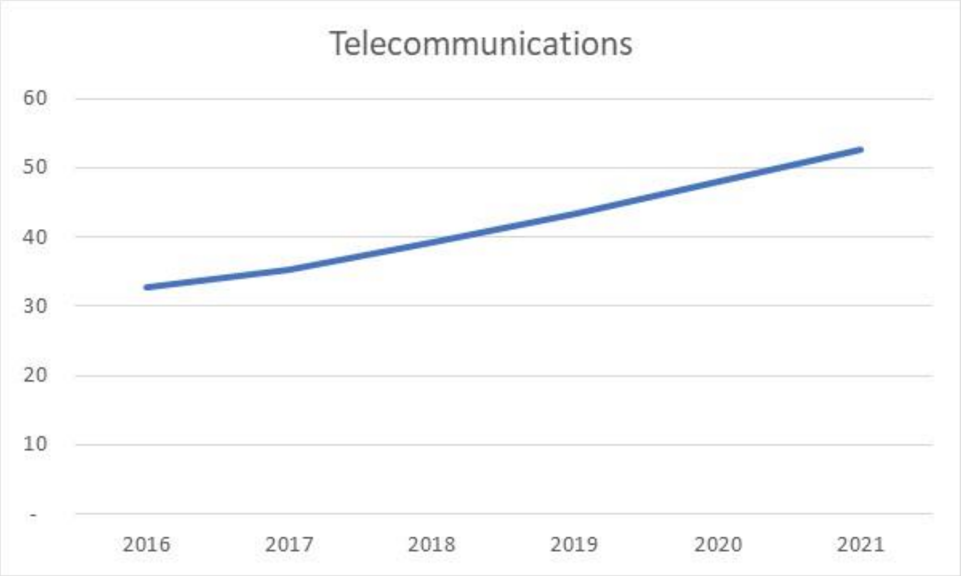
For example, companies that are more advanced in big data solution adoption are more likely to engage in language technologies projects as well. According to IDC, big data and analytics initiatives vary considerably across countries, with Germany and France being the forerunners. and across sub-industries, where it is the Automotive sector that is showing the strongest investment in big data.

Big data initiatives continue moving beyond the IT departments, reaching business analysts across lines of business. In leading industries such as the Automotive sector, big data initiatives predominantly reside within lines of business. The most common and promising use cases are for the analysis of operational data, factory automation, and the analysis of online customer behaviour.

2.4.3.3. Telecommunications

Language technologies spending in the Telecom industry accounts for €39M in 2018 and is expected to grow to €53M by 2021 showing a 2016-2021 CAGR of 10.0%.

Figure 31 EU 28 including Norway and Iceland (in EUR million), LT in telecom industry



Source: IDC 2018 for SMART 2016-0103 Lot 1

Communications service providers are often global organisations with operations located in diverse areas, spanning urban to the most remote areas globally. These companies operate in heavily regulated, compliance-driven, politicised, and market-driven environments.

They must participate in global ecosystems through network-to-network interconnections and IP exchanges to enable delivery of roaming and remote location connectivity to domestic customers.

The use of chatbots and conversational intelligent assistance for customer handling is one of the many use cases for the incorporation of cognitive computing and AI technologies within Telecommunications organisations. Contact centres, customer-facing employees and CRM systems are perfect candidates for these technologies for rapid handling of repetitive inquiries and for the parsing of more complex customer requirements as a tier 1 triage.

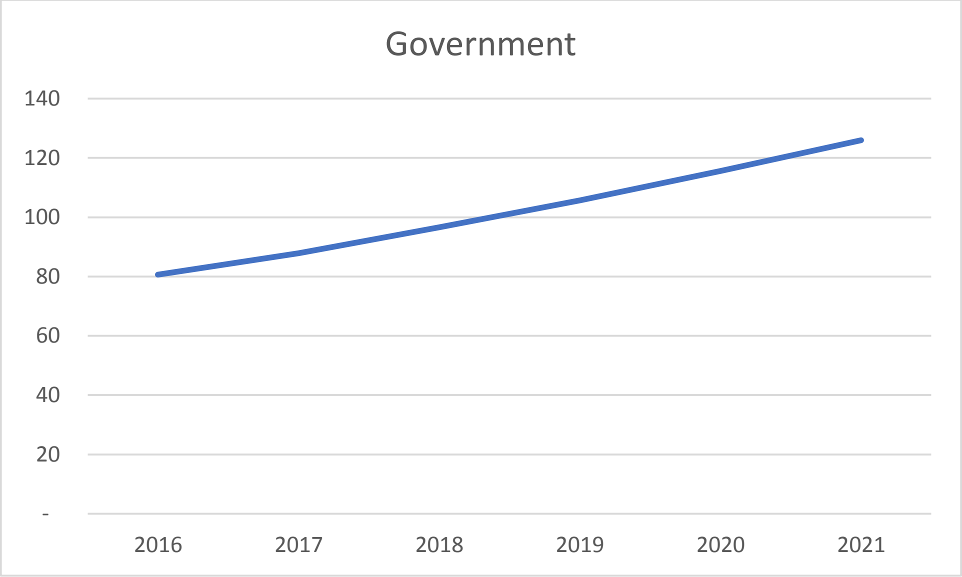
Larger organisations typically lead the way in the adoption of chatbot technologies due to availability of higher resources and the need to pursuit of market differentiation, which is particularly important in the Telecom industry, to try to avoid competing on price only.

Chatbots are therefore becoming increasingly widespread and IDC believes they will become more and more accurate. Adoption will consequently spread rapidly. Telecom operators use language technologies to enhance customer communications, for call center services and to improve search results. Also, chatbot capabilities are no longer limited to consumer applications; vendors such as Ariba have added these kinds of capabilities to their enterprise applications. IDC predicts that this trend will continue and workers whose daily tasks involve the use of enterprise applications will have access to intelligent personal assistants such as chatbots to augment their skills and expertise.

2.4.3.4. Government

Language technologies spending in the Government sector accounts for €97M in 2018 and is expected to grow to €126M by 2021 showing a 2016-2021 CAGR of 9.3%.

Figure 32 EU 28 including Norway and Iceland (in EUR million), LT in government sector



Source: IDC 2018 for SMART 2016-0103 Lot 1

Governments' use of data and analytics is maturing. Information transformation is a process that government organisations need to pursue to bridge gaps between management and services delivery units so to integrate information around the citizen at both local and central level.

The number of organisations focusing IT spending on 3rd Platform technologies to address the need for innovation in service design and delivery is accelerating and will continue to do so.

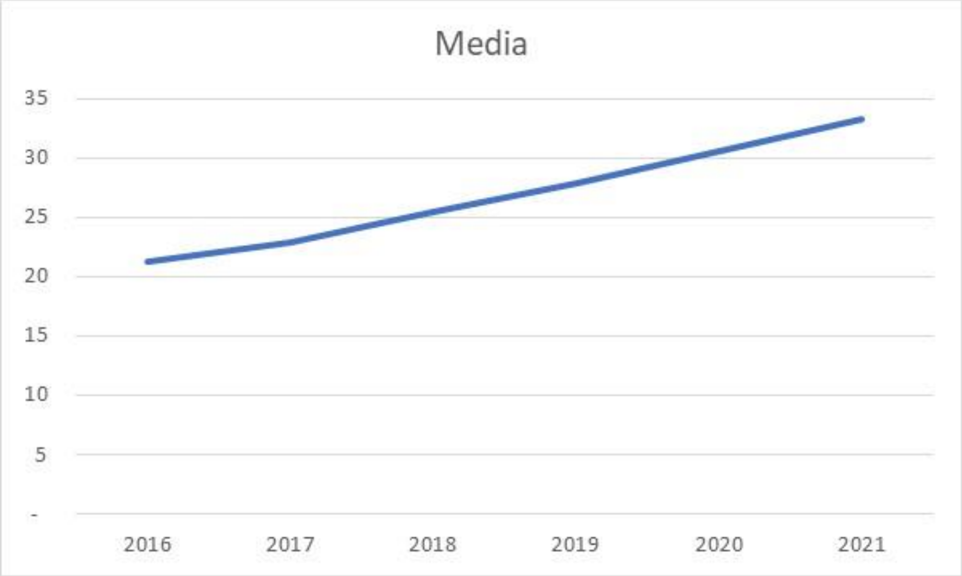
As part of the digitalisation effort of the Public sector, cloud computing is another focus area. Relying on technologies allowing for a fast information exchange and document sharing is essential to speed up procedures and avoid long waiting times. Cloud applications are crucial for governments to implement the European Union's Digital Single Market Strategy, as they encourage innovation through an exchange of services over the internet. Not only does this result in cost reduction and a more efficient and effective administration, but it also allows public officers to rapidly access emails, files, and media contents from anywhere at little or no cost.

Clearly, effective communication and fast and efficient use of data pass through language technologies. Many documents might need translation both to improve communication with citizens and to extract relevant information from data.

2.4.3.5. Media

Language technologies spending in the Media industry accounts for €25M in 2018 and is expected to grow to €33M by 2021 showing a 2016-2021 CAGR of 9.4%.

Figure 33 EU 28 including Norway and Iceland (in EUR million), LT in media sector



Source: IDC 2018 for SMART 2016-0103 Lot 1

Across media companies, digital transformation based on next-generation technologies is revolutionising the way companies provide services and generate revenue streams.

The media industry is responding to consumer demographic and behavioural changes as well as demand for instant access to content anytime and anywhere by embracing innovation and putting digital transformation at the core of this change.

Digital transformation is revolutionising not only the way media companies provide services, but also the way they generate revenues. In fact, many media businesses have embedded technologies in their marketing strategies such as big data and analytics for the creation of smart adverts or promoting content or movies based on personal preferences.

Artificial Intelligence is also gaining ground across media companies, for example transforming the traditional way plots are created and allowing to save time.

Language technologies can be used for example for automatic subtitling, speech recognition is becoming widespread in the newspapers sector, transforming speech into text automatically, with an improvement both on speed and accuracy.

Another LT use case in the Media sector is subtitling in different languages, for example using tools like speech recognition and speech-to-text. This might involve TV programs, movies, etc. This might be to provide citizens living in multilingual countries with the option to choose the language of their preference e.g. in Belgium, French and Flemish, or just to provide more option to an increasingly multilingual population.

2.4.4. Key findings

This section presents the estimate of the size of the European language technologies market in Europe – including Norway and Iceland-, and the forecast to 2021.

Supply side – key findings

- IDC predicts that the language technology market in the EU28 (plus Norway and Iceland) will grow from €706 million in 2017 to €1,040 million in 2021 at a compound annual growth rate (CAGR) of 9.8%. The data shows that the language technology market is growing significantly and will continue to expand over the next three to five years.

- Half of the LT market is represented by search technologies, followed by natural language understanding technologies, showing the highest growth rate (11% CAGR) to 2021 among the categories considered. From a country perspective, Germany holds the largest share of the LT market with a value of €179M in 2017, growing to nearly €270M in 2021, followed by the UK that will grow to €255M in 2021.

- Government, Banking, Telecommunications and Professional Services represent the largest markets for LT technologies. However, although the Public sector is seen as the most important market for LT vendors, it accounts only 20% of their revenues.

- With regards to the languages for which LT services are provided, English, German, French, Spanish and Italian are of greatest importance to the vendors.

- In terms of market trends, natural language processing represents the key emerging trend in terms of adoption of LT, followed by text analytics and speech recognition.

- In terms of marketplace innovation and new entrants, our data shows that vendors are not greatly interested in becoming part of specialised language technology innovation labs or digital hubs. Only 25% have done so. Only 38% of the EU companies in the sample has external / VC funding.

Demand side – key findings

- Our analysis also examined the market from a demand side perspective, with a particular focus on some vertical industries in which language technologies play a key role. The estimate takes into account the spending for these products and services by the companies active in those markets.

- Language technologies spending in the Healthcare industry accounts for €26M in 2018 and is expected to grow to €34M by 2021, showing a 2016-2021 CAGR of 9.8%. Machine translation and speech recognition are the most common applications in this market.

- The second market considered is Manufacturing. Language technologies spending in this industry accounts for €186M in 2018 and is expected to grow to €252M by 2021, showing a 2016-2021 CAGR of 10.4%. The most common and promising use cases are for the analysis of operational data, factory automation, and the analysis of online customer behaviour.

- Language technologies spending in the Telecom industry accounts for €39M in 2018 and is expected to grow to €53M by 2021, showing a 2016-2021 CAGR of 10.0%. The use of chatbots and conversational intelligent assistance for customer handling is one of the many use cases for the incorporation of cognitive computing and AI technologies within organisations.

- The public sector is one of the biggest markets of LT adoption. IDC predicts that language technologies spending in the Government sector accounts for €97M in 2018 and is expected to grow to €126M by 2021, showing a 2016-2021 CAGR of 9.3%.

- Finally, the analysis considered the Media sector. Language technologies spending in the Media industry accounts for €25M in 2018 and is expected to grow to €33M by 2021 showing a 2016-2021 CAGR of 9.4%. The most relevant applications in this sector are automatic subtitling, speech-to-text, speech recognition, and subtitles machine translation.

2.5. Conclusions

Beyond the analysis described and discussed above, IDC conducted selected telephone interviews with a group of relevant LT players. This additional piece of research revealed interesting findings which have been deployed to better define the below conclusions.

- **The LT market is very fragmented and composed by SMEs.** The LT market in the EU is very fragmented and there is a lack of large indigenous players. European players are all SMEs, where SDL is the largest. 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".** One of the positive effects of large players such as Google, Microsoft and Apple from the local vendors' point of view 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.

3. Task 2: Competitiveness analysis

3.1. Preface

Language is a key common denominator that holds Europe together while also keeping it apart. Language equality is a key concept of the European Union, but the fragmentation of the single market by language is a growing issue for the global and internal market competitiveness of Europe. Technology has emerged as a significant tool to maintain the ideals of language equality while diminishing the barriers inherent in a multilingual Europe. New advances in language technologies and specifically machine translation can successfully bring down internal barriers and create new paradigms for European global business development.

The aim of Task 2 is to conduct a competitiveness analysis in three areas of LT – machine translation (MT), cross-lingual search, and speech technology.

The competitiveness analysis is based on an extensive desk research of various studies, policy papers, and online information sources. Its quantitative foundation is based on the surveys and interviews carried out and analysed in Task 1 as well as aggregation and analysis of data collected from previous studies on machine translation and the broader localisation and translation sector, and overall economic indicators.

The qualitative research design is based on consolidation and contrasting validation of findings and views reflected in the large variety of secondary sources analysed in the research, including reports and statistics published by leading expert groups, industry associations and international organisations such as Common Sense Advisory (CSA) (Lommel et al., 2016), TAUS (Massardo, 2016; Seligman, 2017; TAUS, 2017), Slator (Slator, 2018), Eurostat, World Economic Forum (World Economic Forum, 2017), and others.

The report of Task 2 includes the following parts:

- The competitiveness analysis of three LT areas: machine translation, cross-lingual search, and speech technology.
- An analysis of the outcome of the market survey and interview carried out in Task 1.
- Identification of strong points and shortcomings in LT areas under investigation through a ranking and weighting method.
- Identification of external threats as well as opportunities for development of LT areas.
- Identifying and proposing potential effective actions at the EU level.
- The use of in-depth information, facts and figures from available market studies.
- Applying expert knowledge of LT market stakeholders.

3.2. Methodology

The report of Task 2 starts with an in-depth analysis of the selected factors and a comparison of European market performance with competitors. Section 3.3 provides the summarised information about the comparison of markets in the areas of machine translation, cross-lingual search, and speech technology. Sections 3.4 to 3.10 contain detailed information and an in-depth analysis of 7 dimensions broken down by the areas mentioned.

Following from the conducted research, Section 3.11 provides a two-part **SWOT** (Strengths, Weaknesses, Opportunities, and Threats) analysis – the strengths and weaknesses internal to European LT endeavours and external opportunities and threats. Finally, Section 3.12 offers recommendations.

3.2.1. Market dimensions

Two markets were selected to compare Europe with – North America and Asia (as defined in Annex B).

We have identified the following 7 dimensions to decompose the LT markets:

- Research
- Innovations
- Investments
- Market dominance
- Industry
- Infrastructure
- Open data

These seven LT dimensions were analysed in the context of global competitiveness, highlighting particularly the most important achievements and gaps of the LT ecosystem between Europe and its largest global competitors - North America and Asia.

To characterise each dimension, a number of objective criteria have been identified. An in-depth analysis is conducted on each criterion. Using these results, we have ranked the markets within each dimension according to how strong they measure against each other on a scale of 1 to 3:

- Strongest (3 points)
- Average (2 points)
- Weakest (1 point)

The analysis of the 7 dimensions of the LT market allows us to identify effective future policies.

3.2.2. SWOT analysis

The study resulted in a SWOT analysis for the European MT market. A SWOT analysis is a commonly employed framework and strategic planning technique that is used to help identifying Strengths, Weaknesses, Opportunities and Threats and serves to uncover the optimal match between the internal strengths and weaknesses of a given organisation, concept, or market entity and the external trends – opportunities and threats that the entity or concept must face in the marketplace.

The objective of a SWOT analysis is to identify the favourable and unfavourable internal and external factors to support decision making:

- **Strengths:** assess the characteristics of the selected entity that give it an advantage over others. Strengths are internal factors that support an opportunity and may include technological, product, or solution advantages, financial strengths, infrastructure, human resources talent, natural resources, just to name a few.
- **Weaknesses:** are the factors or characteristics that place the entity at a disadvantage when compared to others. Weaknesses are also internal factors, such as financial weakness, inflexible technological stack, shortage of necessary expertise, expensive human and other resources etc.
- **Opportunities:** are elements or trends that the entity or concept could profit from. Opportunities are external factors arising from many sources such as technological innovations, new social trends, or an immature market. Moreover, opportunities may be tangible (e.g. products or solutions) or intangible, such as enhancing reputation or branding.
- **Threats:** elements which may cause trouble for the entity. Threats are external factors that may include restrictive regulation, new competitors, and potential loss of reputation, just to name a few.

The SWOT analysis is based on the results of in-depth dimensions research and on the analysis and assessment of the secondary sources, including reports and statistics published by leading industry associations or organisations, as well as information extracted from the extensive list of online information sources listed in Sources and references (preceding the annexes).

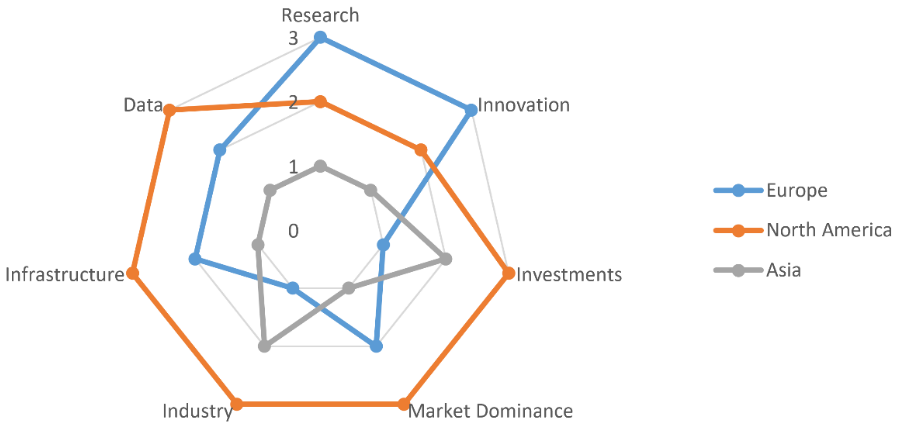
3.3. Comparative position of the European LT market

The graphical summary of the comparative ranking below provides a visual overview of the relative positions (based on a score from one to three) of the major economic regions (markets) within the dimensions we have selected to juxtapose.

3.3.1. Comparative position of the European MT market

Figure 34 provides a summary view of the comparative position of the European MT market versus the regions North America and Asia.

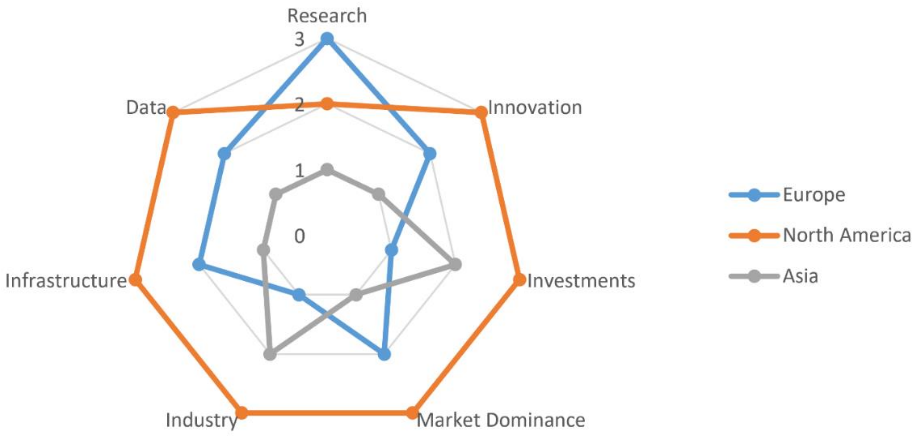
Figure 34 Comparative position of European MT market versus North America, Asia



3.3.2. Comparative position of the European speech technology market

Figure 35 provides a summary view of the comparative position of the European speech technology market versus the regions North America and Asia.

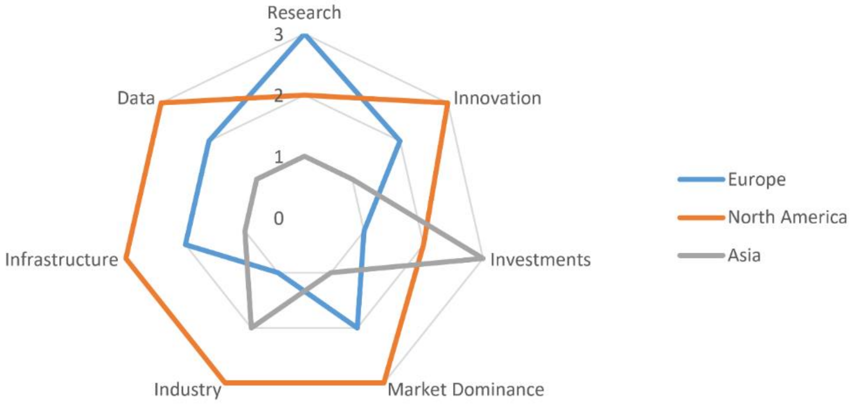
Figure 35 Comparative position of European speech technology market versus North America, Asia



3.3.3. Comparative position of the European search market

Figure 36 provides a summary view of the comparative position of the European search market versus the regions North America and Asia.

Figure 36 Comparative position of European search market versus North America, Asia



3.4. Research

In this section, research activities for all three areas (MT, speech technologies and search technologies) of LT are quantified by reviewing and engaging in a deeper analysis of the number and provenance of the following criteria, which were selected as objective indicators:

- Research centres working on a selected area
- Research publications
- Organisational infrastructure (e.g. associations, networks and research infrastructures)

We analysed publicly available information about research centres in different countries for each of the three areas. Since information about the size of research institutions (e.g. number and qualification of researchers, research budget, number of projects) is not available in public sources, research institutions are not weighted for size of business.

In this study, we performed research on publications in the Scopus database.[2] Research publications describe both academic and industrial research results. However, it could be that industrial research is not revealed completely, since not all industrial research results are made public. The options for information sources of scientific publications that could be used in our study are rather limited. Although research papers in the fields of our study are collected by several online repositories - SCOPUS, Web of Science (WoS), DBPL, Google Scholar, arXiv, CiteSeer – only SCOPUS and WoS provide the information of analytical tools that were needed for this study. Both SCOPUS and WoS are well established academic citation indexes that are widely used to assess the outcome and impact of the scientific work. However, SCOPUS has better coverage for the fields of our study.

To calculate the regional distribution of publications, the methodology used by Scopus to count the distribution of publications between countries was applied, i.e. if authors of the same publication represent different regions, then this publication is counted for each region that the authors represent.

3.4.1. Research in MT

Europe has achieved the highest score between the analysed regions due to the long-term EU policy of multilingualism that encourages rich linguistic diversity and is illustrated by a large number of research centres (60 vs. 28 in Asia and 24 in North America), and a significant number of scientific publications. Particularly notable outcomes of European research include the *Moses*[3] statistical

[2] The Scopus database can be found in https://www.scopus.com.

[3] http://www.statmt.org/moses

machine translation toolkit, and the *Nematus*[4] and *Marian*[5] neural machine translation toolkits that are widely used by the research community and the industry.

Summarised by criteria the results look as following:

- Europe has the biggest number of research centres, almost twice as many as North America, which is second.
- The number of publications in top conferences and journals is very similar for North America and Europe. Moreover, from the top 20 authors half are European, and only 1 is from the US. However it should be noted that the trend of the last two years is an increase in the number of researches in Asia.
- There are rich possibilities in Europe for cooperation and networking, placing it ahead. North America comes second.

The ranking of the markets within this dimension is presented in Table 5.

Table 5 Market relative score in research in MT

Market	Relative Score
Europe	3
North America	2
Asia	1

3.4.1.1. Research centres

The recent Wikipedia article *"List of research laboratories for machine translation"*[6] lists 113 institutions, from which 91 are in the scope of our study. This list includes academic, governmental and corporate sites. This list confirms our findings of strong research in Europe, as it includes 47 academic research centers in Europe and only 18 in America and 9 in Asia (see Table 6).

[4] https://github.com/EdinburghNLP/nematus

[5] https://github.com/marian-nmt/marian

[6] https://en.wikipedia.org/wiki/List_of_research_laboratories_for_machine_translation, retrieved on 12.07.2018

Table 6 Number of research laboratories for MT in different regions

	ACADEMIC	GOVERNMENTAL	CORPORATE	TOTAL
EUROPE	47	1	6	**54**
ASIA	9	4	1	**14**
AMERICA	18	1	4	**23**
TOTAL	**74**	**6**	**11**	**91**

A higher number of European research centres compared to North American research centres is also reflected in the membership of the European Association of Machine Translation (EAMT)[7] that lists 43 R&D groups and 16 corporate members. The American Association of Machine Translation (AMTA) lists 15 academic research organisations and 6 industrial research labs.[8] The Asia-Pacific Association for Machine Translation has 32 corporate members and over 66 individual members.[9]

3.4.1.2. Publications

Research excellence is usually confirmed by the number of publications in top conferences and journals.

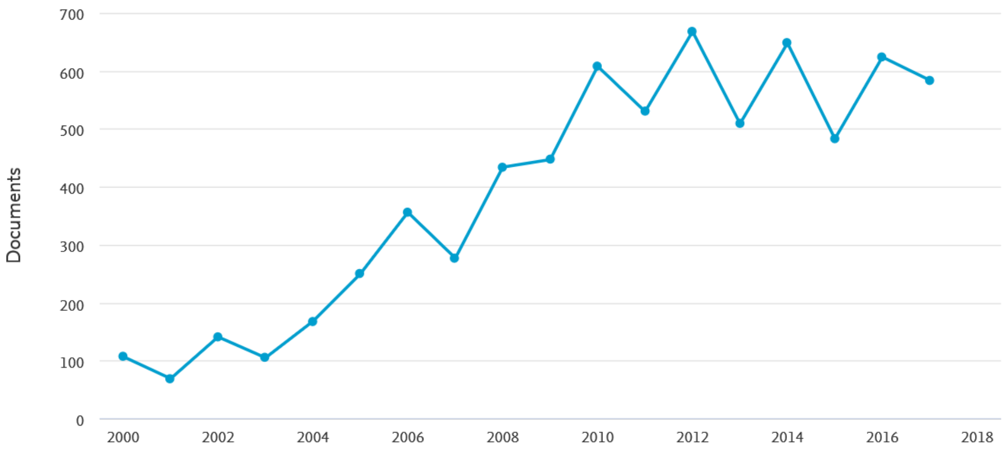
We performed research on publications in the Scopus database, in which we analysed publications retrieved by querying for “machine translation” in title, abstract and keywords. Figure 37 shows the number of publications for the time period 2000-2017 (7008 in total), clearly demonstrating the increase of interest in this topic in the first decade of this century and the relatively stable number of publications in this decade.

[7] http://www.eamt.org/, retrieved on 12.07.2018

[8] https://amtaweb.org/resources, retrieved on 12.07.2018

[9] http://www.aamt.info/english/about/01.php, retrieved on 12.07.2018

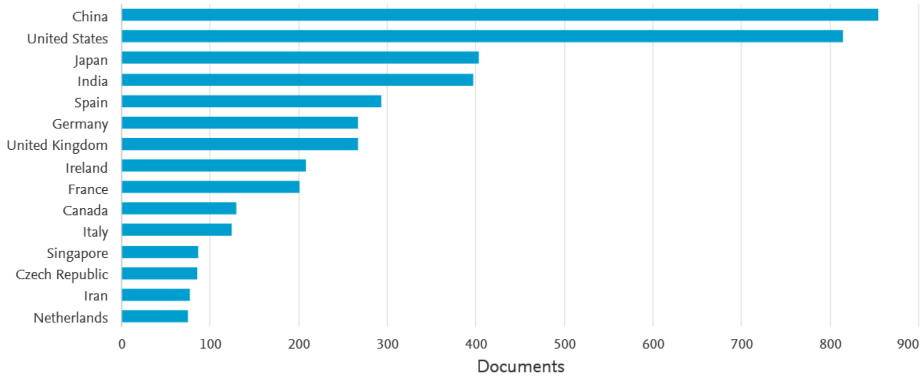
Figure 37 Number of publications for "machine translation" (2000-2017)



When querying "machine translation" for the years 2010-2018, we found 4931 publications, 4723 of these publications are from the countries/regions studied in Task 2 (on July 10, 2018). Publications on CAT tools were not included and analysed in this study, because the number of publications only[10] on CAT tools in Scopus DB for 2010-2018 is very small (only 149 additional publications or about 3% were found).

Figure 38 shows the top 15 countries that have the highest number of publications for the years 2010-2018. We can see that the leader is China (854 publications), followed by the United States (814 publications), and Japan (403 publications). The list of the top 15 countries includes such European countries as Spain (293 publications), Germany (266 publications), UK (266 publications), Ireland (208 publications), France (200 publications), Italy (124 publications), Czech Republic (85 publications), and The Netherlands (75 publications).

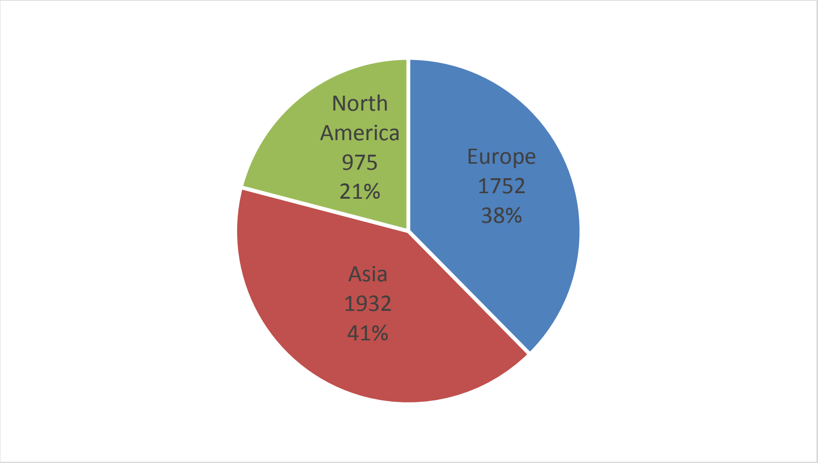
Figure 38 Number of MT-related publications in Scopus: top 15 (2010-June 2018)



[10] These publications do not mention "machine translation" in title, abstract or keywords.

When the number of publications is compared between North America, Asia and Europe, the leader is Asia with 1932 publications, followed by Europe with 1752 publications and North America with 975 publications (Figure 39[11,12]).

Figure 39 Distribution of publications between regions (2010-2018)

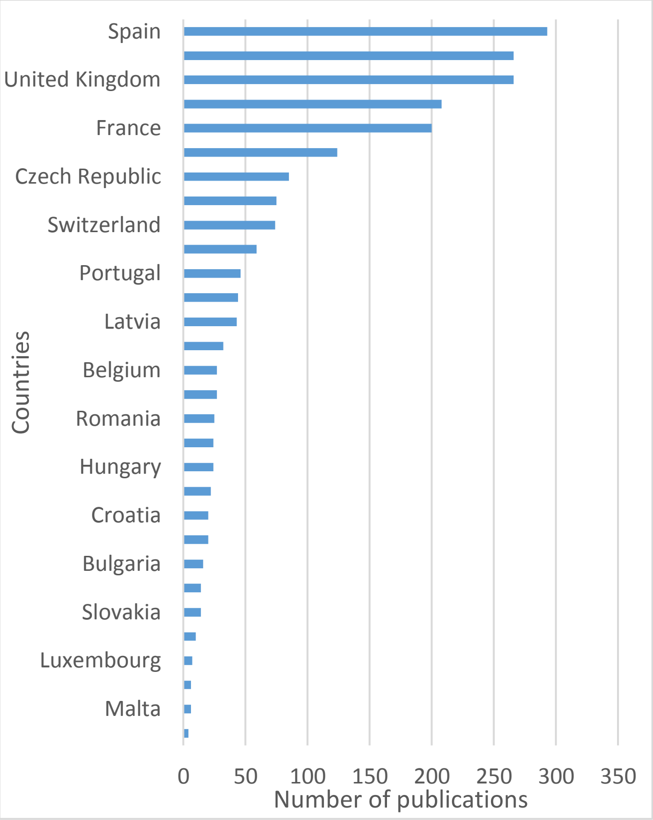


When different European countries are analysed, we see that most of the research publications have been produced by five countries – Spain (293 publications), Germany (266 publications), UK (266 publications), Ireland (208 publications), and France (200 publications), as is illustrated in Figure 40.

[11] When the regional distribution is calculated, we follow the methodology of Scopus DB: if the publication is written by several authors from different regions, the publication is counted for all regions represented by the authors.

[12] Only countries from Annex B included.

Figure 40 Number of MT-related publications from European countries in Scopus (2010-June 2018)



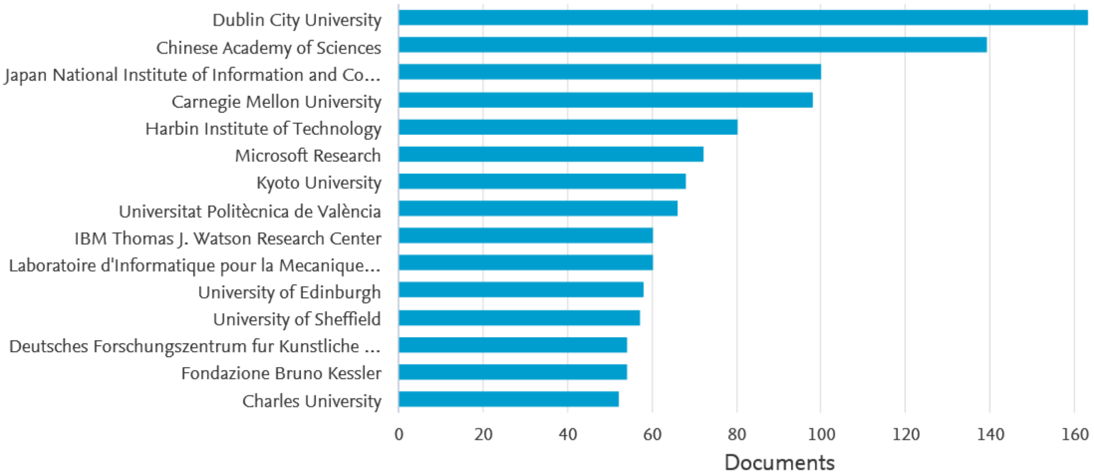
When top 20 authors are compared, half (10) of the most prolific authors are currently working in Europe, 9 in Asia and only one in America (see Table 7).

Table 7 Authors publishing on MT (2010 - June 2018) with +25 publications (top 20) in Scopus

	Author name	Number of publications	Country	Region
1.	Way, A.	75	Ireland	Europe
2.	Sumita, E.	67	Japan	Asia
3.	Liu, Q.	55	Ireland	Europe
4.	Casacuberta, F.	45	Spain	Europe
5.	Specia, L.	44	UK	Europe
6.	Zhao, T.	40	China	Asia
7.	Utiyama, M.	35	Japan	Asia
8.	Xiong, D.	35	China	Asia
9.	Zhang, M.	34	China	Asia
10.	Zhou, M.	34	US	America
11.	Ney, H.	31	Germany	Europe
12.	Yvon, F.	31	France	Europe
13.	Neubig, G.	29	Japan	Asia
14.	Zong, C.	29	China	Asia
15.	Liu, Y.	28	China	Asia
16.	Turchi, M.	28	Italy	Europe
17.	Van Genabith, J.	28	Germany	Europe
18.	Costa-Jussà, M.R.	27	Spain	Europe
19.	Finch, A.	26	Japan	Asia
20.	Toral, A.	26	Netherlands	Europe

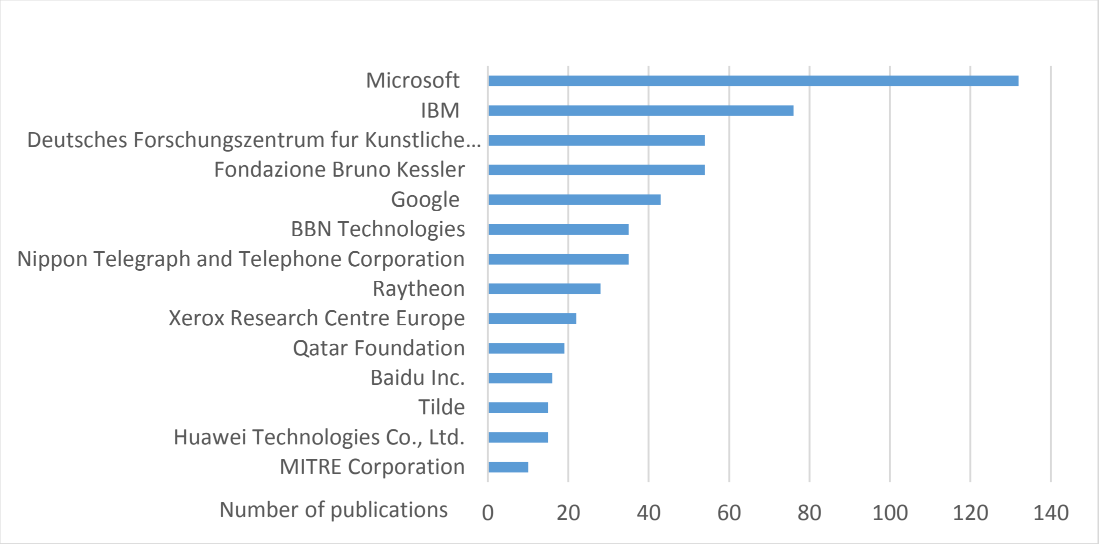
When results are compared by organisations, there are 8 institutions from Europe, 4 from Asia, and 3 from America among the published top 15 (see Figure 41).

Figure 41 Top 15 organisations that published papers on MT (2010-June 2018) in Scopus



When only industry and privately financed organisations are compared, global companies – *Microsoft (132), IBM (76)* and *Google (43)* with headquarters in the US*,* together with DFKI (54) and FBK (54) form the top 5 (see Figure 42).

Figure 42 Industry, privately financed organisations that published on MT (2010-June 2018) in Scopus



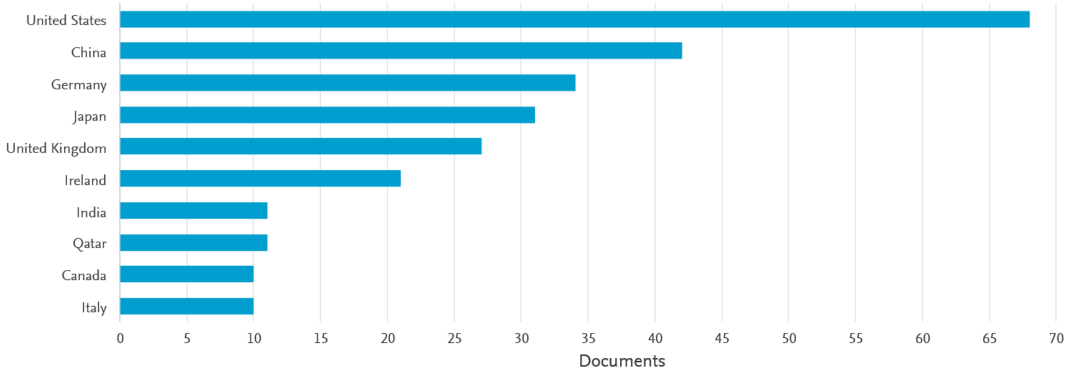
These findings corroborate the conclusions of the META-NET SRIA (Multilingual Europe Technology Alliance - Network Strategic Research Agenda) that "*Europe is the most appropriate place for accomplishing the needed breakthroughs in fundamental and applied research and technology evolution. Europe has more than 2,500 small and medium sized enterprises in language, knowledge and interface technologies, and more than 5,000 enterprises providing language services that can be improved and extended by technology. In addition, it has a long-standing R&D tradition with over 800* *centres performing scientific and technological research on all European and many non-European languages."* (META-NET, 2013: 3).

We also analysed conference proceedings of five important computational linguistics conferences by querying for "machine translation":

- ACL: Annual Meeting of the Association for Computational linguistics (ACL), proceedings of 2010-2017 are included in Scopus
- COLING: Conference on Computational Linguistics, proceedings of 2010, 2012, 2013, 2014, 2016 are included in Scopus
- EACL: the European Chapter of the Association for Computational linguistics, proceedings of 2010, 2012, 2013, 2014, 2017 are included in Scopus
- NAACL: North American Chapter of the Association for Computational Linguistics, proceedings of 2010, 2012, 2013, 2015 and 2016 are included in Scopus
- NIPS: Annual Conference on Neural Information Processing Systems, proceedings of 2010-2017 are included in Scopus

We found more recent (2015-2017[13]) papers from the United States (68) and China (42), but less from Germany (34), the United Kingdom (27), Ireland (21) and other European countries (Figure 43).

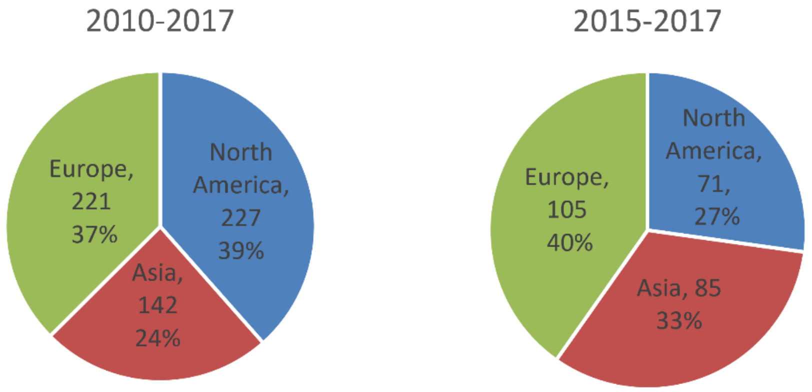
Figure 43 Number of papers/country, "machine translation", ACL/COLING/EACL/NAACL/NIPS (2015-2017)



While US authors have more publications than authors from each single EU or Asian country, European countries are still leaders when the regional distribution of publications is compared (Figure 44).

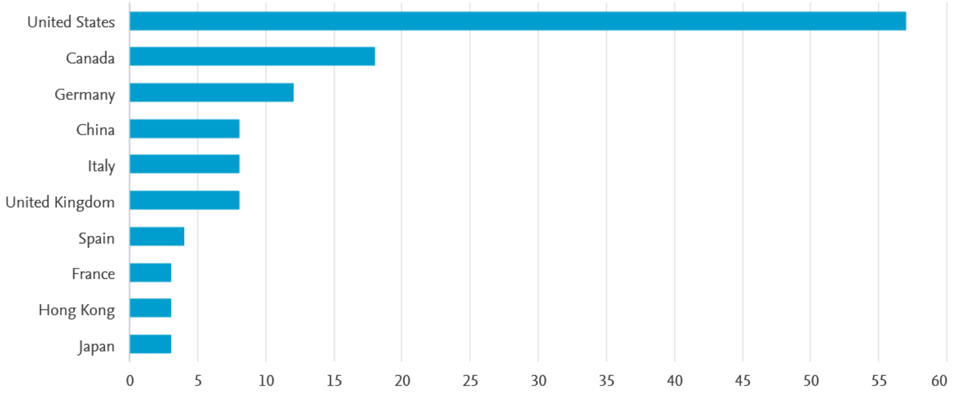
[13] The abovementioned conferences were not indexed for the year 2018 at the time of this study.

Figure 44 Distribution of publications on "machine translation" in ACL/COLING/EACL/NAACL/NIPS



Finally, we analysed the top 100 most cited papers in the Scopus database that are written from 2010 till June 2018 (the list of top 10 publications is included in Annex C) on "machine translation". Figure 45 shows countries with at least 3 publications in the top 100 list. The leader is US (57 publications), followed by Canada (18 publications) and Germany (12 publications).

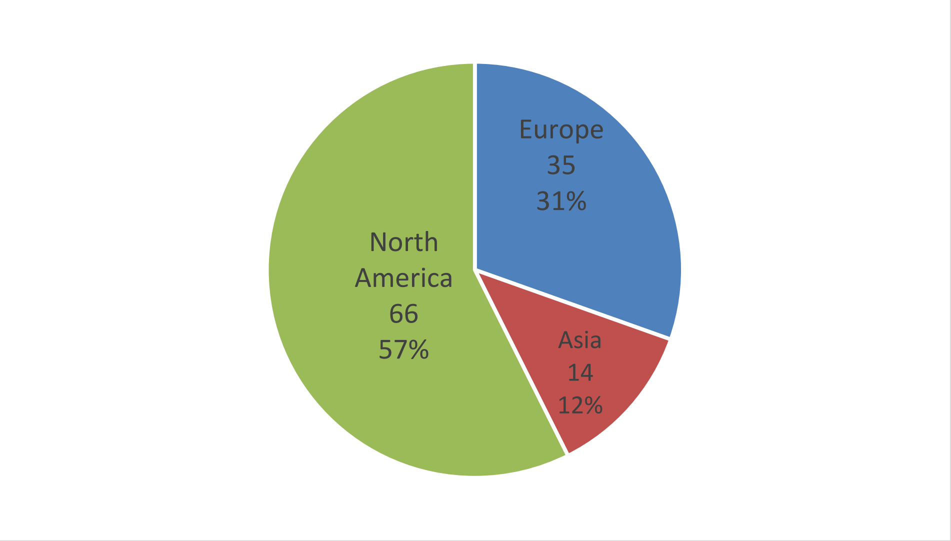
Figure 45 Distribution top 100 most cited MT publications, Scopus 2010-2018, countries >= 3 publications



When the regional distribution of publications is analysed (see Figure 46), we can see that most (57%) of the top 100 publications include authors from North America, while 31% of the publications include authors from Europe and only 12% include authors from Asia.[14]

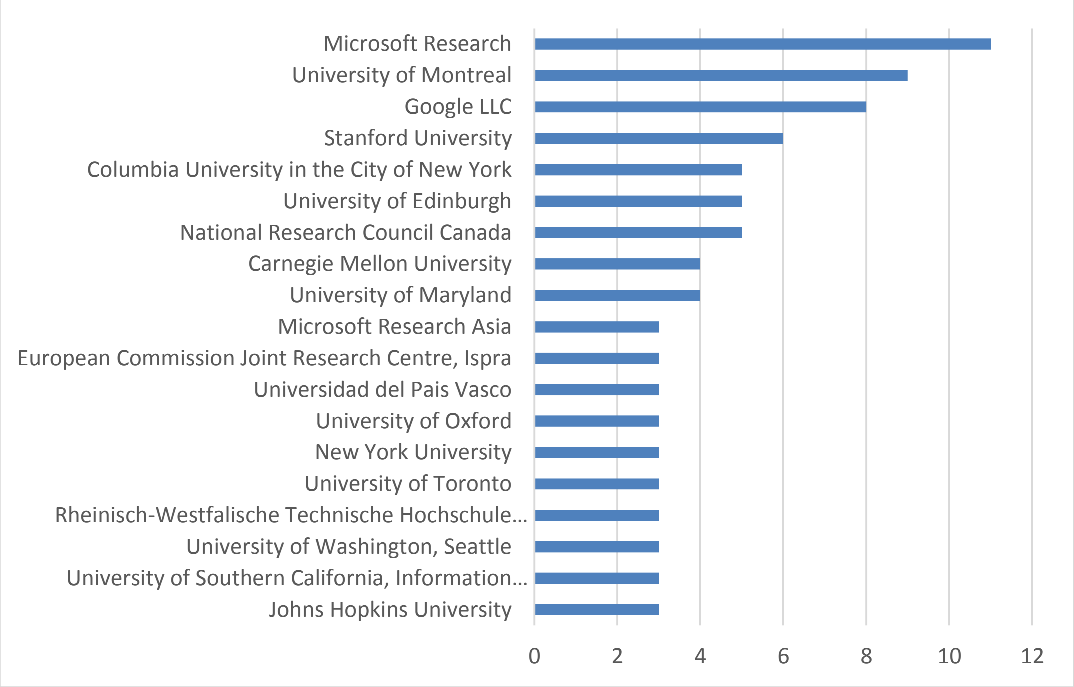
[14] The authors of one publication could be from different countries/regions, therefore the total number of publications in this diagram exceeds 100.

Figure 46 Distribution top 100 most cited MT publications, Scopus 2010-2018, regions



When the authors' affiliations are compared (see Figure 47), the leader is Microsoft (11 publications), followed by University of Montreal (9 publications) and Google (8 publications).

Figure 47 Distribution top 100 most cited MT publications, Scopus 2010-2018, institutions >= 3 publications



The quality of EU research is also demonstrated through shared tasks of WMT (conference/workshop on Machine Translation). For the WMT 2018 news translation task 103 systems were submitted from 32 institutions (Bojar et al., 2018). MT systems were built for translation between English and

Chinese, Czech, Estonian, German, Finnish, Russian and Turkish. Table 8[15] presents simplified results of WMT 2018, showing the top 3 systems for each translation direction. From the 42 systems, 16 are from Europe, 12 from Asia, 2 from North America, 11 are not identified. Among the identified systems, those of Asian MT developers in most cases showed better results for translation from/to Chinese, Russian and Turkish, while those of European MT developers were mostly the best for translation from/to EU languages. For English to German, US companies achieved the best results.

Table 8 Top 3 MT systems for news translation task of WMT 2018

Language pair	Institution	Country	Region
Chinese->English	NiuTrans Co., Ltd.	China	Asia
	Online-B		
	University of Cambridge	UK	Europe
Czech→English	Charles University	Czech Republic	Europe
	University of Edinburgh	UK	Europe
	Online-B		
English→Chinese	Tencent	China	Asia
	Unisound	China	Asia
	Global Tone Communication Technology	China	Asia
English→Czech	Charles University	Czech Republic	Europe
	University of Edinburgh	UK	Europe
	Online-B		
English→Estonian	Tilde	Latvia	Europe
	NICT	Japan	Asia
	Tilde	Latvia	Europe
English→Finnish	NICT	Japan	Asia
	University of Helsinki	Finland	Europe
	University of Edinburgh	UK	Europe
English→German	Facebook AI Research	US	North America
	Online-B		
	Microsoft	US	North America
English→Russian	Alibaba Group	China	Asia
	Online-G		
	Online-B		
English→Turkish	Online-B		
	University of Edinburgh	UK	Europe
	Alibaba Group	China	Asia
Estonian→English	Tilde	Latvia	Europe
	NICT	Japan	Asia
	Tilde	Latvia	Europe
Finnish→English	NICT	Japan	Asia
	University of Helsinki	Finland	Europe
	University of Edinburgh	UK	Europe
German→English	RWTH Aachen	Germany	Europe
	University of Cambridge	UK	Europe
	NTT Corporation	Japan	Asia
Russian→English	Alibaba Group	China	Asia
	Online-B		
	Online-G		
Turkish→English	Online-G		
	Online-A		
	Alibaba Group	China	Asia

[15] https://slator.com/academia/heres-what-happened-at-the-worlds-biggest-machine-translation-conference

3.4.1.3. *Organisational and Research Infrastructure*

The Machine Translation community is well represented and has rich networking opportunities through various associations such as:

- Asia-Pacific Association for Machine Translation:[16] 32 corporate members (MT and MT-related software developers, MT system distributors, MT research institutes, etc.) and more than 65 individual members (researchers, developers, distributors, translators, etc.).
- Association for Machine Translation in the Americas:[17] 15 academic research organisations and 6 industrial research labs (Microsoft Research, IBM Research, SRI Research, Raytheon BBN Techs and SDL Language Cloud).
- European Association for Machine Translation:[18] 43 R&D groups and 16 corporate members (e.g. CrossLang, Kantan MT, Pangeanic, PROMT, Tilde).

Several associations have been established for collaboration on MT from different perspectives:

- The network established by European Language Resource Coordination (ELRC)[19] aims to maintain and coordinate the language resources in official languages of the EU and associated countries that help to improve the quality, coverage and performance of automated translation solutions in the context of current and future CEF digital services.
- LT Innovate[20] is the Language Technology Industry Association. The LT Directory provides information about suppliers, integrators, users, researchers and language service providers.
- TAUS[21] is a language data network with more than 100 members (industry, research, associations) “for sharing knowledge, metrics and data that help stakeholders in the translation industry develop a better service”.
- META-NET[22] is a Network of Excellence dedicated to the technological foundations of a multilingual European information society. META-NET builds the Multilingual Europe Technology Alliance by bringing together researchers, commercial technology providers, private and corporate language technology users, language professionals and other information society stakeholders.

[16] Asia-Pacific Association for Machine Translation

[17] Association for Machine Translation in the Americas

[18] European Association for Machine Translation

[19] http://www.lr-coordination.eu

[20] http://www.lt-innovate.org

[21] https://www.taus.net

[22] http://www.meta-net.eu/front-page?set_language=en

While Europe lacks computing power and internet infrastructure as compared to North America (for details, please, see Section 3.9), it does have a good LT research infrastructure, e.g. CLARIN (European Research Infrastructure for Language Resources and technology[23]), ELEXIS (European Lexicographic Infrastructure[24]), DARIAH (Digital Research Infrastructure for ARTs and Humanities[25]), etc.

Researchers and developers can also benefit from MT development tools and platforms mostly created in Europe. The Moses toolkit (Koehn et al, 2007)[26] is a platform for training statistical machine translation systems for any language pair, it supports inclusion of morphological and syntactic features presented in training data. It has been developed and maintained through several FP6 and FP7 projects (e.g. EuroMatrix, EuroMatrixPlus, LetsMT, etc.) and supported by European Universities. The Nematus[27] toolkit (Sennrich et. al, 2017) is a widely used open source toolkit for neural machine translation (NMT), which was developed with support from the Horizon 2020 projects QT21, TraMOOC, HimL and SUMMA. Recently developed popular alternatives include SOCKEYE[28] (Hieber et al., 2017) by Amazon, fairseq[29] (Gehring et al, 2017a and 2017b) by Facebook, Open-NMT[30] (Klein et al., 2017) by Harvard University and Systran, and Marian[31] (Junczys-Dowmunt et al., 2018), which has mainly been developed at the Adam Mickiewicz University in Poznań and at the University of Edinburgh, supported by the H2020 projects SUMMA, Modern MT, TraMOOC and HimL.

3.4.2. Research in speech technologies

Analysis of research in speech technologies is summarised in Table 9. The EU multilingualism policy and language diversity of Europe are reasons for more research centres in Europe than in other regions. While in general the number of publications is higher for Asia than for Europe, this proportion changes when publications of top conferences are compared, putting Europe in the first place, followed by North America. However it should be noted that the trend of the last two years is an increase in the number of researches in Asia in comparison to Europe and North America.

[23] https://www.clarin.eu

[24] https://elex.is

[25] https://www.dariah.eu

[26] http://www.statmt.org/moses

[27] https://github.com/EdinburghNLP/nematus

[28] https://github.com/awslabs/sockeye

[29] https://github.com/facebookresearch/fairseq

[30] http://opennmt.net

[31] https://marian-nmt.github.io

The state-of-the-art speech recognition toolkit Kaldi[32] is developed internationally, while other popular speech recognition tools are developed in the US. The situation is opposite for speech synthesis tools, popular tools are developed in Europe.

Table 9 Market relative score in research in speech technologies

Market	Relative Score
Europe	3
North America	2
Asia	1

3.4.2.1. Research centres

Research in speech technology occurs in companies and in academic research centres. The International Speech Communication Association (ISCA)[33] lists 176 speech laboratories from 37 countries around the world.[34] Table 10 summarises ISCA listed speech organisations by region. The full list of association members is presented in Annex I.

[32] http://kaldi-asr.org

[33] ISCA Web https://www.isca-speech.org/iscaweb

[34] The list of laboratories unified by the International Speech Communication Association is in http://www.isca-students.org/?q=speechlabs.

Table 10 Number of speech laboratories listed by ISCA association

Region	Number of organisations in countries of this study	Total number of organisations
North America	45	48
Asia	38	44
Europe	70	72
Australia		9
Africa		3

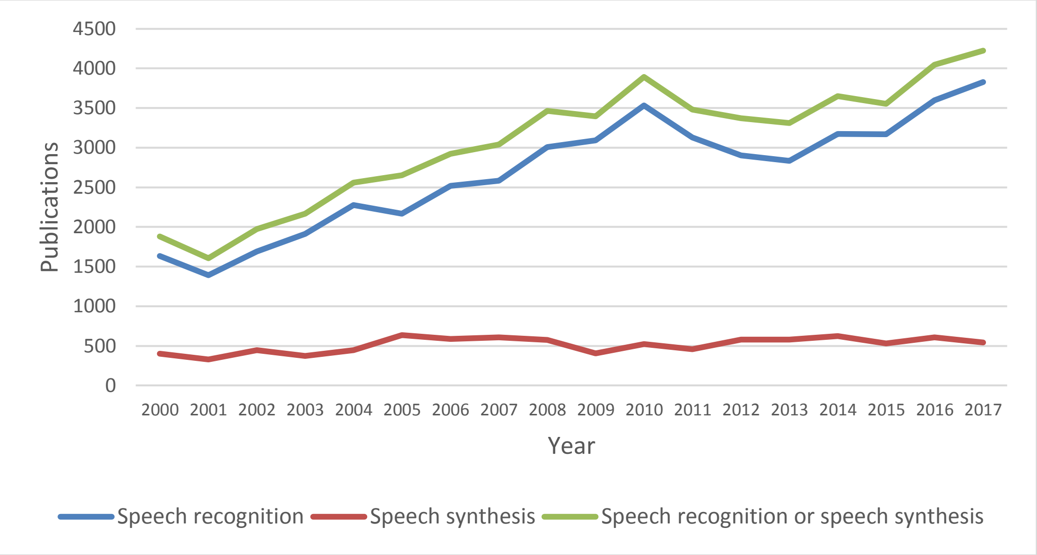
As presented in Table 10, Europe leads with the largest number of laboratories (72 in total out of 176). However, when the 12 UK laboratories, which work on the widely-spoken English language, are excluded, the laboratory per language ratio for the European languages drops to only ~2.72.

3.4.2.2. Publications

The bare numbers of companies, associations and laboratories alone do not present a complete picture of the speech technologies field. The activity of researchers can be revealed from an analysis of their scientific publications (representing the completed work, introduction and dissemination of new ideas) in top conferences or journals. For this reason we have analysed publications found in the Scopus database. The publications were retrieved by querying the database for “speech recognition” OR “text-to-speech” OR “speech synthesis” in the title, abstract and keywords.

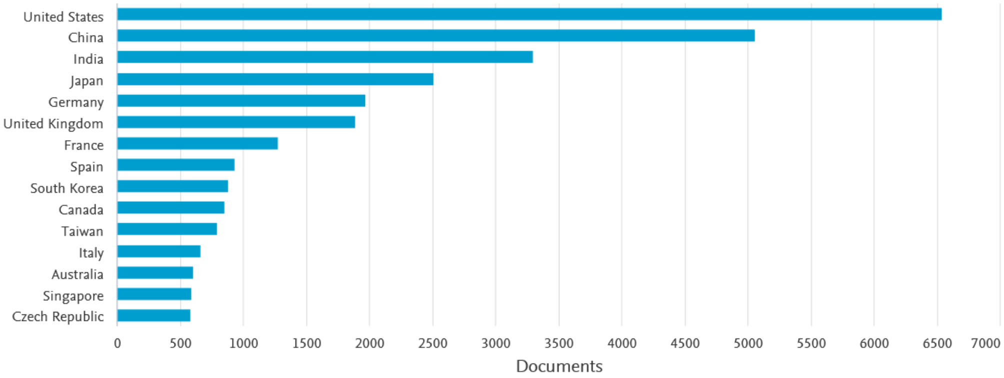
Figure 48 presents the number of publications by year during the period from 2000 to 2017 (55,185 publications in total). The curve clearly demonstrates an increasing interest over the latest years in speech recognition and this gain is mostly due to the recent advances in technology based on more sophisticated/powerful/accurate deep learning methods. At the same time although there is less interest in speech synthesis, this interest remains stable.

Figure 48 Number of publications/year, “speech recognition/text-to-speech/speech synthesis”, Scopus 2010-2017



In this study we analyse publications in the period from 2010 to October 2018 (32,545 publications in total). Figure 49 shows the top 15 countries that have the highest number of publications for this period. The leader is United States (6535 publications), followed by China (5051 publications) and India (3295 publications).

Figure 49 Top 15 countries "speech recognition/text-to-speech/speech synthesis", Scopus 2010-Oct. 2018



When regions are compared,[35] the leader is Asia (41% of the publications), followed by Europe with 11,596 or 34% of the publications, while for 7,811 publications (25%) at least one author is from North America (see Figure 50).

Figure 50 Publications/region, "speech recognition/text-to-speech/speech synthesis", Scopus 2010-Oct. 2018

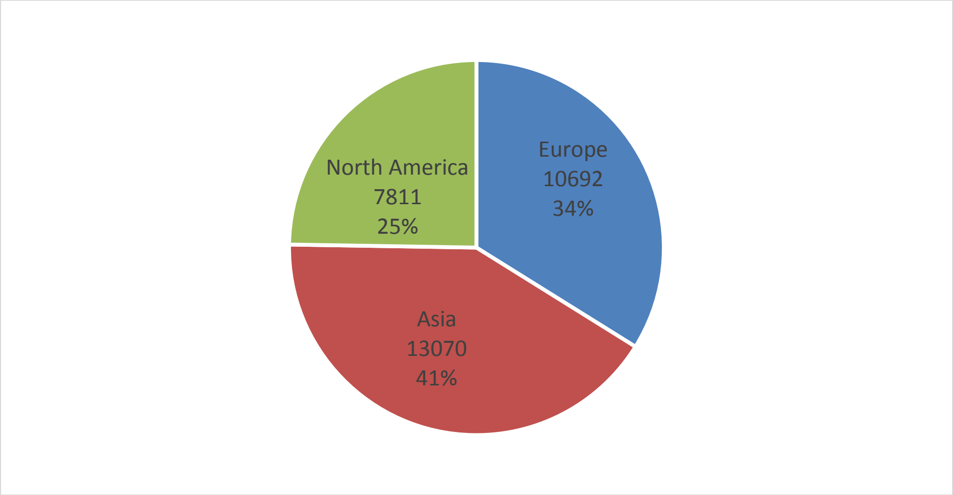
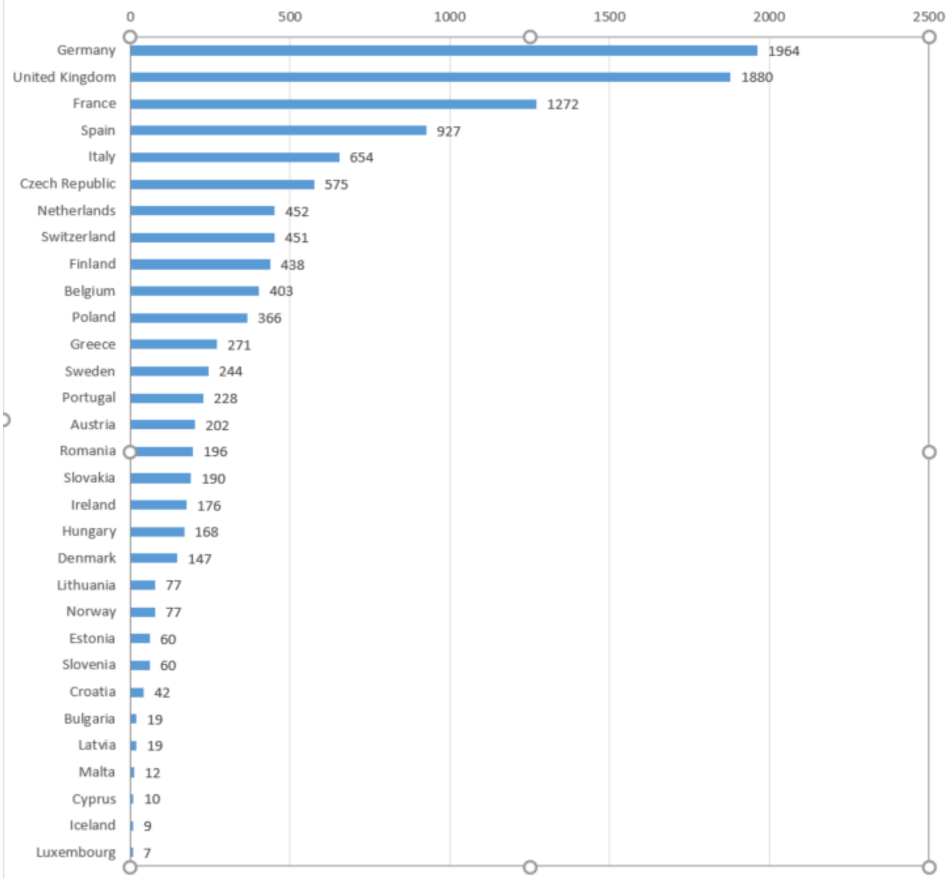


Figure 51 summarises the publications of EU and EFTA countries. The leader is Germany (1,964 publications) followed by UK (1,880 publications) and France (1,272 publications).

[35] Regions and countries included are listed in Annex B.

Figure 51 Publications EU/EFTA, "speech recognition/text-to-speech/speech synthesis", Scopus 2010-Oct. 2018



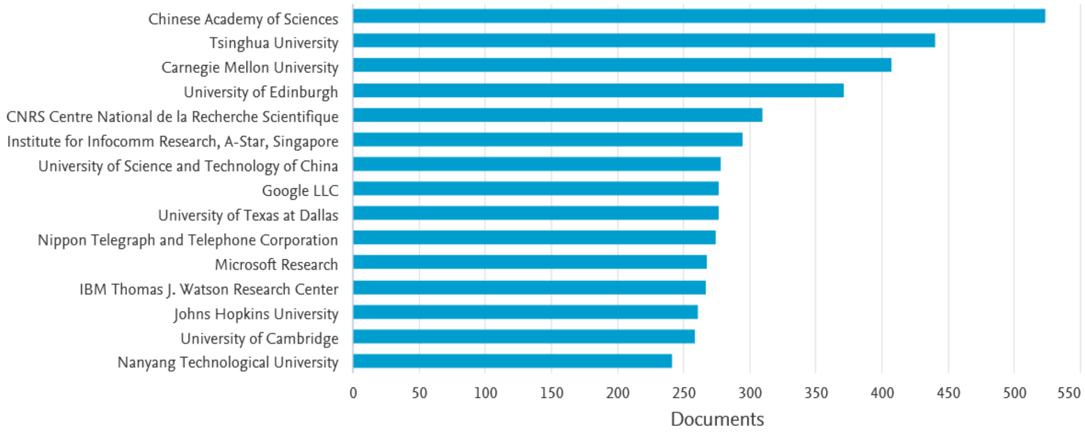
The list of the most productive authors is summarised in Table 11. More than half (11) of the top 20 authors are from Asia, 7 are from Europe and 2 are from America.

Table 11 Top authors "speech recognition/text-to-speech/speech synthesis", Scopus 2010-Oct. 2018

No.	Author	Numb. of publications	Country	Region
1	Li, H.	183	Singapore	Asia
2	Hansen, J.H.L.	161	USA	America
3	Schuller, B.	147	Germany	Europe
4	Gales, M.J.F.	126	UK	Europe
5	Yamagishi, J.	122	Japan	Asia
6	Yan, Y.	122	China	Asia
7	Liu, J.	108	China	Asia
8	Ney, H.	104	Germany	Europe
9	Watanabe, S.	102	Japan	Asia
10	King, S.	99	UK	Europe
11	Rao, K.S.	91	India	Asia
12	Ramabhadran, B.	90	USA	America
13	Chng, E.S.	89	Singapore	Asia
14	Nakatani, T.	89	Japan	Asia
15	Ma, B.	87	Singapore	Asia
16	Patil, H.A.	85	India	Asia
17	Kawahara, T.	80	Japan	Asia
18	Renals, S.	77	UK	Europe
19	Kinnunen, T.	76	Finland	Europe
20	Schultz, T.	72	Germany	Europe

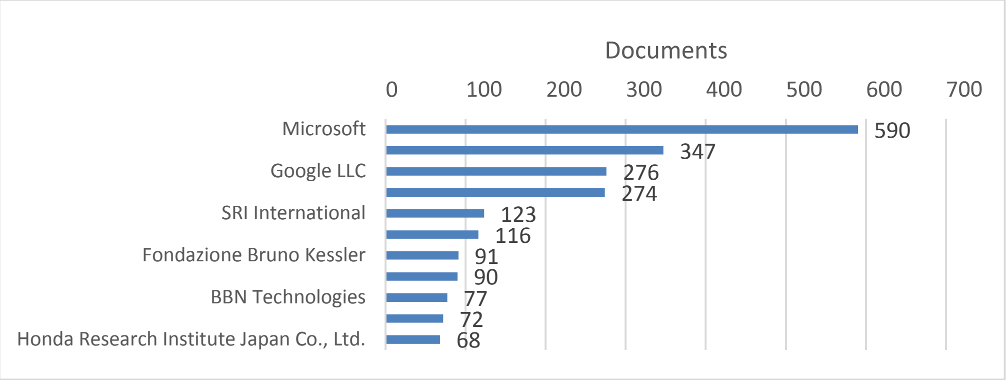
The number of publications from the Scopus database is also compared by institutions (see Figure 52). Among the top 15 institutions, only 3 institutions are from Europe, while 6 are from Asia, and 6 are from North America.[36]

Figure 52 Top organisations "speech recognition/text-to-speech/speech synthesis", Scopus 2010-Oct. 2018



If only companies and private institutions are considered, there are 4 Asian and 6 American and only 1 European institution publishing about speech technologies (see Figure 53).

Figure 53 Companies/private institutions, "speech recognition/text-to-speech/speech synthesis", Scopus 2010-Oct. '18



Finally, we analysed the publications from the three important conferences for speech processing for two periods (2010-2018 and 2016-2018, see Figure 54):

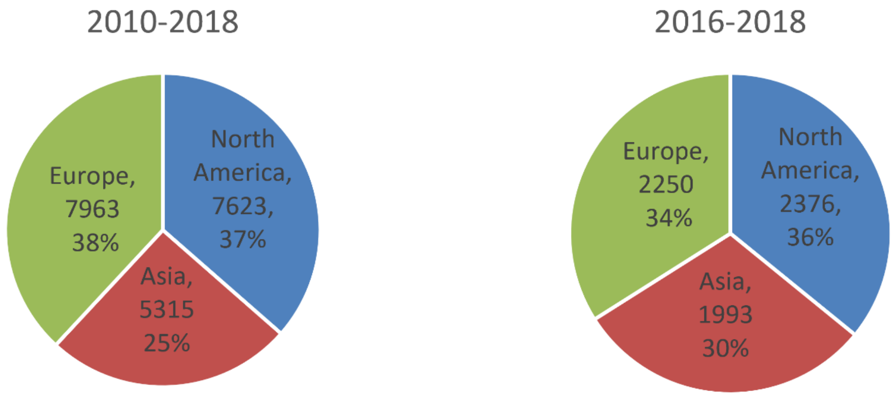
- ICASSP: IEEE International Conference on Acoustics, Speech and Signal Processing, proceedings of 2010-2018 are included in Scopus;
- INTERSPEECH: Conference of the International Speech Communication Association, proceedings of 2010-2017 are included in Scopus;

[36] The global organisations are attached according to their headquarters.

- ASRU: IEEE Automatic Speech Recognition and Understanding Workshop, proceedings of 2011, 2013, 2015, 2017 are included in Scopus.

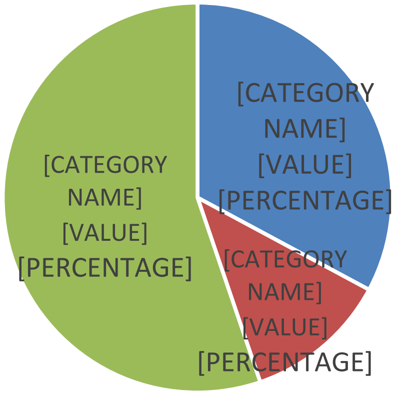
While the proportion of publications from North America has changed a bit, the number of publications from Europe has decreased from 38% to 34%, while for Asia this proportion has increased. It also needs to be mentioned that the proportion of publications presented in Figure 54 is the same for Europe, while for North America and Asia a significant difference is observed.

Figure 54 Publications from INTERSPEECH/ASRU/ICASSP



Finally, we analysed the top 100 most cited papers in the Scopus database written from 2010 till 2018 (the list of top 10 publications is included in Annex D) on automatic speech recognition and speech synthesis (Figure 55). More than half (55%) of these publications have at least one author from North America, while 33% of the publications include contributions from European authors, but only 12% of the publications have authors from Asia.

Figure 55 Authorship, top 100 most cited, "speech recognition/text-to-speech/speech synthesis", Scopus 2010-2018



3.4.2.3. Organisational and research infrastructure

There are two main professional associations for speech communication science and technology:

- The **International Speech Communication Association (ISCA)**[37] is the speech technology research association that promotes the study and application of automatic speech processing. The association organises the annual INTERSPEECH conference. Out of the 20 listed associations collaborating with ISCA[38] (and including the global ones), half are European, i.e. unifying rather small communities working on various European languages, e.g., AISV (research association with the main focus on the Italian language), SFA (focus on French), UAsIPPR (focus on Ukrainian), etc.
- The **IEEE Signal Processing Society**[39] is a professional society for signal processing scientists and professionals. The society was founded in 1948. The network includes signal processing engineers, industry professionals, academics, and students from about 100 countries (more than 19,000 members). Approximately 40% of members are from the US, 27% from Europe, Africa and Middle East, 27% are from Asia, 3% from Canada and 3% from Central/South America[40]. While 54% members are from academia and 46% from the industry, for the US the distribution is different – 35% academic members and 65% from industry.

Where it concerns tools, the state-of-the-art toolkit for speech recognition used in many research laboratories as well as by industry is Kaldi (Povey et al, 2011),[41] other options include HTK (originally developed at Cambridge University, currently Microsoft retains the copyright to the original HTK code),[42] CMUSphinx (Carnegie Mellon University)[43] and EESEN[44] (Miao et al., 2015; Carnegie Mellon University).

For speech synthesis two toolkits developed at the University of Edinburgh - Merlin (Wu et al, 2016)[20] and Festival[21] - as well as proprietary tools, are used (an overview of the latest techniques is presented in Section 3.5.2.3).

3.4.3. Research in search technologies

Information retrieval from text documents is an active and stable research topic in all three regions of our analysis. When the number of research organisations is compared, Europe has a leading

[37] ISCA Web https://www.isca-speech.org/iscaweb

[38] https://www.isca-speech.org/iscaweb/index.php/liaison/professional-organisations

[39] IEEE Signal Processing Society https://signalprocessingsociety.org

[40] State of the Society: https://signalprocessingsociety.org/sites/default/files/uploads/our_story/docs/State_of_the_Society_ICASSP_2018.pdf

[41] http://kaldi-asr.org

[42] http://htk.eng.cam.ac.uk

[43] https://cmusphinx.github.io

[44] https://github.com/srvk/eesen

position, while the number of publications is higher for Asia than for Europe. However, this proportion changes when publications of top conferences are compared, putting Europe in the first place followed by North America. However it should be noted that the trend of the last two years is an increase in the number of researches in Asia in comparison to Europe.

At the same time, most of the industrial research is performed in companies with headquarters in the US (Figure 56). The US also has several initiatives related to cross-lingual search in low-resourced languages. Within this dimension the relative strengths and weaknesses of the markets are presented in Table 12.

Table 12 Market relative score in research in search technologies

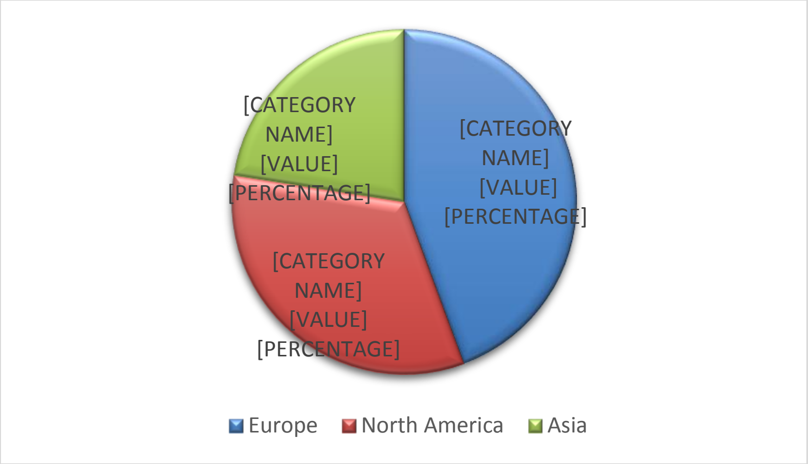
Market	Relative Score
Europe	3
North America	2
Asia	1

3.4.3.1. Research centres

Wikipedia lists only 22 organisations[45] that specify ‘information retrieval’ as category, in addition four organisations are listed: *Alphabet Inc.* (US), *Google* (US)*, Waymo* (US)*,* and *Yandex* (Russia). Therefore, we measure the number of research institutions working in information retrieval by comparing the number of organisations that have published papers on this topic in the field’s most important conferences - SIGIR, WSDM, ICTIR, ECIR and SPIRE - for the time period 2010-2018. In total 160 organisations have been identified (see Annex J). 142 are from the countries included in our analysis. Most of the organisations (63 in total) are from Europe, while there are 47 institutions from North America and 32 from Asia (Figure 56).

[45] https://en.wikipedia.org/wiki/Category:Information_retrieval_organizations

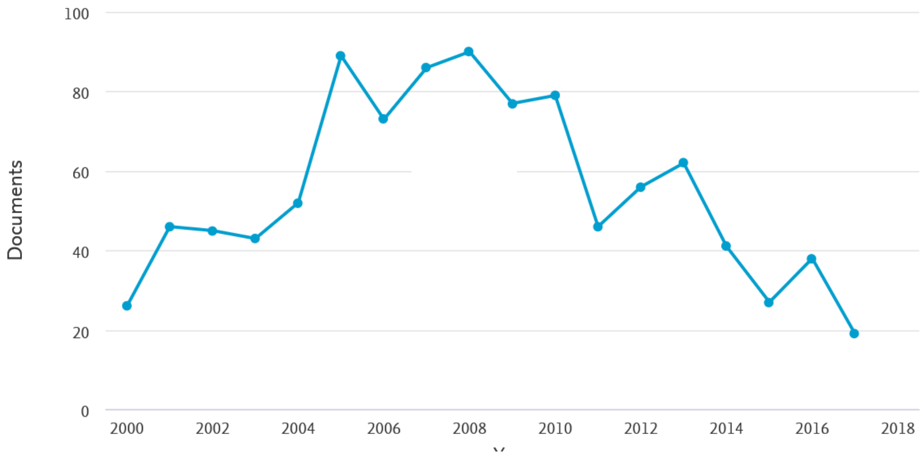
Figure 56 Distribution of research organisations working on information retrieval between regions



3.4.3.2. Publications

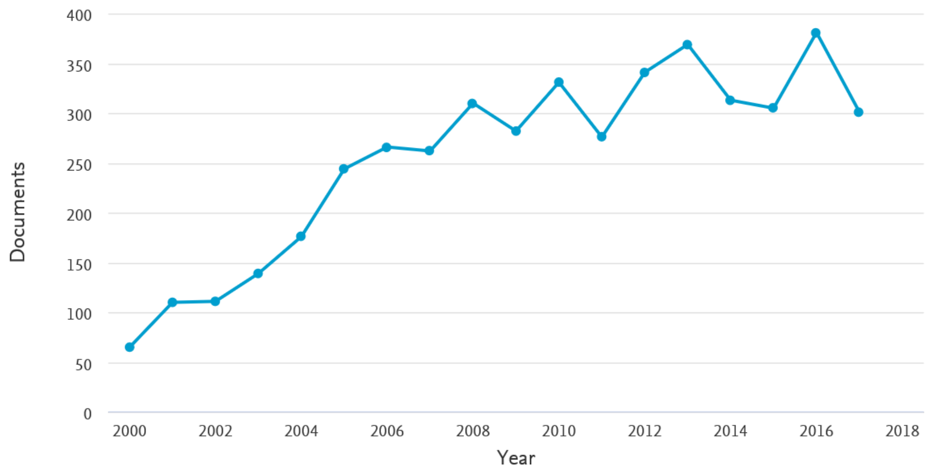
Similarly to the two other LT fields we analysed publications that are indexed in the Scopus database by searching for “cross language information retrieval” or “cross lingual information retrieval” in the title, abstract or keyword field. Figure 57 shows search results for the years 2000-2017, demonstrating increasing research interest in cross-lingual search in the time period 2003-2010, with about 80 indexed publications each year (2005-2010). However, the analysis also demonstrates a constant decrease of publications in 2010-2017, with less than 60 publications per year.

Figure 57 Scopus search “Cross Language/Cross Lingual Information Retrieval”, 2000-2017 (total 995 publications)



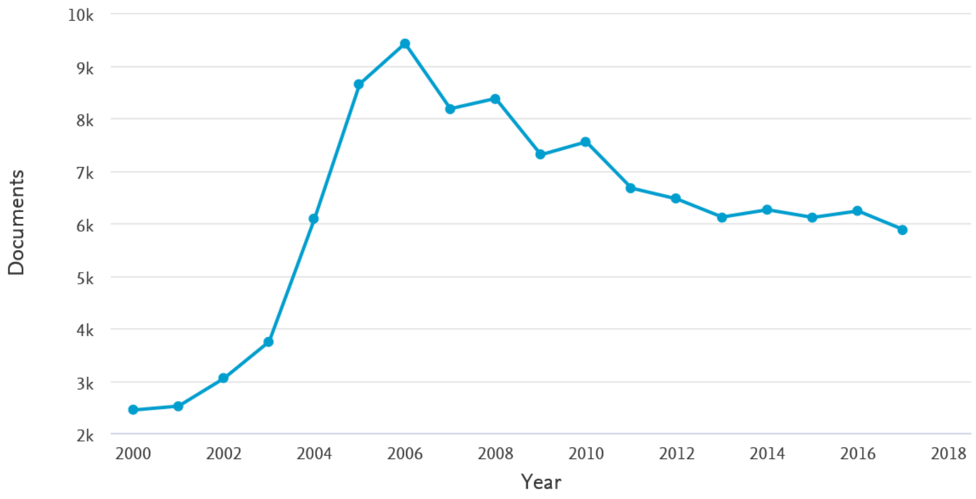
Because of this tendency we widened our query and looked for publications related to the concept of “cross language” solutions, which seems to be a stable and slowly growing research topic. As it is shown in Figure 58, starting from 2010 there are more than 250 publications per year.

Figure 58 Scopus search "Cross Language/Cross Lingual", 2000-2017 (total 4582 publications)



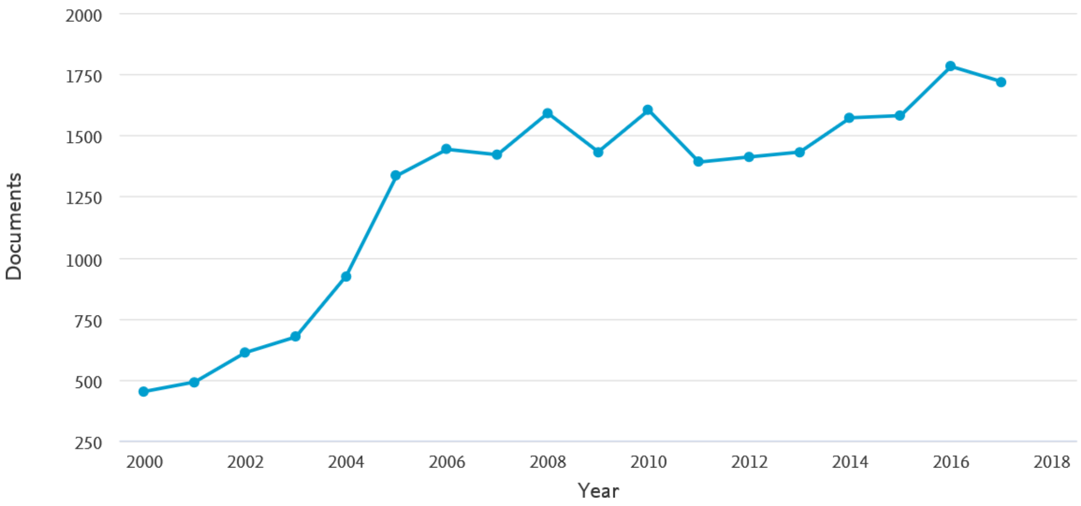
Our analysis of publications related to information retrieval for 2000-2017 also demonstrates more interest in this topic before year 2010 (Figure 59). However, the number of publications after 2010 is rather stable (about 6K a year, compared to 20-60 for CLIR).

Figure 59 Scopus search "Information Retrieval", 2000-2017 (total 111 053 publications)



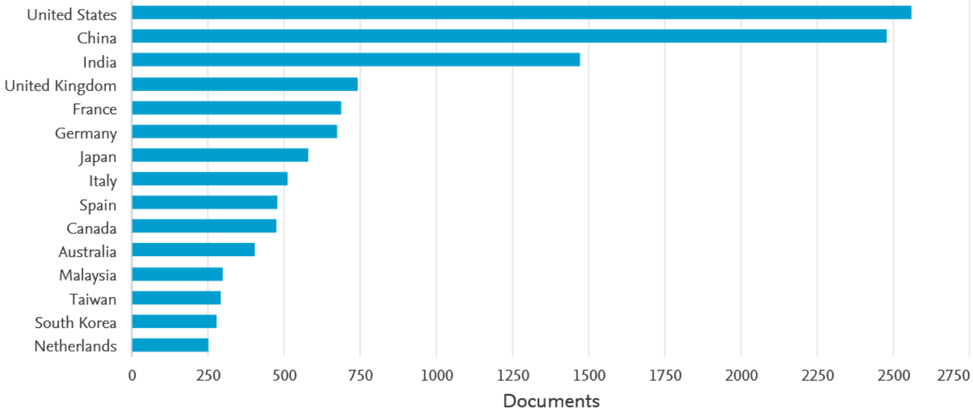
The information retrieval field covers different topics that are not related to search in natural language, thus in this study only publications from Scopus database in which "information retrieval" is mentioned together with "text" or "word" in the title, abstract or keyword field are analysed. Figure 60 shows the result of this query for 2000-2017, demonstrating an increase of interest during the first decade of this century and a rather stable number of publications for 2011-2017.

Figure 60 Scopus search "information retrieval" AND ("text" OR "word"), 2000-2017 (total 22,874 publications)



When querying for "information retrieval" together with "text" or "word" in the title, abstract or keyword field for the years 2010-2018, we found 14,017 publications in total. Figure 61 below shows the top 15 countries that have the highest number of publications for the years 2010-2018. Most of the publications (2557) are from the US, followed by China (2477 publications) and India (1470). The list of the top 15 countries includes such European countries as United Kingdom (741 publications), France (687 publications), Germany (671 publications), Italy (510 publications), Spain (477 publications) and The Netherlands (251 publications). It needs to be mentioned that a significant number of publications are from Australia (403 publications).

Figure 61 Scopus search "information retrieval" AND ("text" OR "word"), number ofpublications top 15, 2010-Nov. '18



When the number of publications is compared between countries of our study in North America, Asia and Europe, the leader is Asia with 4933 publications, followed by Europe with 4394 publications, and North America with 2963 publications (Figure 62).

Figure 62 Scopus search "information retrieval" AND ("text" OR "word"), publications per region, 2010-Nov. '18

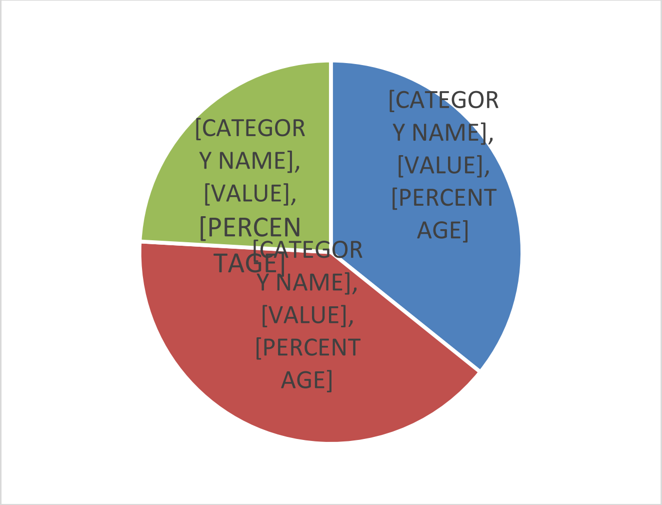
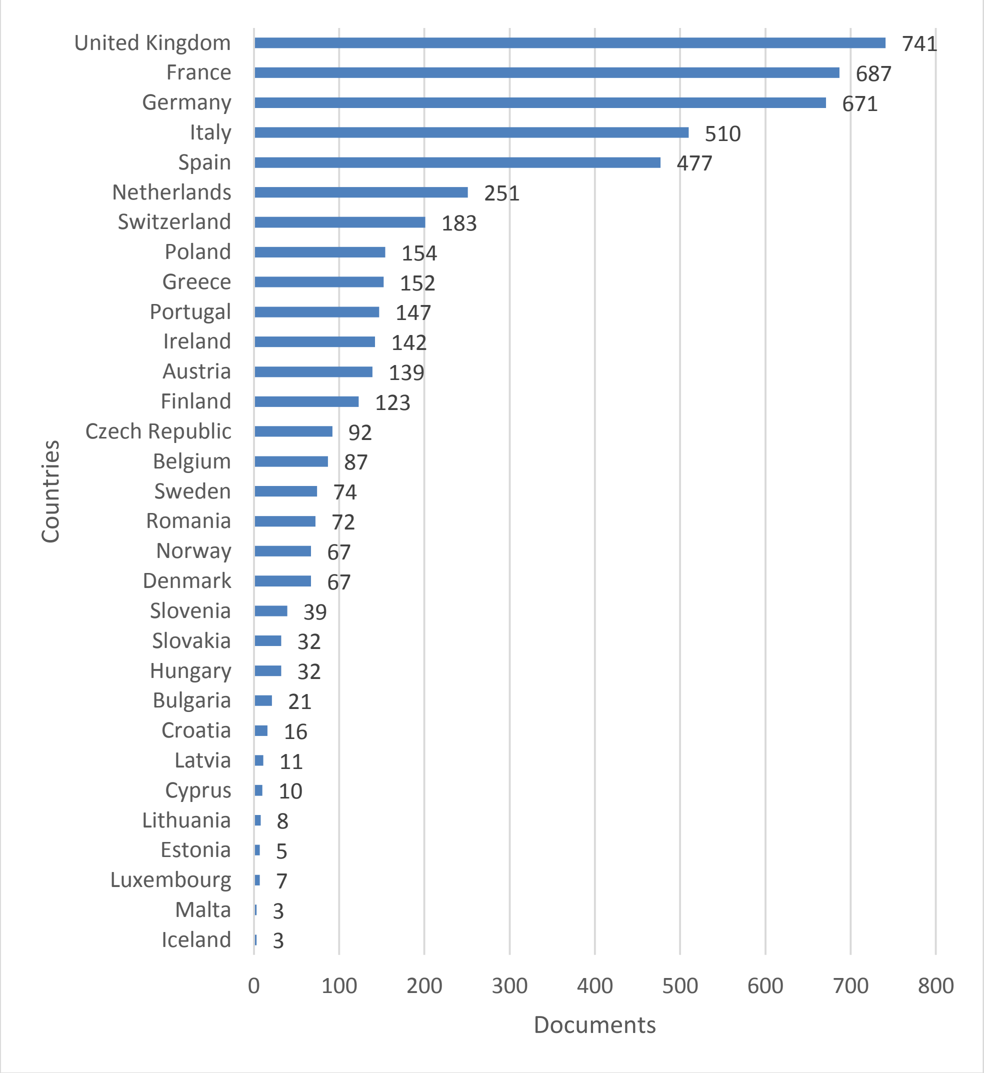


Figure 63 shows the distribution of publications among European countries of this study. United Kingdom, with 741 publications, together with France (687 publications) and Germany (671 publications) are the leaders among the European countries.

Figure 63 Number of Scopus publications from European countries on text-related IR, 2010-Nov. '18



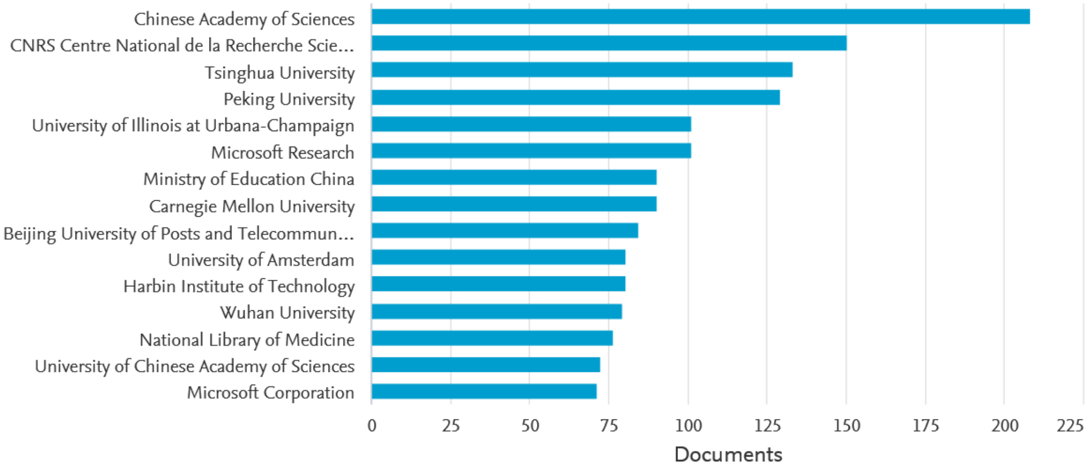
When the top 21 authors are compared, 7 of the most prolific authors are currently working in Europe, 5 are from North America, and 5 are in Asia (see Table 13).

Table 13 Top 20 authors publishing on text-related IR in Scopus, 2010-Nov. '18

Author	Number of publications	Country	Region
1. Jones, G.J.F.	37	Ireland	Europe
2. Rosso, P.	29	Spain	Europe
3. Zuccon, G.	28	Australia	
4. Müller, H.	27	Switzerland	Europe
5. Navarro, G.	27	Chile	
6. Soman, K.P.	27	India	Asia
7. Demner-Fushman, D.	25	United States	North America
8. Anand Kumar, M.	24	India	Asia
9. Croft, W.B.	23	United States	North America
10. Roche, M.	23	France	Europe
11. Kamps, J.	22	Netherlands	Europe
12. De Rijke, M.	22	Netherlands	Europe
13. Gelbukh, A.	22	Mexico	
14. Lin, H.	22	China	Asia
15. Pal, U.	22	India	Asia
16. Thoma, G.R.	22	United States	North America
17. Ganguly, D.	21	Ireland	Europe
18. Varma, V.	21	India	Asia
19. Antani, S.	20	United States	North America
20. Omar, N.	20	Malaysia	
21. Han, J.	20	United States	North America

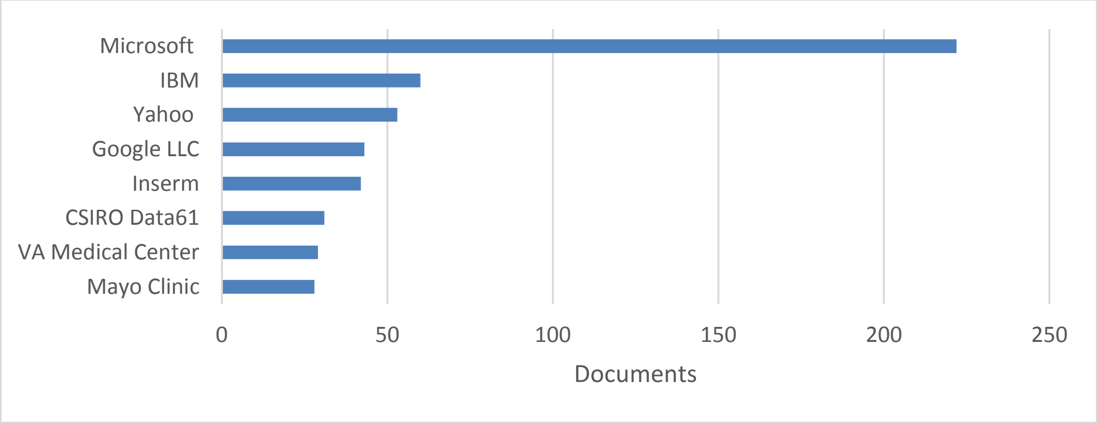
When the results are compared by top 15 organisations of countries of this study, the leader is Asia, while North America has five and Europe only two institutions with a high number of publications (see Figure 64).

Figure 64 Top 15 organisations with papers on text-related IR in Scopus, 2010-Nov. '18



When only companies are compared, the leader is the US, having headquarters for such international companies as *Microsoft, IBM, Yahoo* and *Google* (see Figure 65).

Figure 65 Scopus publications of companies, "information retrieval" AND ("text" OR "word"), 2010-Nov. '18

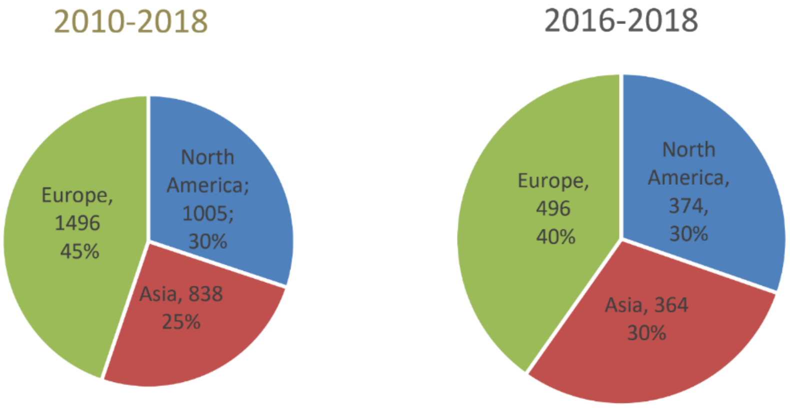


Finally, papers from six representative conferences on information retrieval field were analysed:

- SIGIR: International ACM SIGIR Conference on Research and Development in Information Retrieval, 2010-2018
- WSDM: The ACM International Conference On Web Search And Data Mining, 2010-2018
- ICTIR: The ACM SIGIR International Conference On The Theory Of Information Retrieval, 2011, 2013, 2015, 2016, and 2017
- ECIR: European Conference on Information Retrieval, 2010-2018
- AIRS: Asia Information Retrieval Societies Conference, 2010-2017
- SPIRE: International Symposium on String Processing and Information Retrieval, 2010-2018

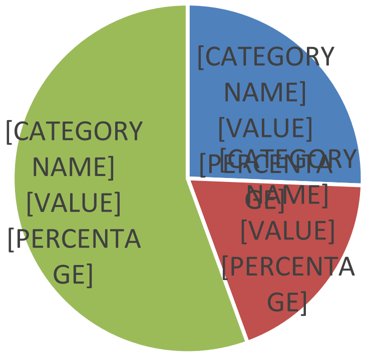
The distribution of publications between regions (for countries of our study) is shown in Figure 66. While in absolute numbers the leader is the US with 877 publications followed by China (511 publications) and the United Kingdom (467 publications), Europe (45% of publications) is a leader when regions are compared.

Figure 66 Distribution of publications between regions published in SIGIR/WSDM/ICTIR/ECIR/AIRS/SPIRE



Finally, we analysed the top 100 most cited papers in the Scopus database that are written from 2010 till 2018 (the list of top 10 publications is included in Annex E) on text-related information retrieval. As it is demonstrated in Figure 67, more than half (55%) of the publications include contributions from North American authors, while 26% of the publications include work of European authors. For 19% of the publications, authorship is attributed to Asia.

Figure 67 Distribution of authorship for top 100 most cited publications on text-related IR in Scopus, 2010-2018



3.4.3.3. Organisational infrastructure

The information retrieval community is united through several special interest groups:

- **SIGIR**[46] is the Association for Computing Machinery's (ACM) Special Interest Group on Information Retrieval. Its focus is information – storage, retrieval and dissemination. SIGIR sponsors or co-sponsors several conferences in the field of information retrieval, including SIGIR, CIKM, JCDL, WSDM, ICTIR and CHIIR;
- **BCS IRSG**[47]**,** BCS, the Chartered Institute for IT, Information Retrieval Specialist Group – "aims include supporting communication between researchers and practitioners, promoting the use of IR methods in industry and raising public awareness". IRSG has a newsletter and supports the European Conference on Information Retrieval (ECIR).

TREC[48] (Text Retrieval Conference) was started with the aim to support research in IR by providing infrastructure for large-scale evaluation of text retrieval methods. For each TREC, NIST provides a test set of documents and questions. The TREC test collections are available to the research community.

The following annual conferences are organised for researchers and the industry of a particular region:

[46] http://sigir.org

[47] https://irsg.bcs.org

[48] https://trec.nist.gov

- AIRS - Asia Information Retrieval Societies Conference
- ECIR – European Conference on Information Retrieval
- SIGIR - International ACM SIGIR Conference on Research and Development in Information Retrieval, 2010-2018

Several initiatives and activities are related to cross-lingual IR. "The CLEF Initiative (Conference and Labs of the Evaluation Forum, formerly known as Cross-Language Evaluation Forum) is a self-organised body whose main mission is to promote research, innovation, and development of information access systems with an emphasis on multilingual and multimodal information with various levels of structure". [49] CLEF provides an infrastructure for:

- Multilingual and multimodal system testing, tuning and evaluation;
- Investigation of the use of unstructured, semi-structured, highly structured, and semantically enriched data in information access;
- Creation of reusable test collections for benchmarking;
- Exploration of new evaluation methodologies and innovative ways of using experimental data;
- Discussion of results, comparison of approaches, exchange of ideas, and transfer of knowledge.

OpenCLIR (Open Cross Language Information Retrieval) evaluation[50] aims to develop techniques that allow to find text (written or spoken) in low-resourced languages by using English queries. Recently the OpenCLIR 2019 challenge[51] was announced by IARPA and NIST, asking researchers to contribute to this problem.

Finally, the MATERIAL programme (Machine Translation for English Retrieval of Information in Any Language[52]) by IARPA aims to develop methods that allows finding written or spoken text in low-resourced languages that is a relevant English query in a particular domain.

[49] http://www.clef-initiative.eu

[50] https://www.nist.gov/itl/iad/mig/openclir-evaluation

[51] https://openclir.nist.gov

[52] https://www.iarpa.gov/index.php/research-programs/material

3.5. Innovation

In our study, we use the definition of innovation as “the implementation of a new or significantly improved product (good or service), or process, a new marketing method, or a new organisational method in business practices, workplace organisation or external relations.” (OECD and Statistical Office of the European Communities, 2005)

As proxies for innovation by region, we analysed the:

- Market of origin of the most popular tools
- Emergence of start-ups in the respective industry across regions
- Known implementation of the latest techniques in the respective area

3.5.1. Innovation in machine translation

Europe is the leader in the development and implementation of translation automation tools (computer-aided translation tools), with North America coming second. The same situation is observed in the area of translation technology start-ups. Based on an analysis of comparative advances, it can be concluded that Europe is the global leader in innovating translation technologies and services.

Table 14 Market relative score in innovation in machine translation

Market	Relative Score
Europe	3
North America	2
Asia	1

3.5.1.1. Market of origin of the translation automation tools

European leadership in the development and implementation of machine translation is supported by the Common Sense Advisory survey of 900 global enterprises, language service companies, and freelance translators (Lommel et al., 2016).

Parallel to MT technologies, we witness a dynamic innovation in computer assisted translation tools (CAT), that play a major role in the automation of professional translation. Despite a huge improvement in quality of MT thanks to the advances in neural MT, recent research has shown that MT systems are still not able to produce translations of sufficient quality on sentence level and even more so on document level and the output still requires post-editing by a human to correct errors and improve translation quality (Läubli et al., 2018).[53] CAT tools incorporate this manual editing stage into the software, making translation an interactive process between human and computer.

Eleven of the twenty-four recognised CAT tools that are used by a majority of LSPs have been developed in Europe. Table 15 provides a summary of the most recognised CAT tools and their regions of origin.[54]

Table 15 CAT tools by region of origin

	CAT TOOL	REGION
1	Across Language Server	Europe
2	CafeTran	USA
3	Déjà Vu	Europe
4	Fluency Now	USA
5	GlobalSight	USA
6	Gtranslator	USA
7	Lokalize – KDE	Europe
8	MateCat	Europe
9	memoQ	Europe
10	Memsource	Europe
11	Meta Texis	Europe
12	NET PROXY	No information
13	OMEGA T	Europe
14	Open Language Tools	USA
15	Pairaphrase	USA
16	Poedit	USA
17	Pootle	South Africa
18	SDL Trados	Europe
19	SmartCAT	USA
20	Star Transit	Europe
21	VIRTAAL	South Africa
22	Wordfast CLASSIC	USA
23	wordfast pro	USA
24	XTM Cloud	Europe

3.5.1.2. Translation technology start-ups

Another indicator of innovation is the emergence of start-up companies that introduce new technologies in the market, innovative ways of addressing business needs and novel business models. For this analysis we collected a list of translation technology start-ups from AngelList database[55] and assigned their regional attribution based on the location of their headquarters. As can be seen in Figure 68, Europe is the leader in the number of emerging innovative start-ups, closely followed by North America, leaving Asia in a distant third position.

[53] Although there has been research suggesting (for just one language pair, Chinese-English, in one domain – news, by one party – Microsoft) human parity on sentence-level translation (Hassan et al., 2018), its correctness has been already disputed by Toral et al. (2018), who claim that human parity has not been reached.

[54] https://en.wikipedia.org/wiki/Comparison_of_computer-assisted_translation_tools

[55] https://angel.co

Figure 68 Geographical distribution of translation technology start-up companies

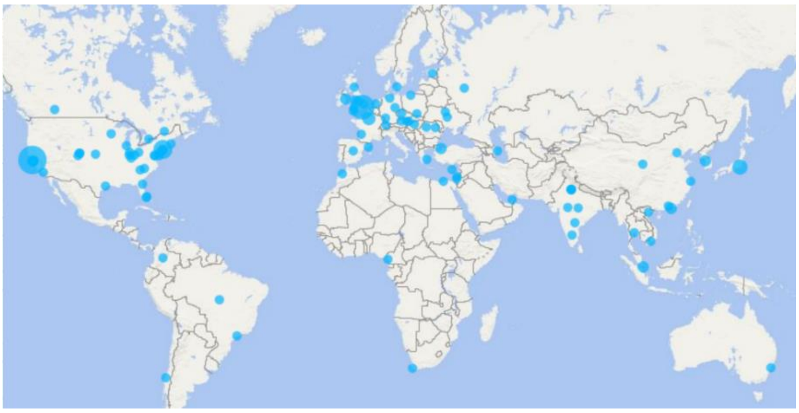
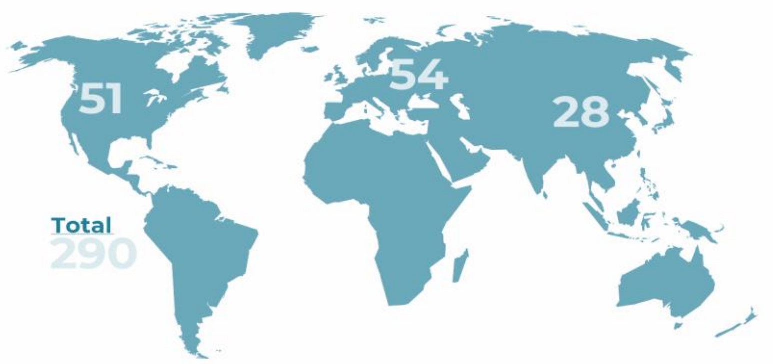


Figure 69 Regional distribution of translation technology start-up companies



3.5.1.3. Adoption of Neural MT

In recent years, neural MT (NMT) has become a global trend in MT development that has created opportunities for new services. Global adoption of neural MT is led by global providers Google and Facebook but European companies and public services are quick to follow. In a few months from the first release of Chinese-English NMT by Google there were numerous NMT systems launched by European players Tilde, KantanMT, SDL, and DeepL. The European Commission also is on a fast track in the adaption of NMT by transferring statistical MT systems of MT@EC translation service to the neural MT systems in the eTranslation platform.

European players are making a particularly remarkable progress in using neural MT to advance the quality of machine translation for complex European languages. Tilde has achieved the best results in the global machine translation competition WMT 2017 for Latvian-English-Latvian and WMT 2018 for Estonian-English-Estonian systems. eTranslation NMT systems for such challenging languages as Hungarian and Finnish have made them suitable for both direct usage in online services and post-editing for publishing quality translation. European NMT systems are rapidly attracting high-profile application areas such as automation of translation work at the EU Council Presidencies in Estonia, Bulgaria and Austria.

Asian headquartered companies are also actively embracing neural MT, with Baidu and Systran having the most visible profile on the innovation scene.

3.5.2. Innovation in speech technologies

Based on an analysis of comparative advances as summarised in Table 16, it can be concluded that North America is the global leader in innovating speech technologies and services. It is the absolute leader in all three criteria chosen for comparison, with Europe coming second.

Table 16 Market relative score in innovation in speech technologies

Market	Relative Score
Europe	2
North America	3
Asia	1

3.5.2.1. Market of origin of the most popular speech recognition tools

In speech recognition technologies, the European market is dominated by multi-national players headquartered in the United States (including Microsoft, Nuance, Amazon, IBM, Google, Apple, and Facebook). Indigenous vendors are predominantly niche players serving local markets. The presence of these large players is deterrent to market entry by local entrepreneurs and innovators. Such conclusion has been corroborated by IDC data (see Task 1).

Commonly used speech technologies provide solutions for automatic transcription, hearing and understanding customers, identifying callers, monitoring agents, getting data on customers, writing letters and e-mails.[56] Some of these features are not new anymore and have been integrated in business operations by companies that think about innovation and cost saving solutions. Amazon's Alexa, Google's Now, Microsoft's Cortana, and Apple's Siri are among the most notable voice recognition solutions globally. Google has made speech recognition a central focus for growing its business.[57]

The major players operating in the speech technologies market are VoiceBox Technologies Corp., Alphabet Inc., Acapela Group SA, Sensor, Hoya, Iflytek Co. Ltd., Nuance Communications Inc., VoiceVault Inc., Cantab Research Limited, Pareteum Corporation, LumenVox, LLC, Microsoft Corporation and other companies, mostly based in North America[58] (see Table 17[59]). The most

[56] https://www.transcribeme.com/blog/8-innovative-ways-to-use-speech-recognition-for-business

[57] https://medium.com/swlh/the-past-present-and-future-of-speech-recognition-technology-cf13c179aaf

[58] Taken from the report presented in https://globenewswire.com/news-release/2018/08/22/1555231/0/en/Speech-and-Voice-Recognition-Technologies-Market-Will-Reach-USD-22-32-Billion-Globally-By-2024-Zion-Market-Research.html.

advanced speech recognition tools of 2018 according to business.com magazine are Dragon Naturally Speaking Individual, Dragon Naturally Speaking Premium, Dragon for Mac, Dragon Naturally Speaking Home, Voice Finger and others.[60]

Table 17 Leading market players in speech recognition

	Company	Country of HQ	Region
1	Acapela Group	Belgium	Europe
2	Alphabet Inc.	US	North America
3	Amazon.Com	US	North America
4	Baidu	China	Asia
5	Cantab Research Limited	UK	Europe
6	CereProc	UK	Europe
7	Facebook	US	North America
8	Google	US	North America
9	IBM	US	North America
10	Iflytek Co., Ltd.	China	Asia
11	iSpeech Inc.	US	North America
12	LumenVox LLC	US	North America
13	Microsoft Corporation	US	North America
14	NeoSpeech	US	North America
15	Nexmo	US	North America
16	NextUp Technologies	US	North America
17	Nuance Communication	US	North America
18	Pareteum Corporation	US	North America
19	Hoya	US	North America
20	rSpeak	The Netherlands	Europe
21	Sensory Inc.	US	North America
22	SESTEK	Turkey	Other
23	TextSpeak	US	North America
24	VoiceBox Technologies Corp.	US	North America
25	VoiceVault Inc.	US	North America

3.5.2.2. Voice and speech recognition start-ups

Another indicator of emerging innovations are start-up companies that introduce new solutions to the market that address business needs and novel business models.

Using the AngelList database, we tracked emerging start-ups and screened the voice and speech recognition services that the new companies offer. According to the Angel List database resources on 11 October, altogether 204 start-ups are operating in the fields of voice/speech recognition.[61] The Majority (113) are located in North America, while 51 are registered in Europe but 25 in Asia, as

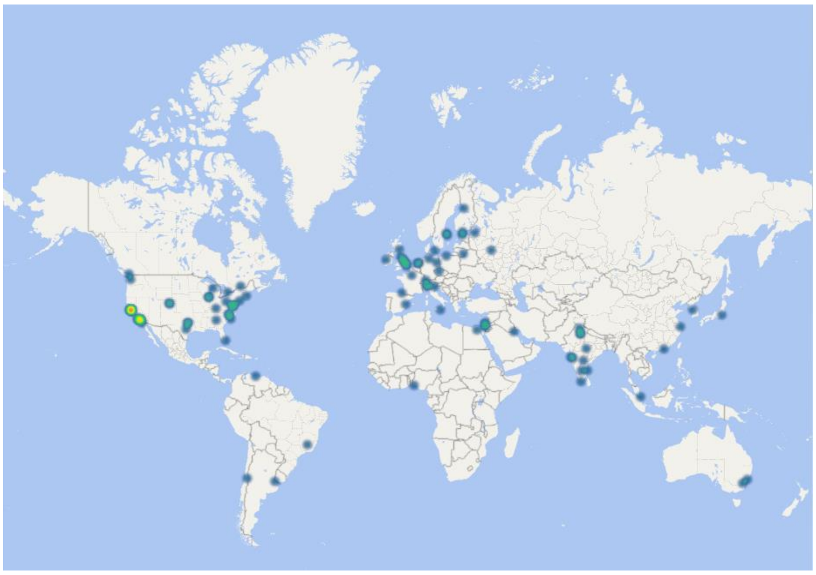
[59] The selection of speech technology companies is based on "Speech and Voice Recognition Market by Technology, Vertical and Geography - Global Forecast to 2023" and "Text-to-Speech Market by Vertical, and Geography – Global Forecast to 2022" by marketsandmarkets.com.

[60] More details on the best speech recognition software of 2018: https://www.business.com/categories/best-voice-recognition-software.

[61] Key words used for search: automated speech recognition (ASR), speech synthesis (text-to-speech, TTS), interactive voice recognition (IVR).

illustrated in Figure 71. 15 companies were registered in other regions (South America, Africa, Australia) or information regarding their location was not provided. Figure 70 shows the regional distribution of the speech recognition start-ups.

Figure 70 Geographical distribution of speech technology start-up companies



By sorting the start-ups by the solutions they provide, it is worth noting that solutions vary by sector, starting from security and emergency services to entertainment. For example, Fluent.AI recognises voices in various languages and accents and transmits the message directly to the internet to manage practical domestic tasks (smart homes). There are also solutions to track cardiac arrest by an emergency caller's voice.[62]

Furthermore, Speakwithme claims that it has introduced a platform that interacts between context, memory, emotion, and personalisation.[63] Ubi provides operating tools to integrate voice with hardware using Alexa provided services.[64] A similar solution has been created by HelixAI, which uses the Alexa platform to perform audio-based specific searches in science laboratories.[65]

There are also efforts to analyse conversational data and recognise human emotions, for example, determining engagement, frustration and politeness by capturing emotions and deciphering how something is being said. The solution is provided by Behavioural Signals.[66] There are also successful

[62] https://www.gearbrain.com/corti-cardiac-arrest-artificial-intelligence-2525654278.html

[63] https://www.speakwithme.com

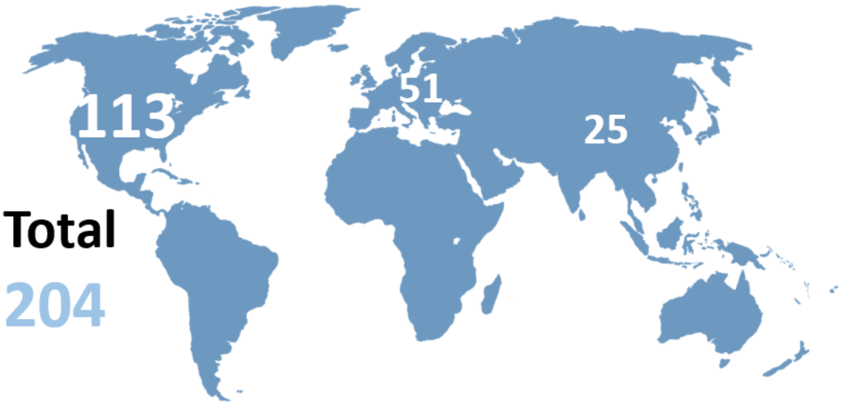
[64] http://www.ucic.io

[65] https://www.askhelix.io

[66] https://behavioralsignals.com

pilot projects to evaluate the credibility of a financial institution’s client before making lending decisions.[67]

Figure 71 Regional distribution of speech recognition start-up companies



Speaker identification and speaker verification are classified as "behavioural biometrics".[68] The private banking division of Barclays was the first financial services firm to deploy voice biometrics as the primary means to authenticate customers for their call centres. In 2016 the UK-based bank HSBC announced that it would offer 15 million customers its biometric banking software to access online and phone accounts using their voice. In addition to other biometric verification, there is hope that it will tackle the issue of forgotten passwords.**[69]** Speaker recognition may also be used in criminal investigations and to track outlaw location.[70]

3.5.2.3. Implementation of the latest techniques in speech technologies

To build a robust speech recognition experience, the machine learning techniques behind it have to become better at handling challenges such as accents and background noise. Today, developments in natural language processing and neural network technology have improved the speech and voice technology, so much that today it is reportedly on par with humans. In 2017, for example, the word

[67] https://toneboard.com

[68] https://en.wikipedia.org/wiki/Speaker_recognition#cite_note-16

[69] https://www.theguardian.com/business/2016/feb/19/hsbc-rolls-out-voice-touch-id-security-bank-customers

[70] https://www.theguardian.com/media/2014/sep/02/steven-sotloff-video-jihadi-john

error rate (WER) for Microsoft's voice technology has been recorded at 5.1 percent by the company, while Google reports that it has reduced its rate to 4.9 percent.[71]

Furthermore, most innovative solutions based on convolutional and LSTM neural networks with the spatial smoothing and lattice-free MII acoustic training (Kiong et al, 2016) achieves human parity in speech recognition. The novel deep generative model of raw audio waveforms is able to generate speech which mimics any human voice and which sounds more natural than the best existing text-to-speech systems, reducing the gap with human performance by over 50% on English and Chinese languages (van den Oord, 2016).[72]

However, although Google supports 119[73] and Nuance over 86[74] languages and dialects, the speech recognition performance among the languages is not equal.

Although quite a few speech recognition toolkits (the most popular is Kaldi[75], others are CMUSphinx,[76] HTK,[77] RWTH ASR,[78] Julius[79] and EESEN[80]) are available today for building speech recognition engines, each language is unique and still requires adaptation (search for the best method, parameter tuning) or other specific language-dependent solutions. Global companies, e.g. *Google* or *Microsoft,* use proprietary tools for development and decoding.

3.5.3. Innovation in search technologies and services

Even though an analysis of comparative advances in search shows that North America is a global leader in innovating search technologies and services by dominating the global market and boosting start-ups, it must be concluded that when it comes to cross-lingual search, Europe's and China's demand for translated information retrieval is fostering the regions seek for solutions.

[71] https://www.techemergence.com/ai-for-speech-recognition

[72] WaveNet methodology and evaluation for text-to-speech is explained in https://deepmind.com/blog/wavenet-generative-model-raw-audio.

[73] The list of languages Google supports for automatic speech recognition is in https://cloud.google.com/speech-to-text/docs/languages.

[74] The list of languages Dragon supports for automatic speech recognition is in https://www.nuance.com/omni-channel-customer-engagement/voice-and-ivr/automatic-speech-recognition/nuance-recognizer/recognizer-languages.html.

[75] http://kaldi-asr.org

[76] https://cmusphinx.github.io

[77] http://htk.eng.cam.ac.uk

[78] https://www-i6.informatik.rwth-aachen.de/rwth-asr

[79] http://julius.osdn.jp/en_index.php?q=en/index.html

[80] https://github.com/srvk/eesen

Table 18 Market relative score in innovation in search technologies

Market	Relative Score
Europe	2
North America	3
Asia	1

3.5.3.1. Market origin of the most popular search tools

There is strong evidence that Google is a global leader in web search technologies covering 92% of the global market. Although globally Google dominates, the regional picture in Asia is more diverse. As an example, in China the dominant search engine with over 82 percent market share is Baidu while Google comes in at 0.61 percent and Bing at 0.37 percent.[81] Moreover, in Russia Yandex is aggressively expanding its ecosystem beyond its core search engine. It offers e-mail and cloud services, a virtual assistant named Alice, an AI-powered recommendations platform, integrated streaming videos on its homepage and a ride hailing and food ordering through Yandex.Taxi, which was merged with Uber's services as a joint venture. As a result, Yandex leads the Russia market with 57.9%, leaving Google in second place with 43,3% of market share.[82] Using the statcounter.com tool, we tracked the market share of web search engines and added language support to each item. The result is reflected in Figure 72.[83]

[81] https://www.searchenginejournal.com/seo-101/meet-search-engines

[82] http://gs.statcounter.com/search-engine-market-share/all/russian-federation/#monthly-201709-201809

[83] http://gs.statcounter.com/search-engine-market-share

Figure 72 Market of origin of most popular search tools

Search engines	Language[84][85]	Region (HQ)	Market Share Worldwide (Sept. 2018)
Google	Multilingual	North America	92.31
Bing	Multilingual[86]	North America	2.27
Yahoo! (powered by Bing)	Multilingual[87]	North America	2.51
Baidu	Chinese	Asia	0.85
YANDEX RU	Multilingual[88]	Other (Russia)	0.61
Shenma	Chinese	Asia	0.18
YANDEX[89]	Multilingual[90]	Other (Russia)	0.31
DuckDuckGo	Multilingual	North America	0.33
Naver	Korean	Asia	0.18
Haosou	Chinese	Asia	0.08
Sogou (runs CLIR platform 'Sogou English')	Chinese/English[91]	Asia	0.1
MSN (powered by Bing)	Multilingual	North America	0.08
Daum	Korean	Asia	0.02
Mail.ru		Other (Russia)	0.04
Seznam	Czech	Europe	0.04
Ask Jeeves/Ask.com	Multilingual	North America	0.01
CocCoc (powered by Google)	Vietnamese/English[92]	Other (Vietnam)	0.02
Other			0.06

When it comes to language usage, multinational companies in North America (Google, Bing, Yahoo[93]) have been thinking about language availability and search in various languages some time ago[94] while emerging Asian tech giants are approaching the issue cautiously. But since Baidu started targeting Chinese tourists traveling overseas by rolling out a talking translator and assistant in 2017 and Sogou has partnered with Microsoft to use Bing to search for English results and get the result translated back into Chinese, Asian internet search providers are also slowly moving in the direction

[84] https://en.wikipedia.org/wiki/List_of_search_engines

[85] http://www.searchengineshowdown.com/language/limits.shtml

[86] https://docs.microsoft.com/en-us/azure/cognitive-services/bing-web-search/language-support

[87] https://developer.yahoo.com/search/languages.html?guccounter=2

[88] https://yandex.com/support/webmaster/robot-workings/supported-languages.html

[89] Although from company perspective Yandex and Yandex.ru are managed by one company 'Yandex', the statcounter.com methodology divides usage of two different sites 'Yanex.ru' which used predominantly used in Russia and 'Yandex' which is targeted outside Russia. https://searchengineland.com/russias-yandex-search-engine-goes-global-42381

[90] https://yandex.com/support/webmaster/robot-workings/supported-languages.html

[91] https://en.wikipedia.org/wiki/Sogou

[92] https://coccoc.com/search

[93] https://developer.yahoo.com/search/languages.html?guccounter=2

[94] http://www.searchengineshowdown.com/language/limits.shtml

of multilingual search.[95] However, language support for small languages is still lagging behind for most of the search providers.[96]

3.5.3.2. Market of origin of the most popular enterprise/website search tools

Based on publicly available resources, we reviewed the evaluation and assessments by experts of enterprise search/website engines. We analysed four lists of popularity measures. (1) Magazine CIO Application has collected information from enterprises and created their list of most reputable tools.[97] (2) Analysts from G2 Crowd have done research of most popular enterprise search software tools, based on three criteria: ease of use, requirements, and ease of doing business, and created the list of companies that provide the most efficient solution.[98] (3) At the same time business review journal Business Online has created the list of top 20 companies that most fit enterprise needs.[99] (4) We also looked at the list which is purely based on reviews of open source tools.[100] Based on the findings, we created one list that reflects the most popular search tools, summarised in Figure 73.[101]

Figure 73 Market of origin of most popular enterprise search tools

Company	**Region**
Elasticsearch/ Elastic.co (based on Apache Lucene)	Europe
Apache Solr (based on Apache Lucene)	North America
Amazon Elasticsearch Service	North America
Sphinx	Europe
Microsoft Azure Search (built on Elasticsearch)	North America
Google Enterprise Search (built on Apache Solr/uses Lucene search library)	North America
Swifttype Site Search	North America
Coveo Solutions	North America
Algolia	North America
Apache Lucene	North America
Lucidworks (based on Apache Solr and Apache Spark)	North America

[95] https://asia.nikkei.com/Business/AC/Baidu-s-talking-translator-gives-tourists-a-hand

[96] http://www.searchengineshowdown.com/language/limits.shtml

[97] https://www.cioapplications.com/vendors/top-10-enterprise-search-solution-providers-2018-rid-75.html

[98] https://www.g2crowd.com/categories/enterprise-search#highest_rated

[99] https://financesonline.com/site-search/#unbxd

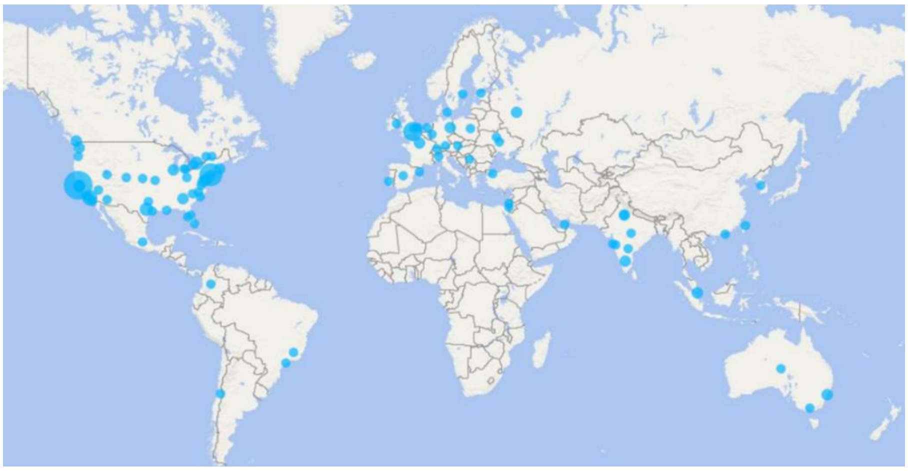
[100] https://greenice.net/elasticsearch-vs-solr-vs-sphinx-best-open-source-search-platform-comparison

[101] Methodology: by reviewing three lists we counted mentions in each popularity list (if the enterprise got mentioned in one review it got one point, if it was mentioned in two reviews it scored '2' etc.). Based on the methodology, we created a list that reflects the most popular enterprise search engines. Note: to avoid subjectivity, we eliminated the tools that have only one mention.

3.5.3.3. Search technology start-ups

Another indicator of innovation is the emergence of start-up companies that introduce new technologies in the market, innovative ways of addressing business needs and novel business models and services. By collecting a list of speech technology start-ups from the AngelList database,[102] the regional belonging has been assigned based on the location of the company headquarters. As it is seen in the resulting Figure 75, North America is the leader in the number of emerging start-ups followed by Europe and Asia in a distant third position.

Figure 74 Geographical distribution of search technology start-up companies



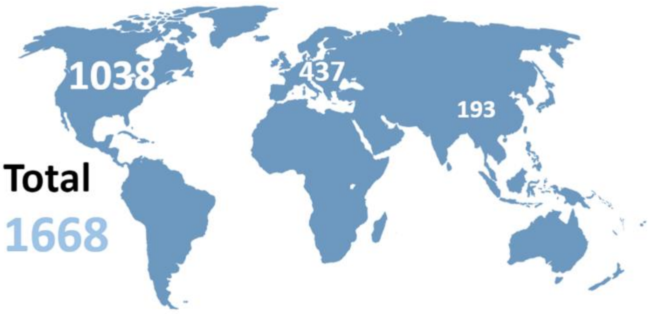
Reviewing services provided by start-ups, it may be concluded that the majority of them are providing or trying to provide search services in specific segments. The services may be divided into two broad categories. First, targeted social networking in a specific category (business, sports, parenthood, research, babysitting, travel or other leisure activities) that bring people together based on common interests. Another segment of companies is providing services and recommendations for individuals seeking to find something useful in a quicker way (gym, restaurant, movie, clothing, hotel recommendations).[103]

Numerous start-ups are targeting the Airbnb and Couch-surfing type services. There is also a health and sports category, where one can find a proper coach or bring together potential teammates. Many provide services for leisure activities such as finding an appropriate restaurant based on the consumer's preferences or best destination. The database also lists start-ups that help to find and purchase goods or services, like bus tickets.

[102] https://angel.co

[103] https://angel.co/companies?keywords=search

Figure 75 Regional distribution of search technology start-ups



Having looked at semantic search as one of the subcategories of search technology and most notable examples, the social network LinkedIn has developed and published their semantic search approach to job search by recognising and standardising entities in both queries and documents, e.g., companies, titles and skills, then constructing various entity-aware features based on the entities. The company has concluded that the search results have slightly improved.[104]

Since semantic search offers a structured way of searching information which leads to improvements in search results, there are start-ups that follow the semantic type search trend. We have analysed the multilingual search category in AngelList. There are 8 start-ups in AngelList under the category of multilingual search, but information about them is scarce. They provide mainly services that retrieve information about events, flights, or other services from webpages.[105] [106] To sum up, it must be concluded that there are no specific start-ups providing cross-lingual search, but multilingual solutions are offered as a by-product to information retrieval services, such as looking up gyms and restaurants or travel destinations.

3.5.3.4. Known implementation of the latest techniques

Site search or enterprise search technologies allow searching for content in certain websites, document libraries, events on computer’s mailboxes and directly in mailbox archive files. They can be offered via APIs to cloud-based solutions as well hostable on premises. To enable different search services and products, separately described search engines can lay under the hood, for example, Amazon Elastic Services relies on the Elasticsearch open source product, which is built upon the

[104] https://www.kdd.org/kdd2016/papers/files/adp0518-liA.pdf

[105] https://angel.co/rankabove

[106] https://www.facebook.com/fiestafy

Apache Lucene search library. Apache Lucene is supported by Apache Software Foundation, registered in the US.[107]

The search technologies have evolved starting from keyword search to semantic search till contextual search and cognitive search where machine learning is involved. The new generation of enterprise search solutions employs AI technologies such as natural language processing and machine learning to ingest, understand, organise, and query digital content from multiple data sources.[108] Moreover, today, search is not just about a text box on an enterprise portal. Enterprises are building search applications that embed search in customer 360 applications, virtual assistants, pharma research tools, and many other business process applications. [109]

By looking at various reviews, we analysed the most popular enterprise search engines and created a list of the most popular and used enterprise/site search engines (Figure 73) and analysed their efficiency based on various views expressed by analysts. Based on reviews there are two leaders in the enterprise search category – Apache Solr and Elasticsearch. Both use Apache Lucene library and therefore have many similarities, then disparities. Despite similarities some differences emerge - while Apache Solr is full of features relating to full-text search with impressively rich features, Elasticsearch relies on single, dedicated suggesters API. In this case, the details of the implementation are not available to users or developers. On the contrary, Solr falls a bit behind from the DevOps point of view as the information that DevOps people need is often fragmented and incomplete. Meanwhile, troubleshooting Elasticsearch is an easier process as developers are able to easily get information such as work statistics, disk usage, memory, usage of thread pools, caching and buffer information. Elasticsearch also takes the lead in terms of tools and features, as its ecosystem is more up to date, it works with constantly updated Kibana and other features. Both engines provide similar machine learning capabilities, but Apache offers it free of charge.[110]

Generally, there are more similarities than differences in language support as both engines use the Lucene library.[111] However, Elasticsearch has strong language support.[112] Elastic's language analysers[113] are made up of two main components: a tokeniser and a set of token filters. The

[107] https://en.wikipedia.org/wiki/The_Apache_Software_Foundation

[108] https://www.prefixbox.com/blog/ecommerce-site-search

[109] https://techbeacon.com/sites/default/files/res136544_forrester_cognative_search.pdf

[110] https://thishosting.rocks/comparing-elasticsearch-with-solr

[111] https://en.wikipedia.org/wiki/Apache_Lucene

[112] https://www.elastic.co/guide/en/elasticsearch/guide/current/languages.html

[113] Elasticsearch, similarly to other search engine providers, has a built-in collection of language analysers that provide good, basic, out-of-the-box support for many of the world's most common languages: Arabic, Armenian, Basque, Brazilian, Bulgarian, Catalan, Chinese, Czech, Danish, Dutch, English, Finnish, French, Galician, German, Greek, Hindi, Hungarian, Indonesian, Irish, Italian, Japanese, Korean, Kurdish, Norwegian, Persian, Portuguese, Romanian, Russian, Spanish, Swedish, Turkish, and Thai.

tokeniser splits text into tokens according to some set of rules, and the token filters each perform operations on those tokens. The result is a stream of processed tokens, which are either stored in the index or used to query results.[114]

[114] https://qbox.io/blog/elasticsearch-english-analyzer-customize

3.6. Investments

Investopedia defines an investment as “the act of committing money or capital to an endeavour (a business, project, real estate, etc.), with the expectation of obtaining an additional income or profit.”[115] Investments in the context of this study are measured by the merger and acquisition, venture capital, and start-up financing of companies that can be identified as being engaged in language services and specifically in machine translation development and implementation.

3.6.1. Investments in machine translation

Although Europe may have a global lead in research, as noted above, it lags in investment capacity. Even though many of the most important MT advances in recent decades have come from Europe and EU-funded projects as illustrated by an analysis of publications, nevertheless, the biggest commercial developers are U.S.-based tech firms (such as Facebook, Google, Amazon, and Microsoft) that have staffed their research programs with European participants or bought European technology. North America has a dominant presence in machine translation developed by the abovementioned technology giants. In addition, North America also dominates the translation sector and by association also the machine translation component. Due to recent investments by mainly Chinese e-commerce entities such as Alibaba, Baidu and Tencent, Asia is a participant to be reckoned with in the MT sector.

Table 19 Market relative score in investments in machine translation

Market	Relative Score
Europe	1
North America	3
Asia	2

The U.S. economy remains the largest in the world in terms of nominal GDP. The $19.42 trillion U.S. economy is 25% of the gross world product. The United States is an economic superpower that is highly advanced in terms of technology. The nominal GDP for the U.S. and China for the year 2022 is estimated at $23.76 trillion and $17.71 trillion respectively while the economies of countries in the European Union account for just over 20% of the world’s total GDP.[116]

The largest global players, investors, and developers of machine translation technologies are currently companies based in the US: Google, Amazon, Facebook, Microsoft, Apple, eBay, and the US

[115] https://www.investopedia.com/terms/i/investing.asp#ixzz5IHnesEyW

[116] https://www.imf.org/external/pubs/ft/weo/2017/01/weodata/index.aspx

government are also the largest investors in language technologies. The main drivers are large internet commerce based entities.[117]

In Asia a similar scenario is emerging with the largest machine translation investments coming from entities such as Alibaba, which after years of growth has realised that language services are crucial to growth[118], and Tencent.[119]

In Europe the single largest investor through various support programs has been the European Commission. The fragmentation of the single market by country and language has resulted in a market size that rivals North America and Asia, but internal fragmentation has held back the emergence of large companies to rival the US or China.

As concluded in the recent Science and Technology Options Assessment report by European Parliament, "*The European HLT industry is mainly made up by innovative smaller companies and micro-enterprises. Although most of them have been established in the market for several years, the fragmentation of the LT market in Europe (local/national companies with expertise in local languages that serve local markets) hamper their growth. The transformation into global players capable of competing with global companies requires financing in all stages of business life cycle, not only in research activities*." (European Parliament, 2017, p. 104)

The language technology business sector as a whole is experiencing a time of acquisitions and investment. Although the industry is fragmented with several thousand service providers of 5-10 employees (as has been demonstrated in Task 1), the quest to provide faster and less expensive services is driving the trend of mergers and investments to take advantage of scale and productivity enhancing technologies.

The market has experienced a number of high-profile investments, buy-outs, and mergers in the sector over the past years, few language service companies are publicly traded, and most transactions are private and information is limited. One of the first high-profile acquisitions of a machine translation technology company outright was the purchase of Language Weaver by SDL for $42.5m in 2010. In many cases however, such as with Hewlett-Packard, whose translation service places them in the top 10 global translation service providers, it is such a small part of their overall business turnover that there is no information available even through publicly available SEC filings regarding their investment in machine translation. Another example is from the publicly listed

[117] http://translation-blog.multilizer.com/why-amazon-alibaba-and-ebay-develop-machine-translation

[118] https://slator.com/technology/alibaba-launches-language-services-unit

[119] https://ai.tencent.com/ailab/paper-list-2.html

Lionbridge, where their 2015 acquisition of CLS Communications has not been divulged even in NASDAQ filings.[120]

Below are some relatively recent high-profile acquisitions/investments in the machine translation industry. Unfortunately, as already noted above, the acquisition amounts are rarely disclosed, so that it is difficult to estimate the size of this acquisition market.

- **Amazon** has recently purchased **Safaba** for an undisclosed amount. Safaba was a Carnegie-Mellon University start-up which acquired three rounds of funding for undisclosed amounts prior to being sold to Amazon.[121]
- **Amplexor** has purchased **Sajan** with a strong machine translation component.[122]
- **Transperfect** has taken a different route and hired a leading industry veteran Eric Blassin, from Lionbridge with the intent to develop and perfect their internal machine translation solutions.[123]
- **eBay** has acquired **AppTek** to strengthen and develop its MT capacity and reach.[124]
- **ULG** (United Language Group) has acquired **Lucy** to strengthen its existing MT offering.[125]
- **Lionbridge**, one of the perennial leaders in the language services industry has just recently itself been acquired by H.I.G. in a $360m equity deal highlighting the interest of private investors in the language services industry.[126]
- **Facebook** has acquired **Mobile Technologies, LLC** (the developer of the speech recognition and machine translation application Jibbigo) in 2013.[127]

Based on data from Common Sense Advisory translation industry research and on Slator 2018 Language Service Provider Index,[128] Table 20 describes a selection of top 20 global translation companies by turnover. Nearly all the top 20 are investing in machine translation, either developing their own system or buying existing MT service providers. Many have the latest NMT technologies, which illustrates how very important cutting edge technologies are in the language services sector.

[120] http://www.annualreports.com/HostedData/AnnualReportArchive/l/NASDAQ_LIOX_2015_a535871e4a5c403fab0937be6e366a9b.pdf

[121] https://slator.com/ma-and-funding/amazon-acquires-mt-vendor-safaba-creates-machine-translation-rd-group

[122] https://www.owler.com/company/sajan; https://www.sajan.com/sajan-enters-merger-agreement-acquired-amplexor-international

[123] http://www.transperfect.com/technology/machine-translation; http://www.transperfect.com/category/blog-tags/machine-translation; http://www.translations.com/about/news/press-release/transperfect-announces-hiring-translation-technology-pioneer-eric-blassin

[124] https://techcrunch.com/2014/06/13/ebay-acquires-machine-translation-capabilities-from-apptek-to-help-expand-international-sales

[125] https://slator.com/ma-and-funding/ulg-buys-german-machine-translation-developer-lucy

[126] https://www.lionbridge.com/en-us/about/news/lionbridge-enters-definitive-agreement-acquired-hig-capital

[127] https://techcrunch.com/2013/08/12/facebook-acquires-mobile-technologies-speech-recognition-and-jibbigo-app-developer

[128] https://slator.com/features/the-slator-2018-language-service-provider-index

Table 20 Top 20 global translation companies: activities and acquisitions

COMPANY	COUNTRY	ACTIVITIES & ACQUISITIONS	TURNOVER[129]
Lionbridge	US	CLS Communication sold to Lionbridge (2014)[130] Bought by H.I.G. (2016), in-house NMT	$590m
TransPerfect	US	Investments in in-house MT	$615m
HPE ACG	FR	In-house to HP, no info available	No info
LanguageLine Solutions	US	Sold to Teleperforma (FR) for $1.5 b (2016)	$451m
SDL	GB	Aquired Language Weaver for $42.5 (2010)[131] In-house NMT 20 billion words/month MT'd	$388.5m $56 m LT turnover
RWS Group	GB	Uses SDL MT[132]	$221.5m
Welocalize	US	Uses 3rd party MT (Microsoft, Iconic MT etc.)	$200 m
STAR Group	CH	In-house MT[133]	$166.2m
Amplexor	LU	Aquired Sajan for $28.5 (2017)	$175.6m
Moravia	CZ	In-house MT[134] Acquired by RWS (2015)	$100m
Hogarth Worldwide	GB	No info	$177m
CyraCom International, Inc.	US	Interpreting, looking for early stage investment[135]	$161m
RR Donnelley Language Solutions	US	In spin-off mode[136]	$93m
Semantix	SE	No info	$107m
Honyaku Center Inc.	JP	Acquired Media Research Inc for $4.8 (2017)[137]	$26m
Pactera Technology International Ltd	CN	Sold for $675m to HNA EcoTech (2016)[138]	$85.2m
Ubiqus	FR	Interpretation, no known MT	$82.6
Keywords Studios	GB	Games, audio[139]	$180.1m
United Language Group (ULG)	US	ULG purchased Lucy MT for an undisclosed amount (2017)[140]	$79m
Logos Group	IT	No information on MT available	No info
Capita Translation and Interpreting	GB	Acquired through merger SmartMate MT[141]	$178m

Publicly available information on mergers and acquisitions and venture capital investment among top MT providers is scarce. Below is a selection of the most recent information:

[129] https://slator.com/features/the-slator-2018-language-service-provider-index

[130] https://www.lionbridge.com/en-us/about/news/lionbridge-completes-acquisition-cls-communication

[131] https://www.sdl.com/about/investors/annual-report-2016.html

[132] https://www.sdl.com/download/rws-ets-nf/123309

[133] http://www.eamt.org/corporate/stargroup.php

[134] https://www.moravia.com/en/news-events/press-releases/moravia-researcher-to-speak-at-machine-translation-marathon

[135] http://interpret.cyracom.com/investment

[136] https://slator.com/financial-results/donnelley-financials-10-k-filing-reveals-size-translation-business

[137] https://slator.com/ma-and-funding/honyaku-pays-usd-4-8m-for-tokyo-rival-and-buys-stake-in-mt-vendor

[138] https://slator.com/ma-and-funding/blackstone-sells-pactera-giant-chinese-conglomerate

[139] https://slator.com/financial-results/is-keywords-studios-now-overvalued

[140] https://slator.com/ma-and-funding/ulg-buys-german-machine-translation-developer-lucy

[141] https://www.capitatranslationinterpreting.com/smartmate

- **Systran** was acquired by **CSLi** (Korea) in 2014 for an undisclosed amount, which has advanced **CSLi** to a leading place globally as an automated translation service provider.[142]
- **Amazon** now has its own machine translation R&D group, after acquiring Pennsylvania-based **Safaba Translation** for an undisclosed amount. Safaba had been an enterprise machine translation provider for clients such as PayPal and Dell.
- **Iconic** (Ireland) acquired investment funding from Enterprise Ireland, Boole Investment, and Bloom Investment.[143]
- **KantanMT** (Ireland), privately held, received investment from two investment funds, Enterprise Ireland and Delta Partners, no financial information disclosed.
- **Unbabel** has raised $23M for its 'AI-powered, human-refined' translation platform.[144]

Slator has summarised recent investments in translation technology start-ups made in 2017 (Table 21[145]). The volume of investments is relatively thin compared to large merger and acquisition deals for both translation and machine translation. A notable exception is 23M USD round B investment raised by Unbabel, a translation automation company headquartered in Lisbon. Annex M provides additional figures on start-up financing in the field of translation technology.

Table 21 Funding for innovation as represented by language technology start-ups (Slator)

STARTUP	COUNTRY	SECTOR	ROUND	AMOUNT (USD)	INVESTOR	DATE
New Tranx	China	AI and MT	Pre-A	7.5m	Kaitai Capital, Bojiang Capital, Meiya Wutong	Oct 2017
UTH International	China	Multi-lingual TM data	B	6.35m	Sogou	Aug 2017
Transfluent	Finland	Translation platform	–	0.82m	Crowdfunding	May 2017
Aylien	Ireland	AI and ML	–	2.35m	Atlantic Bridge University	Nov 2017
Cadence Translate	US	Remote Interpretation	Seed	0.65m	Various	July 2017
Motaword	US	Translation platform	–	0.6m	Undisclosed	May 2017
Qordoba	US	Localisation SaaS	A	5m	Upfront Ventures, Rincon Venture Partners	May 2017
Interprefy	Switzerland	Remote Interpretation	–	0.875m	Undisclosed	May 2017
SpeakUS	Russia	Remote Interpretation	Pre-Seed	59k	Enterprise Ireland	July 2017
Unbabel	Portugal	Translation automation	B	23m	Scale Venture Partners, Microsoft Ventures etc.	Dec 2017

[142] http://english.hani.co.kr/arti/english_edition/e_business/639506.html

[143] http://iconictranslation.com/about/investors-and-partners

[144] https://techcrunch.com/2018/01/11/unbabel

[145] https://slator.com/ma-and-funding/2017-language-industry-startup-funding

3.6.2. Investments in speech technologies

Investment activity in the speech technology field is dominated by North American companies, with Asian companies coming in second. There is relatively little activity in Europe as summarised below in Table 22.

Table 22 Market relative score in investments in speech technologies

Market	Relative Score
Europe	1
North America	3
Asia	2

Between 2008 and 2018 there have been 12 acquisitions[146] in the speech technology field. Table 23 lists the main investment flows by region (Northern America, Asia and Europe). One acquisition has been made by North America to Asia, three have been done in North America internally. At the same time, there have been two internal acquisitions in Asia. Meanwhile, European companies have acquired two North American companies, and vice versa – two European companies have been acquired by Northern American companies. Within the time frame European companies have made one acquisition internally and acquired one company from Asia.

Below are some relatively recent high-profile acquisitions in speech recognition. Unfortunately, as already noted above, the acquisition amounts are rarely disclosed, so that it is difficult to estimate the true size of this acquisition market.

- **Ytica**, a speech recognition company based in Czech Republic was acquired by Twillo. The value of the acquisition was not disclosed.
- An Indian machine learning platform was acquired by India-based Flipkat for an undisclosed amount.
- **Semantic Machines** is a US-based company which in 2018 was acquired by **Microsoft**.
- **Nuance Communications** has purchased **VoiceBoxTechnologies** with a strong speech recognition component. The amount of the acquisition deal was 82 million USD.
- **KITT.AI**, a US-based speech recognition company, was bought by Baidu.
- **DigitaL Roots,** a US-based media monitoring company was bought by **Interactions LLC** in 2017.
- **Baidu** acquired China-based AI and machine learning platform **Raven Tech** in 2017.
- **Vxi Corporation**, a US-based machine learning platform was acquired by **Jabra** for 35 million USD.
- UK-based **Speechstorm** was acquired by **Genesys,** which is based in US.

[146] https://index.co/market/speech-recognition/acquisitions

- **CreaWave**, a France-based speech recognition platform was bought by **Acapela** Group from Belgium.
- In 2011, **24/7ai** acquired US-based **Voxify**.
- A speech capture and recognition company from France (**Telisma**) was acquired by **OnMobile** from India in 2011.

Table 23 Acquisitions by region

Year	Company	HQ	Sector	Acquired by	HQ	Price mUSD
2018	Ytica	Czech Republic	Speech recognition, SAAS, machine learning	Twillo	US	N/A
2018	Liv.AI	India	Speech recognition, Artificial Intelligence, machine learning	Flipkart	India	N/A
2018	Semantic Machines	US	Speech recognition	Microsoft	US	N/A
2018	VoiceBoxTechnologies	US	Speech recognition	Nuance Communications	US	82
2017	KITT.AI	US	Speech recognition, home automation	Baidu	China	N/A
2017	Digital Roots	US	Social media monitoring	Interactions LLC	US	N/A
2017	Raven Tech	China	Speech recognition, Artificial Intelligence, machine learning	Baidu	China	N/A
2016	Vxi Corporation	US	VOIP, communication, machine learning	Jabra	Denmark	35
2015	SpeechStorm	United Kingdom	Speech recognition, machine learning	Genesys	US	N/A
2015	CreaWave	France	Speech recognition	Acapela Group	Belgium	N/A
2011	Voxify	US	Speech recognition, enterprise software	[24]7.ai	US	N/A
2008	Telisma	France	Speech capture and recognition, software	OnMobile	India	N/A

Figure 76 (based on Annex G) illustrates the funding of start-ups and venture capital enterprises. As can be seen, most funding activities are taking place in North America, while Europe is in second place.

Figure 76 Funding by region

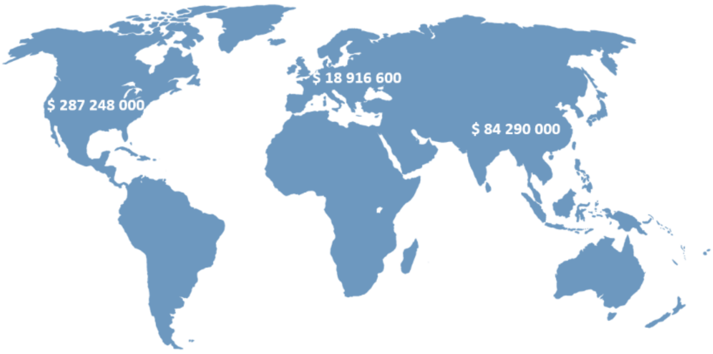


Table 24 Funding by region

Region	***Invested USD***
North America	*287 248 000*
Europe	*18 916 600*
Asia	*84 290 000*
Total (disclosed deals)	***390 454 600***

As can be seen from the extensive list in Annex G and the summary in Table 24, investment funding for developing speech technologies is clearly dominated by companies from North America, where North American companies and start-ups are getting a significant amount of funding from private funds and investors.

However, this analysis raises more issues than it resolves, as it can be observed that Asia-based AI companies (predominantly China) are getting enormous investments from the government and private sector regardless of a company's merits.[147]

At the same time, it must be noted that out of respect for competition not all companies are disclosing the details of deals or exact amounts. Therefore, the true investment amounts (especially in Asia) might be noticeably higher. Analysing the case of China, it must be noted that a significant amount of money is also attracted from public funds, which is not the typical case in the EU and North America, where investors are funds or companies. The latest noticeable case was when the

[147] https://www.usnews.com/news/best-countries/articles/2018-07-27/china-has-too-much-money-for-its-tech-startups-investors-warn

China's Ministry of Finance and China internet Investment Fund[148] financed Beijing-based mobile internet start-up Unisound.[149]

There are also examples in North America where the details of deals are not fully disclosed. For example, Google has recently launched the Assistant Investments program,[150] which invests in start-ups working on voice and assistance technologies, whether hardware or software, and focuses on the travel, games, and hospitality industries.[151]

3.6.3. Investments in search technology and services

Based on an analysis of comparative advances as summarised in Table 25, it can be concluded that Asia is the global leader in investing in search technology and services, while North America takes the second and Europe the third place.

Table 25 Market relative score in investments in search technologies

Market	Relative Score
Europe	1
North America	2
Asia	3

Based on information gathered by Index.co, Asia is dominating the market in terms of attracted investment by search companies. From 2012 to 2018 Asian search companies attracted more than 9 billion of funding, while North American companies have attracted nearly 5 and EU 1.1 billion USD.

It is worth noting that three companies in Asia have attracted more than 1 billion USD. **Baidu** has attracted 3.36 billion USD, while subsidiary of Alibaba **Koubei.com** has received 2.1 billion USD funding and Chinese-language online travel information provider and mainland search engine **Qunar.com** within the respective time period has received investments worth 1.3 billion USD. The abovementioned companies have attracted more than half of all funding in Asia's search companies.

Among Europe's search companies the most attractive in terms of funding has been the travel fare aggregator website and travel metasearch engine **Skyscanner**. The Edinburgh-based travel search firm received investments of 197 million USD, but later was acquired by Chinese travel agency Ctrip.

[148] Public fund founded by government agencies and financed by the largest state owned telecommunications companies.

[149] https://www.crunchbase.com/organization/unisound-beijing#section-overview

[150] https://developers.google.com/actions/assistant-investments

[151] https://www.techemergence.com/ai-for-speech-recognition

In North America the most recent notable investments were received by enterprise search engine **Coveo** (202 million USD), flight and hotel booking company **Hopper** (202.9 million USD) and Q&A platform provider **Quora** (229 million).

Annex L shows the information about the funding in the search category. As can be seen from the extensive list in the annex and the summary in Table 26*,* investments for developing speech technologies are clearly dominated by Asian companies.

Table 26 Investments in search technology companies by region

Region	Investments (USD)
North America	4 806 300 000
Europe	1 107 700 000
Asia	9 032 100 000

Using the Index.co database we found 73 acquisitions that have taken place in the search industry from 2012 to 2018. The investment amount is available for only 20 of the deals, the rest of the deal amounts are not available. Annex K details the information about all the acquisitions in the search category during the respective time period. Below are some relatively high-profile acquisitions in search for the deals where the investment amount is known. Unfortunately, as already noted above, the acquisition amounts are rarely disclosed, so that it is difficult to estimate the true size of this acquisition market.

- In North America the most notable deal in search industry has been closed by **Verizon** which acquired **Yahoo** for 4 480 000 000 USD in 2017. The deal is notable due to Yahoo's 39% of ownership in Alibaba's shares (the leading e-commerce company in China).
- In November 2016, **Ctrip**, the largest travel firm in China, bought **Skyscanner** for 1.75 billion USD.
- The **Priceline Group (Booking Holdings)** acquired **Momondo Group Limited** for a price of 550 million USD in 2017.
- **Randstad Innovation Fund**, an Amsterdam-based human resources and recruitment firm, acquired job hunting portal **Monster Worldwide**, for 429 million USD in 2016.[152]
- Japanese company **Lifull** acquired Spanish search engine **Mitula** in 2018. Lifull owns a portal where one can find all the necessary day-to-day services, such as mobility, leisure or real estate, and Mitula's experience is aimed to improve the portal's usability. The value of the deal was 133 million USD.[153]
- **Cisco** acquired US-based **Mind Meld** for 125 million USD in 2017.
- **Snap Inc.** bought mobile search app **Vurb** for 110 million USD in 2016.
- Spain-based **Trovit** was acquired by **NEXT Co.** The value of the deal is estimated at 90 million USD and it was closed in 2014.

[152] https://thenextweb.com/insider/2016/08/09/job-hunting-portal-monster-is-being-acquired-for-429-million

[153] https://novobrief.com/japanese-company-acquires-mitula/6560

- US-based genealogy site **Ancestry.com** acquired its competitor **Archives.com** for 100 million USD in 2012.[154]
- **Bluefin** was acquired by **Twitter** for 80 million USD in 2013.
- **Flashstock**, a leading custom content creation platform with headquarters based in Toronto, was acquired by **Shutterstock** for 65 million USD in 2017.
- **PROS**, a cloud software company, bought **Vayant Travel Technologies, Inc.**, a privately held company based in Sofia, Bulgaria for 35 million USD.
- An e-commerce platform **Etsy**, which operates marketplaces where people around the world connect sell and buy unique goods, bought US-based **Blackbird**, which provides machine learning technology to deliver search recommendations. The amount of the deal was 32,5 million USD.[155]
- **Care.com** bought a Germany-based company **Betreut.Pflege** for 23,3 million USD in 2012.
- India's e-commerce firm **Infibeam** acquired **Unicommerce eSolutions** for 18 000 000 USD in May 2018.
- **Zillow** has acquired the San Francisco-based rental and real estate search site **HotPads** for $16 million in 2012.
- A California-based search start-up **SphereUp** was acquired by Israeli search company **Zoomd** for 7 million USD in 2015.[156]
- Following the announcement of massive layoffs at **Softonic**, the Barcelona-based company acquired **AppCrawlr** for 6 million USD in 2015 to improve its mobile search engine.[157]
- In 2015 the classified vertical search leader **Mitula** acquired Barcelona-based **Nuroa**. The value of the deals is estimated at 3.3 million USD. The aim of the deal was to strengthen the position of **Mitula** in real estate vertical search. Before the acquisition, **Nuroa** owned 17 real estate vertical search sites in Europe, North America and South America.[158]
- To expand digital advertisement business, **Synacor** acquired the digital advertising platform **Technorati** for 3 million in 2016. By acquiring Technorati, a pioneer in digital advertising, Synacor aims to grow its advertising revenue.[159]

By having revived all the acquisitions in the respective time periods summarised in Table 27, it must be concluded that most investments are going into companies based in North America. This is due to the Yahoo acquisition deal which was worth 4.48 billion USD. Without the mentioned deal, Europe would have been a leader in the acquisitions.

[154] https://techcrunch.com/2012/04/25/ancestry-com-acquires-archives-com-from-inflection-for-100-million

[155] https://venturebeat.com/2016/11/03/etsy-paid-32-5-million-for-ai-startup-blackbird-technologies

[156] http://nocamels.com/2015/11/report-search-company-zoomd-acquires-sphereup-for-7m

[157] https://novobrief.com/softonic-layoffs-appcrawlr/1257

[158] http://www.propertyportalwatch.com/mitula-acquires-the-nuroa-real-estate-vertical-search-group

[159] https://adexchanger.com/platforms/synacor-acquires-technorati-to-expand-its-ad-business

Table 27 Details of acquisition deals in search industry

Name company	Region of company	Acquired by	Acquired on	Acquired amount (USD)
Yahoo	North America	Verizon	June 2017	4 480 000 000
Skyscanner	Europe	Ctrip	Nov 2016	1 750 000 000
Momondo Group Ltd	Europe	Booking Holdings (Priceline Group)	February 2017	550 000 000
Monster	North America	Randstad Innovation Fund	August 2016	429 000 000
Mitula	Europe	Lifull	May 2018	133 000 000
MindMeld, Inc.	North America	Cisco	May 2017	125 000 000
Vurb	North America	Snap Inc.	August 2016	110 000 000
Archives.com	North America	Ancestry	April 2015	100 000 000
Trovit	Europe	NEXT Co	October 2014	90 000 000
Bluefin Labs	North America	Twitter	February 2013	80 000 000
FlashStock	North America	Shutterstock	June 2017	65 000 000
Vayant	Europe	PROS	August 2017	35 000 000
Blackbird Technologies	North America	Etsy	September 2017	32 500 000
Betreut.Pflege	Europe	Care.com	July 2012	23 300 000
Unicommerce eSolutions Pvt. Ltd.	Asia	Infibeam	May 2018	18 000 000
Hotpads	North America	Zillow	Nov 2012	16 000 000
SphereUp	North America	Zoomd Inc.	Nov 2012	7 000 000
AppCrawlr	North America	Softonic	March 2015	6 000 000
Nuroa	Europe	Mitula	March 2016	3 300 000
Technorati	North America	Synacor	February 2016	3 000 000

3.7. Market dominance

Market dominance is defined as a measure of the strength of a brand, product, service, or firm, relative to competitive offerings, including the extent a product, brand or firm controls a product category in a given geographic area.[160] We analysed the market dominance in all three areas by comparing total web traffic (e.g. number of times a unique IP address has entered the webpage of the said company) received by the dedicated web domains of the largest providers of the respective LT services.

We selected two web traffic analysis tools for collecting and analysing data. We used Semrush to analyse market dominance in all categories and subcategories, except web search service providers, where we selected the specialised analysis tool Statcounter, that specifically provides web search tool traffic analytics.

[160] https://en.wikipedia.org/wiki/Dominance_(economics)

3.7.1. Market dominance in machine translation

In this study the main indicator for measuring market dominance is web traffic attracted by MT service providers.

Based on the analysis, North America clearly dominates the market in terms of attracting customers to its services. With their relatively small number, but clearly dominating presence and market penetration, the Asian MT companies are snapping at the heels of the North American companies. There is a greater number of European companies, but their market presence is more fragmented, resulting in a weaker market position.

Table 28 Market relative score in market dominance in machine translation

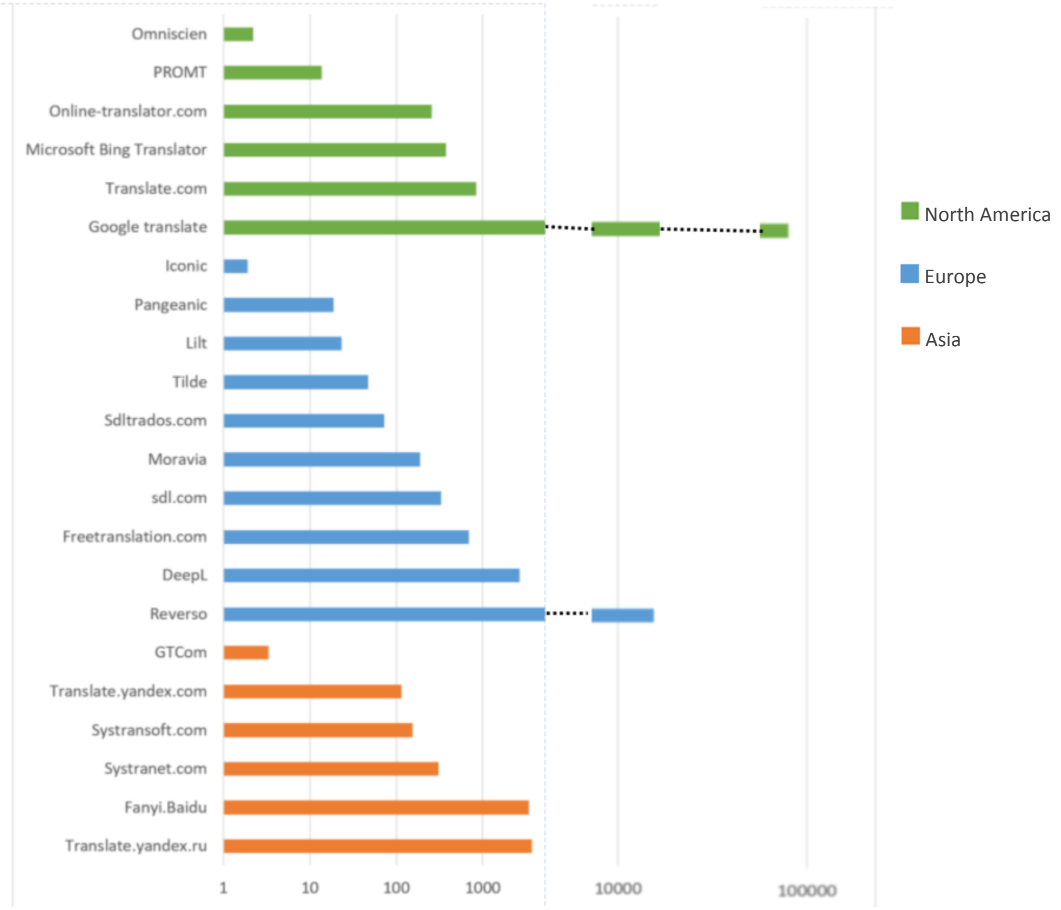
Market	Relative Score
Europe	2
North America	3
Asia	1

In order to get a high-level outlook of the visibility and brand awareness of different MT service providers, the authors gathered and analysed the total web traffic[161] received by web domains of the 22 largest MT companies.[162] Detailed statistics are shown in Annex F.

[161] Number of times a unique IP address has entered the webpage of the said company (i.e. the total number of page access events).

[162] The selection of MT companies is based on the "Slator Neural Machine Translation Report 2018" and "The Top 100 LSPs in 2018" by Common Sense Advisory.

Figure 77 Monthly website visits, average March-Sept. 2018, logarithmic scale



As the largest MT companies (with their respective brands and services) are headquartered in the US, the MT landscape is dominated by North American providers. The North American MT industry clearly outperforms European and also Asian businesses in terms of market power and dominance.

North American MT providers also have strong market positions in Asia and Europe. Asian markets have strong rivals to North America based global MT companies such as Baidu and Yandex.

As the global MT market has a very high degree of concentration – 20% of the market players, a “mix of big Internet, pure-play MT and Large LSP/MLV companies such as Google, Systran, Microsoft, SDL” (TAUS, 2017), accounts for more than 80% of the revenue and a majority of MT companies earn less than a million euros annually – there is in fact a low level of competition overall and even more so in the markets outside North America.

According to TAUS estimations (TAUS, 2017), more than 40% of the global MT market is dominated by “a small set of very big ‘Internet’ companies including Google, Amazon, Microsoft, Yandex, Facebook and Baidu, who offer a free MT service either to all-comers or to their global customers (Amazon), and/or in certain cases a paying service to enterprises and other large-scale users”.

As a result of the dominance by large players both in B2B and B2C markets, smaller MT developers and service providers, including the majority of European based companies, are facing challenges in gaining market visibility and increasing their brand awareness.

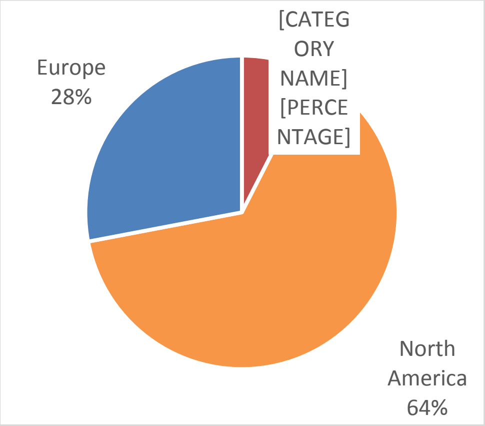
As seen in Figure 77 above, Google Translate is the de facto leader of the largest MT brands, attracting more than 2.6 times the web traffic of its closest competitor. It is worth noting that Google Translate is also natively integrated into the Chrome web browser which similarly leads web browser popularity charts with 56.9% market share.[163] This would likely increase its position as a market leader even further.

Free online MT as a service, e.g. Google, freetranslation.com (powered by Microsoft) and Reverso, does not only have a major impact on the MT market in terms of the perceived value – MT services have been commoditised, even devaluated – but also has a strong impact on the perceived quality expectations of both individual consumers and businesses. As stated in the results of Task 1, "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."

On the other hand, there may be an improved technology response to the constant global demand for more, cheaper content translation with the growth of NMT. "In this case, the current wave of corporate and government interest in machine (deep) learning in general as an effective solution for many enterprise data-related concerns should encourage greater take-up of 'neural' solutions in the translation industry." (TAUS, 2017)

To better compare the European MT company market strength, we propose to compare the total summed traffic from all of them to everyone else in the list.

Figure 78 Web traffic share by region



[163] http://gs.statcounter.com/browser-market-share

Our analysis indicates that the major European MT players are capable of attracting almost a third (28%) of the global targeted web traffic.

3.7.2. Market dominance in speech technologies

In this study, the main indicator for measuring market dominance is web traffic attracted by speech technology service providers.

In order to conduct market dominance analysis for speech technology, we have looked at the two main subcategories and the respective service and technology suppliers – (a) speech and voice recognition technology providers, and (b) voice synthesis and text-to-speech technology providers, analysing the web traffic to the dedicated websites and landing pages of the top industry players.

Our analysis indicates that large North American multinationals have an excessive market dominance in terms of recognition and ability to attract web traffic to their websites. (It has to be noted that we were not able to collect data and analyse domestic traffic for any speech technology provider's website in China, e.g. Baidu, as none of the popular web analysis tools enables traffic analysis of the Chinese domestic internet.)

Table 29 Market relative score in market dominance in speech technologies

Market	Relative Score
Europe	2
North America	3
Asia	1

In order to get a high-level outlook of the visibility and brand awareness of different speech technology service providers, we gathered and analysed the total web traffic[164] received by dedicated web domains of the 25 largest speech technology companies.[165] Detailed statistics are shown in Annex H.

3.7.2.1. Speech and voice recognition market

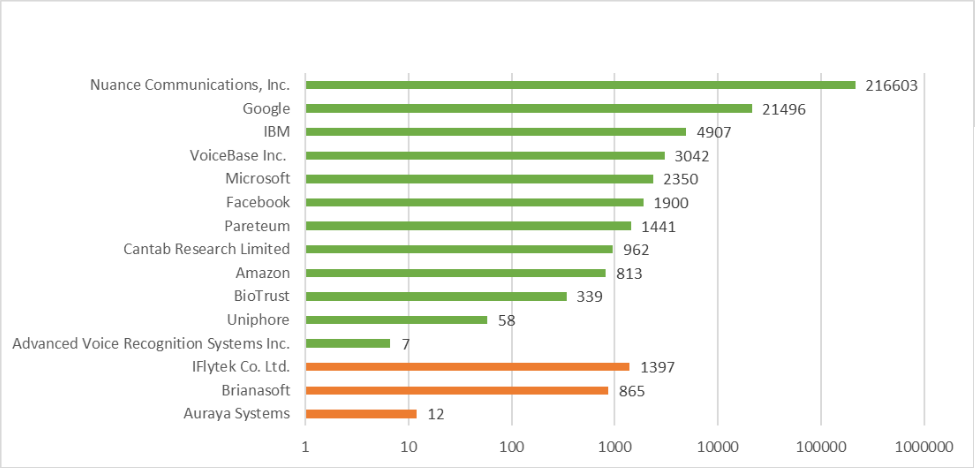
The speech and voice recognition market is almost completely dominated by the large US-based global corporations that are using the speech recognition technology as part of their product or service functionality enhancing closer communication with the end user. The overall popularity of

[164] The number of times a unique IP address has entered the webpage of the said company (i.e. the total number of page access events) during the defined period of time.

[165] The selection of speech technology companies is based on the "Speech and Voice Recognition Market by Technology, Vertical and Geography - Global Forecast to 2023" and "Text-to-Speech Market by Vertical, and Geography – Global Forecast to 2022" by marketsandmarkets.com.

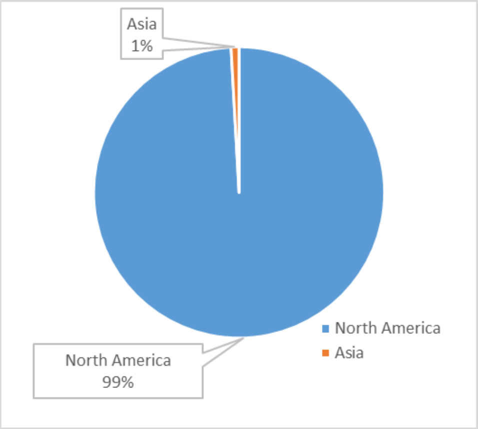
this technology is rapidly increasing as we are moving toward the commoditisation of the natural user language interfaces.

Figure 79 Monthly speech synthesis dedicated website visits, average March-Sept. 2018, logarithmic scale



Our analysis clearly demonstrates the extensive market dominance by the North American players followed by a tiny fraction of the web traffic to the 3 Asia-based company websites (Brianasoft, IFlytek, Auraya Systems). The web traffic of the leading speech recognition service provider Nuance exceeds the closest follower Google by tenfold, and Google surpasses the next in the row eight times. There are no European companies among the 15 largest speech/voice recognition service providers.

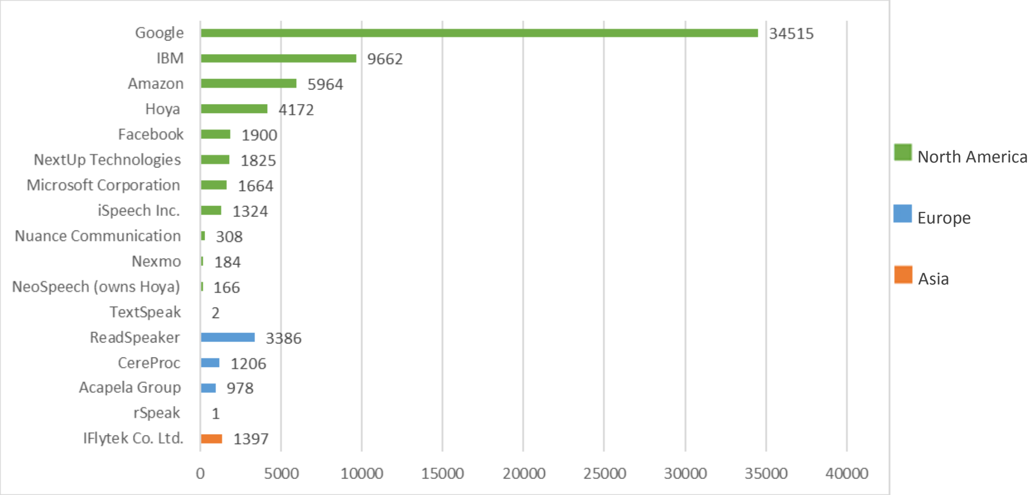
Figure 80 Web traffic share by region



Our analysis indicates that most of the internet traffic to the websites of the speech and voice recognition service providers is aimed at the end product or service rather than the underlying technology itself.

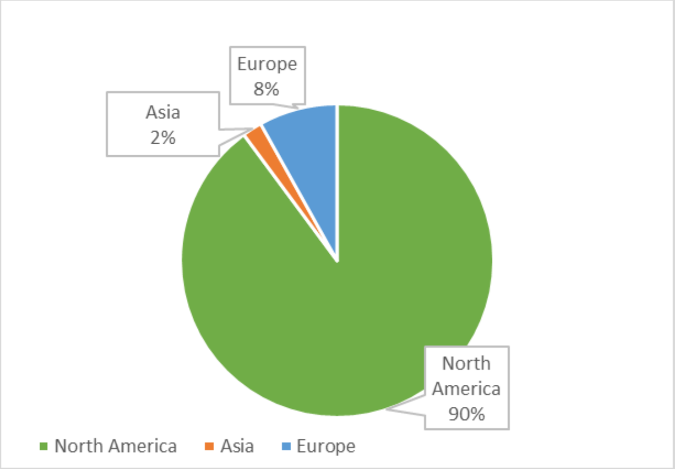
3.7.2.2. Voice synthesis and text-to-speech market

Figure 81 Monthly speech synthesis dedicated website visits, average March-Sept. 2018



In the speech synthesis market (Figure 81), the US-based tech giants outmatch the top companies focussed on speech synthesis alone (exc. Hoya) multiple times, with Google being a clear leader. At the same time, organic web traffic to their speech synthesis dedicated web addresses forms just a tiny fraction of the general traffic to their main websites.

Figure 82 Web traffic share by region



Our analysis of the major speech synthesis market players suggests clear leadership of the North America based service providers, followed by European players, and Asia-based companies. It must be noted though that a large portion of the traffic to the dedicated websites of service providers in e.g. China could be domestic and is not accessible to web analytics tools.

The major factors driving the growth of the speech technology market and affecting the market dominance of certain players in the future are the effective integration of the technologies due to increased demand for voice and speech-based biometric systems, the increase in demand for voice communication in mobile applications, and the use of artificial intelligence to improve the accuracy of speech and voice recognition and synthesis.

3.7.3. Market dominance in search technology and services

In order to conduct a market dominance analysis for search technology, we have looked at the two main subcategories and the respective service and technology suppliers – (a) web search providers, and (b) enterprise search tool providers, analysing the relative market share of the main web search companies and the web traffic to the dedicated websites and landing pages of the top enterprise search tool providers.

Our analysis shows clear dominance of Google Search in the web search market of the respective regions, whereas the enterprise search tool market is led by the European company Elasticsearch.

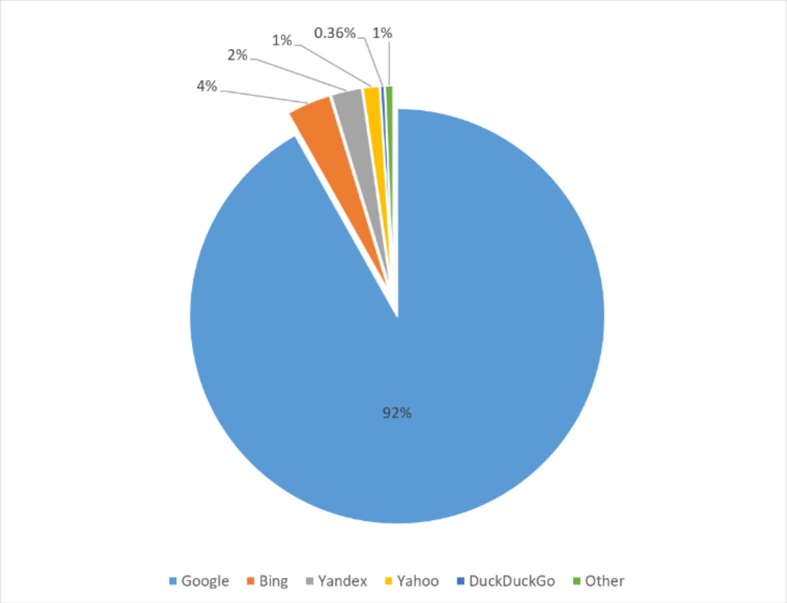
Table 30 Market relative score in market dominance in search technology

Market	Relative Score
Europe	2
North America	3
Asia	1

3.7.3.1. Web search market

We analysed the web search market in Europe, North America and Asia using the information from StatCounter[166], a web analytics service with tracking code installed on more than 2 million sites globally.

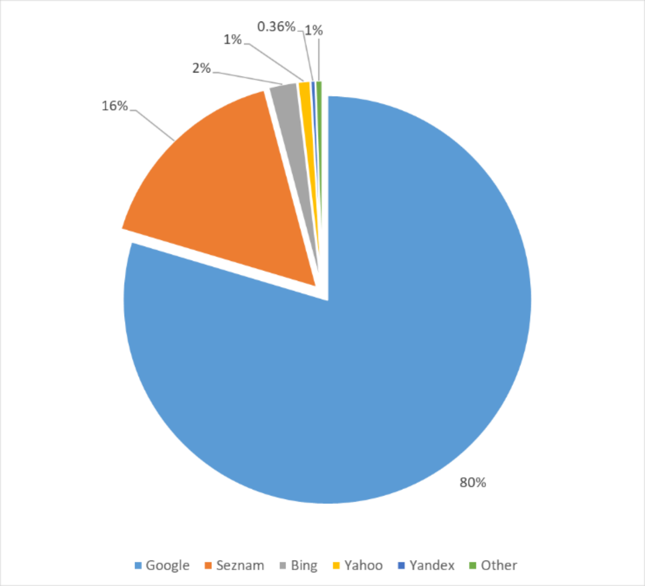
Figure 83 Average monthly search engine market share: Europe, April-Sept. 2018



Our analysis (see Figure 83) confirms the dominance of Google web search services in Europe, leaving other players less than 10% of the total market in Europe. The exception is the Czech Republic, where the local web search service provider Seznam has managed to take a comparatively large stake of the local web search market (Figure 84).

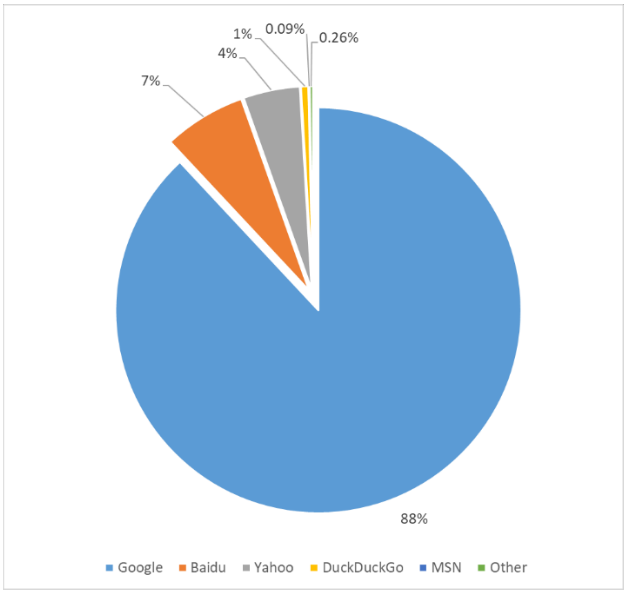
[166] http://gs.statcounter.com

Figure 84 Average monthly search engine market share: Czech Republic, April-Sept. 2018



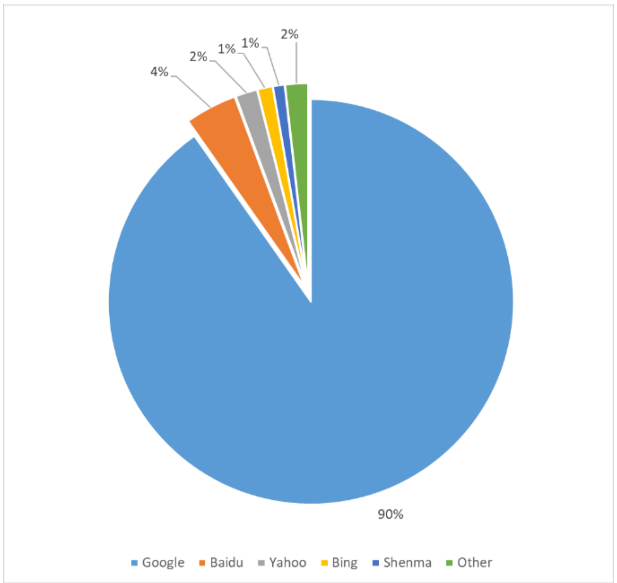
The North American web search market is also strongly dominated by Google (88%) followed by Baidu (7%) and Yahoo (4%).

Figure 85 Average monthly search engine market share: North America, April-Sept. 2018



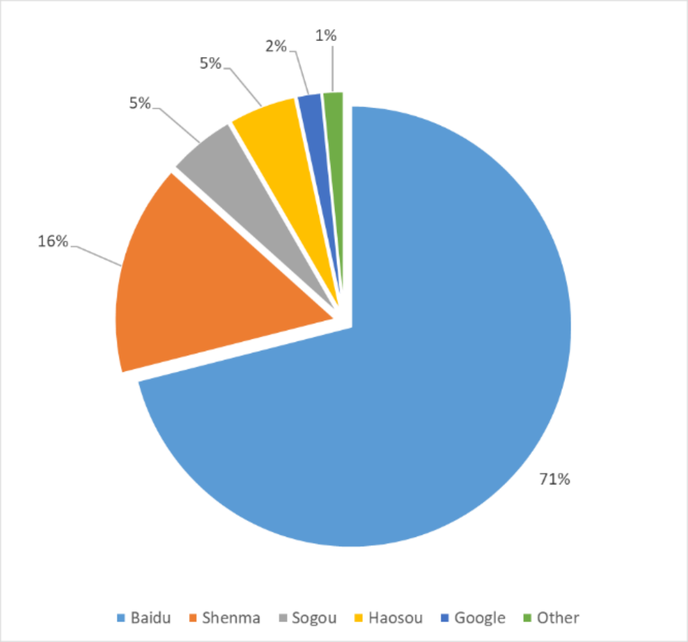
Google web search also has the largest (90%) market share in Asia (Figure 86). The rest is divided between 2 major Chinese players – Baidu (4%) and Shenma (1%), and Yahoo and Bing with 2% and 1% of the total market respectively.

Figure 86 Average monthly search engine market share: Asia, April-Sept. 2018



The exception is the Chinese market (Figure 87), where largely due to the political environment and government control of the internet and its resources a major share of the search market belongs to the local providers Baidu (71%), Shenma (16%), Sogou (5%), Haosou (5%), leaving Google with a tiny 2% of the total market.

Figure 87 Average monthly search engine market share: China, April-Sept. 2018

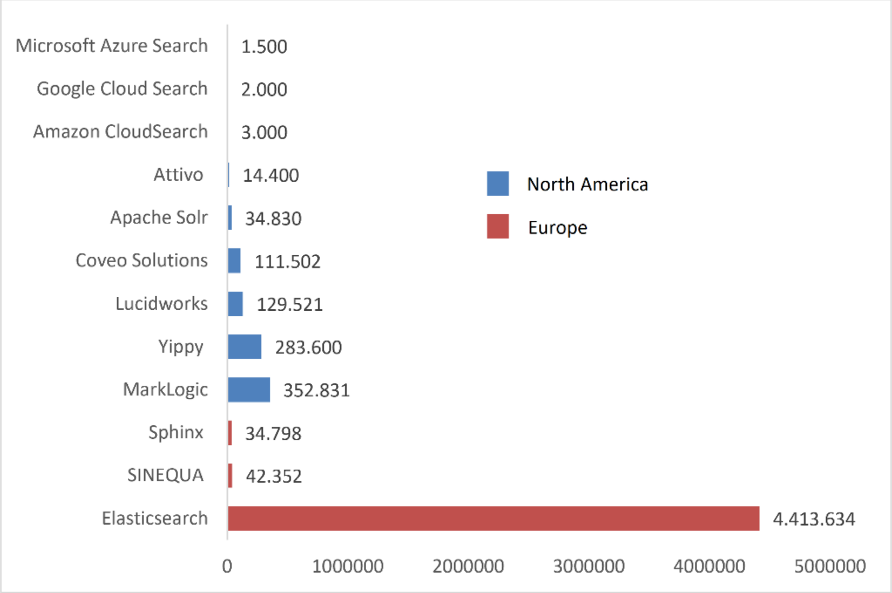


3.7.3.2. Enterprise search tools market

In order to get a high-level outlook of the visibility and brand awareness of different enterprise search technology service providers, the authors gathered and analysed total web traffic[167] received by dedicated web domains of the 12 largest enterprise search technology companies[168].

Enterprise search tool providers have relatively small web traffic to their dedicated websites and/or landing pages. The market is dominated by the European and North American providers specialised in specific application verticals (Figure 88).

Figure 88 Monthly enterprise search tool dedicated website visits, average March-Sept. 2018

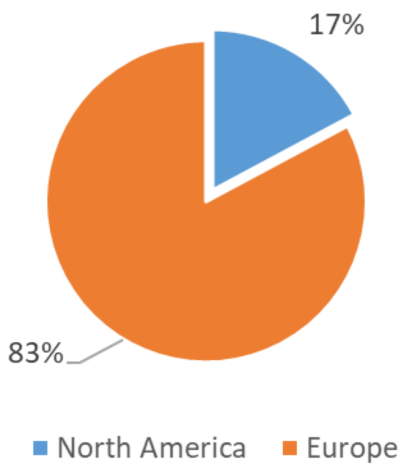


The European company Elasticsearch is dominating the enterprise search tool market, outpacing the closest rivals by tenfold. The North American search engine market is also dominated by specialised search solution providers such as MarkLogic, Yippy and Lucidworks, outperforming the respective service providing units of global giants such as Microsoft, Google and Amazon. None of the Asian companies has been able to attract any significant web traffic, at least outside the Chinese market which has limited access as for web traffic analysis tools.

[167] The number of times a unique IP address has entered the webpage of the said company (i.e. the total number of page access events) during the defined period of time.

[168] Methodology: by reviewing three lists we counted mentions in each popularity list (if the enterprise got mentioned in one review it got one point, if it was mentioned in two reviews it scored '2' etc. Based on the methodology we created a list that reflects the most popular enterprise search engines. Note: to avoid subjectivity, we eliminated the tools that have only one mention.

Figure 89 Web traffic share by region



As Elasticsearch has been able to attract a large part of the dedicated web traffic, the European providers are put in the clear lead of the web traffic comparison.

3.8. Industry

Industry in the context of this study is defined as the commercial machine translation product developers and service providers.

The criteria for measuring the Industry dimension are the market capitalisation and estimates of market revenues of the companies that can be identified as being engaged in language services and specifically in machine translation development and implementation.

3.8.1. Machine translation industry

Table 31 Market relative score in machine translation industry

Market	Relative Score
Europe	1
North America	3
Asia	2

North America exhibits global dominance due to US-based tech giants with the Asia region developing quickly based on several giant Chinese e-commerce companies that have a greater average market capitalisation in comparison to the EU-based companies. The EU lags behind as European companies have a lesser global presence.

As illustrated in Table 32, 13 out of the top 20 global companies by market capitalisation are US-based companies, 5 are Asian, and only 2 are European.

Table 32 Top 20 global companies by market capitalisation and their MT activities, as of March 31, 2018

COMPANY NAME	NATIONALITY	INDUSTRY	MARKET CAP 2018 ($B)	IN-HOUSE MT
Apple	United States	Technology	851	MT
Alphabet	United States	Technology	719	MT
Microsoft	United States	Technology	703	MT
Amazon	United States	Consumer Services	701	MT
Tencent	China	Technology	496	MT
Berkshire Hathaway	United States	Financials	492	
Alibaba	China	Consumer Services	470	MT
Facebook	United States	Technology	464	MT
JPMorgan Chase	United States	Financials	375	
Johnson & Johnson	United States	Health Care	344	
ICBC	China	Financials	336	
Exxon Mobil	United States	Oil & Gas	316	
Bank of America	United States	Financials	307	
Samsung Electronics	South Korea	Consumer Services	298	MT
Walmart	United States	Consumer Services	264	
Royal Dutch Shell	United Kingdom	Oil & Gas	263	
China Construction Bank	China	Financials	259	
Wells Fargo	United States	Financials	256	
Nestle	Switzerland	Consumer Goods	246	
Visa	United States	Financials	246	

It can be clearly seen that North America, and specifically the US, absolutely dominates the market by number of companies, their average market capitalisation, and their top positions in the top global companies. The Asia region, dominated by China, comes in second, and Europe has only two contenders in the global top 20 by market capitalisation.

The top global companies based in North America also dominate the technology sector and by association the machine translation market (as referenced in Table 32 with “MT”), if not directly by revenue, then by market penetration with their global reach and ambition.

As revenue from machine translation is not always the strategic goal of the companies developing technologies, we have extrapolated the monetary size of the markets from various sources.

- The market researcher Common Sense Advisory (CSA) has estimated that the total global market for language services in 2016 is around $40 billion (EUR 35 billion).
- Of the total global market for language services Business wire estimates the global machine translation market to be valued at $306.6m in 2017. It is projected to expand at a CAGR (compound annual growth rate) of 20.42% over the forecast period to reach $934.76m by 2023.

- Globenewswire estimated the global machine translation market to be worth $250m in 2012 and expected it to reach approximately $1 480m by 2022, with a potential CAGR of 19.40% during the forecast period.[169]
- In 2014, TAUS estimated the global machine translation market to be worth around $250 million.
- In its 2017 annual report SDL estimates the global language technology market to be $20b with MT accounting for $386m of the total.

By comparing these independent estimations, we can assume that the global machine translation market in 2017 was worth $300m – $350m with an annual growth rate close to 20%.

According to the study in Task 1, the estimated European market for translation technologies is EUR 67m ($78.3m). This would lead to an estimation of the share for the European MT market in a range of 22%-26% or about a quarter of the global market.

A TAUS study attributes the relatively small size of the market to the disruptive business model of globally dominating online service providers: "*this leads to the paradoxical situation that although the MT market is vibrant and evolving very rapidly, the sector remains relatively small in terms of value, with fierce and often disguised competition. Well-established MT players may find themselves suddenly competing with eBay (which acquired the MT developer AppTek in 2014) or Facebook (which acquired the speech translation application Jibbigo in 2013) or Amazon (which acquired Safaba in 2015). Like Google, Microsoft, Baidu, and Yandex, these Internet giants do not primarily make money from the MT technology itself but their investments and offerings have of course a disruptive and innovative effect on the MT market.*" (TAUS, 2017, p. 27)

Still, MT provides a strong export market for European companies. According to the CSA analysis, "*the majority of current demand for machine translation services comes from North American tech firms, but the overwhelming majority of global supply comes from small and medium enterprises in Europe*" (Lommel et al., 2016).

[169] https://globenewswire.com/news-release/2017/11/21/1197950/0/en/Machine-Translation-Market-Global-Industry-Insights-Trends-Outlook-and-Opportunity-Analysis-2012-2022.html

3.8.2. Speech technology industry

Table 33 Market relative score in speech technology industry

Market	Relative Score
Europe	1
North America	3
Asia	2

Within the speech technology industry there are two distinct segments. The first segment consists of large developers for whom speech technologies are a competitive advantage technology for enhancing, popularising, and marketing other products and services such as Alexa by Amazon. This could be considered a B2C (business to consumer) segment. The second segment consists of developers for whom the technology itself is the product, and who supply speech technologies as a service, for instance, Nuance supplies speech recognition software for use by Daimler in automobiles.[170] This could be considered a B2B (business to business) segment. However, as illustrated by the emergence of examples of the integration of Alexa in third party services such as Logitech, which has built Alexa into its Harmony remote units to control home entertainment systems and smart home devices and "*at the CES 2018 in Las Vegas, Sony, TiVo and Hisense unveiled smart home skills that integrated Alexa, enabling customers to control the TV by voice, the original B2C providers are muscling in on the B2B market. Home appliance makers such as Whirlpool, Delta, LG and Haier have also added Alexa's voice-recognition skills to help people control all aspects of their home, from TVs and microwaves to air conditioning units and faucets.*"[171] The B2C segment is moving beyond its initial focus.

The speech technology market (which includes automatic speech recognition, text-to-speech and similar services) is growing every year and it is forecasted that the market will grow from $3.7 billion a year to about $10 billion a year by 2022,[172] reaching $22.32 billion by 2024.[173] The statistics show

[170] https://www.worldscientific.com/doi/pdf/10.1142/S1363919699000177 Corporate Activities in Speech Recognition and Natural Language: Another "New Science" Based Technology, K. Koumpis and K. Pavitt, International Journal of Innovation Management VOL. 03, NO. 03.

[171] https://www.techemergence.com/ai-for-speech-recognition

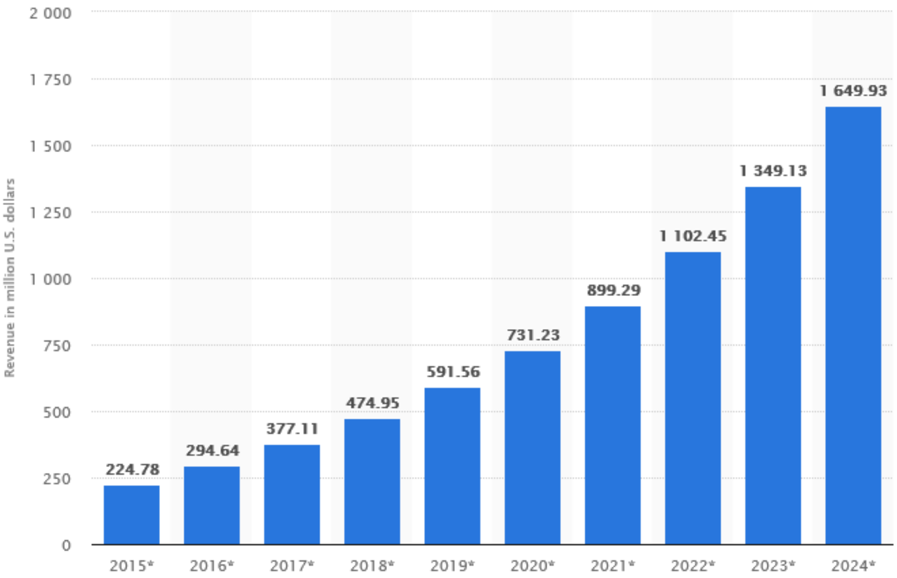
[172] More information is in https://blog.neospeech.com/top-5-open-source-speech-recognition-toolkits.

[173] More information is in https://globenewswire.com/news-release/2018/08/22/1555231/0/en/Speech-and-Voice-Recognition-Technologies-Market-Will-Reach-USD-22-32-Billion-Globally-By-2024-Zion-Market-Research.html.

that the size of the European speech technology market alone will reach over $1.6 billion by 2024 (see Figure 90).[174]

According to Businesswire research, the market for various speech recognition services is forecasted to grow approximately seven times between 2017 and 2025 at a CAGR of approximately 24-27%.[175] [176] [177]

Figure 90 Forecasted European speech technology market growth till 2024



This growth is due to the versatility of the technology. Speech recognition engines are able not only to understand what humans are saying, but also to convert the speech audio signal into text or vice versa very precisely. Developers are integrating speech recognition/synthesis engines into their applications; home appliance users rely on these technologies to accomplish everyday domestic tasks. There are many innovative ways to use speech technologies in business, [178] such as the following:

- Automatic transcription
- Hearing and understanding customers

[174] The statistics about the European speech technology market are taken from https://www.statista.com/statistics/608587/europe-voice-speech-recognition-software-market.

[175] https://www.businesswire.com/news/home/20180423005643/en/North-America-Speech-Voice-Recognition-Market-Analysis

[176] https://www.businesswire.com/news/home/20180417005875/en/European-1.66-Billion-Speech-Voice-Recognition-Market

[177] https://www.businesswire.com/news/home/20180423005631/en/Asia-Pacific-Speech-Voice-Recognition-Market-Analysis

[178] Innovative ways how to use speech recognition for business are described in https://www.transcribeme.com/blog/8-innovative-ways-to-use-speech-recognition-for-business.

- Streamlining support processes
- Identifying callers and mitigating risk
- Monitoring support agents and representatives
- Getting more data on customer demographics
- Writing important e-mails quickly and accurately
- Making use of voice data and continuously optimising business processes

In both of the above noted segments, speech technology has overwhelmingly been developed by North America based companies, followed by their Asian counterparts for whom speech technologies are a means by which to better penetrate consumer markets. In the B2B segment, companies for whom speech technologies are their core business are also overwhelmingly based in North America.

Table 34 illustrates the forecast growth of the speech technology market across the economic regions. Although the growth in demand for speech technologies is similar across the regions, European countries are at a disadvantage in developing and delivering these services. Speech recognition services are overwhelmingly available from North America based companies, with Asian companies actively seeking to penetrate the European market as well.

Table 34 Forecasted growth of speech technologies

REGION	Y 2017	Y 2023/25
NORTH AMERICA	312 m USD	2.0 b USD
EU	287 m USD	1.7 b USD
ASIA	397.5 m USD	2.8 b USD

As can been seen from Table 35[179], leading market players in voice recognition software development are located in North America (predominantly in the US).

[179] The selection of speech technology companies is based on the "Speech and Voice Recognition Market by Technology, Vertical and Geography - Global Forecast to 2023" and “Text-to-Speech Market by Vertical, and Geography – Global Forecast to 2022” by marketsandmarkets.com.

Table 35 Leading market players in voice recognition (listed in alphabetical order)

	COMPANY	COUNTRY	REGION
1	Acapela Group	Belgium	Europe
2	Alphabet Inc.	US	North America
3	Amazon.Com	US	North America
4	Baidu	China	Asia
5	Cantab Research Limited	UK	Europe
6	CereProc	UK	Europe
7	Facebook	US	North America
8	Google	US	North America
9	Hoya	US	North America
10	IBM	US	North America
11	Iflytek Co., Ltd.	China	Asia
12	iSpeech Inc.	US	North America
13	LumenVox LLC	US	North America
14	Microsoft Corporation	US	North America
15	NeoSpeech	US	North America
16	Nexmo	US	North America
17	NextUp Technologies	US	North America
18	Nuance Communication	US	North America
19	Pareteum Corporation	US	North America
20	rSpeak	The Netherlands	Europe
21	Sensory Inc.	US	North America
22	SESTEK	Turkey	Other
23	TextSpeak	US	North America
24	VoiceBox Technologies Corp.	US	North America
25	VoiceVault Inc.	US	North America

North America exhibits global dominance due to US-based tech giants, with the Asia region developing quickly based on several giant Chinese e-commerce companies that have a greater average market capitalisation in comparison to the EU-based companies. The EU lags behind as European companies have a lesser global presence. For example, IFlytek is China's leading voice recognition tech company and dominates the segment with a more than 70% market share. Still, less than 1% of its sales come from overseas, which it plans to change shortly.[180] The company's dominance in China is obvious, but it is a different story outside China. Google, Microsoft and Apple are the global leaders in voice recognition technology, while iFlytek's market share is still far from double digits. A barrier to Flytek's expansion is that it does not receive support from foreign governments like it does from the Chinese government. In addition, many overseas clients remain cautious about Chinese companies.[181]

[180] https://asia.nikkei.com/Business/Company-in-focus/China-s-leader-in-voice-recognition-chases-Google-and-Microsoft

[181] https://asia.nikkei.com/Business/Company-in-focus/China-s-leader-in-voice-recognition-chases-Google-and-Microsoft

As illustrated in Table 36, 13 out of the top 20 global companies by market capitalisation are US-based companies, 5 are Asian, and only 2 are European. As per publicly available information, most of the top 20 global companies by market capitalisation have developed or use third party speech recognition services.

Table 36 Top 20 global companies by market capitalisation and their speech technology activities, as of 31/03/2018

	COMPANY	HQ	INDUSTRY	MARKET CAP 2018 ($B)	ST	IN-HOUSE
1	Apple	US	Technology	851	B2C, B2B	Yes
2	Alphabet	US	Technology	719	B2C, B2B	yes
3	Microsoft	US	Technology	703	B2C, B2B	yes
4	Amazon	US	Consumer Services	701	B2C, B2B	yes
5	Tencent	China	Technology	496	B2C	yes
6	Berkshire Hathaway	US	Financials	492		
7	Alibaba	China	Consumer Services	470	B2C, B2B	yes
8	Facebook	US	Technology	464	B2C	yes
9	JPMorgan Chase	US	Financials	375	B2C	no
10	Johnson & Johnson	US	Health Care	344	B2C	no
11	ICBC	China	Financials	336	B2C	no
12	Exxon Mobil	US	Oil & Gas	316		
13	Bank of America	US	Financials	307	B2C	
14	Samsung Electronics	South Korea	Consumer Services	298	B2C	yes
15	Walmart	US	Consumer Services	264	B2C	no
16	Royal Dutch Shell	UK	Oil & Gas	263		
17	China Construction Bank	China	Financials	259	B2C	no
18	Wells Fargo	US	Financials	256	B2C	no
19	Nestle	Switzerland	Consumer Goods	246	B2C	no
20	Visa	US	Financials	246	B2C	No

An overwhelming proportion of the top 20 global companies by market capitalisation (17 out of 20, as could be learned from publicly available information) are either selling, using or both selling and using in terms of integrating speech technologies in their business workflows.

3.8.3. Search technology and service industry

Table 37 Market relative score in search technology industry

Market	Relative Score
Europe	1
North America	3
Asia	2

We have identified three segments in this market: publicly available B2C (such as Google) and, for internal company use, B2B (such as Amazon) and the technologies underlying both segments.

In all three segments the industry is clearly dominated by the North America based search giants Google and Microsoft and the underlying Apache technology. Search is clearly a market defining and influencing technology for information retrieval and analysis potential. While the giant North America based search companies have the European and Arabic language markets wrapped up,[182] Asian companies are fighting it out in their home markets for dominance. There is a greater number of European search companies offering enterprise services and specific languages. Therefore, their market presence is more fragmented, resulting in a weak position.

Although there appears to be activity on the research side of cross-lingual information retrieval, on the consumer side it does not appear to be a hot feature. While the leading search engines support multiple languages, each search in each language must be completed separately, the exception being the Google update noted below.

In 2013 Search Engine Land reported that "Google has quietly dropped the 'Translated Foreign Pages' search filter from the Google search options menu",[183] but in 2018 Google has resurrected a multilingual search feature. The only other company that has recently announced a CLIR system is Sogyou. Sogyou is a relatively small Chinese online search company. Even the Sogyou system – created in cooperation with Microsoft – is very basic. First it translates from Chinese to English, then it searches and then it translates the English information back to Chinese.[184]

Oracle and IBM, which have significant ERP system businesses, both have versions of CLIR, based on machine translation, built into the database management that underlies their ERP systems.

As illustrated in Table 38, among the top 20 global companies by market capitalisation, North America exhibits global dominance in search using either proprietary software or using the software developed by peers. There are no European companies in the global top 20 that have a market presence in the search market for either B2C or B2C or underlying technology segments.

Table 38 Top 20 global companies by market capitalisation and their activities on search, as of 31/03/2018

	COMPANY NAME	NATIONALITY	INDUSTRY	MARKET CAP 2018 ($B)	IN-HOUSE SEARCH
1	Apple	US	Technology	851	Google, MS
2	Alphabet	US	Technology	719	B2C, B2B
3	Microsoft	US	Technology	703	B2C, B2B
4	Amazon	US	Consumer Services	701	B2C, B2B
5	Tencent	China	Technology	496	Sogou (MS)
6	Berkshire Hathaway	US	Financials	492	
7	Alibaba	China	Consumer Services	470	B2C (inhouse,

[182] https://www.extradigital.co.uk/articles/seo/arabic-search-engines.html

[183] https://searchengineland.com/google-drops-translated-foreign-pages-search-option-due-to-lack-of-use-160157

[184] http://www.chinadaily.com.cn/business/2016-05/19/content_25370528.htm

					Aliyun)
8	Facebook	US	Technology	464	B2C
9	JPMorgan Chase	US	Financials	375	
10	Johnson & Johnson	US	Health Care	344	Google
11	ICBC	China	Financials	336	
12	Exxon Mobil	US	Oil & Gas	316	
13	Bank of America	US	Financials	307	
14	Samsung Electronics	South Korea	Consumer Services	298	
15	Walmart	US	Consumer Services	264	B2C (inhouse, Polaris)
16	Royal Dutch Shell	United Kingdom	Oil & Gas	263	
17	China Construction Bank	China	Financials	259	
18	Wells Fargo	US	Financials	256	
19	Nestle	Switzerland	Consumer Goods	246	
20	Visa	US	Financials	246	

The data indicate that although more than half of the content of the internet currently is in English, currently for Chinese[185] it is only 1.7% as measured by w3techs. It can be reasonably expected that the content in Chinese will grow significantly. The European markets are relatively mature in both content and percentage of users. Additionally, according to Bloomberg, the economic growth of Europe is expected to slow down, while that of China is expected to continue to be robust over the next 5 years.[186] Therefore, it can be expected that the majority of industry developments will be focused on gaining a larger share of the Asian market. Countries with a larger global economic impact have a share of internet content that is larger than the number of their internet users. For instance, German internet content accounts for 6.2% of internet content, but German speaking internet users account for only 2.2% and English language content accounts for 53.6% of content but only 25.4% of global internet users.

[185] https://w3techs.com/technologies/overview/content_language/all

[186] https://www.bloomberg.com/news/articles/2018-03-06/china-s-economy-is-set-to-overtake-combined-euro-area-this-year

3.9. Infrastructure

In this study, infrastructure is defined as technical (computing) infrastructure needed for developing, running and utilising machine translation services.

While organisational infrastructure could be evaluated as rather similar for all three regions (e.g. there is an MT association in each region), the computational infrastructure is much more developed by North American headquartered global players (e.g., *Google, Microsoft,* and *Amazon*). Europe lacks computational resources. This could be an obstacle and result in slower R&D of MT in Europe.

Table 39 Market relative score in infrastructure

Market	Relative Score
Europe	2
North America	3
Asia	1

Availability and access to computing infrastructure is key to developing competitive machine translation services. For this analysis we made a regional comparison of both generic ICT infrastructure as well as cloud computing resources affecting development and usage of machine translation services.

As an indicator of availability and access, the Network Readiness Index, prepared by the World Economic Forum, scores countries on how well they have embraced the factors that make them competitive in the digital world and establishes a quantitative hierarchy of the most ICT-ready countries by 53 various factors. In this index, Europe takes 11 out of the top 20 places, however, except for Germany and the UK, it is mostly represented by Europe's smaller economies. North America is fully represented, and although Asia is well represented, China (Asia's largest economy by GDP) does not make the top 20 and can be found in the 59th place out of 139.[187]

[187] http://reports.weforum.org/global-information-technology-report-2016/networked-readiness-index

Table 40 Top 20 countries by Network Readiness Index (NRI)

Rank	Country	NRI	Region
1	Singapore	6	Asia
2	Finland	6	Europe
3	Sweden	5.8	Europe
4	Norway	5.8	Europe
5	United States	5.8	North America
6	Netherlands	5.8	Europe
7	Switzerland	5.8	Europe
8	United Kingdom	5.7	Europe
9	Luxembourg	5.7	Europe
10	Japan	5.6	Asia
11	Denmark	5.6	Europe
12	Hong Kong SAR	5.6	Asia
13	Korea, Rep.	5.6	Asia
14	Canada	5.6	North America
15	Germany	5.6	Europe
16	Iceland	5.5	Europe
17	New Zealand	5.5	Pacific
18	Australia	5.5	Pacific
19	Taiwan, China	5.5	Asia
20	Austria	5.4	Europe

Rapid development of cloud computing democratises access to high performance computing needed for developing state-of-the-art machine translation systems. Running machine translation services in the cloud also dramatically extends the reach of machine translation.

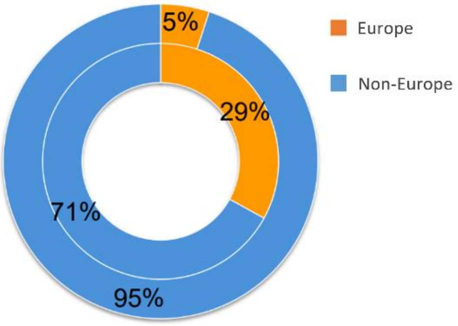
According to estimates by the European Commission, Europe needs to invest close to $800bn in its digital infrastructure to catch up with the United States and China.[188] Although this is a total estimate that includes investments in fiber optics networks, 5G networks and other ICT infrastructure, a substantial part of these investments are needed to meet European demand for high performance computing power.

Europe is lagging behind other global economic powers in providing computing power for computing intensive applications such as machine translation. Although Europe consumes 29% of global HPC resources it supplies less than 5% of them (Figure 91).[189]

[188] https://www.reuters.com/article/us-europe-digitalization-oettinger-idUSKCN1174M9?il=0

[189] Impact assessment. Accompanying the document *Proposal for a Council Regulation on establishing the European High Performance Computing Joint Undertaking*.

Figure 91 Europe's consumption of the global HPC resources (29%) versus HPC resources supplied in Europe (5%)



3.10. Data

3.10.1. Data for machine translation

In the context of machine translation, we analysed the availability of data that is used for the development of machine translation systems. Data is crucial as almost all contemporary machine translation systems are based on data-driven techniques that train computers how to translate based on huge volumes of human-created texts.

As indicators for data availability by region, we analysed the:

1) Availability of open data
2) Access to proprietary data resources
3) Legal regulations of data usage

Table 41 Market relative score in data for machine translation

Market	Relative Score
Europe	2
North America	3
Asia	1

3.10.1.1. Open data

In the context of this analysis, open data are defined as data that can be freely used, re-used and redistributed by anyone - subject only, at most, to the requirement to attribute and share alike.[190]

In our study we come to the conclusion that Europe outperforms North America and Asia in terms of developed and freely accessible language resources that play an essential role in the development of machine translation systems.

EU institutions have released massive volumes of freely available language resources that contain data for more than 24 EU languages and exceed 5 billion words. Table 42 lists several of the resources released by EU institutions.

[190] http://opendatahandbook.org/guide/en/what-is-open-data

Table 42 Translation data provided by EU institutions

CORPORA	LANGUAGES	SIZE	DATA HOLDER
DGT-Translation Memory (DGT-TM)	24 EU languages	2 030 m words	EC DG for Translation
EAC-TM	22 languages of the EU (all except Irish) plus Icelandic, Croatian, Norwegian and Turkish	0.32m words	EC DG for Education and Culture
ECDC-Translation Memory	the 23 languages of the EU plus Norwegian (Norsk) and Icelandic	1.3m words	European Centre for Disease Prevention and Control (ECDC)
JRC-Acquis	22 languages of the EU	636 m words	Acquis Communautaire
DCEP-Digital Corpus of the European Parliament	22 languages of the EU	1 370 m words	European Parliament

The process of opening data by EU institutions was facilitated by introducing Directive 2003/98/EC on the re-use of public sector information. The Directive, also known as the ‘PSI Directive’, entered into force on 31 December 2003 and was revised by the Directive 2013/37/EU, which entered into force on 17 July 2013. This directive recognises that documents produced by public sector bodies of the Member States constitute a vast, diverse and valuable pool of resources that can benefit the knowledge economy. With entry into force of this directive public institutions become a valuable source of language resources.

Apart from translation data released by EU institutions, MT developers can use large national corpora (e.g. Bulgarian National Corpus, Croatian National Corpus, Slovenian National Corpus, National Corpus of Polish, Spanish text corpus, Latvian language corpus) and corpora developed by European universities (e.g. OPUS corpora[191]). The European Open Data Portal provides access to diverse language resources.[192] The European Language Resource Coordination Action (ELRC),[193] funded by the EU Connecting Europe Facility programme, creates and populates a dedicated repository of public sector language resources for machine translation.

In North America and Asia, open data initiatives have been primarily concerned with structured data from registers and databases as well as machine generated data, mostly in numerical format. Open data repositories in North America and Asia (e.g. US Government open data,[194] Japan government open data portal[195]) provide only few, if any, language resources.

[191] http://opus.lingfil.uu.se

[192] https://data.europa.eu/euodp/en/home

[193] http://lr-coordination.eu

[194] https://www.data.gov

[195] http://www.data.go.jp/data/en/dataset

3.10.1.2. Proprietary data resources

In regard to proprietary data and user generated content, global online US and Asia companies have a strong advantage versus European players. Global dominance of companies like Facebook, Google and Amazon in their primary business activities in the fields of social media, internet search and e-commerce allow them to harvest unmatchable amounts of data that they can use in other areas of their activities like machine translation.

This is also true for Chinese firms like Alibaba and Tencent, which have become similarly dominant in their home market (Giles, 2018).

China's biggest data advantage is its 770 million internet userbase. Government-lead centralisation of data is putting China's tech giants in charge of specific types of digital information, tuning them, in effect, into national data champions.[196] This helps Alibaba and Baidu to control huge data assets that, among their primary online activities, also help them to boost their efforts in machine translation. But such large collections of data, held by siloed entities, can become a barrier to new competitors entering the market.

3.10.1.3. Copyright regulation

European copyright regulation is much more restrictive for data usage compared to the United States. Lack of the fair use principle prevents huge volumes of copyright-protected data from being used by European researchers and machine translation developers. At the same time, US businesses and research institutions reap an advantage by using these data based on the fair use exception.

3.10.2. Data for speech technologies

Data is a crucial asset for speech technology development. The availability of data directly correlates with advances in the development of speech technology-based products and services.

Most speech data are available for English and Mandarin, some data are available for German, French, Italian and Spanish. A lack of speech and text resources for less resourced languages (i.e., speech/text corpora, external language-specific tools) for the acoustic and language models, respectively, are among the key reasons for the speech technology quality gap between languages. More diversity in the available speech and text corpora (in terms of age, gender groups, dialects, background noises, speech and language types) leads to potential in creating more capable and general-purpose speech recognition engines. Moreover, languages that are highly inflective and rich in morphology and vocabulary usually require even larger amounts of textual resources to cover all inflective forms. Although the availability of open data sets might initially give an assumption that data is not a problem, it must be noted that crowd-sourced data can only be used for ASR exclusively and are not applicable for TSS.

[196] The Economist, "The Ultimate Walled Garden", June 30th, 2018

Our study shows that the majority of open databases for speech resources originate primarily in America, Europe comes in second. There are no notable open speech databases in Asia. In the databases that are located in Europe, the language coverage surpasses databases located in other regions. The available data in open databases and in catalogues of language resource data distributors shows the global pre-eminence of the English language. The prominence of the English language creates more opportunities for the North America region. This correlates with other sections in our study and confirms that data is a crucial asset for speech technology development.

Table 43 Market relative score in data for speech technologies

Market	Relative Score
Europe	2
North America	3
Asia	1

As indicators for data availability by region, we analysed the availability of open data and the languages represented.

In the context of this analysis, open data are defined as data that can be freely used, re-used and redistributed by anyone - subject only, at most, to the requirement to attribute and share alike. Some open sources are providing publicly available open data sets, like Librispeech ASR Corpus which is a large scale (1 000 hours) corpus of read English speech.[197] Although there are also some efforts to capture speech corpora for languages other than English, there remains a lack of relevant data in other languages that can be used to establish a usable data set.

The available data can also be divided by its applicability to various speech technologies, the most prominent being ASR (Automatic Speech Recognition) and TTS (Text-To-Speech synthesis). The datasets required by these technologies have differing requirements, and are often not directly compatible. For example, ASR training data should represent the acoustic variability of speech (speakers, styles and environments), while TTS training data should in general represent consistent speech of a single speaker or a few speakers in a well-controlled environment (recording studio). This usually precludes the use of crowd-sourced data for TTS purposes.

[197] http://www.openslr.org/12

Notable open TTS datasets include:

- LJ Speech[198] dataset, which is a single-speaker English subset of LibriVox project data. Total amount: approximately 24 hours. The developers are located in America.
- Simple4All Tundra[199] corpus, which consists of recordings for 14 languages: Bulgarian, Danish, Dutch, English, Finnish, French, German, Hungarian, Italian, Polish, Portuguese, Romanian, Russian and Spanish, (single speaker for each language), with typically 4-6 hours for most languages. The developers are located in the UK.

The leading TTS research groups (Google and Baidu) each use their proprietary TTS dataset, spoken by a single speaker in a controlled environment and style. The exact details of these datasets are not implicitly specified, however, they can be estimated from scientific publications. Google's dataset comprises 24.6 hours of North American female voice by a professional speaker, where "the speaker primarily speaks in a neutral prosody, a small subset of the corpus uses more expression (including performing as a game show host, reading jokes and poems, etc)". Baidu's dataset comprises approximately 20 hours of English speech.

The research community often uses the Blizzard[200] challenge dataset, which is a limited proprietary dataset available for research purposes. As a benchmark, several research systems have been trained on 147 hours of single-speaker American English audiobook using the 2013 release of the Blizzard dataset.

3.10.2.1. Open data

Below are examples of notable open databases for ASR. The majority of the data are in English, with a small amount in other languages.

OpenSLR is a site devoted to hosting speech and language resources, such as training corpora for speech recognition, and software related to speech recognition. It aggregates information about large speech corpora sources (e.g. TEDtalks, audiobooks etc.) that are available under different versions of CC licence. The web site has a German IP address.[201]

American English Dialect Recordings: The Center for Applied Linguistics Collection contains 118 hours of recordings documenting North American English dialects. The Library of Congress provides access to these materials strictly for educational and research purposes. The written permission of the copyright owners and/or other holders of rights (such as publicity and/or privacy rights) is

[198] https://keithito.com/LJ-Speech-Dataset

[199] http://tundra.simple4all.org

[200] http://www.cstr.ed.ac.uk/projects/blizzard/data.html

[201] http://www.openslr.org/12

required for distribution, reproduction, or other use of protected items beyond that is allowed by fair use or other statutory exemptions. The web site has an American IP address.[202]

The Common Voice project by Mozilla was launched to help make voice recognition open to everyone. Visitors of the website can donate their voice to help build an open-source voice recognition engine that anyone can use to make apps for devices and the web that make use of voice recognition. The website asks visitors to read a sentence to help the machine system learn how real people speak, and allows validating the sentences read by other people. For English, German and Kabyle languages the website has been successfully localised, and it has collected enough sentences to allow ongoing Speak and Listen contributions. The next language that shows good progress is Chinese, for other languages the progress is slow. It is possible to request that a language be added to the list. Mozilla publishes Common Voice data sets under a CC-0 license. Mozilla is headquartered in America. [203]

Google Audioset is an expanding ontology of 632 audio event classes and a collection of 2,084,320 human-labelled 10-second sound clips drawn from YouTube videos. The dataset is made available by Google Inc. under a Creative Commons Attribution 4.0 International (CC BY 4.0) license. Google is headquartered in America. [204]

LibriVox audiobooks data set of text and speech contains nearly 500 hours of clean speech of various audio books read by multiple speakers, organised by chapters of the book containing both the text and the speech. Most releases are in English, but many non-English works are also available. In total there are datasets in 36 languages (e.g. French 630 books, German 2182 books, Italian 123 books, Spanish 410 books, Ancient Greek 31 books, Bulgarian 8 books, Cantonese Chinese 4 books, Chinese 420 books). The project is worldwide but coordinated and managed from an office based in America.[205]

VoxForge is a free speech corpus and acoustic model database for open source speech recognition engines. The corpus is available under a GPL licence. The speech audio is for use with open source speech recognition engines such as Julius, ISIP, Sphinx and HTK. The speech database working group is located at Carnegie Mellon University in America. [206]

CHIME is a noisy speech recognition challenge dataset. This dataset contains real, simulated, and clean voice recordings. Real being actual recordings of 4 speakers in nearly 9 000 recordings over 4

[202] https://www.loc.gov/collections/american-english-dialect-recordings-from-the-center-for-applied-linguistics/about-this-collection/rights-and-access

[203] https://voice.mozilla.org/en/new

[204] https://research.google.com/audioset/index.html

[205] https://librivox.org

[206] http://www.voxforge.org/home

noisy locations, simulated is generated by combining multiple environments over speech utterances, and clean being non-noisy recordings. The corpora are the result of the 5th CHiME Speech Separation and Recognition Challenge. The challenge is organised by the University of Sheffield (UK), Johns Hopkins University (US) and Inria (France). [207]

In addition to our study on the availability of open data and languages represented we analysed available data sets that are listed in language resource data distributor catalogues, with the headquarters in the regions of our interest. In addition we analysed datasets that are available from Appen, which is headquartered in Australia.

- ELRA (Europe)
- The Linguistic Data Consortium (LDC) (America)
- Speechocean (KingLine Data Center) (Asia)
- Appen (Australia)

Data from the study shows that the overwhelming majority of the available datasets concerns English, next is Chinese. The study shows that databases for European languages are not well represented.

3.10.2.2. Proprietary data resources

Several organisations provide various proprietary speech corpora.

The Linguistic Data Consortium (LDC) is an open consortium of universities, libraries, corporations and government research laboratories. Initially, LDC's primary role was as a repository and distribution point for language resources. Since then, and with the help of its members, LDC has grown into an organisation that creates and distributes a wide array of language resources. The LDC database has a significant amount of speech corpora suitable for ASR. LDC Online contains an indexed collection of Arabic, Chinese and English newswire text, the full text of the Brown corpus, millions of words of English telephone speech from the Switchboard and Fisher collections and the American English Spoken Lexicon.[208] LDC Online is a free service for LDC current year members. The LDC catalogue also contains data sets representing Europe, such as data sets for Bulgarian (2 datasets), Czech (7 datasets), French (7 datasets), German (9 datasets), Hungarian (3 datasets), Italian (4 datasets), Polish (3 datasets), Slovak (2 datasets), Spanish (33 datasets). It must be noted that for some of the languages there is only a small amount of data and it is part of multilingual data sets (e.g. telephone speech CSLU: 22 Languages Corpus). There is no data for smaller languages such as Estonian, Finnish, Latvian, Lithuanian or Croatian.

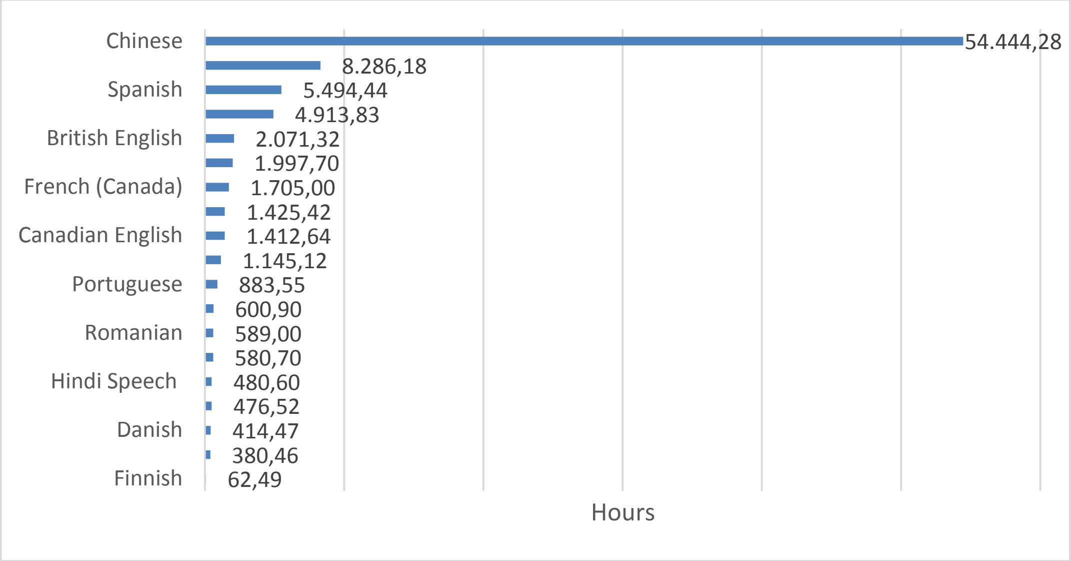
[207] http://spandh.dcs.shef.ac.uk/chime_challenge/index.html

[208] https://online.ldc.upenn.edu/login.php

Speechocean by KingLine Data Center is a language-related resource and service provider in the fields of Human Computer Interaction and Human Language Technology. At present, Speechocean can provide data services with 110+ languages and dialects across the world. Kingline Data Canter is operated and supervised by Speechocean.

Corpora available at KingLine Data Center are of varying sizes and created using different methods to record or create corpora, for example, desktop platform, mobile platform, Smart TV in car microphones. This provides the option to specially design materials for both training and testing of speech recognisers in multiple scenarios of applicability. The database contains a majority of corpora for Chinese, while the second largest available set of corpora is for English. There are no data for smaller languages such as Latvian, Lithuanian or Croatian. KingLine Data Center also has high-quality academic resources to satisfy the experimental and testing needs of scientific research institutions, colleges and individuals around the world. All these corpora can be purchased at a minimised cost for research purposes. Figure 92 illustrates the size of corpora in hours and the languages within the regions covered by this study.

Figure 92 KingLine Data Center ASR speech corpora size



ELRA Catalogue includes language resources in 67 different languages. For some countries, only one language is represented in the catalogue. For some other countries the catalogue contains resources for several languages (e.g. French and English in Canada; Arabic and French in Morocco and Tunisia). A third group of countries has more than 3 languages represented in the catalogue (e.g. Spain, with Spanish, Galician, Basque and Catalan represented, and India, with a total of 11 languages represented, from both the Dravidian and the Indo-Aryan language families).

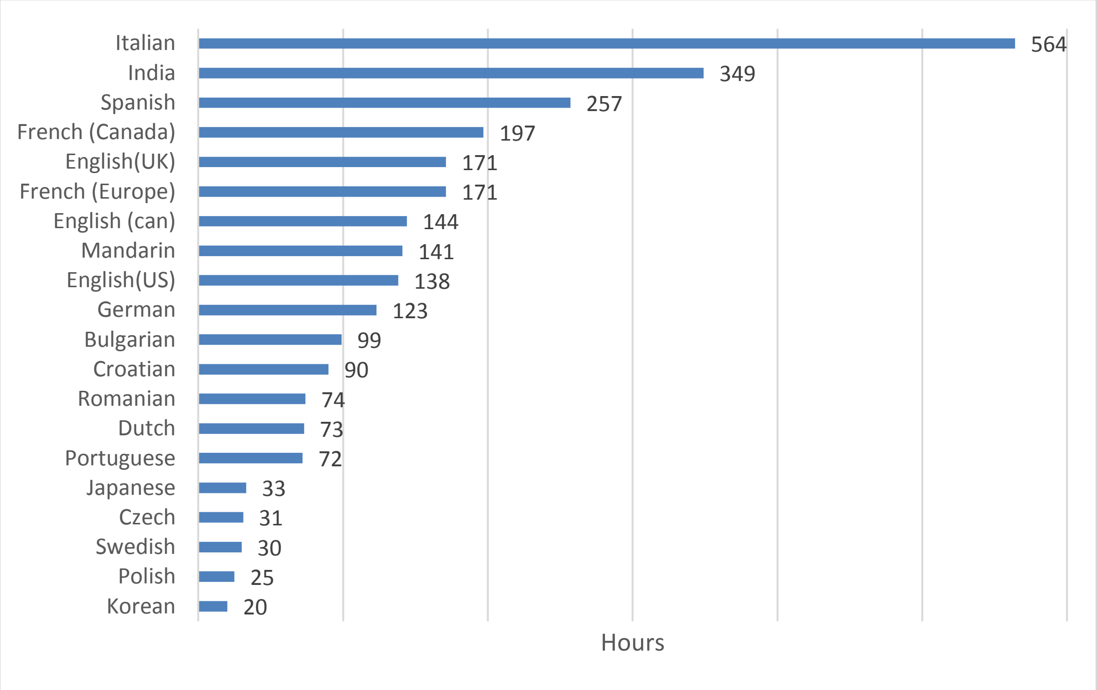
For the overall catalogue of language resources (LRs), ELRA has over 7 TB of data available, out of which over 5.6 TB only for speech and multimodal resources, over 1.7 TB for evaluation LRs, the remaining being shared among less voluminous written and terminology resources.

The majority of the ELDA Catalogue resources may be purchased. The catalogue also includes a set of LRs which are free for research purposes. Most of these LRs are written and speech corpora.

The ELRA catalogue contains a large amount of English language data sets. Of the 5,218 entries, it contains 1,364 LRs related to English (close to 26%, compared to 29% in 2011), followed by French (286), German (265), Spanish (222), Italian (158) and Dutch (110). While still very few exist for Slovak (9), Irish Gaelic (5) or Maltese (4), we note a big increase for Estonian (from 7 to 23), and for regional languages (67 to 103) and for non-EU European languages (63 to 293).[209]

Appen is a publicly traded company listed on the Australian Securities Exchange (ASX) under the code APX. Appen develops high-quality, human-annotated data sets for machine learning and artificial intelligence. With over 20 years of industry experience, Appen work in more than 180 languages and dialects and have access to a global crowd of over 1 million skilled contractors.

Figure 93 Appen ASR speech corpora size



3.10.3. Data for search technologies

Almost all contemporary search systems are based on data-driven techniques that train computers to improve search and information retrieval. In particular, user activity history is the most crucial data for ranking search results by their popularity and relevance. As indicators for data availability by region, we analysed (1) the total visits of top 10 most popular web search sites and (2) usage of language in internet.

[209] http://catalog.elra.info/en-us

By having reviewed both indicators, it must be concluded that North America due to the Google's dominance in web search and the online dominance of the English language, receives the highest ranking in data availability, followed by Europe with its diversity of multilingual data for European languages. Meanwhile Chinese lags behind in the availability of data, although spoken by approximately the same amount of internet users as European languages (Table 46).

Table 44 Market relative score in data for search technologies

Market	Relative Score
Europe	2
North America	3
Asia	1

With no surprise, the top of the total search site visits is headed by Google of which the headquarters are in Mountain View (North America), followed by Baidu (Asia) and Yahoo (North America).

Table 45 Total visits on desktop and mobile web, in the last 6 months

No	**Search site**	**Region of HQ**	**Total Visits**
1	Google	North America	42.57B
2	Baidu	Asia	10.58B
3	Yahoo! Search	North America	4.50B
4	Yandex	Other (Russia)	2.93B
5	Bing	North America	1.22B
6	Naver	Asia	696.84M
7	DuckDuckGo	North America	421.46M
8	Seznam	Europe	227.73M
9	Ask	North America	165.06M
10	Aol Search	North America	59.18K

InternetWorldStats (Table 46) estimates the number of English language internet users at 25.4%, while Chinese is used by 19.30% of internet users.[210] Although Chinese is the second largest language in terms of number of users, the total number of European internet users exceeds it. At the same time, top ten language as criteria used to identify advantages for data must be carefully evaluated.

[210] https://www.internetworldstats.com/stats7.htm

Since there are different languages, lots of resources must be put in to develop solutions for semantic search.

Table 46 Top ten languages used in the web – 31/12/2017

LANGUAGE	World Population for this Language (2018 Estimate)	Internet Users by Language	Internet Penetration (% Population)	Internet Users Growth (2000 - 2018)	Internet Users % of World (Participation)
English	1,462,008,909	**1,055,272,930**	72.20%	649.70%	25.40%
Chinese	1,452,593,223	**804,634,814**	55.40%	2,390.9 %	19.30%
Spanish	515,759,912	**337,892,295**	65.50%	1,758.5 %	8.10%
Arabic	435,636,462	**219,041,264**	50.30%	8,616.0 %	5.30%
Portuguese	286,455,543	**169,157,589**	59.10%	2,132.8 %	4.10%
Indonesian / Malaysian	299,271,514	**168,755,091**	56.40%	2,845.1 %	4.10%
French	412,394,497	**134,088,952**	32.50%	1,017.6 %	3.20%
Japanese	127,185,332	**118,626,672**	93.30%	152.00%	2.90%
Russian	143,964,709	**109,552,842**	76.10%	3,434.0 %	2.60%
German	96,820,909	**92,099,951**	95.10%	234.70%	2.20%
TOP 10 LANGUAGES	5,135,270,101	**3,209,122,400**	62.50%	1,091.9 %	77.20%
Rest of the Languages	2,499,488,327	**950,318,284**	38.00%	935.80%	22.80%
WORLD TOTAL	7,634,758,428	**4,159,440,684**	54.50%	1,052.2 %	100.00%

According to W3Techs[211] (Table 47), the English language content of the internet currently is slightly more than 50% with other language content dropping off sharply.

Table 47 Internet content by language

	Language	**%**
1	English	53.6%
2	German	6.2%
3	Russian	6.0%
4	Spanish	4.9%
5	French	4.1%
6	Japanese	3.5%
7	Portuguese	2.9%
8	Italian	2.4%
9	Persian	2.0%
10	Polish	1.8%
11	Chinese	1.7%
14	Dutch, Flemish	1.3%
15	Turkish	1.2%
16	Czech	1.0%
17	Korean	0.9%
18	Vietnamese	0.6%
19	Arabic	0.6%
20	Swedish	0.5%

[211] https://w3techs.com/technologies/history_overview/content_language

As for enterprise search a language does not play a big role as nearly all search engines are using the same library (Apache Lucene) and are easy adjustable to language preferences.

3.11. SWOT analysis

3.11.1. SWOT analysis for European machine translation

Building on the information gathered above, the SWOT analysis distils the state of the European MT market in the global context.

STRENGTHS

- Europe's researchers have **successfully developed** state-of-the-art MT technologies that have game-changing potential (e.g. the Moses SMT toolkit, the Nematus and Marian NMT toolkits have all shown to push the state-of-the-art in MT).
- EU MT developers have **strong experience** in developing machine translation and other language technologies for smaller and complex languages that has been accumulated thanks to the multilingual internal market in Europe.
- Due to the need to search for niche markets, many European MT developers have accumulated strong experience in **customised and domain-specific** MT solution development.
- European MT industry and research organisations have a strong **cooperation on pan-European level**, which is stimulated by international research and innovation programmes within the EU. The national and pan-European research and innovation funding programmes have shown to be effective means in advancing the state-of-the-art of MT in Europe.
- EU has well-established practices for the creation of **open data** (especially from EU institutions) and **policies fostering** public data **sharing**, which are a strong driving force for MT development in EU.
- European MT developers have been successful in deploying and delivering state-of-the-art MT services for the **public sector** through the support of EU funding programmes (e.g. the CEF eTranslation platform, the EU Council Presidency Translators, the hugo.lv platform for the Latvian government, and many others).
- Europe demonstrates dynamic innovation in translation automation through **start-up formation** and **fast adoption of innovations** in product offerings.

WEAKNESSES

- The EU MT industry is **fragmented** with many small players that struggle to find a place in the market in order to compete with the global online companies providing MT.
- European MT players have **insufficient resources** to invest in innovation, marketing and scaling to stay globally competitive.
- The MT **markets** for most European languages **are small**, limiting business opportunities for MT players focused on particular languages.
- EU MT companies lag behind in crafting successful business models due to **market distortion by global online companies**.
- Europe is losing its MT research advantage over the last two years as large North American and Asian entities have been more successful in **capitalising on research results**. Although Europe

has led the way in statistical MT and has shown strong research capabilities in developing neural MT toolkits, state-of-the-art neural machine translation architectures/models over the last two years (e.g., Transformer, Universal Transformer, and other) have come exclusively from US-based companies that have strong ties to North American research centres.

- EU companies are disadvantaged by **copyright disparities** in data usage compared to US. The explicit permission required by European entities vs. the fair use copyright exception available to US-based companies provides them with a disproportionate competitive advantage in developing and offering innovative products including MT technologies.
- **Insufficient** Europe-based **computing capacity** for deep learning (and the necessity to rely on deep learning infrastructures provided by large US-based corporations) is a slowing factor for EU researchers and developers working on neural machine translation technologies.

OPPPORTUNITIES

- Modern MT technologies can serve as an essential means to **preserve cultural identity** and the very **existence of a language** of the smaller communities in the digital era, thereby supporting a key principle of the European Union, language equality.
- European MT technologies can foster **inclusiveness and equal digital opportunities** for speakers of smaller languages of the European Union, regional languages, and minority languages.
- MT development as part of the Digital Single Market represents tremendous potential to **transform the European economy** and **support** European **SMEs** in their market expansion.
- There is a large potential for European MT companies to provide cost-effective solutions for enabling multilingual **public services**.
- European MT can facilitate the **inclusion process** of newcomers, as well as **aid integration** of internal and external migrants by enabling them access to local information and digital services.
- European MT can provide high security, narrow domain and other **specific MT applications** that are not available from global players.
- European MT can provide services for European **defence applications** such as cyber security, resilience against fake news and hybrid online attack operations.
- MT can be integrated in a **rapid-response infrastructure**, such as messaging and social media systems, in order to speed up the pace by which relief can be provided in crisis situations.

THREATS

- The **availability of free** MT services offered by large (mainly US-based) global companies, whose main business is not MT, disrupts the market; the phenomenon is making it difficult for smaller European MT providers to enter the market and remain in it.
- Global online companies providing MT services have an **overwhelming competitive advantage** both in the accumulated quantity of data and computing resources that are crucial competitiveness factors for MT research and development.
- Despite the vast growth of MT consumption, the **overall level of MT adoption** in business and customer scenarios remains rather low limiting business opportunities for MT providers.
- Innovations created in Europe are quickly **acquired or replicated** by North American and Asian businesses.

- The **relocation of highly qualified talent** in MT, language technologies and artificial intelligence, especially researchers and developers, to other regions, mostly North America.
- European weakness in MT may have strong repercussions for **the development of other** critical **language technologies.**
- Global market dominance by a handful of large online companies may lead to an increased **quality gap** for smaller European languages that are not attractive for investment due to their limited market.
- Market dominance creates a **dependence** of European business and public sector on global monopolies.

3.11.2. SWOT analysis for European speech technologies

Building on the information gathered during this research, the SWOT analysis gives a condensed view of the EU speech recognition market in a global perspective in comparison to Asia and the US.

STRENGTHS

- EU researchers, private companies and public institutions have developed a **distinguished competence** in small languages since nearly each country has its own language.
- European specialists in speech technologies have been **trained and are experienced** in developing technologies for smaller languages.
- Through various investment grants the EU has developed a firm and strong **investment policy** to advance European research, development and innovation.
- The EU has shown **political commitment** to support smaller languages. On 11 September 2018 the European Parliament adopted a resolution on language equality in the digital age. It is a strong signal for a multilingual EU in the digital era.

WEAKNESSES

- Although there are companies and public institutions that are investing in technologies and research, existing **investment volumes are insufficient** and still lag behind competitive regions, where public authorities, private companies and even local governments (China) are major contributors to investment in speech technology.
- Innovative solutions and services created in the EU are **acquired** by companies based in the US and China, thus in later development stage contributing to the US and China economy.
- The European Union is **losing its strength in research**. The leading US and China companies are heavily investing in speech technologies to boost their spoken language enabled services.
- The EU market consists of many smaller markets, which **fragments the market** through various consumer behaviour models, languages and other specifics and makes the EU market difficult to compete in.
- EU companies are at a disadvantage due **to a lack of data for smaller languages**. Little demand for smaller languages and usage of application puts smaller languages at a disadvantage in terms of available data.

- There **is less venture capital** available for the EU speech recognition companies than there is for Asian and US companies.

OPPORTUNITIES

- European multilingualism is giving **drive and motivation** to develop multilingual speech based solutions.
- The EU has already demonstrated successful **experience in multilingual infrastructure** building projects with the aim to reduce digital linguistic fragmentation across the EU.
- Voice-based applications are **significant for people with disabilities and the elderly**, helping them with everyday tasks as well as emergency situations.
- European speech technologies can improve involvement of and equal opportunities for **speakers in smaller languages** and EU regional languages.
- At a time when democratic institutions (elections, government databases) and EU and Member State security in general could possibly be undermined by external cyber threats, voice recognition could provide **solutions to tackle potential cyber threats**.
- Rapid development of **'Internet of things'** gives an opportunity not only to provide a useful money saving solution by connecting physical devices, vehicles, home appliances with software and electronics, but also to collect and exchange a wide range of data across the EU.
- Using voice recognition for **criminal investigation** under reasonable suspicions in EU and national level law enforcement agencies, EU security might benefit from a broader application of voice and speech recognition solutions.
- Europe can leverage its strength in **vertical software markets** by integrating voice-enabled functionality.

THREATS

- Global leading technology companies (Google, Amazon, Facebook and Apple) are **dominating the market** in speech recognition. Market dominance is a threatening and discouraging factor for European speech technology companies.
- While a spoken language interface is not available for smaller European languages, EU **multilingualism is under threat** and the gap for smaller languages is becoming wider.
- Asian and US companies are expanding and dominating the EU market by **acquiring** not only native innovative champions but also making acquisition deals in Europe.
- The EU market is **dependent on non-EU products** in public and private services, predominantly in the mobile application industry, which is a leading industry for voice recognition software.

3.11.3. SWOT analysis for European search technology

Building on the information collected and analysed above, the SWOT analysis distils the state of the European search technology market in the global context.

STRENGTHS

- Europe has strong positions in **vertical search solutions and enterprise search** solutions (Elasticsearch/Sphinx).
- The EU has well-developed **research expertise in specialised search segments and areas** like social networks, medicine, transport and logistics etc.
- EU countries have a **firm base of language resources** to be used in natural language understanding and natural language processing tools to create and train AI-based multilingual search solutions that could be based on semantic search principles.

WEAKNESSES

- The overwhelming and **unreachable dominance of Google** in the search industry. The strong dominance of one market player is forcing consumer behaviour in one single information channel.
- A **fragmented market** by language that leads to a significant cost increase to create new solutions and additional resources to adjust to separate markets in terms of consumer behaviour and other market specifics.
- **Technical gaps** in natural language processing for smaller languages affect the quality of the search solutions for these languages.
- Very **low private investments** in European search companies.
- A **small amount of data** due to the small number of users of smaller languages. Usage history data that is essential for internet search solutions is controlled by the dominant player Google.
- **Start-ups** in the search industry are **being funded less** than in North America, but more than in Asia. It is a signal that new innovative solutions are concentrating in North America.

OPPORTUNITIES

- In the light of information expansion there is a **growing demand** for a new breed of Q&A-based solutions.
- An increase in the **multilingual content** driven by users **demand** to purchase goods and browse content in their native language increases the demand for cross-lingual search tools.
- European **open data initiatives** increase the availability of multilingual data.
- Deep learning gives an opportunity to create **multi-lingual search solutions** that could fit the needs of the European Digital Single Market.

THREATS

- A risk of **information control**. The increased Google market dominance in the search industry is raising information control risks. The search algorithm controls what is being seen and what kind of information is being channelled to the user, undermining the diversity of opinions and freedom of speech.
- The **research potential in the EU can be overtaken** by North America and Asia, and EU is further losing its competitive strength.

- A threat of **losing positions in innovation** not only to North America but also China, which is already overtaking European search companies, such as Skyscanner.

3.12. Recommendations

The results of the extensive desk research have been summarised in the SWOT analysis. It would be sensible to suggest actions which would capitalise on existing strengths and use opportunities, at the same time improving weak areas and mitigating external threats.

The recommendations provided below mostly corroborate the action points earlier identified in such documents as *Empowering a Multilingual Continent. Technologies and Language-Centric AI for Language Equality in Europe* (by LT-Innovate and META-NET*), Strategic Research and Innovation Agenda - Language Technologies for Multilingual Europe* (META-NET), Science and Technology Options Assessment (STOA) *Language equality in the digital age: Towards a Human Language Project*, and finally the European Parliament resolution of 11 September 2018 on Language equality in the digital age. It should be noted that the identified action points put additional emphasis on conclusions provided from the abovementioned documents and should be considered in the same scope.

Europe is traditionally strong in research and innovation but has problems in scaling innovations and conquering market share. Market fragmentation is one of the issues strongly influencing the European LT development. The LT ecosystem should get a boost in order to support further growth. Europe needs a basic European Language Infrastructure for natural language processing, which would provide basic LT services and data sets for all languages. Technology providers, potential customers and research should have the possibility for cooperation.

Language technology is a powerful enabler allowing small and large European businesses to spread out to new geographical markets. The European market by definition is multilingual and needs multilingual solutions. European companies also need efficient multilingual solutions to reach linguistically diverse global markets. Therefore, it is extremely important for Europe to develop its own language technologies in order to avoid dependence on US/Asian providers.

Language technology is essential for human-centric artificial intelligence which drives a major transformation of European economy and society. Currently, almost all developments are focused on English and a few of the largest global languages. This further widens the technology gap in the scope and quality of technological support for European languages, as many of them are not supported in the latest AI-based products and services. European AI should be multilingual to enable all Europeans to benefit from these game-changing technologies.

Public intervention is needed to address market failures. Public procurement is an efficient approach to drive public demand in essential multilingual solutions for Europe. Public procurement of the European multilingual infrastructure should serve as a major driver for the growth and consolidation of the European LT industry, in order to avoid dependence on existing market monopolies. Implementation of corresponding public procurement policies should raise the demand for new products and services, foster the supply of new products, and encourage their faster and more efficient production, and will in general improve competitiveness of the LT sector.

The next frontier in LT development is deep language understanding – systems that can learn, interact and explain themselves, to do to it reliably and across languages. To achieve that, Europe should continue investing into basic and applied research. However, an increase of efficiency of research is necessary, and the next scientific breakthrough is very much awaited. There is a need for a holistic approach on European, national and regional levels for coordination of actions and policies, for lining up research activities and projects. Politics, business, research and society should all participate in the initiative. The European LT industry should reap the benefits from close involvement in the initiative, providing industry-driven challenges, guiding and monitoring research progress, evaluating research results in prototype solutions, and transferring research achievements into innovative applications for the European and global market.

A lot has already been done in the field of the collection of data sets for LT development, in particular by promoting open data, opening public sector information and involving national administrations in the sharing of language data. Still, there are not enough resources to satisfy the needs of all languages equally and to move innovations in LT further. Language resources should be further produced, gathered and provided to industry and researchers, for every language and every domain. Furthermore, intellectual property rights regulation of data use should be carefully reviewed and made more open for the development of language technologies, as it currently hinders the development of language technologies and puts Europe at a disadvantage to other regions where the “fair use” principle is applied to copyrighted data.

The LT European sector, just like other industries, is threatened by a diminishing number of new technology professionals. The increase of demand for qualified professionals from the side of US and Asia only worsens the situation. New incentives should be provided for graduated professionals in order to encourage them to stay in Europe and to avoid brain drain. At the same time, corresponding policies should be introduced to make current students more aware of the technical disciplines related to LT and existing career opportunities in the LT area. Possibly, new study programmes need to be developed and introduced.

4. Task 3: Analysis of LT adoption by public administrations

4.1. Executive summary

4.1.1. General summary

The activities of Task 3 mainly relate to the analysis carried out about Human Language Technology[212]-based services and solutions used by public administrations in the Member States, both those currently in use and the ones planned within the next few years. Possible shortcomings of the European LT market in this regard will also be identified.

Via data collection and the development of a taxonomy and classification of LT tools, we aimed to gain a better understanding of what the Public Sector within the different Member States is using or planning to use as language technologies to render their services.

Task 3 is led by the Evaluations and Language resources Distribution Agency (ELDA) in collaboration with the consortium partners, and has as its main objective to provide a reliable overview of the use of Human Language Technologies (HLT) by the European Public Sector and to draw the perspectives and roadmap of such use as reported by the interested parties.

The outcomes and conclusions presented in the report of Task 3 (either graphics or analysis) result from **79 online survey responses** from top managers of public services within all Member States. The potential targets of this survey were representatives of the administrations that we identified through a number of actions, in particular the lists of attendees (or interested parties) of the European Language Resource Coordination's (ELRC) two rounds of workshops (conducted in 2016 and 2018). We also established contacts through the ELRC Public Service National Anchor Points of the EU Member States, Norway and Iceland and in some cases we used additional contacts to achieve the planned number of responses.

The first phase of this study consisted in designing the online survey, with the support of the consortium partners, in particular IDC. The draft questionnaire was submitted to the EC for comments to better tune the questions to the expectations and requirements. The questionnaire can be found in Annex O. The second phase consisted of some "mass" mailing to all our contacts and the corresponding reminders. The last phase consisted of an analysis of the data collected.

4.1.2. Objectives

As indicated above, the objective is to draw a first picture of HLT being used or planned to be used in the EU Member States. Dedicated sections will give highlights and details about the collected figures while emphasising some of the information when relevant.

[212] In the report of Task 3, we use both Human Language Technologies (HLT) and Language Technologies (LT); HLT is more frequent in our field to distinguish it from Programming Language Technologies.

The main findings we expected to obtain cover some of the following items:

- Which public services are most represented in our survey as early adopters of these technologies or with clear plans for future use?
- Which technologies are cited most? Which ones are being used, which ones are planned to be used and which ones are not needed?
- Which are the suppliers mentioned by the respondents? Where are they located (EU, US, etc.)?
- Which languages are served (official, regional, non EU languages)?
- What is the degree of cooperation with the academic and research communities?
- Can the collected data be used to draw up country profiles?

4.1.3. Very brief overview

From the results we obtained, we see that the major technologies used are related to Automated Translation (or Machine Translation (MT) and its derivatives, 66 out 79 responses). Many aspects of the translation workflow automation (translation memories but also other applications such as terminology extraction and management) are widely used and very often in the planning by a large number of respondents. Optical Character Recognition (OCR) comes next in our list. In the top 5 we also found Speech Technologies (mostly Speech Recognition) and Multilingual and Semantic Search (both with 37/79 responses). Technologies related to localisation seem to be the least popular (5/79). We expected higher interest in and use of localisation tools of web sites, but apparently this is of less concern to the respondents of the survey.

The number of responses per public service is not significant enough to draw conclusions. The administration of the State or Economic and Social Policy of the Community (excluding fiscal administration) is best represented with 16 responses, covering many different ministries, depending on the country. Some respondents choose “Other” as their affiliation to be more specific in the description of their mission. The figures, however, do not allow to draw a country profile with respect to the use of LTs.

Through this survey and the lessons learnt from the various workshops organised in the framework of ELRC (www.lr-coordination.eu), it appears that the adoption of these technologies remains at a very low level and the early-adopters are often translation services, helpdesks or call centers.

Many respondents indicated that the technologies mentioned in our questions are either planned (very often by 2020) or that the needs have been identified and hence we assume planning to be their next step.

A major conclusion that can be drawn from this survey is that the European public services do not incorporate the new HLT tools despite the fact that most of them identified the need for such technologies. At this stage, it is difficult to speculate about the main reasons behind this. Looking at the responses on the integration planning, it seems that many of them feel that these technologies are not mature enough (though administration bodies could be very early adopters). However, machine translation and all related technologies seem to be the focus of many services and hence could not avoid adoption at some point.

4.2. Introduction

4.2.1. About this task

Task 3 deals with the analysis of the HLT use and plans of the Member States' public bodies. Task 3 "Analysis of LT services and solutions currently in use by public administrations in the EU as a whole and for each of the individual Member States, including Norway and Iceland", led by ELDA, in collaboration with the consortium partners, aims to provide a reliable overview of the use of (or of plans to make use of) HLTs by the European administrations.

The objectives of Task 3:

- Provide an analysis of LT services and solutions currently in use by public administrations in the EU as a whole and for each of the individual Member States (and Norway and Iceland, if possible).
- Identify existing solutions by public administrations and services for addressing their needs related to LT.
- Identify (characteristics of) LT providers used and/or degrees of spending.
- Present quantitative information on the uptake of the eTranslation service and a qualitative assessment of the eTranslation service based on information collected from workshops organised in the framework of the SMART 2016/0103 LOT 2 service contract.

The report of Task 3 is structured as follows:

1. An executive summary with the major findings.
2. A section that introduces the work carried out and the objectives setup for the survey.
3. A section with a detailed overview and analysis of the results (at a higher level).
4. In order to improve readability, detailed results are provided in Annex N.
5. The lessons learnt from the survey and the conclusion.

The questionnaire used for the survey can be found in Annex O.

4.2.2. Objectives of this survey and the methodology

The main objective is to understand how Language Technologies are used by the EU Member States (MS) administrations for the benefits of citizens and businesses.

In order to collect such information, the ELDA team designed a very extensive list of questions that were discussed internally within the Consortium and then with the EC officials. A consolidated version was setup in the best possible user-friendly way.

The first part of the questions allowed us to collect information on the institution represented by the person completing the questionnaire for a wide range of public services (from Social Security to Domestic affairs and Utility services, etc.). A second series of questions was related to which Human Language Technologies are of interest to the respondent institution. Specific interest can be reflected by the current use of such technologies (technologies in operation), a plan to use it (deployment or review planned), or no interest expressed either because there is a need but no plans are being discussed or no needs identified.

A detailed list of HLT options was given to allow the respondents to select the appropriate technology:

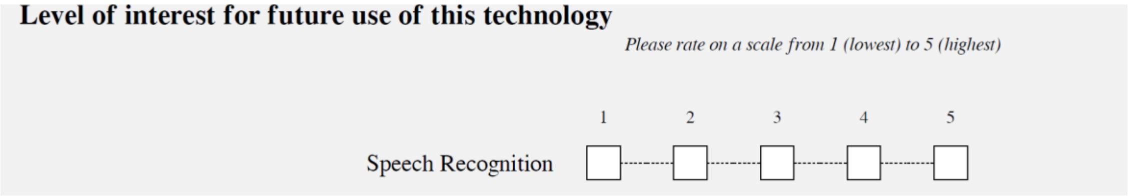
- Speech Technologies
  - For Speech Recognition (Speech-to-text)
  - For Speech Synthesis (text-to-speech)
  - For Speech Translation
- Translation Technologies
  - Machine Translation
  - Computer Aided Translation (CAT) tools
  - Translation Memories
  - Alignment Tools
  - Translation Workflow management
  - Authoring Tools
- Terminology Technologies
  - Terminology Management Systems
  - Terminology Extraction
- Localisation technologies
  - Localisation tools applied to Websites
  - Localisation tools applied to Software
  - Localisation tools applied to Forms
  - Localisation tools applied to Subtitling/Dubbing production
- Natural Language Understanding (NLU) Technologies
  - Chatbot / Virtual Assistant
  - Keyword Extractor
  - Topic Modelling Tools
  - Automatic Summarisation tools
- Text Analytics Technologies
  - Text Mining tools
  - Sentiment Analysis tools
  - Text Prediction tools
  - Authorship Attribution tools
- Multilingual and Semantic Search Technologies
  - Question Answering System
  - Search Engine
- Optical Character Recognition (OCR)

In the analysis of potential services rendered by the public bodies to citizens and/or to businesses, we included all scenarios from phone inquiries to call centers (that may use speech recognition and speech synthesis to understand the query and provide the information), machine translation to supply information in multiple languages and/or to understand queries received in other languages than the ones of the respondent administration, the integration of a number of components into workflows management (alignment of bilingual texts, terminology management, etc.), authoring tools that may be integrated in the information production platform (authoring tools, text prediction tools, etc.), customised search engines for information retrieval in the administration portal and/or archives, the OCR to digitise the archives but also any hardcopy correspondence received.

We also considered technologies that were mentioned by HLT suppliers (see the supply side information in Task 1) as important to public administrations. This allowed us to collect information to corroborate the results of the supply side survey carried out in Task 2.

The online questionnaire was setup in a way that a description of each technology could be seen in a pop-up window during the completion. In addition, we offered to help the respondent through email. No one used that service.

To get a better view of the current and future use of HLT, we also asked the respondents who do not operate any technology, to indicate their level of interest for future use within their administration. We used a scale of 1 to 5 (lowest to highest) as shown herein:



Level of interest for future use of this technology

Please rate on a scale from 1 (lowest) to 5 (highest)

	1	2	3	4	5
Speech Recognition	□	□	□	□	□

We consolidated these figures into some indicators that we listed for each technology in Annex N.

To get more input we did ask the respondents to list their suppliers if possible, but most respondents decided not to disclose that kind of information.

A specific section was devoted to the collaboration between the public administrations and the academic and research community. It is clear that many technologies require a heavy customisation and tuning that can be done by academia.

Last but not least, we took this opportunity to raise the awareness of the Member State representatives of the European Commission's machine translation platform, eTranslation, and the activities of the European Language Resources Coordination (ELRC), which is an essential instrument to support the eTranslation development. This should strengthen the impact of the ELRC information dissemination during its rounds of workshops in each Member State.

4.2.3. Note about the approach to select the "targets"

In order to complete the list of potential targets of our survey, we used all our contacts compiled over the years. The major sources used are the lists of participants to the ELRC workshops. This list comprises over 600 contacts. This survey explicitly (and exclusively) targeted the public sector representatives, at the national, federal and regional levels. We did our best to cover all the public services (see methodology and the profiles of our targets).

The list was first revised to keep public sector representatives only. Some of the workshop attendees, whether free-lancers or LSP employees, were removed from the list.

To achieve a high response rate, we considered the possibility to translate the questionnaire into multiple EU languages. After consulting with some of the EU Member State representatives, we decided to produce two language versions: English and French. We used the French questionnaire only in France and used the English one in the other Member States.

The first invitations to the questionnaires were sent out in July 2018 and responses were collected until after 1 September.

4.2.4. Selection of respondent profiles

In addition to the list received from the ELRC workshop organisers, we wanted to have a selection of public services so as to have a broad view of the services to which language technologies bring value. An *a priori* list of services was drafted and comprised of:

- Justice and Judicial Activities
- Tax and Revenue
- Public order and safety
- Administration of the State or Economic and Social Policy of the Community (excluding fiscal administration)
- Compulsory Social Security Activities
- Defense Activities
- Transport
- Fire Services
- Foreign Affairs
- Healthcare Provider
- Education
- Cultural services
- Utilities (e.g. Gas, Electricity, Telephone, Water...)
- Other

The *a priori* analysis of the email recipients was done internally with the assistance of the ELRC National Anchor Points whenever needed (for instance for information on the profiles of the institutions of the recipients).

4.2.5. Statistics about the responses

Response summary	
Total responses	196
Full responses	79
(These are the ones exploited herein)	
Incomplete responses	117

4.3. Results and analysis: landscaping the use of HLT in Europe

79 full responses, corresponding to a success rate of 13% of the recipients, were received, covering all the “departments” within the Member States. All countries contributed to the survey except the Czech Republic and the UK.

The distribution of the respondents was not balanced throughout the countries. We received more contributions from larger countries like France and Germany (8 each), and many with one or two responses. We decided not to draw any statistics per country as this would not have been significant.

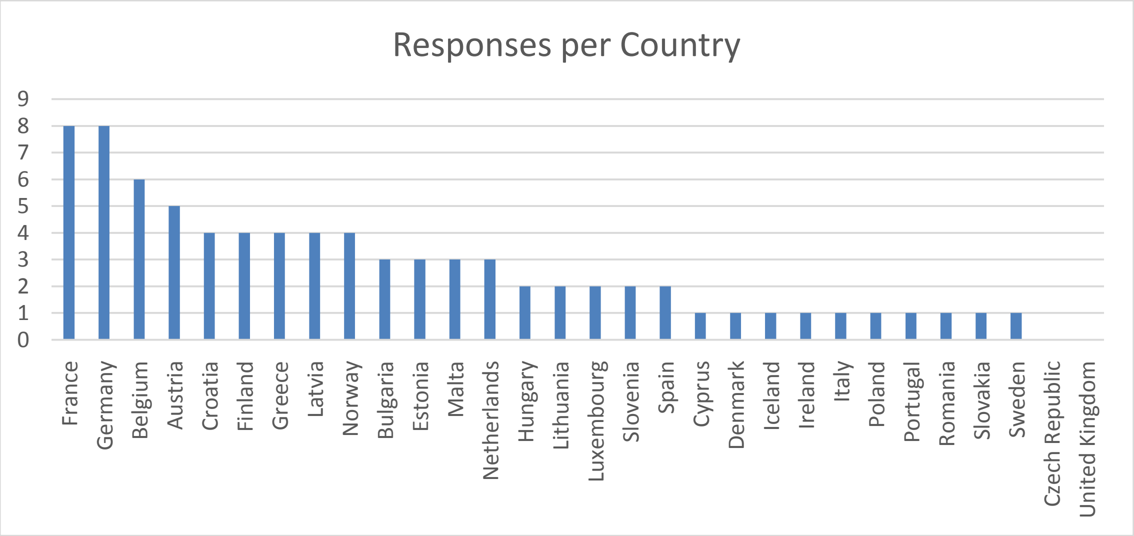
Overall, respondents’ profiles correspond to high management positions (head of, senior, director, manager, etc.). All types of public administrations are represented with a clear predominance of the ministry of economic and finance (16/79), then culture (9/79) and education (5/79). 27 respondents selected "Other" for their affiliations and provided additional details. Some of these 27 administrations belong to the sectors we listed but we decided not to proceed to any adjustment at this stage.

The following sections represent the details of the data resulting from the online survey collected through a group of 79 respondents.

4.3.1. Respondents across the EU Member States

All the countries responded to the survey, except UK and the Czech Republic.

Figure 94 Distribution of the respondents cross the EU countries (+Norway/Iceland)



(N=79)

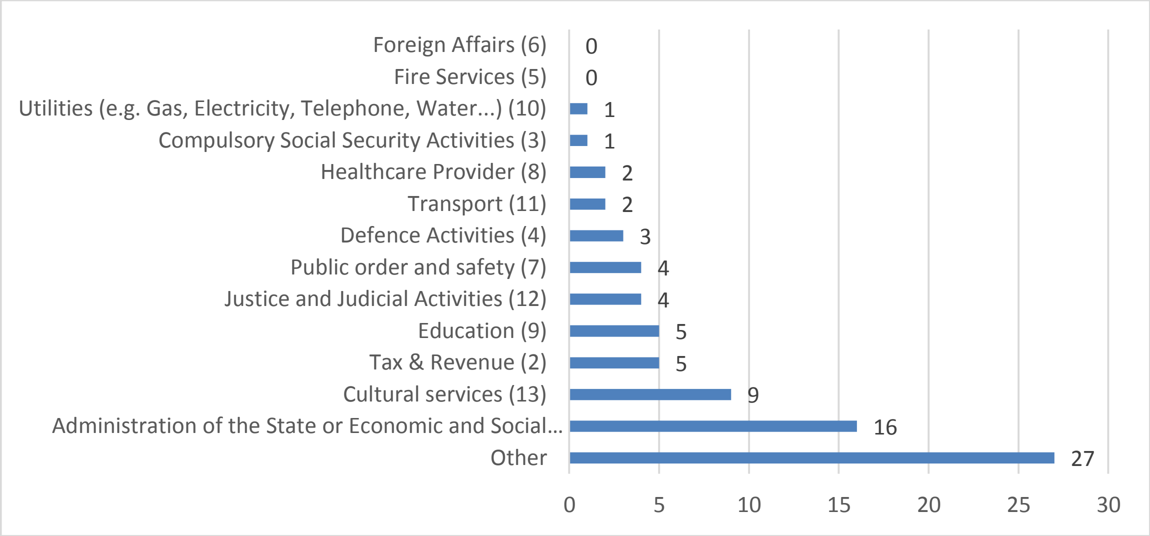
We have received 79 complete questionnaires (incomplete responses could be exploited at a later stage). Unfortunately, we could not secure a balanced representation among the countries. We got very satisfactory response rate from France and Germany (8 each), 6 from Belgium, but there are many countries with 1 response.

All respondents indicated the name of their organisation but less than a half mentioned the Supervisory Authority (about 43%).

4.3.2. Typology of the main users of HLTs in Public Services

4.3.2.1. Which public services (administration, utilities, etc.)?

Figure 95 Public services that fulfilled the questionnaire



(N=79)

34 of the 79 respondents indicated their supervisory administration (Q001) and 27 selected “Other”. Looking at the information provided as “Other”, we can cluster them as:

- Government Communication Service, Translations for the government (incl. Federal) and the public
- Translation and cultural diplomacy, Language Policy
- Telecoms Regulation, Regulation of electronic communications, postal services, railways
- Statistical Offices
- Parliament
- Central banking
- Transparency of Skills and Qualifications
- Tourism
- Press, Freedom of Information
- Standardisation
- Public Service related to Geographic Information
- Social Security Translation Services
- Information Systems
- Health

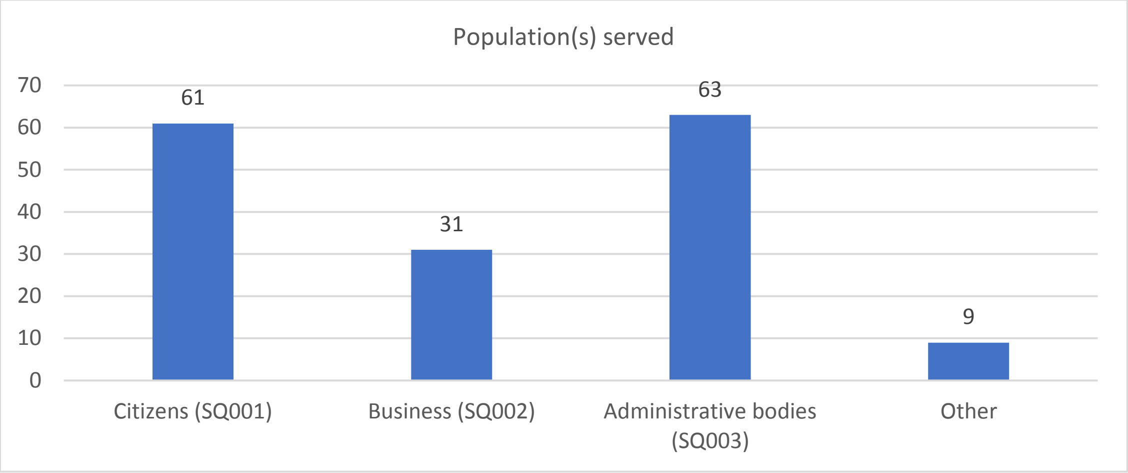
We categorised some of these according to the list of sectors in Section 4.2.4, for instance, “Telecoms Regulation” and “Regulation of electronic communications, postal services, railways” under “utilities”, and “Translation division at Federal Ministry” and “translation” under “Foreign Affairs”. In some cases we found that some respondents operate under multiple ministries.

4.3.2.2. Population(s) served

The services provided by the respondent administrations cover services to citizens, to businesses, and to other administration bodies. Most of the administrations use this technology for internal purposes (63), while a high number serve citizens (61) and businesses (31).

Nine respondents (out of 79) listed other beneficiaries, such as research and academic communities, Parliaments, other stakeholders (N.B. no other details), etc.

Figure 96 Populations served by the public bodies involved in the survey (multiple responses)



4.3.2.3. Languages in which the services are provided

The respondents were asked if they also provide their services in languages other than the state’s official language. 14 (out of 79) stated that they do and some of them specified other languages. In addition to EU languages, some responses also included Arabic, Turkish, Russian, and Hebrew (e.g. City of Vienna or Nicosia). One organisation reported the use of Ladino (Ladin) and one, a social security service, is dealing with 42 languages.

4.3.2.4. Languages

All public services in the countries are provided in their national official language(mandatory), some additional languages are also listed but are non-mandatory (English, French, German).

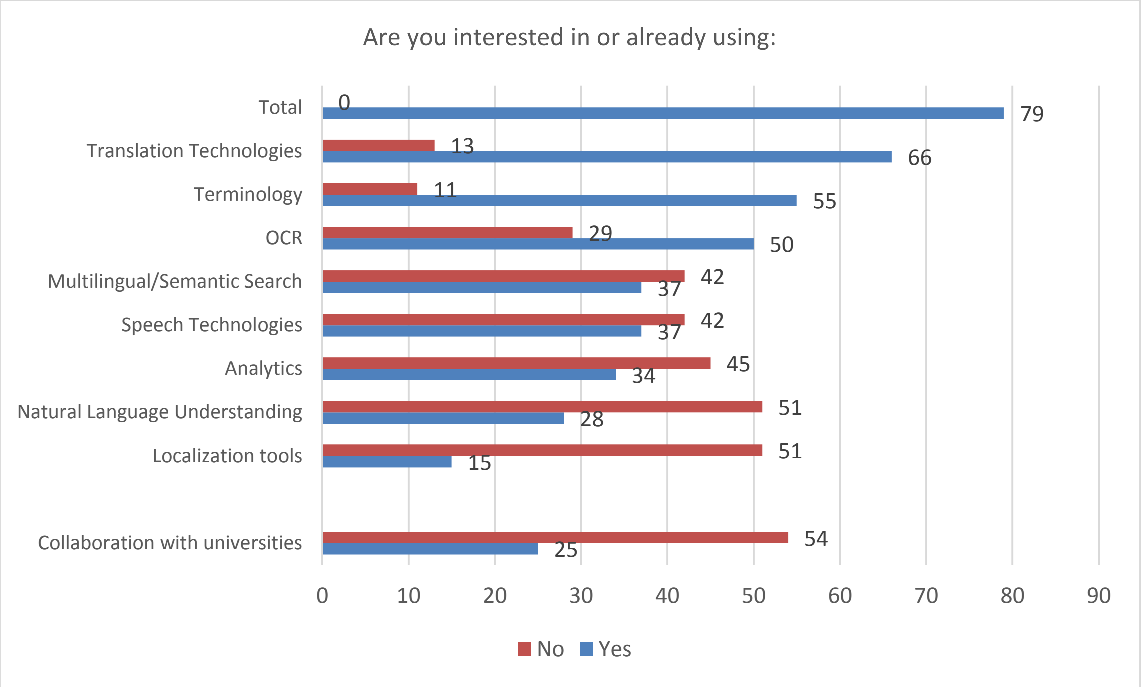
Interestingly, no regional language is mentioned except for Basque and Ladino; unfortunately, regional organisations are not well represented in our sample.

The use of English is mentioned as mandatory in 30 cases, and as non-mandatory in 22 cases (respectively 19, 12 for French and 20, 8 for German organisations). The main findings here are that, in addition to EU languages, there is a need for non-EU languages as well, such as Chinese, Arabic, Russian or Turkish.

4.3.3. Degree of use of Human Language Technologies

"Are you interested in or already using [XXXX] Technologies" and *" the degree of collaboration with academia".*

Figure 97 Status of use of LT in Europe



(N=79)

4.3.4. Speech technologies

These are used or reported to be of very high interest in 46.8% cases (37 out of 79).

4.3.4.1. Speech recognition

14 organisations confirm that they are using this technology. 17 others are planning to incorporate it or have identified their needs. In those cases, speech technology deals mainly with dictation, audio transcriptions of meetings, or automation of call centers (interactive voice response systems). The technology providers are mostly located in the US (e.g. Nuance Technologies) but some are EU companies (Vecsys).

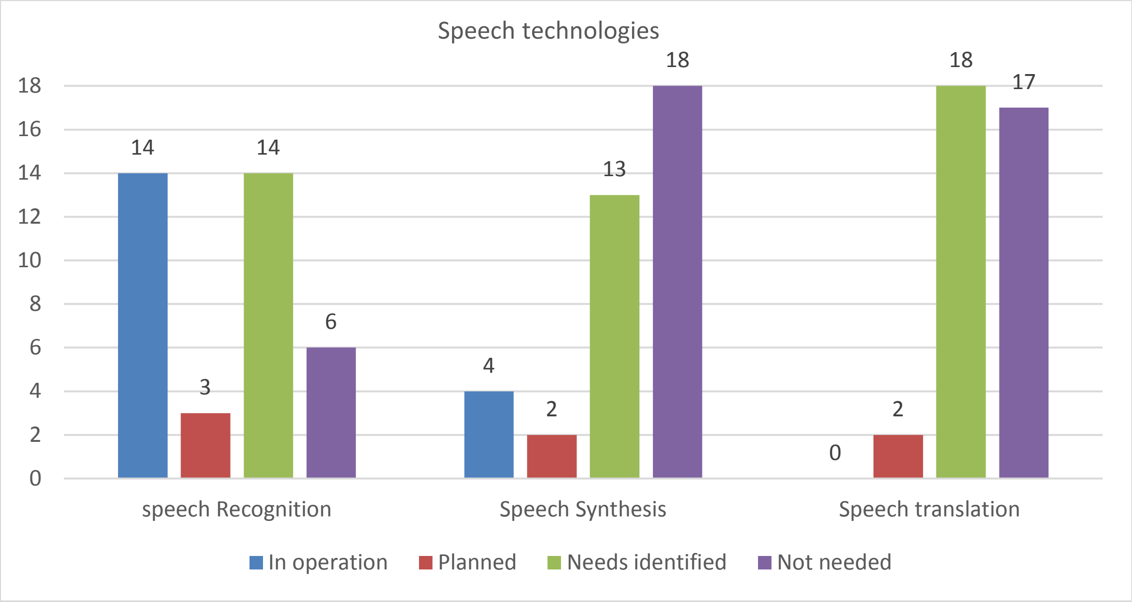
4.3.4.2. Speech Synthesis

It is less used (4/37) but needs have been identified in a high number of cases (13/37), the provider Voxygen is mentioned.

4.3.4.3. Speech translation

This technology and is not in operation yet in our sample. Two organisations have plans to use it, indicating 2019 and 2021 for testing.

Figure 98 Status of use of speech technologies



(N=37/79)

4.3.5. Translation technologies

A large number of respondents (66/79, 83.5%) indicated that these technologies are either in use or of very high interest.

4.3.5.1. Machine Translation

It is used by 17 organisations, 2 have plans to use it and 33 have identified their needs. Only 14 report some interest but without having identified needs. The technology suppliers are global US players or large EU vendors (Google, Microsoft, DeepL, Systran) with some mentioning CEF eTranslation and open source packages, such as Moses.

4.3.5.2. CAT Tools

There is a huge interest (48/66 cases, 72%) in CAT tools and some organisations are long-time users, going back to early 2000s. SDL is the predominant supplier in our sample, also mentioned are MemSource, MemoQ, and Wordbee.

4.3.5.3. Translation Memories

They are used by 32 institutions out of 66 and 13 have identified their needs for it.

4.3.5.4. Alignment tools

These tools are technical features of translation technologies and 27 organisations indicate they have them in operation. Looking at the list of suppliers, it seems that this is often part of the translation packages. However, some CAT users do not use all features (such as alignment, TMX, terminology).

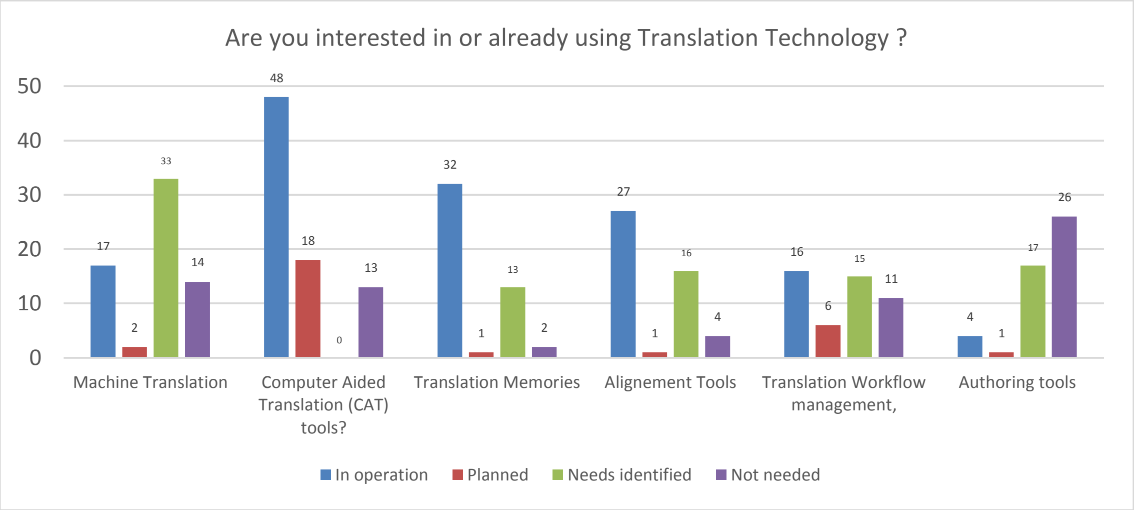
4.3.5.5. Translation Workflow Management

16 respondents operate workflow management tools for their translations. The number is not significant (16/48), however, another 21 respondents indicated they plan to use it or have identified the needs. The major companies supplying these tools in our sample are SDL, Wordbee, Isyde, MemoQ, etc.

4.3.5.6. Authoring tools

Authoring tools are not very specific to translation technologies, but are mostly used in multilingual contexts. The questionnaire focused on the use of controlled vocabularies and similar sophisticated tools). Only 4 respondents indicate it is in operation and 18 indicated that they identified their needs (with high level of interest as 5 respondents out of 17 declared high or very high interest). Only one tool is mentioned (an XML editor called FontoXML).

Figure 99 Status of use of translation technologies in Europe



(N=66/79)

4.3.6. Terminology technologies

This is an important sector for our survey and, as expected 55 of 66 responses for use of translation technologies indicate their interest in it.

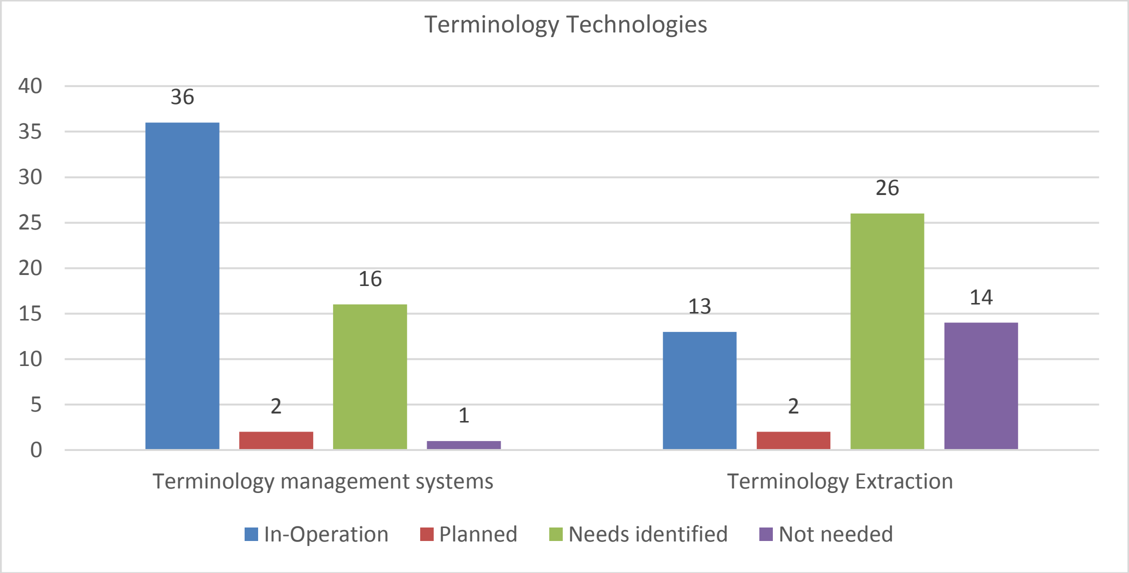
4.3.6.1. Terminology management system

As expected, terminology management systems are extensively used (36 of the 66 replies) or have identified their needs for this technology (16 of 66 replies) and 2 are planning its use already. A large number of suppliers are listed, tools from SDL being the most used ones.

4.3.6.2. Terminology extraction systems

Extraction systems are in operation in 13 cases, 2 have planned to use it and 26 have their needs identified. Again, SDL and Wordbee are widely quoted as providers but there are also some open source packages mentioned, like Sketch Engine.

Figure 100 Status of use of terminology technologies in Europe



(N=55/66)

4.3.7. Localisation technologies

This is an important category because it aims to meet the needs of citizens and other potential users (civil servants) living or working in a multilingual environment, especially for services carrying out cross-border activities. Again, we decided to keep this as a separate item (and not as a sub-item of translation technologies).

4.3.7.1. Localisation tools applied to web sites

15 respondents out of 66 indicated their interest in web site localisation tools. For 5 of them, the tools are operational, 2 have planned their use and 8 identified the needs. SDL and also Google are cited as suppliers.

4.3.7.2. Localisation tools applied to software

We did not expect this item to be useful within public sector environments. Nevertheless, we found that 2 services have it in operation, 1 is planning its integration and 8 have identified needs. The ones that have it in operation are language institutions and archiving services.

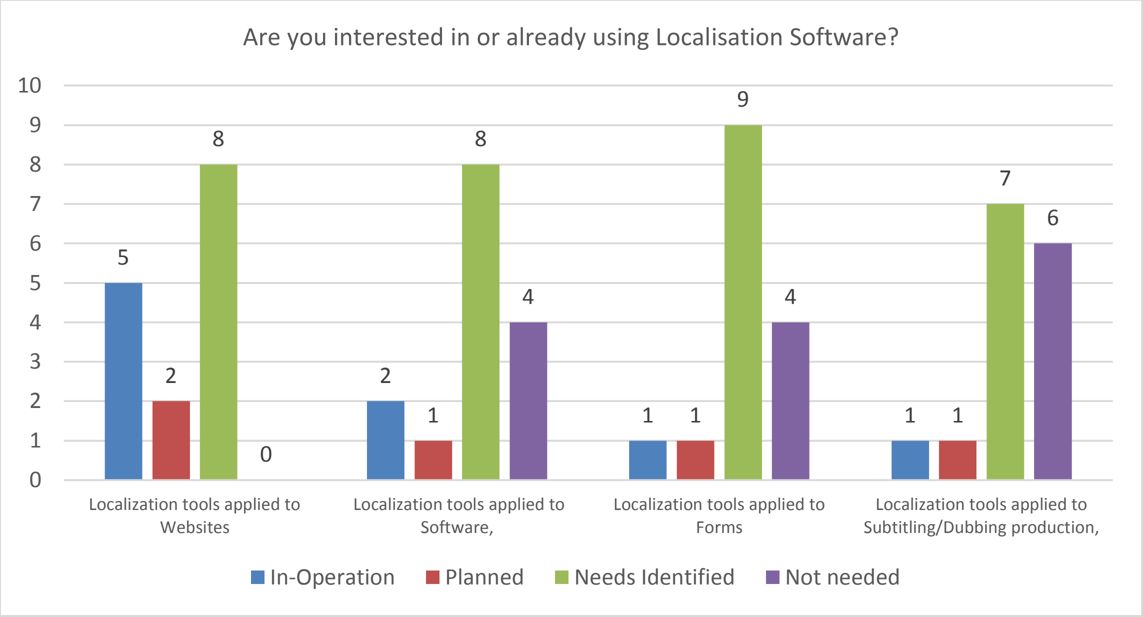
4.3.7.3. Localisation tools applied to Forms

Contrary to the previous item, we expected this to be in higher use, as many administrations work with online forms to be filled. Interestingly, only 1 institution indicated that such tool was in use, 1 is planning it and 9 have identified needs. Only SDL is mentioned as supplier.

4.3.7.4. Localisation tools applied to Subtitling/Dubbing

Here we targeted audiovisual and broadcasting services. These tools are in operation only at one university, while it is planned at another organisation and 7 have their needs identified (out of the 15 expressing some interest).

Figure 101 Status of use of localisation technologies in Europe



(N=15/79)

4.3.8. Natural Language Understanding.

This technology covers a large number of human-machine interactions and it is used or is of interest to 28 respondents (out of the 79).

4.3.8.1. Chatbot/Virtual Agents

This is an innovative application that adds value to traditional call centers. It is in operation at 2 respondents' organisations, planned by another one and 18 have their needs identified.

4.3.8.2. Keyword extraction

We discuss keyword extraction here because the responses in the supply side survey mention it as one of the top-used applications. In the 79 responses, 2 stated it is in operation, 3 that it is planned and 18 that they identified their needs for it.

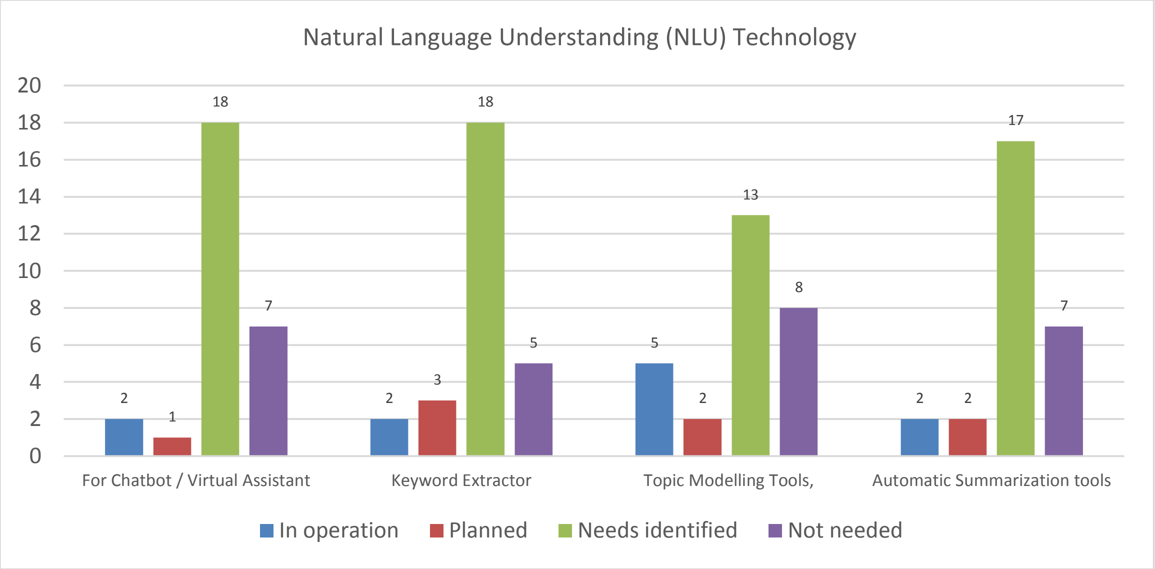
4.3.8.3. Topic Modelling Tools

This technology allows to filter documents according to their specific topics, and is in use by 5 respondents, planned for use by 2 and 13 have identified their needs. The suppliers are mostly academic partners but also software houses.

4.3.8.4. Automatic Summarisation tools

This is a very useful and advanced technology for those dealing with large documents, but it is in use only at 2 respondents (language institutions), while it is planned at another 2 and 17 have signaled they identified their needs. The respondents report using academic prototypes.

Figure 102 Status of use of NLU in Europe



(N=28/79)

4.3.9. Text analytics

Analytics is one of the very hot topics today and covers a large set of applications from text and data mining to sentiment/opinion analysis and detection, plagiarism, etc. 34 of the 79 completed questionnaires indicate they are using it or highly interested.

4.3.9.1. Text Mining tools

As expected, 9 responses indicate it is in use, 2 that it is planned, and 19 that they identify their needs. The users are national administration and utility services. Most of the applications are developed in-house or by software vendors, such as MemSource.

4.3.9.2. Sentiment Analysis

Tools are in operation at two sites (national libraries), planned at a third one and 15 indicate having identified their needs.

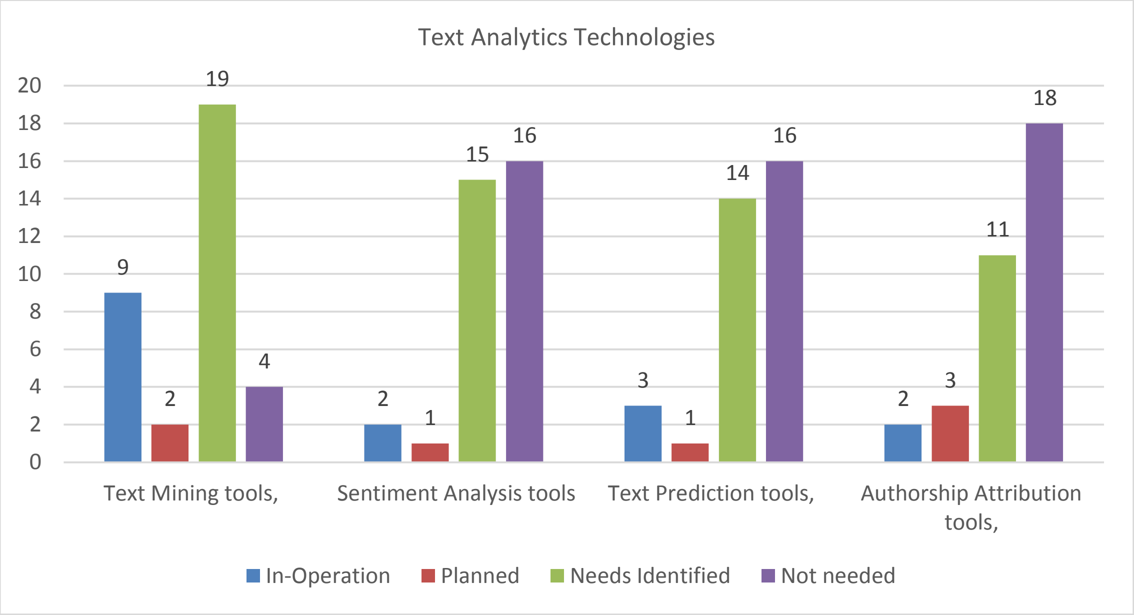
4.3.9.3. Text prediction tools

It is in operation at 3 sites, planned in an additional one, and 14 have signaled they have identified their needs.

4.3.9.4. Authorship attribution

We did not expect this to be widely used outside the academic world, as it is mostly used for plagiarism detection. It is used at 2 sites and 3 have plans to use it while 11 have identified needs. The applications are still based on research prototypes.

Figure 103 Status of use of text analytics technologies in Europe



(N=35/79)

4.3.10. Multilingual and semantic search technologies

Given the questionnaire's focus on public administrations' and services' HLT use, this was expected to be highly used. We see that 37 out of 79 are interested or using it and 42 that are not.

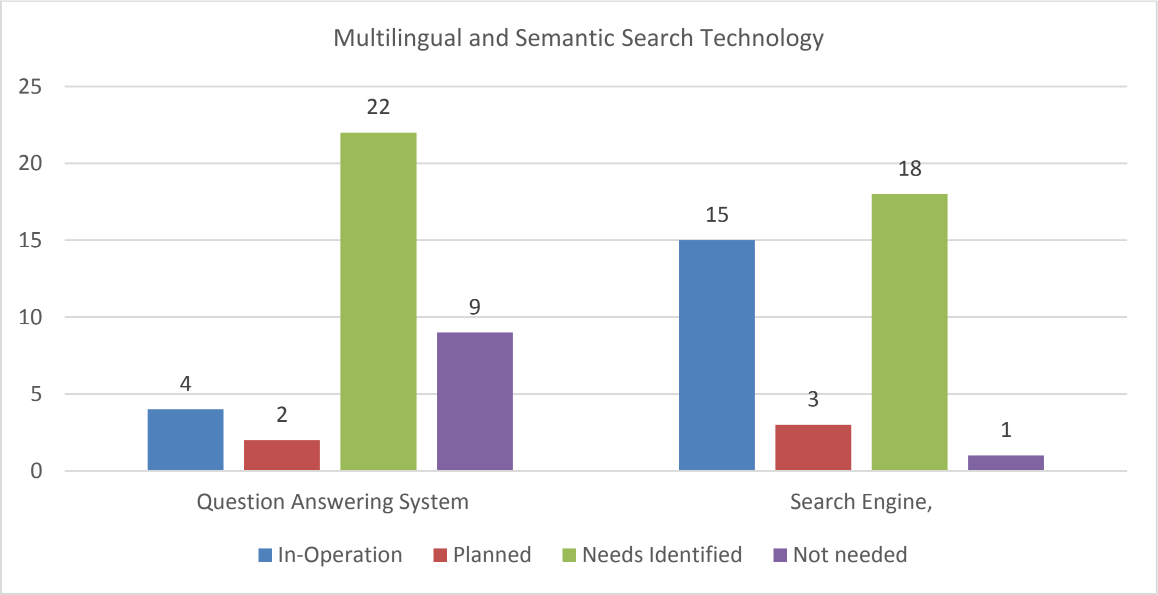
4.3.10.1. Question answering systems

These are sophisticated search applications and we see that 4 respondents have it in operation while 24 have their needs clearly identified. Among these, 2 have plans to incorporate it in their operations. No commercial suppliers are listed, rather prototypes from research and open-source packages.

4.3.10.2. Search engine

Interestingly, only 15 respondents signal that they are using a search engine on their websites, with three that have concrete plans and 18 that have identified their needs. In addition to the major suppliers (Google, Bing, Qwant, etc.), several open source packages have also been listed.

Figure 104 Status of use of multilingual and semantic search technologies in Europe

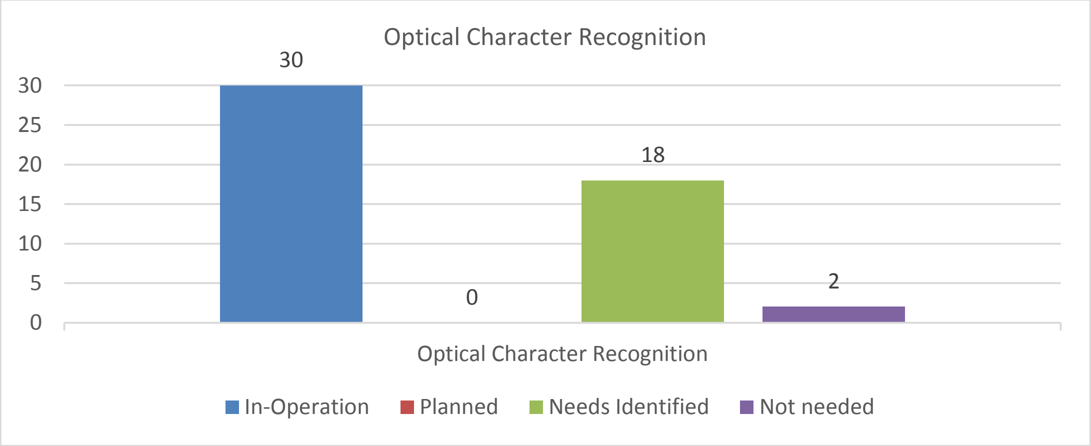


(N=37/79)

4.3.11. OCR

We also added Optical Character Recognition technology to our list as we expected it to be widely used by public services. Of the 50 respondents, 30 indicated it is in use, 18 that they have identified their needs, and 29 that they do not use it. The main suppliers for OCR are different from those providing NLP technologies, the most cited in the replies are Adobe, Nuance Technologies, Abby, and Jouve.

Figure 105 Status of use of OCR technologies in Europe



(N=50/79)

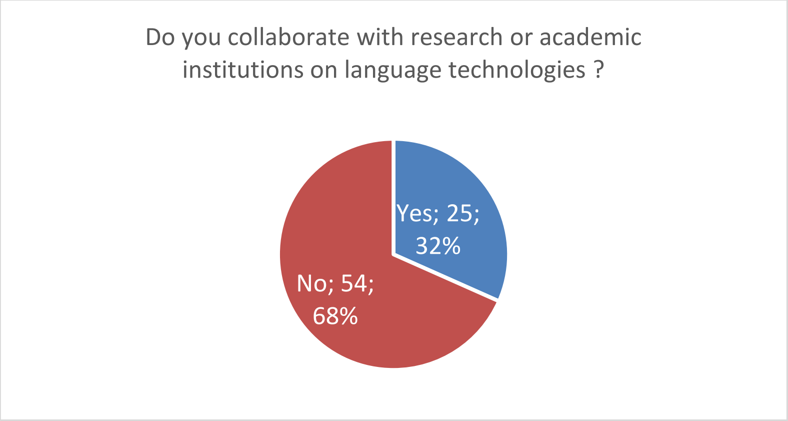
4.3.12. Other aspects of our survey

4.3.12.1. Collaboration with the research community

This set of questions focused on partnerships and collaboration between the public bodies in Member States and the research and academic world.

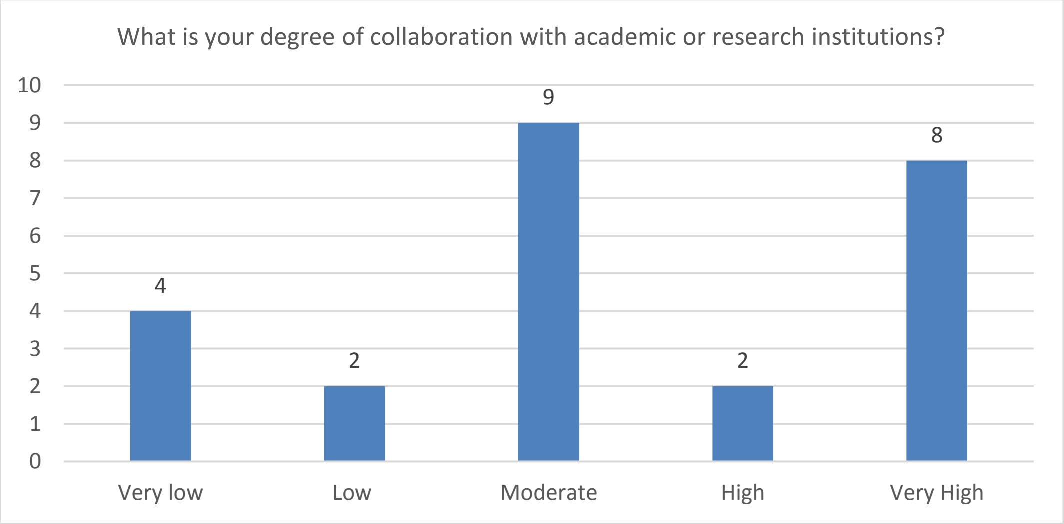
Only a third of the 79 respondents indicated some level of collaboration with R&D bodies and academia. Such collaboration takes primarily place with local and national universities, international collaborations are much less reported. In several responses, up to 5 research/academic institutions were listed (the maximum possible within the questionnaire).

Figure 106 Collaboration with the R&D/academic community



All in all, the degree of collaboration is rated high, where it is established.

Figure 107 Collaboration level with academia



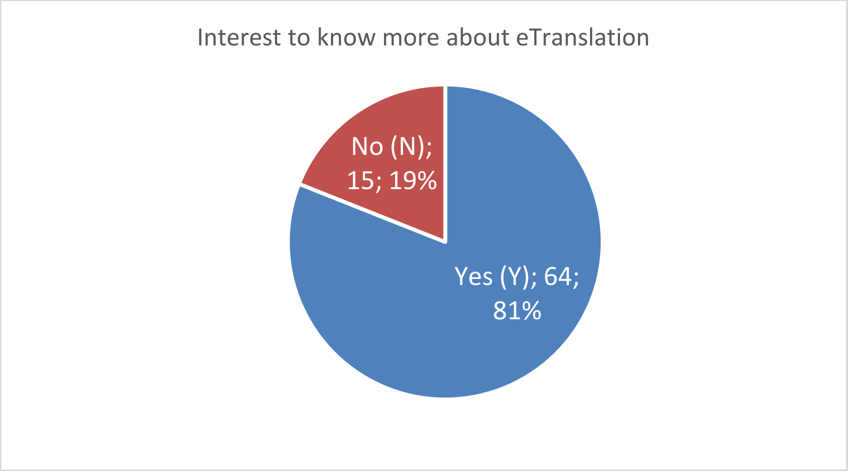
(N=25/79)

4.3.12.2. Expressed interest for "eTranslation" and ELRC

A set of questions were asked to collect information about the use of eTranslation platform. The first question was: "EU Member States' public administrations are allowed and encouraged to use the European Commission's machine translation platform eTranslation (previously known as MT@EC). Would you like to know more about it?"

A large number of respondents expressed their wish to receive more information.

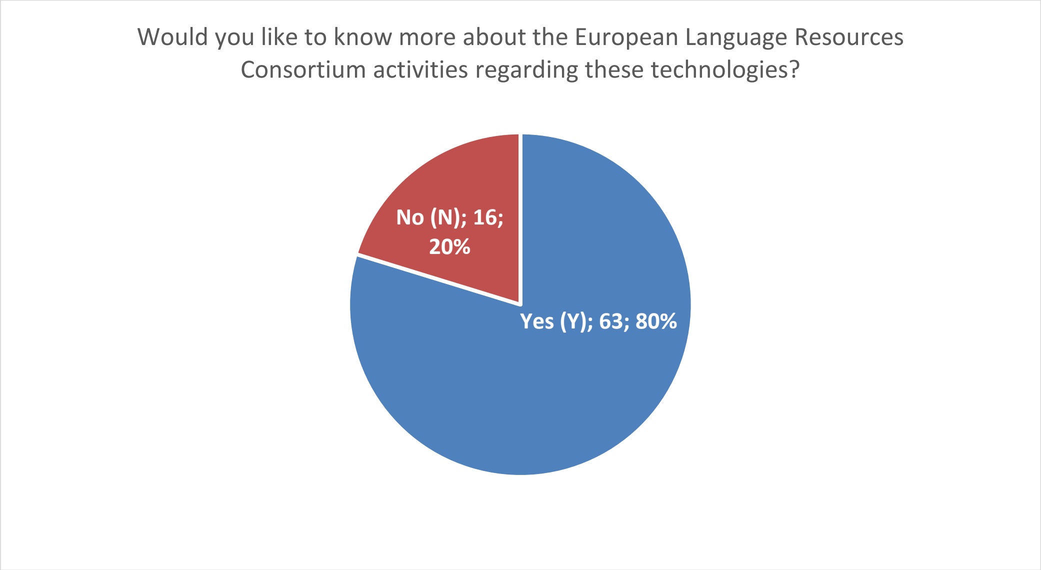
Figure 108 Request for more information about eTranslation



(N=79)

Similar answers were given relating to ELRC: "Would you like to know more about the European Language Resources Consortium activities regarding these technologies?"

Figure 109 Request for more information about ELRC



4.4. General summary and lessons learnt

The preliminary review of the survey shows that not a wide range of applications exploiting LT are used within EU Member States public administrations in our survey (as illustrated by the diagram of Section 4.3.3). The most frequently used technology remains automated translation (MT and its derivatives, 66 out 79 responses), followed by applications related to terminology (management and extraction with 55/66 positive responses), and then OCR. In the top 5 we also find speech technologies (mostly speech recognition) and multilingual and semantic search (both with 37 /79 responses). Technologies related to localisation appear to be low in demand, except for the localisation of web sites. Even for the localisation of web sites, only 5 respondents indicate that it is in use, out of the 15 that carry out some localisation activities.

The number of responses per public service is not high enough to draw far reaching conclusions about the adoption in the areas of activities of the participants. The area “administration of the State or Economic and Social Policy of the Community (excluding fiscal administration)” was the most relevant adopter of LT in our sample (16 responses). This are may cover multiple ministries, depending on the country. Some respondents choose “other” to be more specific in the description of their mission. These figures did not allow to draw up country profiles with respect to the use of LTs.

In all countries participating to the survey, public administrations use all their respective official languages. Some offer services or information in other EU languages and some go beyond EU languages and include e.g. Arabic, Turkish or Chinese, spoken by tourists, immigrants, business partners.

When reviewing the use of technologies and the reported plans, we observed that some have had a long lifetime: some speech technologies applications for train timetable access, for example, are operational since 1994 and machine translation platforms have been in use for almost a decade. It is notable that many participants of the survey are optimistic on the subject of deploying speech translation by 2021, multilingual and semantic search by 2020, etc.

The relationships between public administrations and academia are rather strong, one third of the respondents signal some collaboration with local universities. Out of these, 10 rated their collaboration very close.

Regarding eTranslation, 64 of 79 replies indicate that public administrations could be interested in the service.

Last but not least, when looking at the main suppliers, we found very few European large vendors as language technology providers. The major players are based in the US or are affiliates of US companies located in Europe. Nevertheless in some specific applications of machine translation (e.g. translation memories), European companies are often cited as well.

5. Task 4: Identification of value proposition of CEF AT

5.1. Summary

The purpose of Task 4 is twofold. On the one hand, it identifies the value proposition of the CEF AT building block, which involves describing the position of the building block in the European language technology (LT) market/ecosystem and identifying its qualitative and quantitative impacts. On the other hand, Task 4 aims at assessing potential future avenues of the building block in terms of development, provision of services, sustainability, and pricing (realistic service delivery model).

In order to identify the value proposition and assess potential avenues, we look to CEF AT as a business case and apply the business canvas methodology. A business model is the rationale of how an organisation creates, delivers and captures value. While this type of model typically applies to companies, it can also be used for other types of organisations, like public administrations, by taking care of specific constraints (policy-related, financial and operational). A business model can best be described through nine basic blocks, such as Customer segments and Key resources, which cover the main areas of a business and are described by answering a set of questions.

Based on a number of information sources, we provide answers to the questions associated to the blocks mentioned above and describe CEF AT's current business model. The sources consist of EU policies related to multilingualism and LT, information on the current CEF AT implementation, and meetings of CrossLang representatives with DSIs (focused meetings in the framework of Smart 2016/0103 Lot 2) and with CEF AT staff. Further sources consist of the outcome of three tasks performed in the Smart 2016/0103 project: the analysis of the LT supplier industry in the EU, the analysis of the LT demand of public administrations in the EU, and a competitiveness analysis in three areas of LT, i.e. machine translation (MT), speech technology and cross-lingual search. The latter analysis compares the areas across three regions (Europe, US and Asia).

The business model we describe based on the information sources shows that CEF AT's value proposition is currently focused, on the one hand, on eTranslation, an asynchronous MT service which is offered to the DSIs for the multilingual deployment of their services and guarantees information security, and, on the other hand, on a language resource collection effort (ELRC-SHARE) through which CEF AT publicly provides MT training data.

The European Commission Directorate General "Communications Networks, Content & Technology" unit G3, "Accessibility, Multilingualism & Safer Internet"[213] is the solution owner responsible for the drafting and adoption of CEF AT annual work programmes and focuses on the implementation of the language resources strand (1); the European Commission Directorate General "Translation" unit R3,

[213] https://ec.europa.eu/info/sites/info/files/organisation_charts/organisation-chart-dg-connect_en.pdf

"Informatics"[214], is the first solution provider focusing on the implementation of the automated translation service strand (2) and on the engines factory strand (3); while the European Commission Directorate General "DIGIT" unit C1 " Cloud & Service Management Capabilities"[215] is the second solution provider focussing on technical guidance and infrastructure strand (4). CEF AT budget is annually devised from the CEF Telecom work programme[216].

Through its yearly work programmes, CEF Telecom funds the implementation of the DSIs' Core Service Platforms (central hubs which enable trans-European connectivity) and the development of the DSIs' Generic Services (which link national infrastructures to the Core Service Platforms, provide integration, etc.). Work on Core Service Platforms takes place using service contracts with private sector contractors that have been selected through public procurement procedures; the totality of the cost is covered. Generic Services are developed in projects funded via CEF calls; a maximum funding of 75% of eligible costs is granted. [217]

Based on the information sources mentioned above, we distinguish two potential future business models, which we call the MT Business Model and the LT Business Model. They extend the current business model not only on the level of MT but also LT in general, which is of paramount importance for CEF AT in order to be able to support DSIs requiring cross-lingual functionality.

Whereas Tasks 1 to 3 are a mere reflection of facts and figures, Task 4 expresses the opinion of the consortium and should therefore be considered as a mere suggestion towards CEF AT. In no way, CEF AT can be held liable or accountable for any of the ideas expressed in the sections of the study related to Task 4.

The MT Business Model extends the scale of the MT service of the current business model, offers real-time translation, and makes CEF AT an instrument facilitating the customisation of MT engines. An increase in eTranslation demand is likely given the rising interest from DSIs, and the interest from public administrations in MT shown in the demand side analysis. The inclusion of real-time translation follows from DSIs' interest in chat translation and from the speed expectancies shaped by the online service of global players. Facilitating customisation through projects involving specialised companies allows eTranslation to distinguish itself from the service of global players on the level of domain adaptation, security and under-resourced languages. These are added values which match the deficiencies in the market identified in the competitiveness analysis.

The MT Business Model has clear implications on the level of physical and financial resources, cost structure and revenue streams. Scaling up the MT service requires a larger infrastructure and staff

[214] https://ec.europa.eu/info/sites/info/files/organisation_charts/organisation-chart-dgt_en.pdf

[215] https://ec.europa.eu/info/sites/info/files/organisation_charts/organisation-chart-digit_en.pdf

[216] https://ec.europa.eu/inea/en/connecting-europe-facility/cef-telecom

[217] The CEF-TC-2016-3 call made available €6.5 million, and the CEF-TC-2017-3 and CEF-TC-2018-2 calls €5 million each.

than in the current business model. Under certain conditions, CEF AT may have to charge its customers for the service. Budgets for calls for Generic Services should be sufficiently high in order to enable customisation of MT engines in various environments.

Calls for Generic Services should pay substantial attention to the stakeholders of DSIs, i.e. to the issues of public administrations in Member States. This implies the need for strong awareness raising of the eTranslation service among public administrations. In order to allow for valorisation of results, the calls for Generic Services should also value the business aspect of potential projects.

The LT Business Model goes beyond MT and also involves customisation of LT components in a broader sense. DSIs not only show a vivid interest in MT, but also in LT in general. The interest is not at the same level in public administrations, as they do not consider LT as mature enough, thus not ready to be invested in. However, the possibility to customise LT components in collaboration with specialised companies instead of using off-the-shelf tools to create components could be a strong motivation for public administrations to start using LT tools.

5.2. Introduction

The mission of CEF Automated Translation is **to provide multilingual support to the other pan-European DSIs so that individuals, administrations and companies in all EU Member States and EEA countries participating in the CEF Telecom Programme can use public services in their own language.**

The purpose of Task 4 is to identify the value proposition of CEF AT, as one of the aims of Lot 1 of the SMART 2016/0103 project is to develop a business case for CEF AT. The value proposition involves describing CEF AT's position in the European LT market/ecosystem and identifying its qualitative and quantitative impacts. Furthermore, Task 4 aims at assessing potential avenues for the future, in terms of development, provision of services, sustainability, and pricing (realistic service delivery model).

In order to identify the value proposition, we apply the business model canvas methodology. We select this methodology rather than an approach like functional review of public administrations, as we focus on the business case of CEF AT. A business model is the rationale of how an organisation creates, delivers and captures value. This model can best be described through nine basic blocks that cover the main areas of a business and are described by answering a set of questions. In the business model canvas methodology, these blocks are referred to as building blocks, but we will avoid this term, in order to avoid confusion with the terminology of the CEF programme (see Section 5.3.1). Instead, we will use the term model block. While a business model typically applies to companies, it can also be used for other types of organisations, like public administrations, by taking care of specific constraints (policy-related, financial and operational).

In order to describe the model blocks in the context of CEF AT and to identify future avenues, we take into account different types of model input (i.e. sources of information): the EU policies related to multilingualism and language technology (LT), the current CEF AT implementation, the meeting of CrossLang with CEF AT on 16 October 2018, the results of Task 1 to 3 of Lot 1 (supply side analysis, competitiveness analysis, demand side analysis), and the results of Task 3 of Lot 2 (focused meetings with DSIs).

The report of Task 4 is structured as follows. Section 5.3 details the model input based on which we describe the blocks of the business model and the future avenues. In Section 5.4, we describe the nine model blocks in the context of CEF AT by providing answers based on the information in Section 5.3. Based on these descriptions, we show the current business model of CEF AT. Sections 5.5 and 5.6 provide suggestions for potential future avenues, in the form of business models that extend the current business model. Finally, Section 5.7 lists our conclusions. Details on the methodology are provided in Annex P.

5.3. Model input

This section details the information sources upon which we base the business model described in Section 5.4. These sources consist of information on EU policies, results of other tasks in Lot 1 (supply side analysis, demand side analysis, and competitiveness analysis), and results of Task 3 of Lot 2 (focused meetings with DSIs).

5.3.1. EU policies

The Connecting Europe Facility (CEF)[218] is a key EU funding instrument to promote growth, jobs and competitiveness through targeted infrastructure investment at European level. CEF Telecom[219] is a key instrument that facilitates cross-border interaction between public administrations, businesses and citizens by deploying digital service infrastructures (DSIs) and broadband networks. Some of these DSIs are building blocks, i.e. they belong to the set of generic and reusable DSIs and provide basic functionality, such as e.g. secure communication between IT infrastructures. Among these building blocks is CEF AT, discussed below.

Through its yearly work programmes, CEF Telecom funds the implementation of the DSIs' Core Service Platforms (central hubs which enable trans-European connectivity) and the development of the DSIs’ Generic Services (which link national infrastructures to the Core Service Platforms, provide integration, etc.). Work on Core Service Platforms takes place using service contracts with private sector contractors that have been selected through public procurement procedures; the totality of the cost is covered. Generic Services are developed in projects funded via CEF calls; a maximum funding of 75% of eligible costs is granted. [220]

The mission of CEF AT is to provide multilingual support to DSIs so that individuals, administrations and companies in all EU Member States and EEA countries participating in the CEF Telecom Work Programme can access public services in their own language. From an operational point of view, CEF AT currently provides automated translation services to DSIs and public administrations through the eTranslation service. The primary targets of the service are DSIs.

[218] See Regulation (EU) No 1316/2013 of the European Parliament and of the Council of 11 December 2013 establishing the Connecting Europe Facility, amending Regulation (EU) No 913/2010 and repealing Regulations (EC) No 680/2007 and (EC) No 67/2010 Text with EEA relevance.

[219] See Regulation (EU) No 283/2014 of the European Parliament and of the Council of 11 March 2014 on guidelines for trans-European networks in the area of telecommunications infrastructure and repealing Decision No 1336/97/EC.

[220] The CEF-TC-2016-3 call made available €6.5 million, and the CEF-TC-2017-3 and CEF-TC-2018-2 calls €5 million each.

While CEF is oriented towards infrastructure development and integration, the H2020 programme funds research and innovation (through grants). [221]

5.3.2. Current CEF AT implementation

Currently, CEF AT provides or is in the process of constructing three services, i.e. the eTranslation service, a service desk and a language resources repository.

The eTranslation service is offered to DSIs and public administrations. It is developed by a team at DGT, based on DGT's translation memory (Euramis). The service provides asynchronous secured translation (with a short delivery time) using a SaaS model. As opposed to the MT service of global US players, eTranslation guarantees information security, i.e. it guarantees the confidentiality of the information being exchanged. The eTranslation service is currently being used by four DSIs: e-Justice, eProcurement, Online Dispute Resolution, and Public Open Data. The BRIS DSI (Business Registers Interconnection System) has committed to analysing the adoption of the service. The service can be integrated with the customer's platform (using an API, i.e. Application Programming Interface), or be accessed as a stand-alone service by a DSI or public administration through a web interface in which files can be uploaded (documents or files with text snippets). In the latter case, the user must have an ECAS account (i.e. he or she must log in through the EC's authentication service).

The second service provided by CEF AT is a Customer Service Desk, which is currently being created in the framework of SMART 2016/0103 Lot 2. It targets DSIs as well as public administrations seeking information on or assistance with the integration of the eTranslation service into their workflows and with the use of the service.

The third service provided by CEF AT is ELRC-SHARE, a language resources repository constructed by the ELRC (European Language Resource Coordination) action under the SMART 2014/1074 and SMART 2015/1091 LOT 2 contracts. The collected resources will be used to train the MT systems made accessible through the eTranslation service. The resources are governed by various licences (e.g. CC, Public Domain, …). DSIs and public administrations can also upload new corpora themselves. In addition, DGT makes a part of the data that it uses for training its MT system publicly available through the JRC (Joint Research Centre), as the DGT-TM corpus.[222] For legal reasons, not all of the data can be released publicly.

[221] H2020 supported research and innovation on MT through HLT topics ICT17-2014 and ICT-17 ('Cracking the language barrier'), amounting to €15 million of funding.

[222] https://ec.europa.eu/jrc/en/language-technologies/dgt-translation-memory

5.3.3. Supply side analysis

Task 1 of Lot 1, led by IDC in collaboration with the other consortium partners, had the ambitious objective to provide a complete overview of the European market of LT together with a description of the emerging trends and an estimate of the growth in the revenues.

As a first step, an exhaustive list of companies active in each of the EU member states in the domain of language technology was created. After a thorough desk research, a list of 473 companies was eventually produced that fully qualified as LT vendors. All relevant company details were collected in a database that can serve potential future business models (see Section 5.5.3).

Based on further desk research and on IDC's Worldwide Software Tracker, the total size of the LT industry within the EU26 plus Iceland and Norway in 2017 was estimated at approximately 800 million euro of which it is fair to say that this is a relatively small market. Germany holds the largest share of the LT market followed by the UK. Forecasts predict this market to grow at an average rate of 10% between now and 2021. From the online questionnaire it became clear that as of today, profitability is quite low, competition intense and margins are compressed. One of the main reasons for this low overall vendor profitability is the need to keep innovating and the cost related to this need. However, as AI will be increasingly part of LT, most players within the industry are quite positive when looking forward and expect the LT to be a growing market. In that respect, Natural Language Understanding (NLU) in general and chatbot applications in particular were most often mentioned as the emerging technology to look out for and are expected to become increasingly widespread.

The LT market in Europe is very fragmented and composed of Small and Medium-sized Enterprises (SMEs). The EU does not benefit from one global and leading player. Local players provide local solutions. Not surprisingly, English, German, French, Spanish and Italian are of most importance to the LT vendors. As LT markets for most European languages are small, this limits business opportunities for those vendors that focus on particular languages.

Although it became clear from the survey that the Public sector is seen as the most important customer segment for LT vendors, it accounts for only 20% of their revenues. In terms of profitability, the public sector lags behind the private sector.

Translation technology is considered as the biggest revenue contributor 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 have experienced lately in automatic translation accuracy.

5.3.4. Competitiveness analysis

Task 2 of Lot 1, dealing with the analysis of competitiveness in three LT areas, i.e. machine translation, speech technology and cross-lingual search, compares the market between the EU, the US and Asia. This resulted in a SWOT analysis for the EU. In this section, we discuss the most relevant points of the analysis in the context of the value proposition.

As for weaknesses and threats, the EU industry is fragmented with many small players that struggle to find a place in the market in order to compete with the global players, which dominate the market and upon which European businesses and public sector have become dependent. While the research position of the EU in the three areas is weakening, the global US players have a large competitive advantage in terms of research capacities, computing resources and available data. They distort the market, for instance by providing a free MT service, though MT is not their core business. They also have larger amounts of data at their disposal, because of copyright disparities between the EU (explicit permission required by European entities) and the US (fair use copyright exception), and because the intensive use of their systems allows them to collect a lot of user data. While EU industry is experienced with small and complex languages, the market for these languages is limited and provides restricted business opportunities, and the amount of accessible data for these languages is low.

As for strengths and opportunities, European MT developers have been successful in deploying services for the public sector through the support of EU-funded programmes. In the area of speech, the EU has demonstrated successful experience in multilingual infrastructure building projects which aim at reducing digital linguistic fragmentation across the EU. In the market for MT, speech technology and cross-lingual search as a whole, three deficiencies can be observed. Below, we discuss these deficiencies, which provide opportunities for the EU.

First, there are gaps in the offering for small and complex languages. The multilingual internal market in Europe has given developers in the EU the possibility to build strong experience in developing systems for smaller and complex languages, While there are limited business opportunities for these languages, as stated above, and the quality gap between those languages and the ones dealt with by global players increases, support for such languages is an essential means to preserve cultural identity, foster inclusiveness, and guarantee equal digital opportunities for speakers of smaller languages, thereby supporting a key principle of the EU, language equality.[223] Support for small and complex languages is also important in shaping a Digital Single Market.

Secondly, there is lack of domain-specific and application-specific MT. Due to the need to search for niche markets, many European developers have accumulated strong experience in customised and domain-specific solution development in the areas of MT and speech. This experience could be helpful to meet the lack of customised systems.

[223] See European Parliament resolution of 11 September 2018 on language equality in the digital age.

Thirdly, the MT market pays little attention to security and privacy. Global players do not provide high security MT applications. As for privacy, the EU has well-established practices for the creation of open data and policies fostering public data sharing.

5.3.5. Demand side analysis

Task 3 of Lot 1, dealing with the analysis of LT services and solutions currently in use by public administrations or planned for use in the next few years, came to the clear conclusion that automated translation is the most used technology. Also, tools and systems closely related to machine translation like CAT tools, translation memories or terminology management systems are most often used. In that respect, it does not come as a surprise that many respondents are interested in receiving more information on the current MT offering of eTranslation.

Although many of the respondents of the online survey gave an optimistic view on their future needs for 2020 and beyond to deploy other language technologies than MT, it seems that these technologies are not considered mature enough today to be used in their working environment. In that respect, optical character recognition (OCR) and speech technology (primarily speech recognition) appear to be on many administrations' radar for implementation, but today's adoption of these forms of LT remains rather low.

In terms of vendors, EU-based players are often cited when referring to some specific applications like translation management systems or translation memories. In all other domains, major players – when cited– are predominantly US-based.

The collaboration between public administrations and academia appears to be strong with a third of the respondents pointing out some collaboration with mostly local or national universities.

5.3.6. Focused meetings with DSIs

As part of Task 3 of SMART 2016/0103 Lot 2, CrossLang held a round of focused meetings with nine DSIs in order to get insight into their activities and needs in the area of MT and in the broader area of LT. The DSIs in question are the following: Business Registers Interconnection System, Cybersecurity, eHealth, e-Justice, eProcurement, Europeana, Online Dispute Resolution, Public Open Data, and Safer Internet.

As regards MT, the meetings show that all DSIs either already apply MT or have an interest in doing so in the future. In some cases, use of an MT service is not possible due to confidentiality reasons (preference for local installation of a system, e.g. Cybersecurity) or due to a need for extremely high quality standards for translations (eHealth). There also appears to be a need for increased interaction with MT developers, for instance in the context of MT customisation (e.g. Europeana) or for the avoidance of erroneous terminology (e.g. eProcurement). Concerning the speed of translation delivery, some DSIs would like to see the possibility of real-time translation in the future (e.g. e-Justice is interested in the translation of chat).

As regards other types of LT, there is also a clear interest from DSIs. For instance, multiple DSIs have an interest in anonymisation of texts before publishing them, or in some type of classification (for instance, classification of documents).

5.3.7. Meeting with CEF AT

On 19 October 2018, a meeting was held in Luxembourg between representatives of CrossLang (the leader of Task 4 in SMART 2016/0103 Lot 1) and CEF AT, in order to exchange ideas about the value proposition. Before this meeting, a draft version of the report of Task 4 was provided to CEF AT. The model block questions (see Section 5.4) in this draft version were discussed with the participants. The draft version was updated based on the outcome of the meeting.

5.4. Business model

This section describes nine building blocks in the context of CEF AT by providing answers to the building blocks' questions based on the information in Section 5.3. At the end of this section, we provide a figure with the business model of CEF AT, based on the answers provided in the section.

5.4.1. Customer segments

CEF AT has customers at different levels:

1) DSIs. CEF AT is in direct contact with them and provides customisation services to them (use cases). It does not provide manual translation services to them, as those are performed by DGT. Neither does it provide post-editing services.

2) Public administrations. CEF AT asks DSIs to promote the eTranslation service towards public administrations in the Member States. As for MT-related contacts of the EC with other EU institutions, contacts pass through DGT.

3) The area of public interest. For instance, museums that are involved in cross-border collaboration may use the eTranslation service. In this respect, EU citizens also indirectly benefit from CEF AT.

4) CEF AT does not have companies as its customer.

DSIs and public administrations are the most important customers of CEF AT. Currently, four DSIs are using the eTranslation service: e-Justice, eProcurement, Online Dispute Resolution, and Public Open Data. Business Registers Interconnection System has committed to analysing the adoption of eTranslation.

5.4.2. Value proposition

CEF AT helps realising the ambition of the different DSIs in making their service/content multilingual, so that all organisations and individuals in the different Member States can benefit from the DSIs' expertise in the different domains they are active in.

CEF AT allows for its customers to reduce their costs, by automating the translation activities that are required for their service/content to become multilingual. Moreover, CEF AT provides indirect benefit to its customers by coordinating MT effort (as they do not need to install MT systems themselves). The CEF Telecom programme also provides indirect benefit by increasing the cohesion between Member States and shaping the Digital Single Market.

CEF AT provides its customers with eTranslation, a service for translating documents or text snippets, within a short delivery time, i.e. asynchronously. The service guarantees information security (confidentiality of the information being exchanged) and can be integrated with the customer's platform or be accessed as a stand-alone service through an interface in which documents can be uploaded. Through a Customer Service Desk (currently under development), CEF AT provides its customers with information on or assistance with the integration of the eTranslation service into their workflows and information on the use of the service.

As for products, CEF AT publicly provides a subset of the data used for training its MT system (this subset is the DGT-TM corpus). The training data are taken from the translation memory Euramis of DGT. For legal, privacy and security reasons, not all of Euramis can be released publicly. CEF also provides corpora through ELRC-SHARE, a language resources repository, as well as the possibility for customers to upload corpora themselves. These corpora are considered useful for feeding the system accessible through the eTranslation service and have different licences (e.g. public domain).

5.4.3. Channels

The AMB (Architectural Management Board) coordinates architectural activities of building block DSIs and is in contact with the DSIs using those building blocks. CEF AT receives feedback from DSIs through the AMB. End users provide technical feedback to DGT. In the framework of projects implementing Generic Services, they also provide feedback to CEF AT. There are also memoranda of understanding between entities working for CEF (legal framework). Information is provided via channels like mailing lists and the SMO (Stakeholder Management Office), the entity responsible for dissemination of information on CEF activities to stakeholders.

5.4.4. Customer relationships

There is communication in one direction: between CEF AT and DSIs: CEF AT asks DSIs what they need. While the initial talks with them were mostly about MT, the mandate (mission) of CEF AT is to make DSIs multilingual. This means providing them also with other services than MT.

The relationship of CEF AT with public administrations passes through ELRC-SHARE, as administrations can download or upload resources through this website.

5.4.5. Revenue streams

Currently, the MT service is free.

5.4.6. Key resources

CEF AT needs physical resources in the form of infrastructure and data for training the MT system. It requires infrastructure for creating engines, running them, and updating them. It makes use of cloud services, but this has to be organised in collaboration with DIGIT, as there is a framework contract. As for data, it makes use of Euramis, the translation memory of DGT, as well as potential other corpora from ELRC-SHARE.

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, etc. The activities are performed in several entities (CNECT, DGT, DIGIT), both by internal staff and by external staff with specific expertise.

As for financial resources, the budget for the Core Service Platform is provided by CEF. This budget covers the costs of the core service, including development, hosting, consultancy, business requirements analysis, technology assessment, resource repository and collection, support, etc. The Generic Services are also funded by CEF. CEF AT drafts the objectives of these services, which have to be cross-border. These services are not owned by the EC, and hence the IPR is not acquired, contrary to the case of the Core Service Platform.

5.4.7. Key activities

CEF AT focuses on operational development and deployment of engines. However, to remain state of the art, technology watching is organised.

Research is not in the scope of CEF.

5.4.8. 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 cloud service broker.

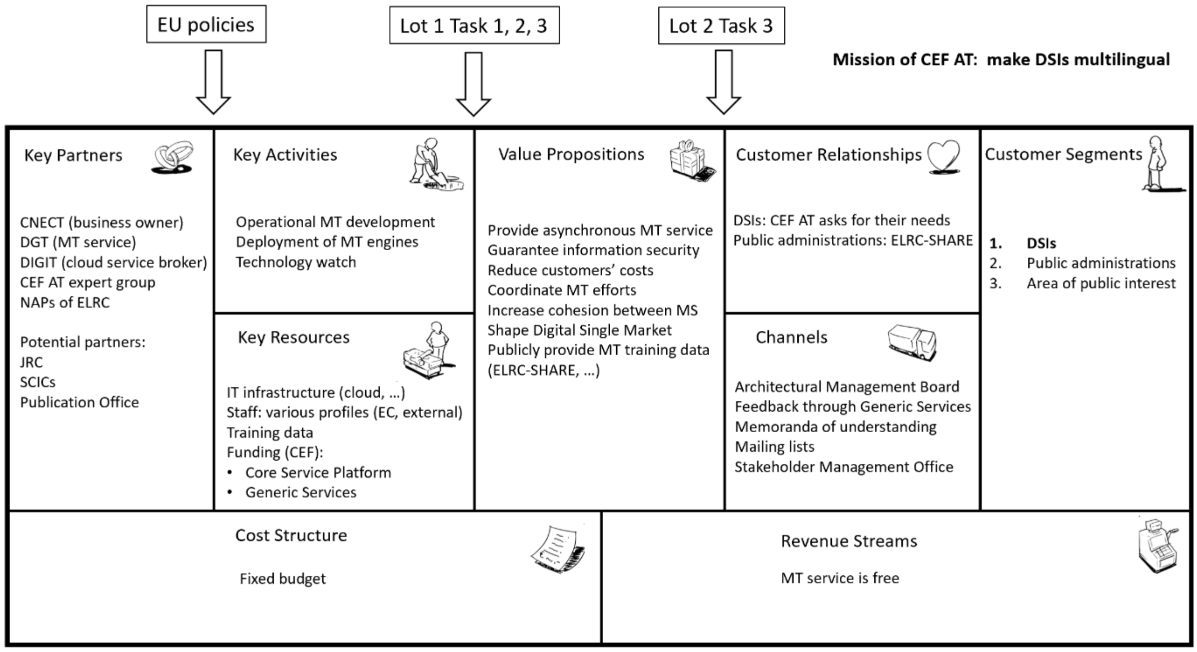
5.4.9. Cost structure

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

5.4.10. CEF AT business model

Based on the answers to questions in building blocks, the current CEF AT business model can be constructed. It is shown in Figure 110.

Figure 110 Current CEF AT business model



EU policies

Lot 1 Task 1, 2, 3

Lot 2 Task 3

Mission of CEF AT: make DSIs multilingual

Key Partners

CNECT (business owner)
DGT (MT service)
DIGIT (cloud service broker)
CEF AT expert group
NAPs of ELRC

Potential partners:
JRC
SCICs
Publication Office

Key Activities

Operational MT development
Deployment of MT engines
Technology watch

Key Resources

IT infrastructure (cloud, …)
Staff: various profiles (EC, external)
Training data
Funding (CEF):

- Core Service Platform
- Generic Services

Value Propositions

Provide asynchronous MT service
Guarantee information security
Reduce customers' costs
Coordinate MT efforts
Increase cohesion between MS
Shape Digital Single Market
Publicly provide MT training data (ELRC-SHARE, …)

Customer Relationships

DSIs: CEF AT asks for their needs
Public administrations: ELRC-SHARE

Channels

Architectural Management Board
Feedback through Generic Services
Memoranda of understanding
Mailing lists
Stakeholder Management Office

Customer Segments

1. **DSIs**
2. Public administrations
3. Area of public interest

Cost Structure

Fixed budget

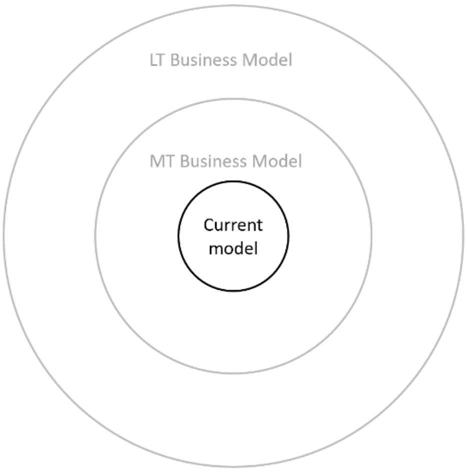
Revenue Streams

MT service is free

5.5. Extension 1: MT Business Model

The business model in Section 5.4 describes the current situation of CEF AT. Based on the information sources for the construction of the business model (Section 5.3), we identified a number of potential future avenues. We present these avenues in the form of two business models that extend the current model to various degrees, as shown in Figure 111. The two business models should be thought of as suggestions. We present the first one in the present section, and the second one in Section 5.6. Many more business models are conceivable, based on specific extensions in the models described here. When describing the two business models, we link the extensions they implement to the information sources in Section 5.3.

Figure 111 Extensions of current business model



Like the current business model, this model is focused on the provision of an MT service. It implements the following extensions with respect to the current business model, in order to make the eTranslation service sustainable (see Section 5.3.1):

1. It scales up the MT service level.
2. It offers real-time translation to DSIs and public administrations.
3. It facilitates customisation of MT engines, especially for under-resourced languages.
4. It provides extra promotion/integration of CEF AT's MT service offering through calls for Generic Services.

These extensions, and their implications in terms of resources, costs and revenue, are discussed in the below sections. Section 5.5.5 visualises the MT Business Model.

5.5.1. Scaling up of MT service level

The motivation for scaling up the service is the observation that the demand for the eTranslation service is very likely to increase. The focused meetings (Section 5.3.6) have shown that the service is being used by a number of DSIs and will be used by other DSIs in the future. The demand may further rise because of the creation of new DSIs and the use of the service by public administrations. The demand side analysis (Section 5.3.5) has shown that MT is the type of LT that raises the highest interest in such organisations.

A further reason to scale up the service is that the quality of MT has substantially improved in recent years, with the advent of neural machine translation. This has increased the usability of MT output.

5.5.2. Real-time translation

As for the delivery speed of translations, the focused meetings have shown that some DSIs are not only interested in asynchronous but also real-time translation. For instance, Safer Internet sees a benefit in the translation of chat in the context of helplines. Cybersecurity equally shows such a benefit, in order to increase cross-border accessibility to information.

Apart from a specific interest in real-time translation, there is also an expectancy from nowadays MT users in terms of translation speed, as the global MT players offer high-speed engines. The intensive use of such engines generates a lot of user data (see Section 0), which are helpful for improving an MT system. Therefore, the provision of a real-time translation service is also highly likely to increase the level of adoption of the eTranslation service.

5.5.3. Customisation of engines

Customisation allows DSIs and administrations to adapt MT resources to their domains. The facilitation of MT engine customisation is motivated by several observations. Some DSIs would like to make use of an MT system that is more fit for their domain or that can be kept local to their environment. For instance, Europeana is interested in an MT system adapted to their domain. eProcurement would like to avoid erroneous translation of their terminology. In case of Cybersecurity, there is an interest in running systems locally in order to guarantee confidentiality. The need for customised and domain-specific MT systems, as well as the need of security and privacy, are confirmed by the competitiveness analysis (Section 0).

Customisation is highly relevant for under-resourced languages. The supply side in the market focuses on a limited number of languages and the competitiveness analysis shows that there are limited business opportunities for most European languages, i.e. small and complex languages. There is also a lack of MT training data for such languages. Therefore, customisation of MT systems for such languages provides an important value.

As the global US players, offering a general-purpose MT service, do not satisfy users’ needs in terms of domain adaptation, security and under-resourced languages, customisation can become an important added value for CEF AT, which it can promote towards public administrations in order to distinguish its service from the well-known online translation engines.

From a technical point of view, CEF AT could facilitate customisation by providing MT resources like engines or data for training engines to DSIs and public administrations (some resources may even be made public, as is currently the case for a part of the MT training data, through the DGT-TM corpus and other language data resources collected through ELRC-SHARE). Based on these resources, DSIs and public administrations could create customised engines. DSIs and public administrations can continue the training of a neural MT engine using new, domain-specific data, and pre- and post-processing modules can be added to an MT system.

Customisation could involve collaboration of the CEF AT consumer with specialised companies. Collaboration can take place in the framework of Generic Services, or the consumer may consult the supplier database created during the supply side analysis (Section 5.3.3) and directly contact a supplier. In addition to performing customisation activities, the specialised company may also provide information on best practices in MT.

5.5.4. Promotion of the calls for Generic Services

The eTranslation service should receive extra promotion through an increase in the number of calls for Generic Services, or in budgets for such services. Through these calls, the MT service can be customised for a specific environment or integrated with it.

A bottom-up approach for calls for Generic Services initiated by CEF AT would be helpful, as the need for services offered by DSIs is often located downstream, i.e. at the level of public administrations. This requires the latter to be sufficiently aware of the potential of CEF AT’s calls for Generic Services, as well as to provide DSIs with information on the specific multilingualism issues they encounter. Based on this information, CEF AT can initiate calls for Generic Services that can eventually lead to solutions for these issues, and that may tackle common issues encountered by the DSIs’ stakeholders. In this respect, the ecosystem graph with DSI stakeholders, developed in the framework of SMART 2016/0103 Lot 2, provides interesting information. It does not only contain coordinates of EC staff and other people involved in the DSIs, but also of national contact points and staff of public administrations in the Member States.

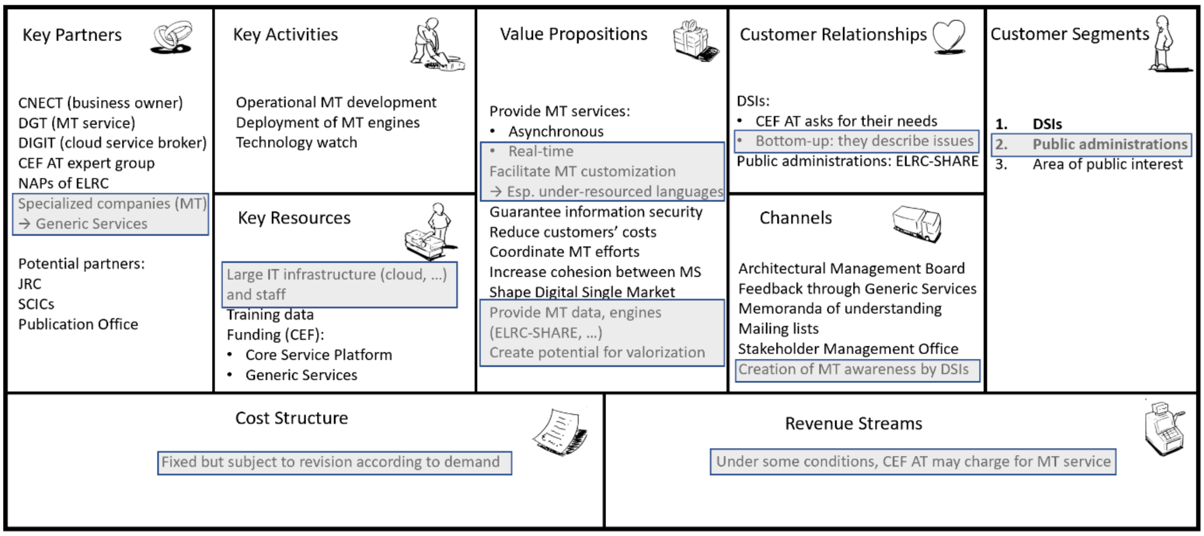
It may be interesting to give a stronger emphasis to the business aspect in calls for Generic Services, i.e. to valorisation. In Task 2, one of the findings was that Europe is strong in research and innovation but not successful in scaling innovations and capturing the market. Preferably, solutions elaborated in projects are reusable by other DSIs and public administrations and also constitute a value for the industry and lead to business growth. In this respect, the idea of a repository such as ELRC-SHARE could help in providing reusable resources.

5.5.5. Extended business model

The extended business model is shown in Figure 112. Changes with respect to the current business model are shown in grey.

The figure shows not only the extensions discussed above, but also their impact on certain model blocks. As for key resources, a larger infrastructure is needed (especially given the computing power required to build and run neural MT engines) as well as a larger staff. As for cost structure and revenue streams, CEF AT may request a fee from public administrations under certain conditions (e.g. large throughput), or its budget may have to be revised according to the size of the demand for the MT service.

Figure 112 MT Business Model



Key Partners

CNECT (business owner)
DGT (MT service)
DIGIT (cloud service broker)
CEF AT expert group
NAPs of ELRC
Specialized companies (MT)
→ Generic Services

Potential partners:
JRC
SCICs
Publication Office

Key Activities

Operational MT development
Deployment of MT engines
Technology watch

Key Resources

Large IT infrastructure (cloud, ...) and staff
Training data
Funding (CEF):

- Core Service Platform
- Generic Services

Value Propositions

Provide MT services:

- Asynchronous
- Real-time

Facilitate MT customization
→ Esp. under-resourced languages
Guarantee information security
Reduce customers' costs
Coordinate MT efforts
Increase cohesion between MS
Shape Digital Single Market
Provide MT data, engines (ELRC-SHARE, ...)
Create potential for valorization

Customer Relationships

DSIs:

- CEF AT asks for their needs
- Bottom-up: they describe issues

Public administrations: ELRC-SHARE

Channels

Architectural Management Board
Feedback through Generic Services
Memoranda of understanding
Mailing lists
Stakeholder Management Office
Creation of MT awareness by DSIs

Customer Segments

1. **DSIs**
2. **Public administrations**
3. Area of public interest

Cost Structure

Fixed but subject to revision according to demand

Revenue Streams

Under some conditions, CEF AT may charge for MT service

5.6. Extension 2: LT Business Model

This business model is an extension of the MT Business Model described in Section 5.5. In the LT Business Model, CEF AT extends its scope beyond MT.

The LT Business Model implements the following extensions with respect to the MT Business Model:

1. Collect and provide new LT resources and tools (in addition to the current resources in ELRC-SHARE)
2. Offer LT services
3. Facilitate customisation of LT resources, tools and services, especially for under-resourced languages
4. Provide promotion of CEF AT's LT offering through calls for Generic Services

These extensions, and their implications in terms of resources, costs and revenue, are discussed in the below sections. Section 5.6.3 visualises the LT Business Model.

5.6.1. Customisation of components

As mentioned in Section 5.3.1, the mission of CEF AT is to provide multilingual support to DSIs, which is not necessarily restricted to MT. Based on the demand side analysis (Section 5.3.5), it appeared that the adoption of LT other than MT by public administrations is rather low, and that they perceive LT as insufficiently mature. However, as for DSIs, focused meetings showed that the latter have a clear interest for LT other than MT.

The possibility of supporting the dissemination of LT has similar motivations as in case of the MT Business Model. The competitiveness analysis of Section 0 shows that market deficiencies related to domain adaptation, customisation, security and under-resourced languages are not restricted to MT, but also apply to other LT types, i.e. speech technology and cross-lingual search. Therefore, the LT Business Model generalises the possibility of customising components from MT to LT, especially for under-resourced languages. This gives both DSIs and public administrations the possibility to improve LT components in their context and makes it more likely that public administrations will adopt LT tools. Customisation may take place in a local environment, in order to guarantee confidentiality.

From a technical point of view, CEF AT could facilitate customisation by extending the data currently provided through ELRC-SHARE with LT resources like components, data for training components, and tools required for training. Examples of LT components are the ones that CEF AT is using or planning to use in its MT system, such as components for Named Entity Recognition (NER), Quality Estimation (QE), protection of tags, combination of speech and translation, etc. Other examples relate to classification and anonymisation (see Section 5.3.6).

As in the case of MT engine customisation, any technical assistance would originate from a specialised company in the framework of Generic Services rather than from CEF AT itself, or from a supplier in the database mentioned in Section 5.3.3. As in case of the MT Business Model, the company may not only perform customisation activities but also provide information on best practices in LT.

5.6.2. Promotion of calls for Generic Services

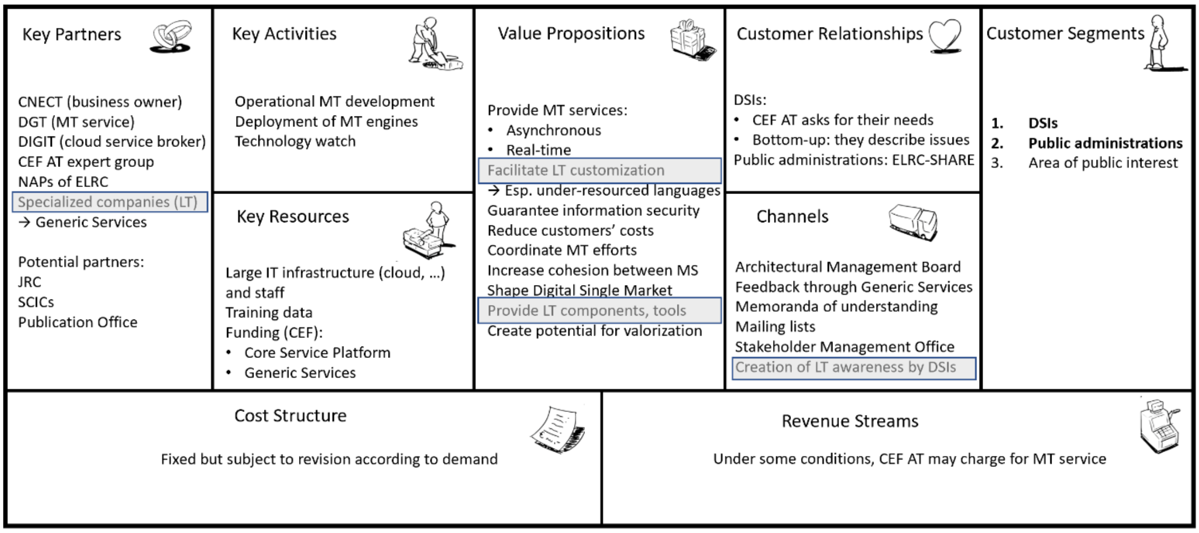
Similarly to the MT Business Model, the possibility of customising LT components should be promoted through calls for Generic Services, and a bottom-up approach for identifying the LT needs of public administrations is helpful for shaping the calls for Generic Services. Valorisation aspects in the calls should receive due attention.

5.6.3. Extended business model

The extensions mentioned above are not expected to significantly impact the key resources, cost structure and revenue streams present in the MT Business Model.

The LT Business Model is shown in Figure 113. Changes with respect to the MT Business Model are shown in grey.

Figure 113 LT Business Model



Key Partners

CNECT (business owner)
DGT (MT service)
DIGIT (cloud service broker)
CEF AT expert group
NAPs of ELRC
Specialized companies (LT)
→ Generic Services

Potential partners:
JRC
SCICs
Publication Office

Key Activities

Operational MT development
Deployment of MT engines
Technology watch

Key Resources

Large IT infrastructure (cloud, …) and staff
Training data
Funding (CEF):
- Core Service Platform
- Generic Services

Value Propositions

Provide MT services:
- Asynchronous
- Real-time

Facilitate LT customization
→ Esp. under-resourced languages
Guarantee information security
Reduce customers' costs
Coordinate MT efforts
Increase cohesion between MS
Shape Digital Single Market
Provide LT components, tools
Create potential for valorization

Customer Relationships

DSIs:
- CEF AT asks for their needs
- Bottom-up: they describe issues

Public administrations: ELRC-SHARE

Channels

Architectural Management Board
Feedback through Generic Services
Memoranda of understanding
Mailing lists
Stakeholder Management Office
Creation of LT awareness by DSIs

Customer Segments

1. **DSIs**
2. **Public administrations**
3. Area of public interest

Cost Structure

Fixed but subject to revision according to demand

Revenue Streams

Under some conditions, CEF AT may charge for MT service

5.7. Conclusions

In the report on Task 4, we presented the value proposition of CEF AT, as well as potential avenues for the future. We identified the value proposition using the business model canvas methodology, which is based on nine building blocks and a set of questions. Taking into account EU policies related to multilingualism and LT, as well as the current implementation of CEF AT, the results of tasks in Lot 1 (supply side, competitiveness and demand side analysis), the findings of Task 3 in Lot 2, and discussions with CEF AT, we applied the methodology to CEF AT. This results in a description of its current business model, illustrating its position in the European LT market/ecosystem and identifying its qualitative and quantitative impacts. In the process of applying the methodology, we identified potential future avenues, which we presented as two possible future business models that extend the current model to various degrees.

The current business model of CEF AT shows that the value proposition is currently focused, on the one hand, on eTranslation, an asynchronous secured MT service which is offered to the DSIs for the multilingual deployment of their services and guarantees information security, and, on the other hand, on ELRC-SHARE, a language resource collection effort through which CEF AT publicly provides MT training data. While CEF AT has several key partners, its main partners are the MT team at DGT, which deploys the eTranslation service, and DG DIGIT, which acts as a cloud service broker. CEF AT's activities are geared towards operational development and deployment. The prime customers of CEF AT are DSIs, but it also serves public administrations and the area of public interest. CEF AT operates on a fixed budget.

We distinguish two potential future business models, which we call the MT Business Model and the LT Business Model. They extend the current business model not only on the level of MT but also LT in general, which is of paramount importance for CEF AT in order to realise its overall mission statement of becoming a truly multilingual enabler. We would like to stress that these potential models should be considered as suggestions by the consortium towards CEF AT, rather than the only possible tracks to follow.

The MT Business Model extends the scale of the MT service of the current business model, offers real-time translation, and makes CEF AT an instrument facilitating the customisation of MT engines. An increase in eTranslation demand is likely given the rising interest from DSIs, and the interest from public administrations in MT shown in Task 3 (demand side analysis). The inclusion of real-time translation follows from DSIs' interest in chat translation and from the speed expectancies shaped by the online service of global players. Facilitating customisation through Generic Services, involving specialised companies, allows eTranslation to distinguish itself from the service of global players on the level of domain adaptation, security and under-resourced languages. These are added values which match the deficiencies in the market identified in Task 2 (competitiveness analysis).

The MT Business Model has clear implications on the level of physical and financial resources, cost structure and revenue streams. Scaling up the MT service requires a larger infrastructure and staff than in the current business model. Under certain conditions, CEF AT may have to charge its customers for the service. Budgets for Generic Services should be sufficiently high in order to enable customisation of MT engines in various environments.

Calls for Generic Services should pay substantial attention to the stakeholders of DSIs, i.e. to the issues of public administrations in Member States. This implies the need for strong awareness raising of the eTranslation service among public administrations. In order to allow for valorisation of results, the calls for Generic Services should also value the business aspect of potential projects.

The LT Business Model goes beyond MT and also involves customisation of LT components in a broader sense. DSIs not only show a vivid interest in MT, but also in LT in general. The interest is not at the same level in public administrations as they do not consider LT as mature enough, thus not ready to be invested in. However, the possibility to customise LT components in collaboration with specialised companies instead of using off-the-shelf tools to create components could be a strong motivation for public administrations to start using LT tools.

6. Final conclusions

This study presented the outcome of Lot 1 of the SMART 2016/0103 project. It positioned the CEF AT building block in the European market for language technologies (LT) and described the building block’s value proposition. The building block's mission is to provide multilingual support to DSIs so that individuals, administrations and companies in all countries of the European Economic Area that participate in the CEF Telecom Work Programme can access public services in their own language.

Methodology

The study proceeded according to a four-step methodology. The first step involved an analysis of the European LT market in terms of supply and demand together with a description of the emerging trends and an estimate of the growth in the revenues. The second step consisted of a competitiveness analysis of the LT market, based on a selection of three areas (MT, speech technology and cross-lingual search) and three regions (EU, US and Asia). The third step involved the analysis of the adoption of LT by public EU-level and national administrations. The fourth and final step consisted of the description of CEF AT's current business model and potential future models through a definition of CEF AT’s value proposition.

Each of the steps corresponded to a task performed by one of the consortium members. The steps were interwoven to a certain extent, as some of them made use of the findings of previous steps. Each step applied its own specific methodology, though there are some commonalities among the different steps. The first step proceeded through desk research based on public sources and in-house IDC databases, and through primary research based on an online questionnaire and telephone interviews. The second step made use of the findings of the market analysis performed in Step 1, as well as various studies, policy papers and online information sources, and produced a SWOT analysis for the EU. The third step, like the first one, made use of a questionnaire and also applied the same LT taxonomy as the first step. The fourth step made use of the business model canvas, in which questions related to basic blocks are answered. The input consisted of information originating from the other three steps, as well as EU policies and meetings with staff of DSIs (in the framework of Smart 2016/0103 Lot 2) and CEF AT staff.

Findings

The first three steps and the other information sources of the fourth step, such as meetings with DSIs and CEF AT staff, resulted in a number of findings related to the LT market and its deficiencies, to the strengths of the EU industry, to the needs of DSIs and public services concerning MT and LT in general, and finally to the current value proposition of CEF AT.

The total size of the **LT market** within the EU26 plus Iceland and Norway is estimated at approximately 800 million euro, which is a relatively small market in IT terms. Germany holds the largest share of the market, followed by the UK. Forecasts predict this market to grow at an average rate of 10% between now and 2021. The LT market in Europe is very fragmented and composed of 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. Translation technology is considered as the biggest revenue contributor followed by speech technology. In terms of customer segments, vendors consider the public sector as the most important segment, though this sector accounts for only 20% of their revenues. Most LT suppliers expect the LT market to grow, as Artificial Intelligence will be increasingly part of LT. In that respect, Natural Language Understanding (NLU) in general and chatbot applications in particular were often mentioned as the emerging technology to look out for.

There are three major **deficiencies** in the LT market, when viewed from the angle of three of its areas (MT, speech technology and cross-lingual search). On the one hand, global US players have a large competitive advantage with respect to the fragmented EU industry because of their research capacities, computing and data resources, and market-distorting strategies, for instance the provision of a free MT service. On the other hand, the global LT offering is limited in three ways: there are gaps as regards small and complex languages, there is a lack of domain-specific and application-specific MT, and there is a lack of attention for security and privacy. The global players and the EU industry generally focus on larger languages like English, German and French. While EU industry has also built strong experience for small and complexes languages thanks to the multilingual market, these languages involve a limited market and restricted business opportunities, and the amount of accessible data for these languages is low.

The analyses carried out in the study have pinpointed the **strengths of the LT industry in the EU**. It has built strong experience for small and complexes languages, as mentioned above. It also has a strong track record concerning pan-European projects involving research organisations and concerning the deployment of services for the public sector through the support of EU-funded programmes. Furthermore, many European developers have accumulated strong experience in customised and domain-specific solution development in the areas of MT and speech, due to their need to search for niche markets. As for privacy requirements, the EU has well-established practices for the creation of open data and policies fostering public data sharing.

Further findings of the study relate to **needs of DSIs and public services** concerning MT and LT in general. DSIs show a rising interest for MT, some of them having a need for real-time translation, similarly to the one provided by the global US players. DSIs also show a vivid interest for LT in general. As for public services, MT is clearly the type of LT most frequently used by them. They also have a strong interest in related tools, like translation memories and terminology management systems. They strongly collaborate with academia, which points towards a need for customisation and tuning of technologies. Many public services are optimistic about their future needs for 2020 and beyond to deploy other LT types than MT, but it seems to them that these technologies are not considered mature enough today to be used in their working environment. As for small and complex languages, public services should also take into account the EU's commitment to preserve cultural identity, foster inclusiveness (guarantee equal digital opportunities across languages) and shape the Digital Single Market.

Final findings involve the **current value proposition of CEF AT** and the business model through which it operates. CEF AT is currently focused, on the one hand, on eTranslation, an asynchronous secured MT service which is offered to the DSIs for the multilingual deployment of their services and guarantees information security, and, on the other hand, on ELRC-SHARE, a language resource collection effort through which CEF AT publicly provides MT training data. Its prime customers are DSIs, but it also serves public administrations and the area of public interest. It operates on a fixed budget.

Opportunities

Based on the findings, a number of opportunities related to LT in the EU can be identified. The strong experience of European LT companies concerning small and complex languages as well as customised and domain-specific MT meets two of the major market deficiencies mentioned earlier. This experience is very helpful to meet the needs of public services. Moreover, EU LT companies have a strong track record in collaborating with the public sector and with academia. The EU is well positioned to meet security and privacy requirements (the third market deficiency): it has ample experience with practices for the creation of open data and policies fostering public data sharing and provides a strong level of security through CEF AT's eTranslation service.

Meeting the three market deficiencies provides a great potential to CEF AT and the European LT suppliers to distinguish their offering from the global US players. Therefore, the consortium suggests two extensions the current business model, the *MT Business Model* and the *LT Business Model*.

The MT Business Model extends the scale of the MT service, provides real-time translation, and makes CEF AT an instrument facilitating the customisation of MT, especially for under-resourced languages. It is motivated by the increasing interest in eTranslation of DSIs and public administrations, on DSIs' interest in chat translation, and on the speed expectancies shaped by the online service of global MT players. Customisation takes place through Generic Services projects involving specialised companies in order to avoid market distortion and focuses on valorisation (business aspects) and reusability of results. The MT Business Model has clear implications on the level of physical and financial resources, cost structure and revenue streams.

The LT Business Model extends the MT Business Model, by widening the scope from MT to LT: it involves collecting and providing new LT resources and tools (in addition the current ELRC-SHARE resources), offering LT services, and facilitating the customisation of LT resources, tools and services, especially for under-resourced languages. It is motivated by the mission of CEF AT, which is oriented towards multilingual support, and by the fact that DSIs not only show a vivid interest in MT, but also in LT in general. The interest is not at the same level in public administrations, as they do not consider LT as mature enough, thus not ready to be invested in. However, the possibility to customise LT components in collaboration with specialised companies instead of using off-the-shelf tools to create such components could be a strong motivation for public administrations to start using LT tools. In that respect, the supplier database containing 473 LT companies created in Task 1 of the study can be a valuable asset for public administrations.

As a final note, the consortium would like to stress that the extended business models should be considered as suggestions rather than the only possible tracks to follow. Many variants of the proposed business models may be conceived. In no way, CEF AT can be held liable or accountable for any of the ideas expressed in the sections of the study related to Task 4. Moreover, the LT market and the technologies themselves are progressing fast, which makes it difficult to predict future changes of the business model.

Sources and references

Baller, S., Dutta, S., and Lanvin, B. (2016). *The Global Information Technology Report 2016. Innovating in the Digital Economy.* Geneva: World Economic Forum, Cornell University, INSEAD.

Barrachina, S., Bender, O., Casacuberta, F., Civera, J., Cubel, E., Khadivi, S., Lagarda, A.L., Ney, H., Tomás, J., Vidal, E., Vilar, J.M., et al. (2009). Statistical Approaches to Computer-Assisted Translation. Computational Linguistics. 35 (1): 3–28.

Bojar, O., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Koehn, P., Monz, C. (2018) Findings of the 2018 Conference on Machine Translation (WMT18). WMT (shared task) 2018: 272-303.

Common Sense Advisory (2017). The Top 100 LSPs in 2017. Extract from "Who's is Who in Language Services and Technology: 2017". Cambridge, Massachusetts: Common Sense Advisory.

CRACKER and LT-Observatory (2015). *Strategic Agenda for the Multilingual Digital Single Market: Technologies for Overcoming Language Barriers towards a Truly Integrated European Online Market.*

Crego, J., Kim, J., Klein, G., Rebollo, A., Yang, K., Senellart, J., et al. (2016). SYSTRAN's Pure Neural Machine Translation Systems. arXiv preprint arXiv:1610.05540

European Parliament resolution of 11 September 2018 on language equality in the digital age (2018/2028(INI)).

Gehring, J., Auli, M., Grangier, D., and Dauphin, Y.N. (2017a). A Convolutional Encoder Model for Neural Machine Translation. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 123-135.

Gehring, J., Auli, M., Grangier, D., Yarats, D., and Dauphin, Y.N. (2017b). Convolutional Sequence to Sequence Learning. arXiv preprint arXiv: 1705.03122.

Giles, M. (2018). It's Time to Rein in the Data Barons. MIT Technology Review, June 19, 2018.

Hassan, H., Aue, A., Chen, C., Chowdhary, V., Clark, J., Federmann, C., et al. (2018). Achieving Human Parity on Automatic Chinese to English News Translation. arXiv preprint arXiv:1803.05567.

Hieber, F., Domhan, T., Denkowski, M., Vilar, D., Sokolov, A., Clifton, A., and Post, M. (2017). Sockeye: A Toolkit for Neural Machine Translation. In eprint arXiv:cs-CL/1712.05690.

Hugenholtz, P.B. (2013). Fair Use in Europe. Communications of the ACM, 56(5), 26-28.

Joscelyne, A., and Lockwood, R. (2003). *The EUROMAP Study. Benchmarking HLT Progress in Europe.* Copenhagen: EUROMAP Language Technologies, Center for Sprogteknologi.

Joscelyne, A. (2017). *TAUS Machine Translation Market Report.* TAUS.

Junczys-Dowmunt, M., Grundkiewicz, R., Dwojak, T., Hoang, H., Heafield, K., Neckermann, T., Seide, F., Germann, U., Fikri Aji, A., Bogoychev, N., Martins, A., Birch, A. M. (2018). Fast Neural Machine Translation in C++. Proceedings of ACL 2018, System Demonstrations, 116-121.

Klein, G., Kim, J., Deng, Y., Senellart, J., Rush, A. (2017). OpenNMT: Open-Source Toolkit for Neural Machine Translation. Proceedings of ACL 2017, System Demonstrations, 67-72.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., Herbst, E. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session, Prague, Czech Republic, June 2007.

Lehmann, J., Stodulka, T., and Huber, E. (2018). *H2020 Project K-PLEX: WP4 Report on Data, Knowledge Organisation and Epistemics*. Berlin: Freie Universtät Berlin.

Lommel, A.R., and DePalma, D.A. (2016). *Europe's Leading Role in Machine Translation: How Europe is Driving the Shift to MT.* Produced for the CRACKER project. Cambridge, Massachusetts: Common Sense Advisory.

LT-Innovate and META-NET (2018). *Empowering a Multilingual Continent. Technologies and Language-Centric AI for Language Equality in Europe.*

Massardo, I., van der Meer, J., and Khalilov, M. (2016). *TAUS Translation Technology Report*. TAUS.

META-NET (2013). *Strategic Research Agenda for Multilingual Europe 2020.* Heidelberg: Springer.

META-NET (2015). *Strategic Research Agenda for the Multilingual Digital Single Market.* http://www.meta-net.eu/projects/cracker/multimedia/mdsm-sria-draft.pdf

Miao, Y., Gowayyed, M., and Metze, F. (2015). EESEN: End-to-End Speech Recognition using Deep RNN Models and WFST-based Decoding. Proceedings of Automatic Speech Recognition and Understanding Workshop (ASRU), Scottsdale, AZ, U.S.A., December 2015. IEEE.

OECD and Statistical Office of the European Communities (2005). *Oslo Manual – Guidelines for Collecting and Interpreting Innovation Data – Third edition.* OECD Publishing.

Osterwalder, A., and Pigneur, Y. (2010). *Business Model Generation*. John Wiley & Sons.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., and Vesely, K. (2011). The Kaldi Speech Recognition Toolkit. Proceedings of IEEE 2011 Workshop on Automatic Speech Recognition and Understanding.

Ruopp, A., and van der Meer, J. (2015). *Moses MT Market Report.* TAUS.

Schwab, K. (2017). *The Global Competitiveness Report 2017-2018*. Geneva: World Economic Forum.

Science and Technology Options Assessment (STOA). European Parliament (2017). *Language Equality in the Digital Age: Towards a Human Language Project*. Brussels: STOA.

Seligman, M., Waibel, A., and Joscelyne, A. (2017). *TAUS Speech-to-Speech Translation Technology Report*. TAUS.

Sennrich, R., Firat, O., Cho, K., Birch, A., Haddow, B., Hitschler, J., Junczys-Dowmunt, M., Läubli, S., Barone, A., Mokry, J., and Nadejde, M. (2017). Nematus: a Toolkit for Neural Machine Translation. In Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics, Valencia, Spain, 65-68.

Slator (2018). *Slator 2018 Neural Machine Translation Report.* Switzerland: Slator.

Strategic Research and Innovation Agenda (2017). *Language Technologies for Multilingual Europe: Towards a Human Language Project*.http://cracker-project.eu/wp-content/uploads/SRIA-V1.0-final.pdf

Toral, A., Castilho, S., Hu, K., and Way, A. (2018). Attaining the Unattainable? Reassessing Claims of Human Parity in Neural Machine Translation. *arXiv preprint arXiv:1808.10432*.

Van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A.W., and Kavukcuoglu, K. (2016). WaveNet: A Generative Model for Raw Audio. CoRR, abs/1609.03499. http://dblp.uni-trier.de/db/journals/corr/corr1609.html#OordDZSVGKSK16

Vasiļjevs, A., Kalniņš, R., Pinnis, M., and Skadiņš, R. (2014). Machine Translation for e-Government - the Baltic Case. Proceedings of AMTA 2014, vol. 2: MT Users, 181-193.

Von Lohmann, F. (2017). Fair Use as Innovation Policy. In Copyright Law, Routledge, 169-205.

Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., et al. (2016). Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. arXiv preprint arXiv:1609.08144.

Xiong, W., Droppo, J., Huang, X., Seide, F., Seltzer, M., Stolcke, A., Yu, D., and Zweig, G. (2016) Achieving Human Parity in Conversational Speech Recognition. CoRR, abs/1610.05256. http://dblp.uni-trier.de/db/journals/corr/corr1610.html#XiongDHSSSYZ16a

Wu, Z., Watts, O., and King, S. (2017). Merlin: An Open Source Neural Network Speech Synthesis System. Proceedings of 9th ISCA Speech Synthesis Workshop (SSW9), September 2016, Sunnyvale, CA, USA.

Zia, M. (2018). *Early Access | Watson Neural Machine Translation*. https://developer.ibm.com/dwblog/2018/watson-language-translator-neural-machine-translation

A. Task 1: details of primary research

Selection approach to identify and qualify the supply-side target

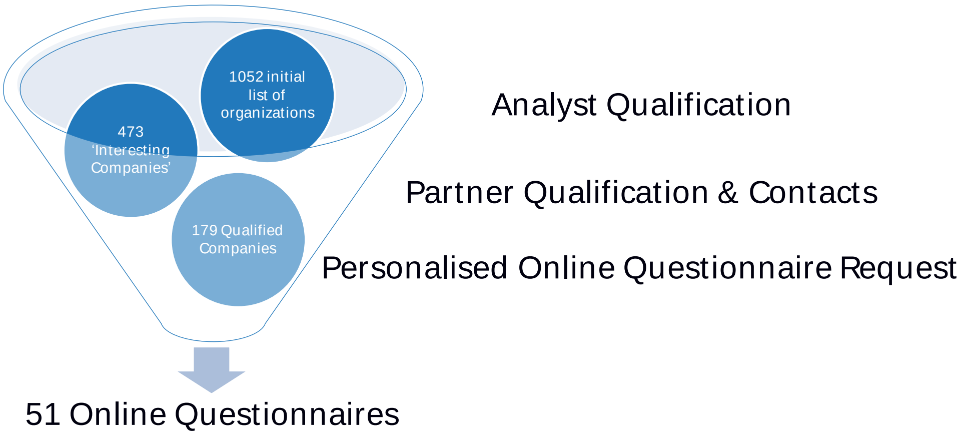
In order to carry out the extensive primary research targeting the supply-side of the language technology market, the study team applied a rigorous research to identify and qualify the relevant target and validate a list of language technology vendors active across Europe.

The first step was to narrow down a list of 1052 of organisations active in the language technology domain, by excluding entities outside the commercial area (such as research organisations and academia stakeholders) as these do not represent the primary target of the market survey. This reduced the number to 473 language technology vendors.

This list of 473 companies was further qualified to validate each individual vendors' direct presence as a language technology vendor, rather than as a language service provider (LSP), language technology services provider, system integrator, translators, consultants, or language technology reseller, which are excluded from the scope of this study. The final output of this selection process produced a short list of 179 technology vendors.

Figure 114 shows the approach to shortlist and qualify the initial group of organisations.

Figure 114 Selection approach for the supply-side survey


1052 initial list of organizations
473 'Interesting Companies'
179 Qualified Companies
Analyst Qualification
Partner Qualification & Contacts
Personalised Online Questionnaire Request
51 Online Questionnaires

Supply-side Online Questionnaire Script

1. Company profile Data

   (pre-filled)

   - Organisation Name:
   - Year Established:
   - Name of Respondent:
   - Job Title:
   - Email:
   - Tel:
   - City / Town location of headquarters:
   - No. of other offices:
   - Other offices location:
   - No. of employees (FTEs):

2. Which type of language technologies products/services does your company offer?

   (multiple answers possible)

   - **Speech technology** (e.g. software for recognising, identifying, and extracting information from audio, voice, and speech data as well as speech identification and recognition plus converting sounds into useful text)
   - **Translation technology** (e.g. automated language translation tools)
   - **Natural language understanding**
   - **Analytics** (e.g. text mining; recognising, understanding, and extracting value from text or by using similar technologies to generate human readable text; language analysers, text clustering and categorisation tools; search applications)
   - **Multilingual and semantic search technology**
   - **Other (please specify)**

3. What were your **total** revenues in your last financial year?

   (Euros) (banded)

3.1 Could you make a rough estimate of how your revenue breaks into the 5 below areas?

   - **Speech technologies** (%)
   - **Translation technologies** (%)
   - **Natural language understanding technologies** (%)
   - **Analytics** (%)
   - **Multilingual and semantic search technology** (%)
   - **Other (%)**

   Programming note: Sum must be 100%

4. Please provide us the % revenue growth your company experienced over the previous fiscal year (banded)

4.1 Please provide the expected growth rate of your revenues for the next three years (up to 2021) (banded)

2019 --%--

2020 --%--

2021 --%--

5. What is your rough mix of your annual revenues? (banded)

% Revenues related to language technology **products**

% Revenues related to language technology **services**

% Revenues related to other non-language technology areas

Don’t know

6. Could you please tell us the profitability (%) of your company? (banded)

7. Which of the following languages do you offer **in production** in your language technology offering?
(translated user interfaces are not considered to be language technology)

- Bulgarian
- Croatian
- Czech
- Danish
- Dutch
- English
- Estonian
- Finnish
- French
- German
- Greek
- Hungarian
- Icelandic
- Irish
- Italian
- Latvian
- Lithuanian
- Maltese

- Norwegian
- Polish
- Portuguese
- Romanian
- Slovak
- Slovenian
- Spanish
- Swedish
- Other (please specify)

8. <u>Which of the following industry sectors do you serve?</u>
(select all that apply)
   - Banking
   - Construction
   - Discrete Manufacturing (e.g. Automotive, aerospace, industrial machinery, High-tech and Electronics)
   - Education
   - Government (including Central and local)
   - Healthcare Provider
   - Insurance
   - Media
   - Personal and Consumer Services (e.g. gambling and betting, sports activities and amusement and recreation activities, etc.)
   - Process Manufacturing (e.g. chemicals, pulp and paper, rubber and plastics, food/beverage/tobacco, pharma)
   - Professional Services (e.g. engineering, legal, accounting, real estate, staffing, IT software developers.) excluding LSPs
   - LSP's (Language Service Providers)
   - Resource Industries
   - Retail/Wholesale
   - Securities and Investment Services
   - Telecommunications
   - Transportation
   - Utilities

8.1 <u>What percentage of your revenues are sourced from (i) public sector bodies (ii) private customers?</u>
(banded)

8.2 <u>What percentage of your revenues are sourced from (i) SMEs or (ii) big companies?</u>
(banded)
(*Small and Medium-sized Enterprise definition: enterprise with < 250 employees and a turnover of ≤ €50M, or a balance sheet total ≤ €43M*)

9. <u>What are the key application areas in which the technologies you provide are being currently used by your customers?</u>

(choose top 3)

- Marketing content services
- Technical documentation
- Traditional media publishing
- Online media publishing
- Web site construction / development
- Social media
- Other (specify)

10. Which types of applications / services do you offer?

    (multiple answers possible)

    - Speech recogniser (speech-to-text)
    - Speech synthesiser (text-to-speech)
    - Speech translation
    - Direct Speech Translation
    - Machine Translation
    - CAT tools (Translation Memories, TMS, ...)
    - Alignment tool (e.g. sentence aligner)
    - Localisation tool
    - Website
    - Subtitling production
    - Dubbing
    - Software
    - Games
    - Authoring tool (e.g. technical writing, controlled language)
    - Terminology Management Systems
    - Term candidate extractor
    - Chatbot (virtual assistant)
    - Keyword extractor
    - Topic modelling tool
    - Text mining tool (e.g. mine financial information in business data)
    - Tools for sentiment analysis (social listening, opinion mining, …)
    - Text prediction tool (e.g. language model, autocompletion, interactive prediction, …)
    - Authorship attribution tool
    - Optical character recognition
    - Question-Answering system
    - Search engine
    - Workflow Management (e.g. translation workflow)
    - Other (please specify)

11. What is your language technologies delivery model?

    - On-premises
    - Cloud instance
    - Both

12. What is your language technologies licensing model?
    - Perpetual license
    - Annual license
    - Software-as-a-Service (SaaS)
    - Other (please specify)

13. External funding (e.g. venture capital funded)?

    YES / NO

    (Start-ups only: *A start-up is a growth-oriented small enterprise, up to 3 years old, searching for a scalable business model or innovative product/service, and open for alternative financing*) *Condition*: *only if answer to Q1 Year established is 2015 or later*

14. Is your company a participant in a specialised language technology innovation lab or digital hub?

    YES / NO

    (Start-ups only: *A start-up is a growth-oriented small enterprise, up to 3 years old, searching for a scalable business model or innovative product/service, and open for alternative financing*) – *Condition: only if answer to Q1 Year established is 2015 or later*

    IF 14 = YES

    Please provide the name of innovation lab or digital hub: (Start-ups only)

15. What percentage of your revenues are generated?
    (for everybody)
    - From inside EU
    - From outside EU

16. Of revenues sourced from inside EU, what percentage is sourced from the following countries?
    - Austria
    - Belgium
    - Bulgaria
    - Croatia
    - Cyprus
    - Czech Republic
    - Denmark
    - Estonia
    - Finland
    - France
    - Germany
    - Greece
    - Hungary
    - Iceland
    - Ireland
    - Italy
    - Latvia
    - Lithuania
    - Luxembourg
    - Malta
    - Netherlands
    - Norway
    - Poland
    - Portugal
    - Romania
    - Slovakia
    - Slovenia
    - Spain
    - Sweden
    - UK

17. What is your future level of interest in these new high growth language technology areas? (scale 1-5 where 1 is "not at all" and 5 is "very high")
    - Machine Translation
    - Speech Translation
    - Automatic Summarisation
    - Search engine
    - Cross-Lingual search
    - Question-Answering
    - Robotic process automation
    - Social listening / sentiment analysis
    - Text analytics
    - Chat Bots
    - Natural language processing

- Speech recognition
- Text to speech
- Speech to Text
- Predictive Text
- Workflow Management

18. To what degree do you collaborate with academic and research institutions?
(scale 1-5 where 1 is "not at all" and 5 is "very high")

19. Can you specify which?
(optional)

20. What are for your customers the most important parameters in service delivery?
Rate the needs of your clients (1-5 scale where 1 not relevant and 5 is very relevant)
    - Data security
    - Safeguarding Intellectual Property
    - Provisioning regional variants (localisation)
    - High-volume (size or number) throughput
    - Provisioning many languages
    - Speed of delivery
    - 100% accuracy

21. To what degree do your clients require industry-specific expertise in your language technology services (i.e. how familiar do your customers expect you to be with their business?)
(scale 1-5 where 1 is "not at all" and 5 is "very high")

22. Do you sell data externally (e.g. to agencies and external customers)?

    YES/NO

    If YES, what type of data?

    - Modality (spoken, written, …)
    - Size
    - Period covered
    - Accessibility (local, cloud, …)

23. Would you like to be contacted by our analysts for a more in-depth interview?
    - Yes (OPTIONAL please leave your contact information)
    - No

Supply-side In-depth Interview Guidelines

- What is your view of the changes in the market over the past 2 years? (telephone interviews only)
- How would you describe the current state of play of the market? (telephone interviews only)
- Are larger vendors taking a more dominating role in this market? (telephone interviews only)
- Who are your top 3 competitors? (telephone interviews only)
- How do you see the market developing over the next 5 years? (telephone interviews only)

More of . . .

Less of . . .

- Will you be incorporating more software into your services portfolio? (telephone interviews only)
- How do you feel your clients will respond to you using more software in your services? (telephone interviews only)
- In what year do you think we will reach 100% accuracy in language translation? And why? (telephone interviews only)

B. Task 2: list of countries by region

	REGION	**GDP in USD (IMF 2017)**
	EUROPE	
	EU Total	17 309 000
1	Austria	416 845
2	Belgium	494 733
3	Bulgaria	56 943
4	Croatia	54 516
5	Cyprus	21 310
6	Czech Republic	213 189
7	Denmark	324 484
8	Estonia	25 973
9	Finland	253 244
10	France	2 584 000
11	Germany	3 685 000
12	Greece	200 690
13	Hungary	152 284
14	Ireland	333 994
15	Italy	1 938 000
16	Latvia	30 319
17	Lithuania	47 263
18	Luxembourg	62 393
19	Malta	12 011
20	Netherlands	825 745
21	Poland	524 886
22	Portugal	218 064
23	Romania	211 315
24	Slovakia	95 938
25	Slovenia	48 868
26	Spain	1 314 000
27	Sweden	538 575

28	United Kingdom	2 625 000
29	Norway	396 457
30	Iceland	23 909
31	Switzerland	678 575

	NORTH AMERICA	
1	United States of America	19 390 000
2	Canada	1 652 000

	ASIA	
1	China	12 015 000
2	Japan	4 872 000
3	India	2 611 000
4	South Korea	1 538 000
5	Singapore	323 900

C. Task 2: most cited Scopus publications – MT

Publications written in the period 2010-2018

1. Vinyals, O., Toshev, A., Bengio, S., Erhan, D. Show and tell: A neural image caption generator. (2015) Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 07-12-June-2015, art. no. 7298935, pp. 3156-3164. **Cited 627 times.** AFFILIATIONS: Google, United States.
2. Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y. Learning phrase representations using RNN encoder-decoder for statistical machine translation. (2014) EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, pp. 1724-1734. **Cited 598 times.** AFFILIATIONS: Université de Montréal, Canada; Jacobs University, Germany; Université du Maine, France.
3. Xu, K., Ba, J.L., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R.S., Bengio, Y. Show, attend and tell: Neural image caption generation with visual attention (2015) 32nd International Conference on Machine Learning, ICML 2015, 3, pp. 2048-2057. **Cited 544 times.** AFFILIATIONS: Université de Montréal, Canada; University of Toronto, Canada; CIFAR, Canada.
4. Navigli, R., Ponzetto, S.P. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. (2012) Artificial Intelligence, 193, pp. 217-250. **Cited 418 times.** AFFILIATIONS: Dipartimento di Informatica, Sapienza University of Rome, Italy
5. Bohnet, B. Very high accuracy and fast dependency parsing is not a contradiction. (2010) Coling 2010 - 23rd International Conference on Computational Linguistics, Proceedings of the Conference, 2, pp. 89-97. **Cited 217 times.** AFFILIATIONS: University of Stuttgart, Institut Für Maschinelle, Sprachverarbeitung, Germany.
6. Luong, M.-T., Pham, H., Manning, C.D. Effective approaches to attention-based neural machine translation. (2015) Conference Proceedings - EMNLP 2015: Conference on Empirical Methods in Natural Language Processing, pp. 1412-1421. **Cited 215 times.** AFFILIATIONS: Computer Science Department, Stanford University, Stanford, CA 94305, United States
7. Navigli, R., Ponzetto, S.P. BabelNet: Building a very large multilingual semantic network. (2010) ACL 2010 - 48th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, pp. 216-225. **Cited 214 times.** AFFILIATIONS: Dipartimento di Informatica, Sapienza Università di Roma, Italy; Department of Computational Linguistics, Heidelberg University, Germany
8. Meisnery, D., Sadlerz, C.M., Barrosoz, L.A., Weberz, W.-D., Wenischy, T.F. Power management of Online Data-Intensive services. (2011) Proceedings - International Symposium on Computer Architecture, pp. 319-330. **Cited 209 times.** AFFILIATIONS: University of Michigan, United States; Google, Inc., United States
9. Androutsopoulos, I., Malakasiotis, P. A survey of paraphrasing and textual entailment methods. (2010) Journal of Artificial Intelligence Research, 38, pp. 135-187. **Cited 181 times.** AFFILIATIONS: Department of Informatics, Athens University of Economics and Business, Patission 76, GR-104 34 Athens, Greece.
10. Agirre, E., Cer, D., Diab, M., Gonzalez-Agirre, A. SemEval-2012 Task 6: A pilot on semantic textual similarity. (2012) *SEM 2012 - 1st Joint Conference on Lexical and Computational Semantics, 2, pp. 385-393. **Cited 176 times.** AFFILIATIONS: University of the Basque Country,

Donostia Basque Country, 20018, Spain; Stanford University, Stanford, CA 94305, United States; Center for Computational Learning Systems, Columbia University, United States; University of the Basque Country, Donostia, Basque Country, 20018, Spain

Publications written in the period 2015-2018

1. Vinyals, O., Toshev, A., Bengio, S., Erhan, D. Show and tell: A neural image caption generator. (2015) Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 3156-3164. **Cited 627 times.** AFFILIATIONS: Google, United States.
2. Xu, K., Ba, J.L., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R.S., Bengio, Y. Show, attend and tell: Neural image caption generation with visual attention. (2015) 32nd International Conference on Machine Learning, ICML 2015, 3, pp. 2048-2057. **Cited 544 times.** AFFILIATIONS: Université de Montréal, Canada; University of Toronto, Canada; CIFAR, Canada.
3. Luong, M.-T., Pham, H., Manning, C.D. Effective approaches to attention-based neural machine translation. (2015) Conference Proceedings - EMNLP 2015: Conference on Empirical Methods in Natural Language Processing, pp. 1412-1421. **Cited 215 times.** AFFILIATIONS: Computer Science Department, Stanford University, Stanford, CA 94305, United States.
4. Chorowski, J., Bahdanau, D., Serdyuk, D., Cho, K., Bengio, Y. Attention-based models for speech recognition. (2015) Advances in Neural Information Processing Systems, 2015-January, pp. 577-585. **Cited 154 times.** AFFILIATIONS: University of Wrocław, Poland, United States; Jacobs University Bremen, Germany; Université de Montréal, Canada; Université de Montréal, CIFAR Senior Fellow, Canada.
5. Sordoni, A., Galley, M., Auli, M., Brockett, C., Ji, Y., Mitchell, M., Nie, J.-Y., Gao, J., Dolan, B. A neural network approach to context-sensitive generation of conversational responses. (2015) NAACL HLT 2015 - 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, pp. 196-205. **Cited 98 times**. AFFILIATIONS: DIRO, Université de Montréal, Montréal, QC, Canada; Microsoft Research, Redmond, WA, United States; Facebook AI Research, Menlo Park, CA, United States; Georgia Institute of Technology, Atlanta, GA, United States.
6. Jean, S., Cho, K., Memisevic, R., Bengio, Y. On using very large target vocabulary for neural machine translation. (2015) ACL-IJCNLP 2015 - 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, Proceedings of the Conference, 1, pp. 1-10. **Cited 93 times.** AFFILIATIONS: Universite de Montreal, Canada; Universite de Montreal, CIFAR Senior Fellow, Canada.
7. Sennrich, R., Haddow, B., Birch, A. Neural machine translation of rare words with subword units. (2016) 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Papers, 3, pp. 1715-1725. **Cited 84 times.** AFFILIATIONS: School of Informatics, University of Edinburgh, United Kingdom.
8. Bengio, S., Vinyals, O., Jaitly, N., Shazeer, N. Scheduled sampling for sequence prediction with recurrent neural networks. (2015) Advances in Neural Information Processing Systems, 2015-January, pp. 1171-1179. **Cited 76 times.** AFFILIATIONS: Google Research, Mountain View, CA, United States.
9. Luong, M.-T., Sutskever, I., Le, Q.V., Vinyals, O., Zaremba, W. Addressing the rare word problem in neural machine translation. (2015) ACL-IJCNLP 2015 - 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, Proceedings of the Conference, 1, pp. 11-19. **Cited 76 times.** AFFILIATIONS: Stanford, United States; Google, United States; New York University, United States.

10. Cho, K., Courville, A., Bengio, Y. Describing Multimedia Content Using Attention-Based Encoder-Decoder Networks. (2015) IEEE Transactions on Multimedia, 17 (11), pp. 1875-1886. **Cited 56 times.** AFFILIATIONS: Information and Operational Research Department, Université of Montréal, Montréal, QC H3T 1J4, Canada; Department of Computer Science, New York University, New York, NY 10012, United States.

D. Task 2: most cited Scopus publications – speech technology

Publications written in the period 2010-2018

1. Lecun, Y., Bengio, Y., Hinton, G. Deep learning (2015) Nature, 521 (7553), pp. 436-444. **Cited 5723 times.** AFFILIATIONS: Facebook AI Research, United States; New York University, United States; Department of Computer Science, Operations Research Université de Montréal, Canada; Google, United States; Department of Computer Science, University of Toronto, Canada.
2. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.Dropout: A simple way to prevent neural networks from overfitting. (2014) Journal of Machine Learning Research, 15, pp. 1929-1958. **Cited 3728 times.** AFFILIATIONS: Department of Computer Science, University of Toronto, Canada.
3. Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A.-R., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T., Kingsbury, B. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups (2012) IEEE Signal Processing Magazine, 29 (6), pp. 82-97. **Cited 2754 times.** AFFILIATIONS: Computer Science, Univ. Toronto, Toronto, Canada; Microsoft Research, United States; Google, United States; IBM T. J. Watson Research Center, United States.
4. Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P. Natural language processing (almost) from scratch (2011) Journal of Machine Learning Research, 12, pp. 2493-2537. **Cited 1762 times.** AFFILIATIONS: NEC Laboratories America, United States; Idiap Research Institute, Switzerland; Google, United States; Microsoft, United States; New York University, United States; Rutgers University, United States.
5. Dehak, N., Kenny, P.J., Dehak, R., Dumouchel, P., Ouellet, P. Front-end factor analysis for speaker verification (2011) IEEE Transactions on Audio, Speech and Language Processing, 19 (4), art. no. 5545402, pp. 788-798. **Cited 1455 times.** AFFILIATIONS: Spoken Language Systems Group, Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, United States; Centre de Recherche Informatique de Montréal (CRIM), Canada; Laboratoire de Recherche et de Développement de l'EPITA, France; École de Technologie Supérieure (ÉTS), Montreal, Canada.
6. Graves, A., Mohamed, A.-R., Hinton, G. Speech recognition with deep recurrent neural networks (2013) ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, art. no. 6638947, pp. 6645-6649. **Cited 1339 times.** AFFILIATIONS: Department of Computer Science, University of Toronto, Canada.
7. Dahl, G.E., Yu, D., Deng, L., Acero, A. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition (2012) IEEE Transactions on Audio, Speech and Language Processing, 20 (1), art. no. 5740583, pp. 30-42. **Cited 1300 times.** AFFILIATIONS: Department of Computer Science, University of Toronto, Canada; Speech Research Group, Microsoft Research, Redmond, United States.
8. Mikolov, T., Karafiát, M., Burget, L., Jan, C., Khudanpur, S. Recurrent neural network based language model (2010) Proceedings of the 11th Annual Conference of the International Speech Communication Association, INTERSPEECH 2010, pp. 1045-1048. **Cited 998 times.** AFFILIATIONS: Speech at FIT, Brno University of Technology, Czech Republic; Department of Electrical and Computer Engineering, Johns Hopkins University, United States.
9. Mohamed, A.-R., Dahl, G.E., Hinton, G. Acoustic modeling using deep belief networks (2012) IEEE Transactions on Audio, Speech and Language Processing, 20 (1), art. no. 5704567, pp. 14-22. **Cited 797 times.** AFFILIATIONS: University of Toronto, Canada.

10. Eyben, F., Wöllmer, M., Schuller, B. OpenSMILE - The Munich versatile and fast open-source audio feature extractor. (2010) MM'10 - Proceedings of the ACM Multimedia 2010 International Conference, pp. 1459-1462. **Cited 730 times.** AFFILIATIONS: Institute for Human-Machine Communication, Technische Universität München, Germany.

Publications written in the period 2015-2018

1. Lecun, Y., Bengio, Y., Hinton, G. Deep learning (2015) Nature, 521 (7553), pp. 436-444. **Cited 5723 times.** AFFILIATIONS: Facebook AI Research, United States; New York University, United States; Department of Computer Science, Operations Research Université de Montréal, Canada; Google, United States; Department of Computer Science, University of Toronto, Canada.
2. Sainath, T.N., Vinyals, O., Senior, A., Sak, H. Convolutional, Long Short-Term Memory, fully connected Deep Neural Networks (2015) ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 2015-August, art. no. 7178838, pp. 4580-4584. **Cited 205 times**. AFFILIATIONS: Google, Inc., United States.
   Publication excluded as unreliable. ITAKURA F, SAITO S. ANALYSIS SYNTHESIS TELEPHONY BASED ON MAXIMUM LIKELIHOOD METHOD. (2017) 2, pp. 17-20. **Cited 193 times.**
3. Chorowski, J., Bahdanau, D., Serdyuk, D., Cho, K., Bengio, Y. Attention-based models for speech recognition (2015) Advances in Neural Information Processing Systems, 2015-January, pp. 577-585. **Cited 182 times.** AFFILIATIONS: University of Wrocław, Poland; Jacobs University Bremen, Germany; Université de Montréal, Canada.
4. Barker, J., Marxer, R., Vincent, E., Watanabe, S. The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines (2016) 2015 IEEE Workshop on Automatic Speech. Recognition and Understanding, ASRU 2015 - Proceedings, art. no. 7404837, pp. 504-511. **Cited 180 times.** AFFILIATIONS: University of Sheffield, United Kingdom; Inria, France; MERL, United States.
5. Severyn, A., Moschittiy, A. Learning to rank short text pairs with convolutional deep neural networks (2015) SIGIR 2015 - Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 373-382. **Cited 159 times.** AFFILIATIONS: Google Inc., Qatar; Qatar Computing Research Institute, Qatar; University of Trento, DISI, Italy.
6. Liu, W., Wang, Z., Liu, X., Zeng, N., Liu, Y., Alsaadi, F.E. A survey of deep neural network architectures and their applications (2017) Neurocomputing, 234, pp. 11-26. **Cited 148 times.** AFFILIATIONS: Department of Computer Science, Brunel University London, United Kingdom; Department of Instrumental and Electrical Engineering, Xiamen University, Xiamen, China; Department of Mathematics, Yangzhou University, China; Communication Systems and Networks (CSN) Research Group, Faculty of Engineering, King Abdulaziz University, Saudi Arabia.
7. Xiao, T., Li, H., Ouyang, W., Wang, X. Learning deep feature representations with Domain Guided Dropout for person re-identification (2016) Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2016-January, pp. 1249-1258**. Cited 143 times.** AFFILIATIONS: Department of Electronic Engineering, Chinese University of Hong Kong, Hong Kong.
8. Miao, Y., Gowayyed, M., Metze, F. EESEN: End-to-end speech recognition using deep RNN models and WFST-based decoding (2016) 2015 IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2015 - Proceedings, art. no. 7404790, pp. 167-174. **Cited 135 times.** AFFILIATIONS: Language Technologies Institute, School of Computer Science, Carnegie Mellon University, United States.
9. Sainath, T.N., Kingsbury, B., Saon, G., Soltau, H., Mohamed, A.-R., Dahl, G., Ramabhadran, B. Deep Convolutional Neural Networks for Large-scale Speech Tasks (2015) Neural Networks, 64,

pp. 39-48. **Cited 133 times.** AFFILIATIONS: IBM T. J. Watson Research Center, United States; Department of Computer Science, University of Toronto, Canada.

10. Wu, Z., Evans, N., Kinnunen, T., Yamagishi, J., Alegre, F., Li, H. Spoofing and countermeasures for speaker verification: A survey (2015) Speech Communication, 66, pp. 130-153**. Cited 131 times.** AFFILIATIONS: Nanyang Technological University, Singapore, Singapore; EURECOM, France; University of Eastern Finland, Finland; National Institute of Informatics, Japan; University of Edinburgh, United Kingdom; Institute for Infocomm Research, Singapore, Singapore.

E. Task 2: most cited Scopus publications – IR from text

Publications written in the period 2010-2018

1. Uijlings, J.R.R., Van De Sande, K.E.A., Gevers, T., Smeulders, A.W.M. Selective search for object recognition (2013). International Journal of Computer Vision, 104 (2), pp. 154-171. Cited **1488 times.** AFFILIATIONS: University of Trento, Trento, Italy; University of Amsterdam, Amsterdam, Netherlands
2. Dai, X., Zhao, P.X. PsRNATarget: A plant small RNA target analysis server (2011) Nucleic Acids Research, 39 (SUPPL. 2), pp. W155-W159. **Cited 837 times.** AFFILIATIONS: Plant Biology Division, Samuel Roberts Noble Foundation, 2510 Sam Noble Parkway, Ardmore, OK 73401, United States
3. Bakshy, E., Mason, W.A., Hofman, J.M., Watts, D.J. Everyone's an influencer: Quantifying influence on twitter (2011) Proceedings of the 4th ACM International Conference on Web Search and Data Mining, WSDM 2011, pp. 65-74**. Cited 815 times.** AFFILIATIONS: University of Michigan, United States; Yahoo Research, NY, United States
4. Liu, B. Sentiment analysis and subjectivity (2010) Handbook of Natural Language Processing, Second Edition, pp. 627-666**. Cited 622 times.** AFFILIATIONS: Department of Computer Science, University of Illinois at Chicago, Chicago, IL, United States
5. Savova, G.K., Masanz, J.J., Ogren, P.V., Zheng, J., Sohn, S., Kipper-Schuler, K.C., Chute, C.G. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): Architecture, component evaluation and applications (2010) Journal of the American Medical Informatics Association, 17 (5), pp. 507-513. **Cited 587 times.** AFFILIATIONS: Division of Biomedical Statistics and Informatics, Mayo Clinic, College of Medicine, Rochester, MN, United States; Computer Science Department, University of Colorado, Denver, CO, United States
6. Aronson, A.R., Lang, F.-M. An overview of MetaMap: Historical perspective and recent advances (2010) Journal of the American Medical Informatics Association, 17 (3), pp. 229-236. **Cited 583 times.** AFFILIATIONS: Lister Hill National Center for Biomedical Communications (LHNCBC), US National Library of Medicine, National Institutes of Health, Bethesda, MD, United States
7. Zhao, W.X., Jiang, J., Weng, J., He, J., Lim, E.-P., Yan, H., Li, X. Comparing twitter and traditional media using topic models. (2011) Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 6611 LNCS, pp. 338-349. **Cited 486 times.** AFFILIATIONS: Peking University, China; Singapore Management University, Singapore
8. Deng, L., Yu, D. Deep learning: Methods and applications. (2013) Foundations and Trends in Signal Processing, 7 (3-4), pp. 197-387**. Cited 478 times.** AFFILIATIONS: Microsoft Research, One Microsoft Way, Redmond, WA 98052, United States
9. Chatr-Aryamontri, A., Breitkreutz, B.-J., Oughtred, R., Boucher, L., Heinicke, S., Chen, D., Stark, C., Breitkreutz, A., Kolas, N., O'Donnell, L., Reguly, T., Nixon, J., Ramage, L., Winter, A., Sellam, A., Chang, C., Hirschman, J., Theesfeld, C., Rust, J., Livstone, M.S., Dolinski, K., Tyers, M. The BioGRID interaction database: 2015 update (2015) Nucleic Acids Research, 43 (D1), pp. D470-D478. **Cited 475 times.** AFFILIATIONS: Institute for Research in Immunology and Cancer, Université de Montréal, Montréal, QC H3C 3J7, Canada; Lunenfeld-Tanenbaum Research Institute, Mount Sinai Hospital, Toronto, ON M5G 1X5, Canada; Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ 08544, United States; School of Biological Sciences, University of Edinburgh, Edinburgh, EH9 3JR, United Kingdom; Centre Hospitalier de l'Université Laval (CHUL), Québec, QC G1V 4G2, Canada

10. Srivastava, N., Salakhutdinov, R. Multimodal learning with Deep Boltzmann Machines (2012) Advances in Neural Information Processing Systems, 3, pp. 2222-2230. **Cited 465 times.** AFFILIATIONS: Department of Computer Science, University of Toronto, Canada; Department of Statistics and Computer Science, University of Toronto, Canada

Publications written in the period 2015-2018

1. Chatr-Aryamontri, A., Breitkreutz, B.-J., Oughtred, R., Boucher, L., Heinicke, S., Chen, D., Stark, C., Breitkreutz, A., Kolas, N., O'Donnell, L., Reguly, T., Nixon, J., Ramage, L., Winter, A., Sellam, A., Chang, C., Hirschman, J., Theesfeld, C., Rust, J., Livstone, M.S., Dolinski, K., Tyers, M. The BioGRID interaction database: 2015 update (2015) Nucleic Acids Research, 43 (D1), pp. D470-D478. **Cited 475 times.** AFFILIATIONS: Institute for Research in Immunology and Cancer, Université de Montréal, Montréal, QC H3C 3J7, Canada; Lunenfeld-Tanenbaum Research Institute, Mount Sinai Hospital, Toronto, ON M5G 1X5, Canada; Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ 08544, United States; School of Biological Sciences, University of Edinburgh, Edinburgh, EH9 3JR, United Kingdom; Centre Hospitalier de l'Université Laval (CHUL), Québec, QC G1V 4G2, Canada
2. Amberger, J.S., Bocchini, C.A., Schiettecatte, F., Scott, A.F., Hamosh, A. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an Online catalog of human genes and genetic disorders (2015) Nucleic Acids Research, 43 (D1), pp. D789-D798. **Cited 337 times.** AFFILIATIONS: McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University, School of Medicine, Baltimore, MD 21287, United States; FS Consulting, LLC, Salem, MA 01970, United States
3. Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., Hovy, E. Hierarchical attention networks for document classification (2016) 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2016 - Proceedings of the Conference, pp. 1480-1489. **Cited 251 times.** AFFILIATIONS: Carnegie Mellon University, United States; Microsoft Research, Redmond, United States
4. Jaderberg, M., Simonyan, K., Vedaldi, A., Zisserman, A. Reading Text in the Wild with Convolutional Neural Networks (2016) International Journal of Computer Vision, 116 (1), pp. 1-20. **Cited 223 times.** AFFILIATIONS: Department of Engineering Science, University of Oxford, Oxford, United Kingdom
5. Kusner, M.J., Sun, Y., Kolkin, N.I., Weinberger, K.Q. From word embeddings to document distances (2015) 32nd International Conference on Machine Learning, ICML 2015, 2, pp. 957-966. **Cited 217 times.** AFFILIATIONS: Washington University in St. Louis, 1 Brookings Dr., St. Louis, MO 63130, United States
6. Severyn, A., Moschittiy, A. Learning to rank short text pairs with convolutional deep neural networks (2015) SIGIR 2015 - Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 373-382. **Cited 178 times.** AFFILIATIONS: Google Inc., Qatar; Qatar Computing Research Institute, Qatar; University of Trento, DISI, Italy
7. McGowan, J., Sampson, M., Salzwedel, D.M., Cogo, E., Foerster, V., Lefebvre, C. PRESS Peer Review of Electronic Search Strategies: 2015 Guideline Statement (2016) Journal of Clinical Epidemiology, 75, pp. 40-46. **Cited 160 times.** AFFILIATIONS: School of Epidemiology, Public Health and Preventive Medicine, University of Ottawa, 85 Primrose Avenue, Ottawa, Ontario K1N 6M1, Canada; Cochrane Information Retrieval Methods Group, Canada; Children's Hospital of Eastern Ontario, 401 Smyth Road, Ottawa, Ontario K1H 8L1, Canada; 1003 Pacific Street, Ste. 1106, Vancouver, British Columbia V6E 4P2, Canada; 55 Livingston Road, Ste. 1014, Scarborough, Ontario M1E 1K9, Canada; Porter Road, Oxford Station, Ontario, K0G 1T, Canada; Lefebvre Associates Ltd, Manor Farm Cottage, Thrupp, Kidlington OX5 1JY, United Kingdom

8. Shen, W., Wang, J., Han, J. Entity linking with a knowledge base: Issues, techniques, and solutions (2015) IEEE Transactions on Knowledge and Data Engineering, 27 (2), art. no. 6823700, pp. 443-460. **Cited 134 times.** AFFILIATIONS: College of Computer and Control Engineering, Nankai University, Tianjin, 300071, China; Department of Computer Science and Technology, Tsinghua University, Beijing, 100084, China; Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL 61801, United States
9. Marrie, R.A., Reingold, S., Cohen, J., Stuve, O., Trojano, M., Sorensen, P.S., Cutter, G., Reider, N. The incidence and prevalence of psychiatric disorders in multiple sclerosis: A systematic review (2015) Multiple Sclerosis Journal, 21 (3), pp. 305-317. **Cited 132 times**.
   AFFILIATIONS: Department of Community Health Sciences, University of Manitoba, Canada; Department of Internal Medicine, University of Manitoba, Canada; Scientific and Clinical Review Assoc., LLC, United States; Mellen Center for MS Treatment and Research, Cleveland Clinic, United States; Department of Neurology and Neurotherapeutics, University of Texas Southwestern, United States; Department of Basic Medical Sciences, Neurosciences and Sense Organs, University of Bari, Italy; Department of Neurology, Copenhagen University Hospital, Rigshospitalet, Denmark; Department of Biostatistics, University of Alabama at Birmingham, United States
10. Derczynski, L., Maynard, D., Rizzo, G., Van Erp, M., Gorrell, G., Troncy, R., Petrak, J., Bontcheva, K. Analysis of named entity recognition and linking for tweets (2015) Information Processing and Management, 51 (2), pp. 32-49. **Cited 108 times.** AFFILIATIONS: University of Sheffield, Sheffield, S1 4DP, United Kingdom; EURECOM, Sophia Antipolis, 06904, France; VU University Amsterdam, HV Amsterdam, 1081, Netherlands; Università di Torino, Turin, 10124, Italy

F. Task 2: MT provider web traffic

HQ REGION	NAME	MAR	APR	MAY	JUN	JUL	AUG	AVERAGE
Asia	Translate.yandex.ru	3 771 880	3 530 960	3 662 440	3 682 960	2 963 240	4 610 160	3 703 607
Asia	Fanyi.Baidu	3 205 680	2 280 000	5 177 120	3 688 280	2 011 720	3 929 960	3 382 127
Asia	Systranet.com	272 000	275 100	366 900	333 000	258 100	336 900	307 000
Asia	Systransoft.com	155 700	165 200	135 500	122 100	132 000	209 000	153 250
Asia	Translate.yandex.com	126 750	111 150	115 050	111 150	101 400	120 900	114 400
Asia	GTCom	8 300	3 400	412	239	864	6 900	3 353
Europe	Reverso	32 300 000	27 100 000	26 600 000	22 700 000	17 400 000	20 800 000	24 483 333
Europe	DeepL	2 700 000	2 600 000	2 900 000	2 700 000	2 400 000	2 700 000	2 666 667
Europe	Freetranslation.com	905 400	880 600	909 200	737 600	373 800	335 100	690 283
Europe	sdl.com	293 000	293 100	322 900	373 300	374 100	329 100	330 917
Europe	Moravia	175 000	115 800	170 300	226 900	207 200	237 700	188 817
Europe	Sdltrados.com	67 000	71 000	75 100	61 000	72 300	85 000	71 900
Europe	Tilde	64 500	62 000	51 500	35 300	37 000	32 900	47 200
Europe	Lilt	7 400	13 900	10 000	12 700	31 000	64 500	23 250
Europe	Pangeanic	20 200	18 900	20 700	15 800	13 500	23 500	18 767
Europe	Iconic	2 600	2 500	89	2 700	221	3 400	1 918
NA	Google translate	62 980 000	62 040 000	71 440 000	67 680 000	57 340 000	63 920 000	64 233 333
NA	Translate.com	863 400	871 300	946 500	814 900	776 600	735 300	834 667
NA	Microsoft Bing Translator	430 040	384 280	431 340	378 040	290 940	334 620	374 877
NA	Online-translator.com	286 000	335 500	275 300	245 100	198 500	207 100	257 917
NA	PROMT	11 700	14 900	16 200	10 800	20 600	8 500	13 783
NA	Omniscien	436	3 400	5 900	555	235	2 700	2 204

Source: Semrush.com, visits per month for respective web domain (year 2018).

G. Task 2: recent start-up financing/venture capital – speech

	USD	COMPANY	HQ	INVESTOR	SECTOR
2018	4 018 400	AudioTelligence	UK	Cambridge Innovation Capital, Cambridge Enterprise	Audio
2018	8 000 000	Observe.AI	India	Nexus Investment, Y Combinator, Nexus, +3	Voice
2018	75 000 000	AISpeech	China	MediaTek, Oriza Holdings, Foxconn Technology Group, +2	Voice
2018	12 000 000	Wnor.ai	USA	Madrona Venture Group, Catapult Ventures, Autotech Ventures, +1	Machine learning
2018	15 000 000	Suki	USA	First Roung, Venrock, Social Capital, +1	Speech recognition, Virtual assistant
2018	1 240 000	Slang Labs	India	Endiya Partners	Voice
2018	8 000 000	Voci Technologies	USA	Grotech Ventures, Harbert Growth Partners	Speech-to-text
2018	1 154 800	Biometric Vox	Spain	Serban Biometrics, Murcia Emprende, InnoCapital	Speech recognition, Robotics
2018		Neosapience, Inc.	South Korea	Chaster Roh	Speech recognition, Machine learning
2018	4 000 000	Babblabs	USA	Jerry Yang, Cognite Ventures	Audio
2018	2 425 000	Soapbox Labs	Ireland	Enterprise Ireland, EU Commission, European Innovation	Speech recognition, Machine learning
2018	10 000 000	ObEN	USA	K11 Art Foundation	Speech recognition, Machine learning
2017	4 500 000	Gridspace	USA	Undisclosed funding	Speech recognition
2017	10 000 000	AISense	USA	Draper Associates, Horizons Ventures, Bridgewater Associates, +2	Speech recognition, Machine learning
2017		SoundAI	China	Baidu, Linekong, FreeS Fund, Aplus Capital, Bank of Beijing	Speech recognition, Machine learning
2017	5 200 000	Aiqudo	USA	Atlantic Bridge Capital	Voice
2017	5 000 000	ObEN	USA	Tencent Holdings, Ruigang LI, Fengshion Capital	Speech recognition, Machine learning
2017		Gridspace	USA	Santander Innoventures	Speech recognition
2017	200 000	Voicefox	UK	No information	Speech recognition, Productivity software
2017	200 000	Fluent.ai	Canada	500 startups Canada	Home automation
2017		Velmai	UK	No information	Messaging
2017	1 800 000	Fluent.ai	Canada	Danhua Capital, BDC Catpital, Maple Leaf Angels,	Home automation
2017		obEN	USA	Softbank Ventures Korea	Machine learning
2017	8 000 000	Voysis	Ireland	Polaris Partners, Discovery	Machine learning
2017	8 000 000	Sense.ly	USA	Fenox Venture Capital, Babylon, Mayo Clinic Rochester, Chendwei Capital, Bioved Ventures, Standford SmartX Fund	Enterprise Software
2017	2 600 000	Xnor.ai	USA	Mandrona Venture Group, Allen Institute for Artificial Intelligence	Machine learning
2017	1 400 000	Soapbox Labs	Ireland	Asita Angels, Enterprise Ireland HPSU, Elkstone	Education, machine learning
2016	1 800 000	Ava	USA	Lerer Hippeau Ventures, SV Angel, Boost VC, Crosslink	Mobile, voice

				Capital, Seraph Group, Partech Ventures, Dorm Room Fund, Blake Mycoskie, Pierre Valade, Tim Draper, Steve Blank, FJ Labs	
2016	7 700 000	ObEN	USA	Shenzhen Leaguer Venture Capital, Tird Wave digital, Gordon Cheng, Dream Maker Entertainment, Cybernaut Westlake Partners, CrestValue Capital, NewDo Venture, E3 Capital Partners	Machine learning
2016	14 000 000	Sage Devices	USA	Shell Technology Ventures, Energy Impact Partners	Consumer electronics
2016	56 000 000	Interactions LLC	USA	Comcast Ventures, Revolution, SoftBank Capital, Sigma Partners, NewSprong Capital, Revolution Growth	Enterprise software
2016	1 000 000	*gram Labs	USA	No information	Machine learning
2016	6 710 000	TranscribeMe	USA	No information	Monetisation, machine learnng
2016	50 000	Witlingo	USA	No information	SAAS, machine learning
2016	30 000 000	AISpeech	China	No information	Machine learning
2016		KITT.AI	USA	Founders' Co-op, Alexa Fund	Home automation
2015	12 300 000	Semantic Machines	USA	No information	Speech recognition
2015	12 380 000	Semantic Machines	USA	No information	Speech recognition
2015	250 000	Gridspace	USA	Wells Fargo Startup Accelerator	Speech recognition
2015	1 000 000	Groupe-Allomedia	France	Kima Ventures, Bpifrance	Advertising, machine learning
2015	2 200 000	Sense.ly	USA	Fenox Venture Capital, Launchpad Digital Health, TV Ventures	Enterprise Software
2015	500 000	Fluent.ai	Canada	Tandem Launch	Home automation
2014	1 250 000	Sense.ly	USA	Eastlink Capital	Enterprise Software
2014	1 000 000	Voci Technologies	USA	No information	Enterprise Software
2014		ObEN	USA	Idealab,PreAngel	Machine learning
2014	25 000	Conversat Labs	USA	AlphaLab, Innovation Works	Mobile
2014		Voxware	USA	Cross Atlantic Capital Partners	Audio
2014	577 400	Groupe-Allomedia	France	Kima Ventures	Advertising, machine learing
2014	116 000	Groupe-Allomedia	France	Kima Ventures	Advertising, machine learing
2014	25 000	MonoLibre	Spain	No information	Education, machine learning
2013	435 000	Koemei	USA	500 startups, ChinaRock Capital Management	Search, machine learning
2013	2 000 000	Fluential	USA	Patrick Soon-Shiong	Artificial intelligence, machine learning
2013	28 000	Sense.ly	USA	Alchemist Accelerator	Enterprise Software
2013	700 000	Robin Labs	USA	Altair Capital, Esther Dyson, Arkady Borkovsky	Search, machine learning
2013	40 000 000	Interactions LLC	USA	SoftBank Capital, North Hill Ventures, Sigma Partners, Cross Atlantic Capital Partners	Enterprise Software
2013	50 000	RealSpeaker	Russia	Microsoft	Artificial intelligence,

					machine learning
2012	900 000	TranscribeMe	USA	Sand Hill Angels, Tech Coast Angels, ICE Angels, Keiretsu Forum, TEC Ventures, TA Ventures, Sierra Angels, Maverick Angels	Monetisation
2012		RealSpeaker	Russia	Undisclosed funding	Artificial intelligence, machine learning
2012	3 120 000	Voci Technologies	USA	No information	Enterprise Software
2012		RealSpeaker	Russia	Undisclosed funding	Artificial intelligence, machine learning
2012		Sonalight	USA	Y Combinator	SMS, mobile, machine learning

Data source: Index.co by TNW[224]

[224] https://index.co/market/speech-recognition/investments

H. Task 2: speech recognition/synthesis company web traffic

REGION	COMPANY NAME	MAR	APR	MAY	JUN	JUL	AUG	AVERAGE
Asia	Brainasoft	63 300	58 300	71 100	47 200	43 100	48 100	55 183
Asia	Iflytek	135 900	100 400	131 200	109 000	141 700	126 800	124 167
EU	Cantab Research Limited	25 900	33 700	31 300	28 100	29 300	24 100	28 733
EU	Neurotechnology	55 900	62 900	67 600	65 200	40 900	53 900	57 733
EU	Acapela Group	342 500	283 300	263 000	264 200	233 400	260 400	274 467
NA	Raytheon	781 900	727 900	662 500	645 400	698 900	655 800	695 400
NA	Pareteum	21 600	13 000	4 500	1 900	9 600	4 500	9 183
NA	Sensory	18 500	10 300	9 200	7 500	8 800	11 100	10 900
NA	Fulcrum Biometrics	29 200	28 400	25 200	21 100	22 600	19 000	24 250
NA	M2SYS Biometrics	50 700	31 800	35 300	35 300	24 400	31 100	34 767
NA	LumenVox	77 800	65 300	57 300	42 900	43 400	83 500	61 700
NA	VoiceBase	158 200	129 500	75 600	88 900	86 100	94 300	105 433
NA	Nuance Communications	1 400 000	1 300 000	1 200 000	1 100 000	1 200 000	1 400 000	1 266 667
Asia	SESTEK	38 900	35 800	34 400	31 600	21 100	24 600	31 067
EU	CereProc	91 800	78 100	77 200	54 500	56 600	76 300	72 417
EU	Acapela Group	342 500	283 300	263 000	264 200	233 400	260 400	274 467
US	Hoya	1 300 000	1 100 000	936 200	781 000	911 700	1 100 000	1 021 483
EU	welocalize	43 600	44 100	87 400	185 100	224 000	183 000	127 867
NA	Sensory	18 500	10 300	9 200	7 500	8 800	11 100	10 900
NA	LumenVox	77 800	65 300	57 300	42 900	43 400	83 500	61 700
NA	NeoSpeech	125 600	91 200	109 800	100 800	91 300	109 200	104 650
NA	NextUp Technologies	176 900	169 000	159 200	124 700	134 100	133 500	149 567
NA	iSpeech	232 600	232 100	222 500	219 800	208 200	225 600	223 467
NA	Nuance Communications	1 400 000	1 300 000	1 200 000	1 100 000	1 200 000	1 400 000	1 266 667
NA	Nexmo	1 500 000	1 800 000	2 700 000	2 900 000	2 500 000	2 100 000	2 250 000

Source: Semrush.com, visits per month for respective web domain (year 2018).

I. Task 2: nr. of speech laboratories unified by ISCA association

The countries of this study are emphasised in bold.

REGION	NUMBER OF ORGANISATIONS	COUNTRY	NUMBER OF ORGANISATIONS
AFRICA	3	South Africa	3
AMERICA	48	Brazil	2
		Canada	6
		Chile	1
		USA	39
ASIA	44	**China**	16
		India	6
		Iran	2
		Japan	12
		South Korea	2
		Singapore	2
		Thailand	2
		Vietnam	2
AUSTRALIA	9	Australia	9
EUROPE	72	**Austria**	1
		Belarus	1
		Belgium	2
		Czech Republic	2
		Denmark	1
		Finland	4
		France	9
		Germany	6
		Greece	2
		Hungary	2
		Ireland	2
		Italy	6
		Netherlands	4
		Norway	1
		Portugal	1
		Romania	1
		Russia	1
		Slovakia	1
		Slovenia	1
		Spain	8
		Sweden	2
		Switzerland	2
		UK	12

J. Task 2: research organisations working in IR

The list includes research organisations which have published research papers indexed by the Scopus database.

AFFILIATION	REGION	COUNTRY
Aalborg Universitet	Europe	Denmark
Aalto University	Europe	Finland
Arizona State University	North America	US
Bar-Ilan University		Israel
Bauhaus-Universitat Weimar	Europe	Germany
Beijing Institute of Technology	Asia	China
Ben-Gurion University of the Negev		Israel
Bilkent Universitesi		Turkey
Birkbeck University of London	Europe	UK
Carnegie Mellon University	North America	US
Centrum Wiskunde & Informatica	Europe	The Netherlands
Chinese Academy of Sciences	Asia	China
Chinese University of Hong Kong	Asia	Hong Kong
CNRS Centre National de la Recherche Scientifique	Europe	France
Commonwealth Scientific and Industrial Research Organization		Australia
Consiglio Nazionale delle Ricerche	Europe	Italy
Cornell University	North America	US
CSIRO Data61		Australia
David R. Cheriton School of Computer Science	North America	Canada
Delft University of Technology	Europe	The Netherlands
Dublin City University	Europe	Ireland
East China Normal University	Asia	China
eBay, Inc.	North America	US
Emory University	North America	US
ETH Zurich	Europe	Switzerland
Facebook, Inc.	North America	US
Florida International University	North America	US
Forschungszentrum L3S	Europe	Germany
Fudan University	Asia	China
Georgetown University	North America	US
Georgia Institute of Technology	North America	US
GESIS - Leibniz Institute for the Social Sciences	Europe	Germany
Google LLC	North America	US
Gottfried Wilhelm Leibniz Universitat	Europe	Germany
Harbin Institute of Technology	North America	US
Helsingin Yliopisto	Europe	Finland
Helsinki Institute for Information Technology	Europe	Finland
Hong Kong Polytechnic University	Asia	Hong Kong
IBM Research	North America	US
IBM Thomas J. Watson Research Center	North America	US

Indian Institute of Technology, Kharagpur	Europe	India
Institute of Automation Chinese Academy of Sciences	Asia	China
Institute of Computing Technology Chinese Academy of Sciences	Asia	China
International Institute of Information Technology Hyderabad	Asia	India
IRIT Institut de Recherche Informatique de Toulouse	Europe	France
Istituto di Scienza e Tecnologie dell'Informazione A. Faedo	Europe	Italy
Johannes Kepler Universitat Linz	Europe	Germany
Johns Hopkins University	North America	US
Karlsruhe Institute of Technology	Europe	Germany
King's College London	Europe	UK
Kobenhavns Universitet	Europe	Denmark
KU Leuven	Europe	Belgium
Kyoto University	Asia	Japan
Kyushu University	Asia	Japan
Laboratoire d'Informatique de Grenoble	Europe	France
Laboratoire d'informatique de Paris 6	Europe	France
Lehigh University	North America	US
Max Planck Institut fur Informatik	Europe	Germany
Microsoft Corporation	North America	US
Microsoft Research	North America	US
Microsoft Research Asia	Asia	China
Microsoft Research Cambridge	Europe	UK
Nankai University	Asia	Asia
Nanyang Technological University	Asia	Singapore
Nanyang Technological University School of Computer Engineering	Asia	Singapore
National Taiwan University		Taiwan
National University of Ireland Galway	Europe	Ireland
National University of Singapore	Asia	Singapore
New York University	North America	US
Norges Teknisk-Naturvitenskapelige Universitet	Europe	Norway
Northeastern University	North America	US
NYU Tandon School of Engineering	North America	US
Open University	Europe	UK
Peking University	Asia	China
Pennsylvania State University	North America	US
Purdue University	North America	US
Qatar Computing Research Institute		Qatar
Queen Mary, University of London	Europe	UK
Queensland University of Technology QUT		Australia
Radboud University Nijmegen	Europe	The Netherlands
Renmin University of China	Asia	China
Research Organization of Information and Systems National Institute of Informatics	Asia	Japan
RMIT University		Australia
Robert Gordon University	Europe	UK
Rutgers, The State University of New Jersey	North America	US
Shandong University	Asia	China

Shanghai Jiao Tong University	Asia	China
Simon Fraser University	North America	Canada
Singapore Management University	Asia	Singapore
Sorbonne Universite	Europe	France
Stanford University	North America	US
Tampereen Yliopisto	Europe	Finland
Technical University of Berlin	Europe	Germany
Technion - Israel Institute of Technology		Israel
Technische Universitat Wien	Europe	Austria
Texas A and M University	North America	US
The University of British Columbia	North America	Canada
The University of North Carolina at Chapel Hill	North America	US
Tianjin University	Asia	China
Tsinghua National Laboratory for Information Science and Technology	Asia	China
Tsinghua University	Asia	China
UCL	Europe	UK
Universidad Autonoma de Madrid	Europe	Spain
Universidad de Chile	Asia	Chile
Universidad Nacional de Educacion a Distancia	Europe	Spain
Universidade da Coruña	Europe	Spain
Universidade Federal de Minas Gerais		Brazil
Universita degli Studi di Padova	Europe	Italy
Universita degli Studi di Roma La Sapienza	Europe	Italy
Universita degli Studi di Trento	Europe	Italy
Universita degli Studi di Udine	Europe	Italy
Universita della Svizzera italiana	Europe	Italy
Universita di Pisa	Europe	Italy
Universitat Duisburg-Essen	Europe	Germany
Universitat Heidelberg	Europe	Germany
Universitat Pompeu Fabra Barcelona	Europe	Spain
Universite de Toulouse	Europe	France
Universite Grenoble Alpes	Europe	France
Universite Paul Sabatier Toulouse III	Europe	France
Universitetet i Stavanger	Europe	Norway
University of Amsterdam	Europe	The Netherlands
University of California, Los Angeles	North America	US
University of California, Santa Cruz	North America	US
University of Chinese Academy of Sciences	Asia	China
University of Delaware	North America	US
University of Essex	Europe	UK
University of Glasgow	Europe	UK
University of Haifa		Israel
University of Illinois at Chicago	North America	US
University of Illinois at Urbana-Champaign	North America	US
University of London	Europe	UK
University of Maryland	North America	US

University of Massachusetts	North America	US
University of Melbourne		Australia
University of Michigan, Ann Arbor	North America	US
University of Montreal	North America	Canada
University of Pittsburgh	North America	US
University of Queensland		Australia
University of Science and Technology of China	Asia	China
University of Sheffield	Europe	UK
University of Southern California	North America	US
University of Southern California, Information Sciences Institute	North America	US
University of Strathclyde	Europe	UK
University of Tehran		Iran
University of Texas at Austin	North America	US
University of Tsukuba	Asia	Japan
University of Twente	Europe	The Netherlands
University of Washington, Seattle	North America	US
University of Waterloo	North America	Canada
Uniwersytet Warszawski	Europe	Poland
Waseda University	Asia	Japan
Wayne State University	North America	US
Wuhan University	Asia	China
Xerox Research Centre Europe	Europe	France
Yahoo Inc.	North America	US
Yahoo Research Barcelona	Europe	Spain
Yahoo Research Labs	North America	US
Yandex		Russia
York University	North America	US
Zhejiang University	Asia	China

K. Task 2: acquisition deals in search industry from 2012 to 2018

NAME	COUNTRY	ACQUIRED BY	ACQUIRED ON	ACQUIRED AMOUNT, USD
Yahoo	US	Verizon	June 2017	4 480 000 000
Skyscanner	UK	Ctrip	Nov 2016	1 750 000 000
Momondo Group Limited	UK	Booking Holdings (Priceline Group)	February 2017	550 000 000
Monster	US	Randstad Innovation Fund	August 2016	429 000 000
Mitula	Spain	Lifull	May 2018	133 000 000
MindMeld, Inc	US	Cisco	May 2017	125 000 000
Vurb	US	Snap Inc.	August 2016	110 000 000
Trovit	Spain	NEXT Co	October 2014	90 000 000
Archives.com	US	Ancestry	April 2015	100 000 000
Bluefin Labs	US	Twitter	February 2013	80 000 000
FlashStock	Canada	Shutterstock	June 2017	65 000 000
Vayant	Bulgaria	PROS	August 2017	35 000 000
Blackbird Technologies	US	Etsy	September 2017	32 500 000
Betreut.Pflege	Germany	Care.com	July 2012	23 300 000
Unicommerce eSolutions Pvt. Ltd.	India	Infibeam	May 2018	18 000 000
Hotpads	US	Zillow	Nov 2012	16 000 000
SphereUp	US	Zoomd Inc	Nov 2012	7 000 000
AppCrawlr	US	Softonic	March 2015	6 000 000
Nuroa	Spain	Mitula	March 2016	3 300 000
Technorati	US	Synacor	February 2016	3 000 000
Blekko	US	IBM	March 2015	N/I
Cloud Sherpas	US	Accenture	September 2015	N/I
Doodle	Switzerland	Tamedia AG	January 2014	N/I
Moodstocks	France	Google	June 2016	N/I
Swiftype	US	Elastic	Nov 2017	N/I
Totems		Stripe	February 2015	N/I
buyt.in	India	NewsHunt	June 2015	N/I
DocDoc	Russia	Sberbank	May 2017	N/I
Indeed.com	US	Recruit Holdings	October 2012	N/I
Nestpick	Germany	Rocket Internet	December 2014	N/I
Jobspotting	Germany	SmartRecruiters	January 2017	N/I
Conductor	US	WeWork	March 2018	N/I
DataPop	US	Criteo	February 2015	N/I
FanSnap	US	SeatGeek	December 2013	N/I
Simply Hired	US	Recruit Holdings	June 2016	N/I
iProperty Indonesia		REA_Group	Nov 2015	N/I
ZoomInfo	US	Great Hill Partners	August 2017	N/I
Teleport	US	MOVE Guides	April 2017	N/I
Clickable	US	Syncapse Corp.	June 2012	N/I
Moat	US	Oracle	April 2017	N/I
MinHash	US	Salesforce	December 2015	N/I
Desti	US	Nokia	May 2014	N/I

QPID Health	US	eviCore healthcare	February 2016	N/I
Baynote	US	Kibo software	September 2016	N/I
Wink	US	i.am+	June 2017	N/I
Connectivity	Spain	SweetIQ	September 2016	N/I
Linkdex	UK	Authoritas	June 2018	N/I
FanTV	US	Rovi Corporation	Nov 2014	N/I
goHoppit	US	XO Group Inc.	September 2013	N/I
Hyperpublic	US	Groupon	February 2012	N/I
Totaljobs	UK	Axel Springer Digital Ventures	April 2012	N/I
Kooaba	Switzerland	Qualcomm	January 2014	N/I
Doctoralia	Spain	DocPlanner.com	June 2016	N/I
Guidebox	US	Reelgood	October 2018	N/I
Kngine		Samsung	March 2018	N/I
ClipMine	US	Twitch	August 2017	N/I
Vizibility	US	aslegal.com	August 2013	N/I
Publicis Hawkeye	US	Publicis Groupe	March 2014	N/I
BestParking	US	ParkWhiz	January 2016	N/I
Mekanist	Turkey	Zomato	January 2015	N/I
Refined Labs	Germany	Visual IQ	October 2016	N/I
Pocketin	India	OneLoyalCard	February 2017	N/I
Plasmyd	US	Academia	October 2013	N/I
Reach App	India	ixigo.com	January 2017	N/I
Snagr.io	US	WbSrch	March 2016	N/I
Acclaim IP	US	ANAQUA	April 2016	N/I
Market Metrics	US	ASSET INTERNATIONAL	May 2016	N/I
Corrigon	Israel	eBay	October 2016	N/I
Branded3		St Ives Group	May 2013	N/I
AddStructure	US	Bazaarvoice	February 2018	N/I
Econsultancy		Centaur Partners	June 2012	N/I
Mundi		KAYAK	August 2017	N/I
The Echo Nest	US	Spotify	March 2014	N/I
KAYAK	US	Priceline.com	Nov 2012	N/I
PriceArea.com	Indonesia	Yello Mobile	May 2014	N/I
TheFind	US	Facebook	March 2015	N/I

Data source: Index.co by TNW[225]

[225] https://index.co/market/search/acquisitions

L. Task 2: recent start-up financing/venture capital – search

COMPANY	REGION	ATTRACTED FUNDING (USD)[226]
Baidu	Asia	3 366 200 000
Koubei.com	Asia	2 100 000 000
Qunar.com	Asia	1 389 100 000
NetEase Youdao	Asia	1 100 000 000
23andMe	North America	728 750 000
Sogou	Asia	448 000 000
Mobvoi	Asia	252 203 530
Hortonworks	North America	248 000 000
KAYAK	North America	235 024 070
Quora	North America	226 000 000
Hopper	North America	202 923 709
Coveo	North America	202 200 000
Skyscanner	Europe	197 207 611
Quixey	North America	164 900 000
SeatGeek	North America	159 970 000
Momondo Group Limited	Europe	151 987 409
Giphy	North America	150 950 000
Wikia	North America	145 800 000
Yidian Zixun	Asia	112 100 000
ReachLocal	North America	107 450 000
Elastic	Europe	104 000 000
Kensho	North America	97 800 000
Truecaller	Europe	94 243 640
Apptus	Europe	88 000 000
Justdial	Asia	85 000 000
NetBase	North America	84 600 000
Attivio	North America	83 900 000
Algolia	North America	73 700 000
MAANA	North America	68 195 000
Syapse	North America	68 000 000
SpotHero	North America	67 610 000
Moat	North America	67 500 000
TourRadar	Europe	66 500 000
The Zebra	North America	63 000 063
Cloud Sherpas	North America	62 600 000
BrightEdge	North America	61 900 000
Conductor	North America	60 648 126

[226] Values are expressed in USD or converted to USD, if the deal was done in other value.

Blekko	North America	60 200 000
ixigo.com	Asia	53 500 000
Tripping	North America	52 000 000
ChaCha	North America	52 000 000
Skyword	North America	50 050 000
Wego	Singapore	46 500 000
Searchmetrics	Europe	44 800 000
JIBE	North America	40 875 000
Clarifai, Inc	North America	40 000 000
Simply Hired	North America	34 300 000
Cortica	North America	33 400 000
AlphaSense	North America	33 000 000
Zig Bang	Asia	33 000 000
Twiggle	Israel	32 800 000
Baynote	North America	32 495 730
Kenshoo	Israel	32 000 000
Inbenta	North America	31 875 264
Entefy	North America	30 300 000
Cheche365.com	Asia	30 000 000
Navent	Argentina	30 000 000
Moz	North America	29 250 000
HomeToGo	Europe	26 741 573
Sarcos	North America	26 130 000
The Echo Nest	North America	25 609 989
BitClave	North America	25 500 000
SevenFifty HQ	North America	24 720 000
Swiftype	North America	22 200 000
SkillPages	Europe	22 071 429
AnyClip Media	North America	21 000 000
Atlas Informatics	North America	20 700 000
Bluefin Labs	North America	20 350 000
Trapit	North America	20 116 592
SHR	North America	20 000 000
propertyfinder.ae	United Arab Emirates	20 000 000
Vurb	North America	19 500 000
CCT Marketing	Europe	19 344 000
Superfish	North America	19 300 000
iGola	Asia	19 000 000
iCapital Network	North America	18 805 858
Jobcase, Inc.	North America	18 500 000
Simpli.fi	North America	18 300 000
TrialReach	Europe	17 900 000
Jobbio	Europe	17 240 949
AdStage	North America	16 750 000
Adzuna	Europe	16 336 490
QPID Health	North America	16 300 000

RealMatch		14 700 000
Twenga SA	Europe	14 590 548
Audioburst		14 400 000
RealScout	North America	14 100 000
DRIVIN	North America	14 000 000
National Computational Infrastructure	Australia	14 000 000
CỐC CỐC	Vietnam	14 000 000
Doctrine	Europe	13 847 191
DocDoc	Russia	13 600 000
Shift	North America	13 550 000
DuckDuckGo	North America	13 000 000
Nestpick	Europe	13 000 000
Stratajet	Europe	12 615 385
Diffbot	North America	12 500 000
Trumpet Search	North America	12 040 154
Cartasite	North America	12 000 000
Connectifier	North America	11 700 000
Stardog Union	North America	11 300 000
Loci.io	North America	11 250 000
Talent.io	Europe	11 074 797
Evie	North America	11 000 000
ABODO	North America	10 708 000
Barnebys	Europe	10 310 526
StatMuse	North America	10 120 000
Barnebys	Europe	10 100 000
Adeptmind	North America	10 000 000
Constructor.io	North America	10 000 000
Brain, LLC	North America	10 000 000
Unicommerce eSolutions Pvt. Ltd.	Asia	10 000 000
SeoPult	Russia	10 000 000
Linkdex	Europe	9 309 959
crealytics GmbH	Europe	9 300 000
DataPop	North America	9 200 000
Hulbee		9 000 000
Relcy	North America	9 000 000
FanTV	North America	8 393 798
LogDNA	North America	8 300 000
asap54.com	Europe	8 248 371
biNu	Australia	8 050 000
Zorroa	North America	8 000 000
Syte.ai	Israel	8 000 000
ayfie, Inc.	North America	8 000 000
CamFind	North America	8 000 000
DataSphere		8 000 000
TalkLocal	North America	7 900 000
TripleMint	North America	7 890 000

Judicata	North America	7 800 000
Tellius	North America	7 500 000
The Real Goby	North America	7 500 000
Dohop Flight Search	Europe	7 488 372
findo.io	North America	7 000 000
Locality	North America	6 725 000
Goxip	Asia	6 620 000
Niche	North America	6 600 000
Campus Society	Europe	6 503 272
Plenummedia	Europe	6 500 000
MangoPlate	Asia	6 100 000
Loverly	North America	6 000 000
Tipbit	North America	5 950 000
Panjiva	North America	5 600 000
PinMeTo	Europe	5 500 000
Holidu	Europe	5 434 783
BrandYourself	North America	5 300 000
Ark	North America	5 250 000
Sagoon	North America	5 200 000
Mocavo	North America	5 100 000
Cake Technologies	North America	5 000 000
Converseon	North America	5 000 000
Noodle Education	North America	5 000 000
Campanja	Europe	5 000 000
Sophia Search	North America	4 900 000
Trint	Europe	4 857 401
Eversport	Europe	4 595 124
cielo 24	North America	4 570 000
Repositive.io	North America	4 459 459
Edison Software	North America	4 300 000
Adthena	Europe	4 131 579
YaSabe	North America	4 101 314
HeyStaks	Europe	4 100 373
OnPage.org	Europe	3 734 940
Siren Solutions	Europe	3 703 704
PrismaStar	North America	3 674 567
LivingLens	Europe	3 621 250
Mozio	North America	3 250 000
OwlTing	Asia	3 130 000
Frograms	Asia	3 109 276
mrUsta.com	United Arab Emirates	3 101 000
FlashStock	North America	3 100 000
Leap.it	Europe	3 080 000
Wikimedia	North America	3 020 000
TraceAir Technologies	North America	3 000 000
BioMARC	North America	3 000 000

Unsilo	North America	3 000 000
Particle News	North America	3 000 000
VNTrip	Vietnam	3 000 000
Vayant	Europe	3 000 000
Nestigator	North America	3 000 000
ZEEF.com	Europe	2 903 119
Seez_it	United Arab Emirates	2 800 000
Apartum	Europe	2 777 271
rome2rio	Australia	2 732 790
Valossa	Europe	2 650 000
Q-Sensei Corp.	North America	2 580 000
StyleLounge.de	Europe	2 527 473
Surfingbird	Russia	2 525 000
FindWork	Asia	2 500 000
blinkfire labs, inc.	North America	2 500 000
Findyr	North America	2 500 000
Teleport	North America	2 500 000
ALLYKE	North America	2 480 000
Recommend	Europe	2 441 418
Seva	North America	2 400 000
RentCheck	North America	2 400 000
Iris AI	Europe	2 350 000
Accentium Web	Asia	2 300 000
Telectic	Europe	2 272 727
Slab	North America	2 200 000
Spotzot	North America	2 200 000
realla.co	Europe	2 173 913
Darius Cheung		2 160 000
Search'XPR	North America	2 159 091
KlikkaPromo	Europe	2 051 813
Concourse Global	North America	2 000 000
TripChamp	North America	2 000 000
Rechat	North America	2 000 000
Tabulate	North America	2 000 000
Posse	Australia	2 000 000
Desti	North America	2 000 000
Evature		2 000 000
Favbuy	Asia	2 000 000
Fisgo	Brazil	2 000 000
Svetlana Kuznetsova	North America	1 955 000
WizeNoze BV	Europe	1 944 444
syte-vc.com	Israel	1 900 000
Edamam	North America	1 900 000
Newronika	Europe	1 888 889
Trade Machines FI GmbH	Europe	1 868 867
Context Scout	Europe	1 824 323

RankScience	North America	1 800 000
Deepgram	North America	1 800 000
SpazioDati	Europe	1 800 000
JobHive	North America	1 738 275
Careereye.se	Europe	1 730 743
Diagnosia	Europe	1 700 000
Much Better Adventures	Europe	1 631 863
CheckMyBus	Europe	1 627 907
Zodio Philippines		1 600 500
107room	Asia	1 600 000
Jobspotting	Europe	1 596 386
Reve	Europe	1 578 560
SRCH2	North America	1 510 000
me.me	North America	1 500 000
Momlife	North America	1 500 000
Linkapedia	North America	1 500 000
[Partpic]	North America	1 500 000
OMNI Retail Group	North America	1 500 000
SiteWit Corp		1 500 000
Musikki	Europe	1 416 595
AddStructure	North America	1 400 000
Cardihab	Australia	1 350 000
Fligoo	North America	1 339 000
Invajo	Europe	1 318 633
Slash	North America	1 300 000
Gbooking	Israel	1 300 000
ColorModules	North America	1 300 000
CarSnip.com	Europe	1 285 955
Standard Analytics		1 280 000
Proximity Grid	North America	1 250 000
Drop Messages	North America	1 250 000
3D Industries Ltd.	Europe	1 227 174
Knil (Benigo)	Israel	1 200 000
EverWrite		1 100 001
ClipMine	North America	1 100 000
Balakam		1 079 000
Totems		1 070 220
ScholarPro	North America	1 065 000
Meiya	Asia	1 025 904
Blink	North America	1 025 000
Tilofy	North America	1 020 000
XpertDox	North America	1 000 000
iSearchPlant	South Africa	1 000 000
Ajira Digital	Kenya	1 000 000
Findo	North America	1 000 000
Rent College Pads	North America	1 000 000

3Dprintler	North America	1 000 000
RecomN.com	Malaysia	1 000 000
Inpher	North America	1 000 000
Streamoid Technologies Inc	North America	1 000 000
WIV Labs	Asia	1 000 000
buyt.in	Asia	1 000 000
MedTel	North America	900 000
Storyzy	Europe	900 000
Bitext		900 000
Conference Hound	North America	900 000
AutoRef.com	North America	875 000
Allstay	Asia	864 000
Intento	North America	860 000
Tire Agent	North America	850 000
Nooklyn	North America	825 000
Wyndow	North America	800 000
Narratif	Europe	794 109
Puzl_me	Europe	780 000
PHIND	North America	755 000
Stay22	North America	750 000
Enlyton	North America	750 000
Meddik	North America	750 000
Robin Labs	North America	740 000
SHOT & SHOP	Europe	737 349
Venturocket	North America	700 000
HipFlat		670 000
LabWorthy	North America	650 000
AddSearch	Europe	650 000
Walkby	North America	650 000
uCastMe Agency	Europe	634 064
Psykosoft	Europe	618 000
snopes.com	North America	612 690
Jellow	Europe	602 410
Gymtrekker	Asia	600 000
PriceMapApp	Asia	600 000
Loop54	Europe	600 000
MedWhat	North America	560 000
The Venue Report		550 000
Onfan	Europe	537 691
eyesFinder		520 000
Videoly	Europe	500 620
Baarb, Inc.	North America	500 000
Wongnai	Thailand	500 000
Reach App	Asia	500 000
YogaTribes	North America	500 000
Amberjack	North America	500 000

ebindle.com	North America	500 000
Skylight	North America	500 000
Jobstore.com	North America	500 000
goHoppit	North America	500 000
Clever PPC	Europe	463 855
Ttwick	North America	455 000
LogFuze	North America	450 000
SoleTrader.com	Europe	440 061
Storific	Europe	436 730
Koemei	North America	435 000
App-A-Minute	North America	425 000
Underhood	Europe	404 348
BookThatBook	North America	400 000
Course Match	Europe	392 857
Matthias Zeitler	Europe	386 881
Anpro21		386 483
Vioozer	North America	350 000
Seeker-Industries	Europe	277 122
Geliyoo	Turkey	260 000
Taste Filter	North America	250 308
FindURClass	Asia	250 000
Geevv	Indonesia	220 000
Sawerly	Saudi Arabia	211 000
Letme.ai	North America	200 000
Koobee	Australia	200 000
Wherefor	North America	200 000
ST Booking	Asia	200 000
CitySpade	North America	200 000
Bujbu	Australia	200 000
WNNA	North America	195 000
Private.Me	North America	180 000
Looklist	North America	174 000
Choister	Russia	165 000
Sidekick	North America	160 000
Get@	North America	156 000
FindTheRipple	Europe	152 961
axle ai	North America	150 000
Pricebook.co.id	Asia	150 000
Extreme Seo Internet Solutions	Sri Lanka	150 000
Subease	North America	150 000
Pocketin	Asia	150 000
Skoov	Asia	150 000
Hit Labs	North America	150 000
Globehook		150 000
Discoapi		150 000
SocialMart	Russia	150 000

Wikisway.com	North America	140 000
Grr-ithm	North America	130 700
Omniref	North America	120 000
Backstitch	Europe	120 000
OP3Nvoice	Europe	120 000
Weave.ai	Europe	118 000
Liftiee	Asia	108 000
Onyougo	Europe	107 474
WikiReviews	North America	100 000
Your Style Unzipped	North America	100 000
Cantalop	Egypt	100 000
Quickly	North America	100 000
OkCopay		100 000
Seesearch	Europe	96 037
GRAVIDI		88 235
ProfitSourcery	Europe	86 585
FirmPlay	North America	85 000
Pollarize		78 283
Nestd	Australia	73 900
Spacelet, Inc.	North America	70 000
Avtozaper	Russia	70 000
GeniusMatcher		56 874
TheParty.Net		50 000
Carweez		50 000
TheSeaApp	North America	50 000
AI Patents		50 000
WiiiWaaa		50 000
SportCentral	Europe	50 000
LocalSort	South Africa	50 000
WhereInFair	Europe	40 000
Jobyal	Chile	40 000
Localisto	North America	40 000
Roundrate	North America	40 000
Rosters	North America	35 000
Openplay	Europe	34 833
Terrapattern	North America	34 000
Xendo	North America	28 000
Hypecal	Europe	26 001
Qwalytics	North America	25 000
RoommateFit	North America	25 000
Bink! Scan Logos&Win	North America	20 000
Trakstream		15 000
Pximity	North America	15 000
Plasmyd	North America	15 000
Findersfee	Europe	13 568
ClubUp	Asia	10 217

RiteKit	Europe	3 800
JobGator	North America	2 600

Data source: Index.co by TNW[227]

[227] https://index.co/market/search/investments

M. Task 2: recent start-up financing/venture capital – transl. tech.

	USD	COMPANY	HQ	INVESTOR	SECTOR
2018	1 468 429	Jiaoliutong	China	Wuhan Gaoling Capital	Online translation platform
2017	650 000	Cadence Translate	US	500 Startups, Blacktop Capital	Streaming real-time translation
2017	Undisclosed	Agencija INT d.o.o.	Slovenia	Undisclosed	Translation services
2017	Undislosed	Mirai Translate, Inc.	Japan	Honyaku Center	Translation services
2016	2 800 000	SmartCAT	Cyprus	Undisclosed	Translation services SaaS
2016	Undisclosed	United languages Group	US	Yukon Partners and Northern Pacific Group	Translation services
2016	2 500 000	Alugha	Germany	Greinert Verwaltungsgesellschaft	Video translation service
2016	Undisclosed	Translate Now, Inc.	US	Undisclosed	Translation services
2015	100 000	RTT Mobile Interpretation	US	Undisclosed	Translation services
2015	6 000 000	Straker Translations	US	Scobie Ward and Bailador Investments	Translation services
2015	5 000 000	TextMaster	France	Alven Capital and Serena Capital	Translation services
2015	1 086 957	LiveWords	The Netherlands	Paranza and Bram Polak	Translation services
2014	3 445 812	TextMaster	France	Alven Capital	Translation services
2014	1 050 000	Alugha	Germany	Greinert Verwaltungsgesellschaft	Video translation service
2015	300 000	RTT Mobile Interpretation	US	Undisclosed	Translation services
2014	10 000 000	One Hour Translation	US	Fortissimo Capital	Translation services
2013	20 000	TurboTranslations	Poland	Innovation Nest	Translation services
2012	40 000	Lexplique	US	Undisclosed	Translation services
2012	1 500 000	RTT Mobile Interpretation	US	Undisclosed	Translation services
2008	300 000	Straker Translations	US	Undisclosed	Translation services

Data source: Index.co by TNW[228]

[228] https://index.co/market/translation-services/investments

N. Task 3: details of analysis of LT adoption by public services

This annex provides a detailed presentation of the data from the online survey collected through the 79 respondents.

The list of technologies is given in Section 4.2.2.

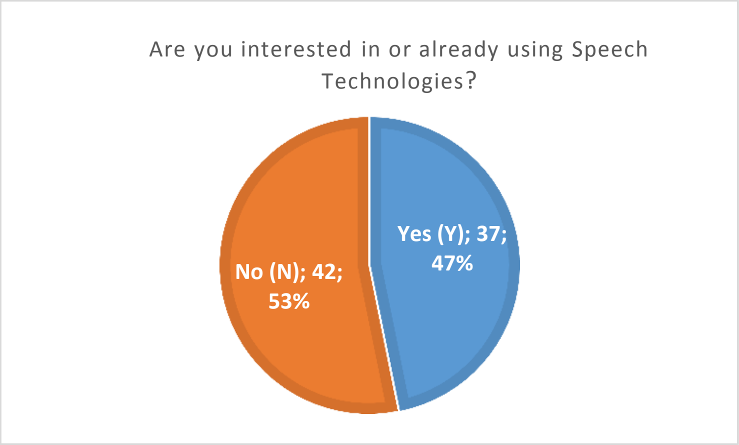
Speech technologies

In the present section, we focus on technologies incorporating speech input/output and list the following applications/components:

- Speech recognition
- Speech synthesis (text-to-speech)
- Speech translation

A number of innovative applications were not included in the survey as they are rarely used at present, in particular those involving biometric technologies such as speaker identification/verification.

Figure 115 Use or interest to use speech technologies

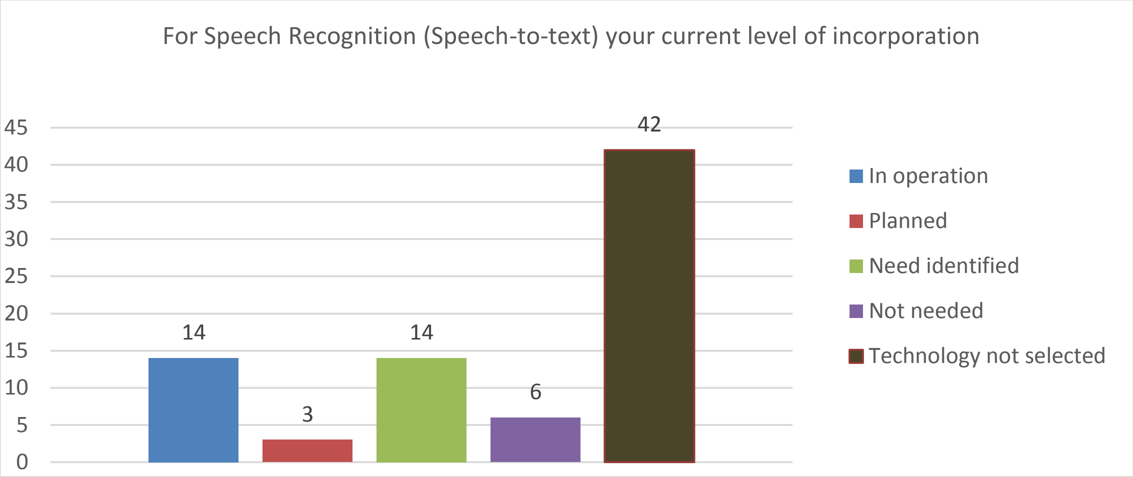


(N=79)

There were 37 responders indicating their use or interest in speech technologies. Our statistics are based on these.

1. Speech recognition

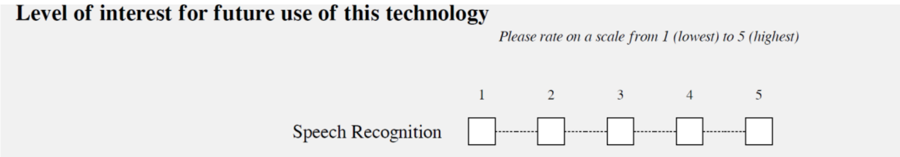
Figure 116 Status of use of speech recognition



(N=37/79)

We see that 14 respondents stated that some automatic speech recognition application is already in operation, while 3 are under planning and 14 with identified needs. Only 6 stated that the technology is not needed. When asked about their suppliers, several answers list the Nuance Dragon Naturally Speaking (a dictation software/application with multiple variants), Wordbee, Vecsys, Onmobile, Accenture, and several universities. Interestingly, some are using the Conference application Zoom with its meeting transcription add-on provided by AISense. The information shows very early adopters (SNCF since 1994, projects with VECSYS) and also recent users (2018).

Some responses about the plans to integrate and deploy such technology, are highlighted herein. As indicated above, this question was asked for all technologies as:



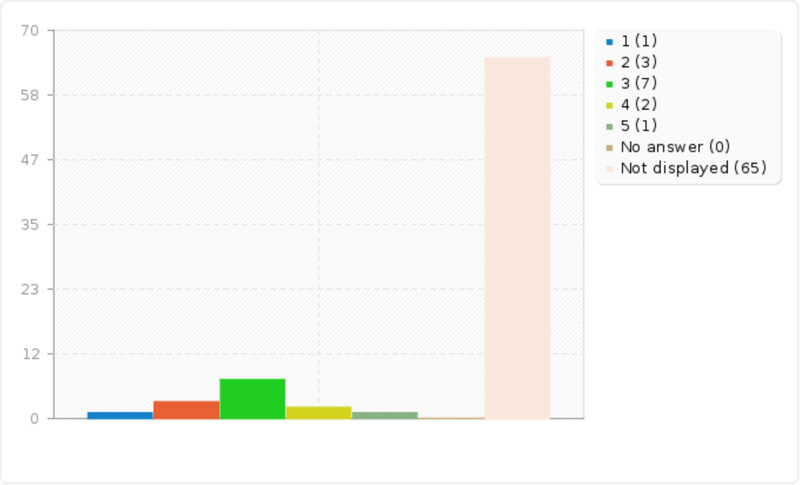
Level of interest for future use of this technology

Please rate on a scale from 1 (lowest) to 5 (highest)

	1	2	3	4	5
Speech Recognition	□	□	□	□	□

We see that for speech recognition, 3 out of 14 indicated high/very high interest (4 or 5). Another 7 indicate a moderate interest (3).

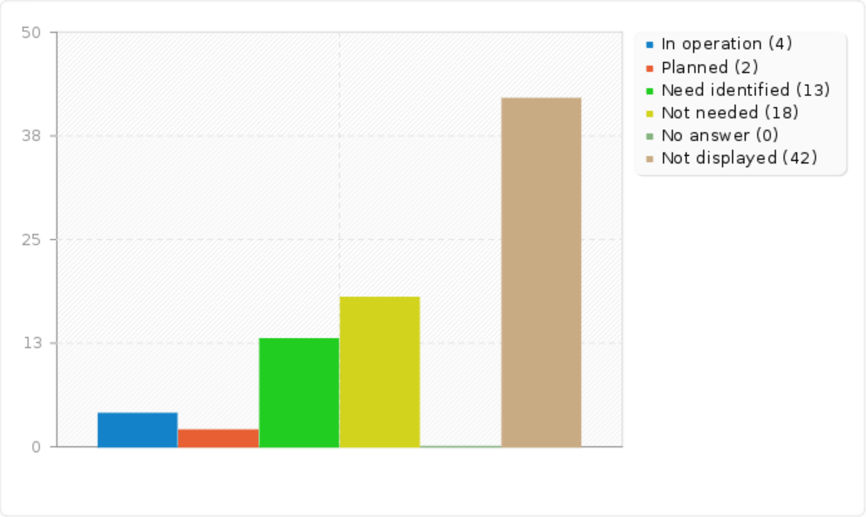
Figure 117 Interest to use speech recognition



2. Speech synthesis (text to speech or TTS)

In the case of speech synthesis (text-to-speech) tools, the current level of incorporation is indicated here:

Figure 118 Status of use of speech synthesis



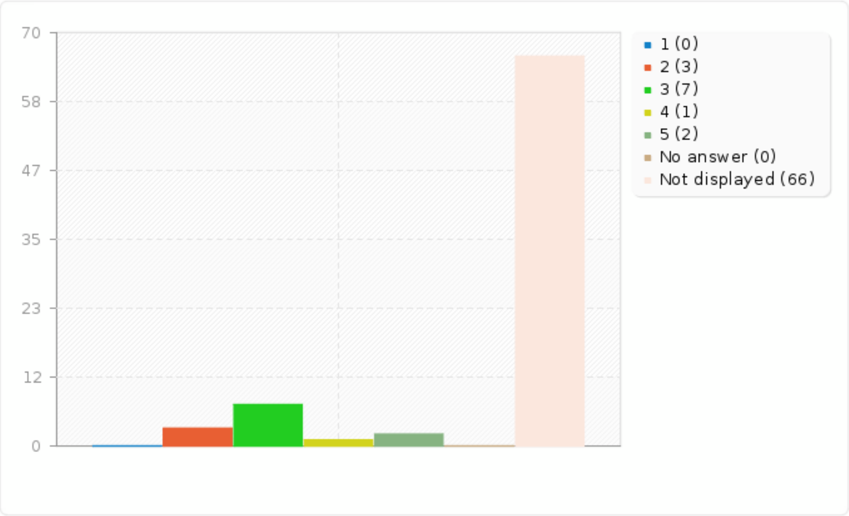
The suppliers mentioned are:

- Readspeaker
- Institute of the Estonian Language
- Voxygen
- K-Pro Informatics Ltd.

The Institute of the Estonian Language in the above list refers to a web page where the institute lists a number of speech synthesis technologies or components such as Festival, eSpeak, Mbrola, etc. The second supplier, K-Pro Informatics Ltd., provides services to customers, including public administrations.

3 out of 13 respondents expressed high to very high interest on future deployment, while 7 of the 13 indicated moderate interest.

Figure 119 Interest to use speech synthesis

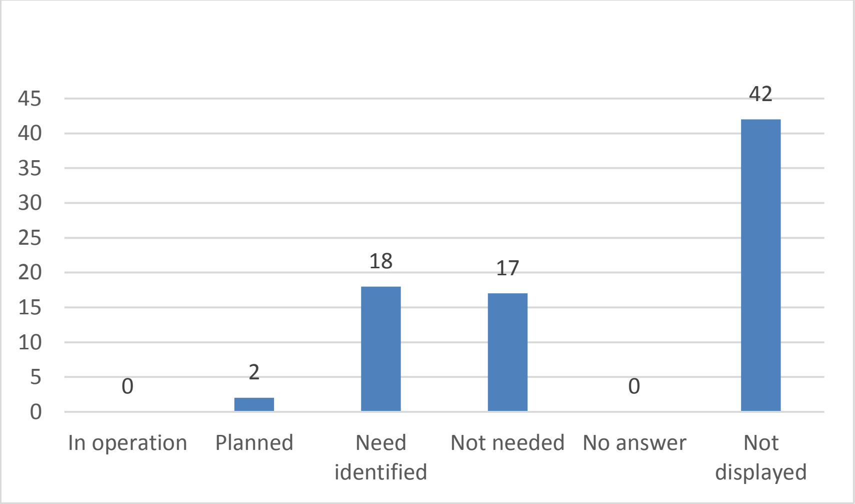


3. Speech translation

Among the 37 positive responses related to speech technologies, there were no indications that the technology is currently in use (we expected some universities to have such tool at least at a prototyping stage).

Surprisingly, 2 responses indicate that it is in their plans while 18 mentioned that they identified the need for such technology. 42 did not get the questions of this section, having indicated that they were not interested in speech technologies.

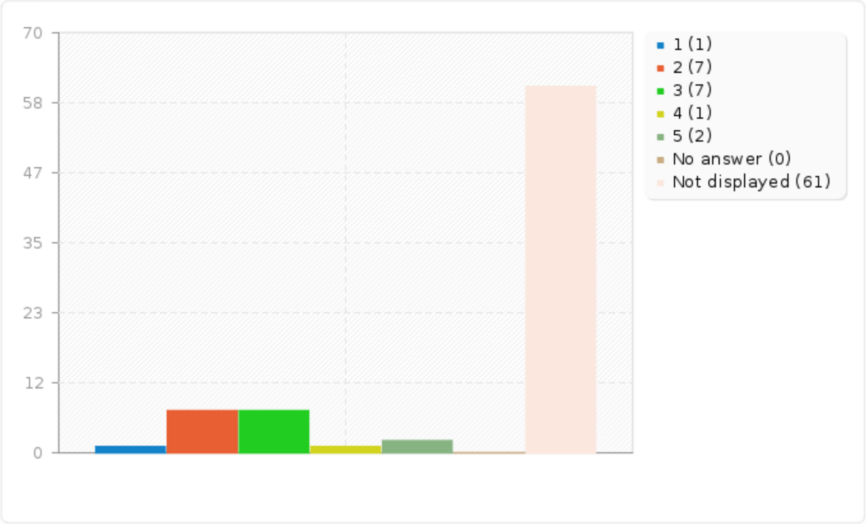
Figure 120 Status of use of speech translation technologies



(N=37/79)

The level of interest is quite low; out of the 18 who expressed their interest, only 3 expressed strong interest (4 or 5).

Figure 121 Interest to use speech translation in the future



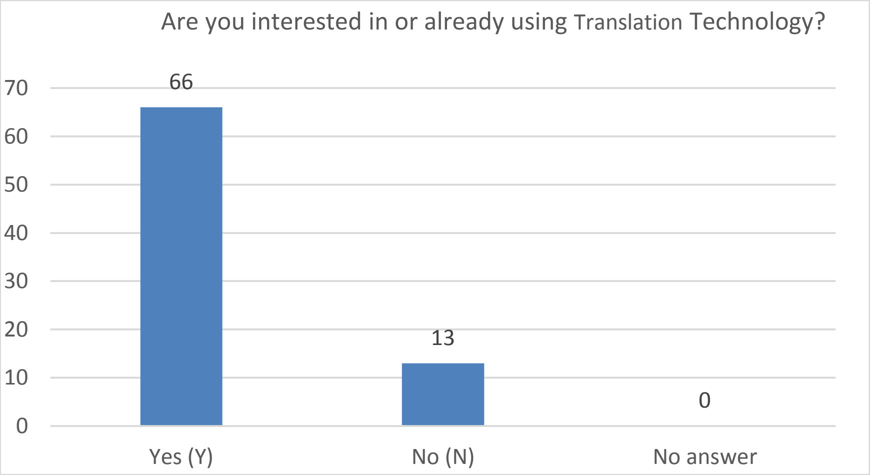
Translation technologies

In this section about translation technologies, we aimed at collecting information about all components of text translation applications and systems. We tried to be exhaustive in selecting the following items:

- Machine translation
- Computer-aided translation (CAT) tools
- Translation memories
- Alignment tools
- Translation workflow management
- Authoring tools

66 of the 79 completed questionnaires (83.5%) indicated that the administration participating to the questionnaire is interested in or already using translation technology, and 13 ticked the NO box.

Figure 122 Use or interested to use translation technologies

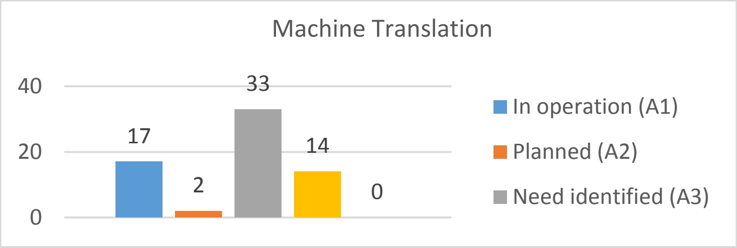


(N=79)

Given the scope of the survey, it is understandable that a very large number of Member States public administrations and services are using or are interested by translation technologies.

1. Machine translation

Figure 123 Status of use of MT

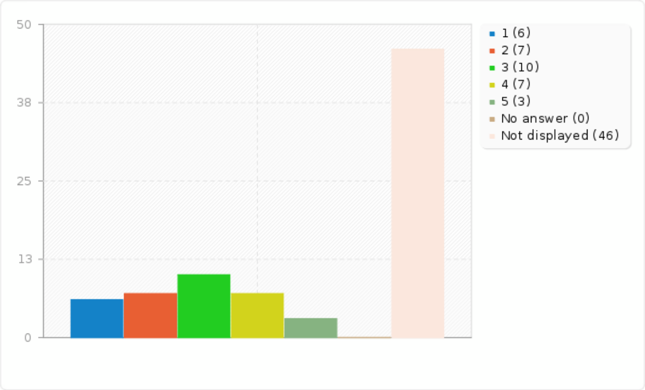


(N=66/79)

The replies show that 17 services already use machine translation and a high number (33) have identified the need for the technology. The suppliers of the 15 respondents who provided an answer include both commercial products and research prototypes and systems:

- Google Translate
- Microsoft
- Moses
- Systran
- DeepL
- European Commission: eTranslation, MT@EC
- EU Presidency Translator
- Wordbee
- Alkonas
- Tartu University Translator
- Tilde

Figure 124 Interest to use MT in the future

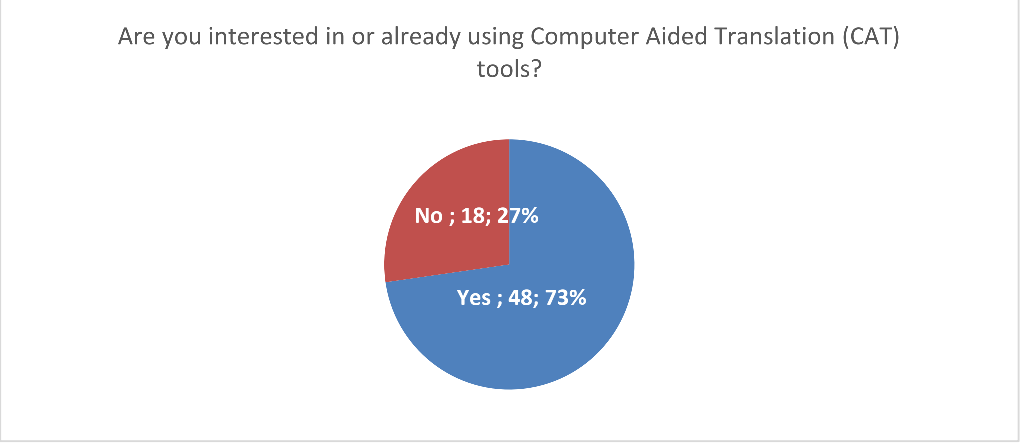


10 out of 33 respondents indicated a strong level of interest in machine translation (4 or 5), another 10 indicated an interest of 3/5.

2. Computer-aided translation (CAT) tools

As expected, CAT tools are widely used by the administrations in our sample. Out of the 66 users/interested respondents, 48 use CAT tools and only 18 do not.

Figure 125 Use or plans for the use of CAT tools



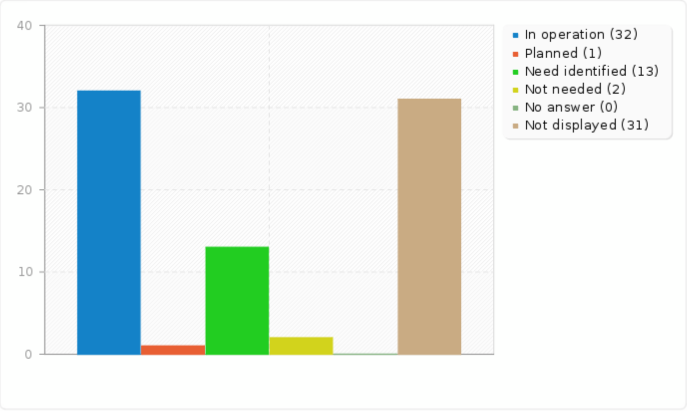
(N=66/79)

3. Translation memories

Translation memories are also widely used, according to our 32 respondents where these are already in operation. One indicated it is planned and 13 have identified their needs for it.

It is interesting that this does not seem consistent with the 48 positive responses on CAT tools use. We assume that the main use of CAT tools concerns translation memories.

Figure 126 Status of use of CAT tools



(N=48/79)

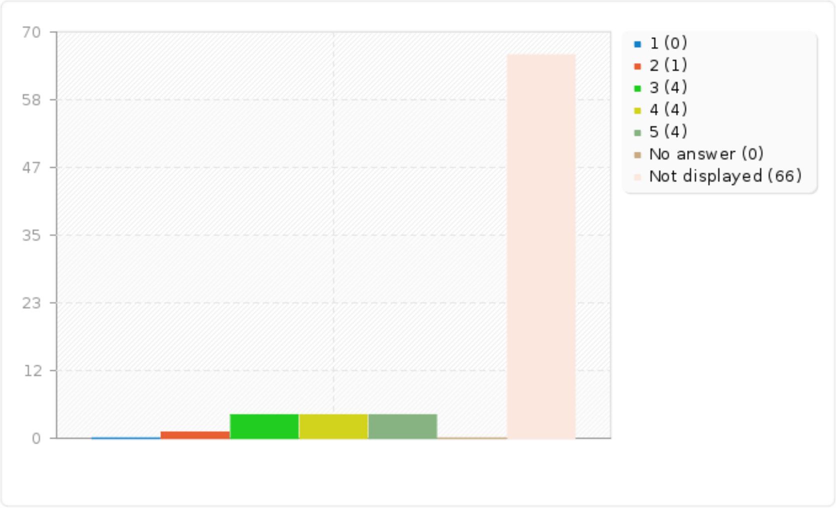
The lists of suppliers cited SDL more than 20 times (out of 30). The list comprises:

- SDL (SDL Trados/Studio/Workbench), in addition to partners and resellers (e.g. Amplexor)

- Atril
- Wordbee
- MemoQ (Kilgray MemoQ)
- Memsource
- Terminotix
- Wordfast

The level of interest for future use is very high, 8 out of 13 indicated high and 4 moderate interest.

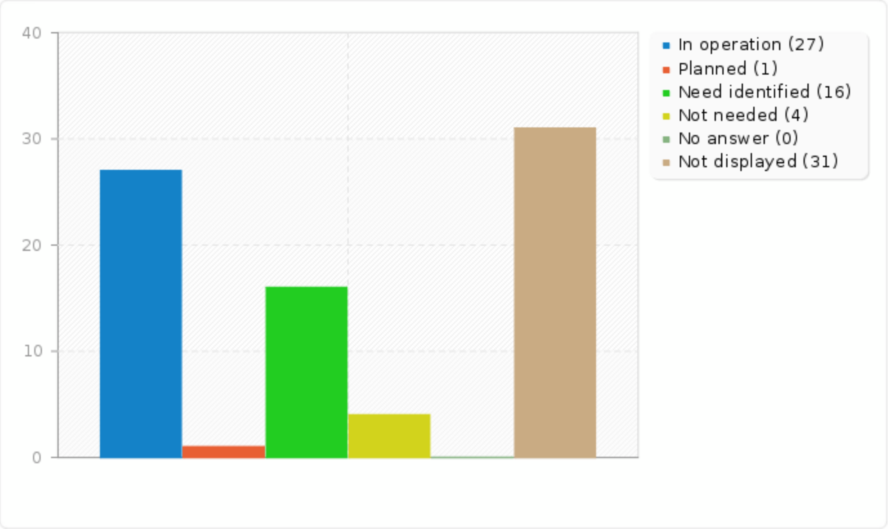
Figure 127 Interest to use translation memories in the future



4. Alignment tools

Many respondents use this technical component within other CAT and MT tools. 27 responded that they have something in use and one that it is planned, 16 that they have identified their needs. 4 stated that it is not needed.

Figure 128 Status of use of alignment tools

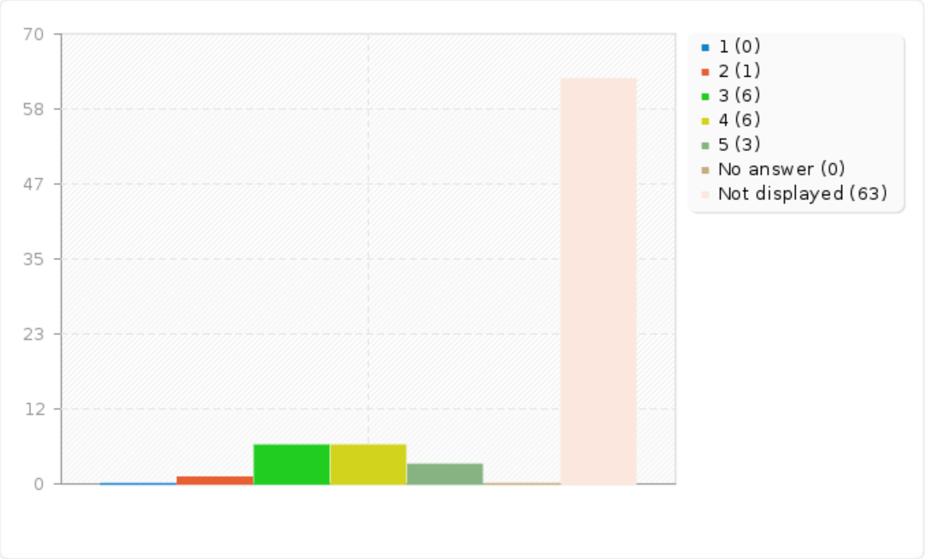


(N=48/79)

The lists of suppliers mention major suppliers of CAT tools (SDL, MemoQ, Terminotix) but also free open source packages, such as LF Aligner.

Despite the technical facet of this module, 9 respondents out of 16 expressed a high level of interest and 6 a moderate interest.

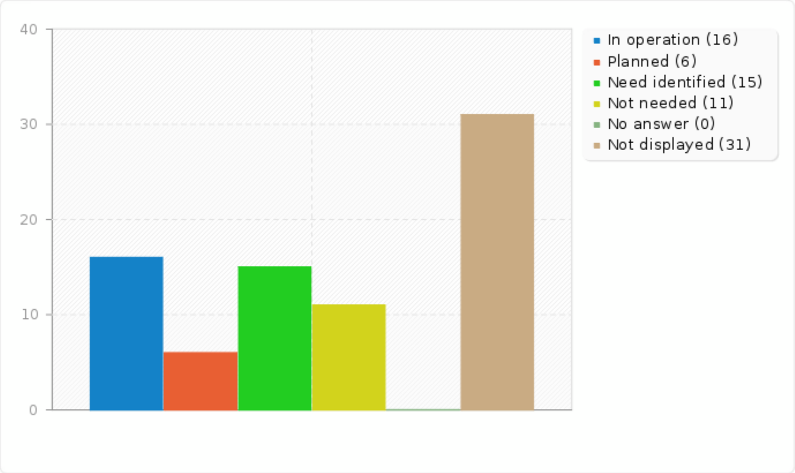
Figure 129 Interest to use alignment tools in the future



5. Translation workflow management

Although this is not a technology *per se*, we collected information on the use of structured management of the translation workflow, as it is highly relevant to activities in language resources collection for future use within the training of various technologies, such as MT.

Figure 130 Status of use of translation workflow management

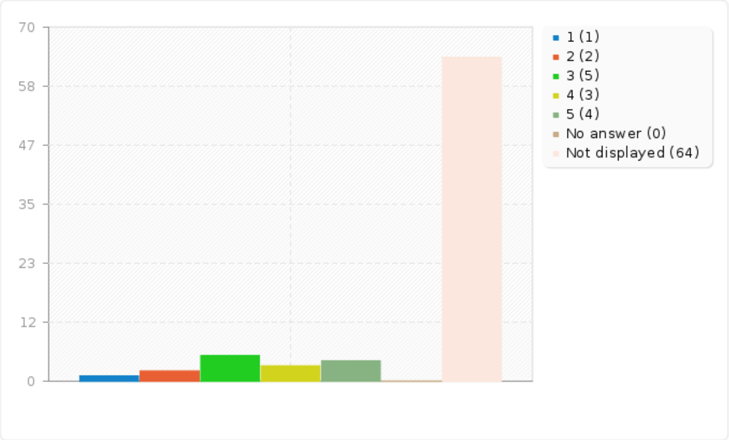


(N=48/79)

16 respondents have already implemented translation workflow management systems, 6 are planning to do so and 15 have identified their needs. Only 11 of the 48 responses stated that they have no need for it. Regarding the list of suppliers, in addition to “in-house” platforms, we found the same companies as in the previous sections above related to MT components (SDL, MemoQ, Wordbee), but also software houses (e.g. Isyde).

The level of interest for future use of management workflows is moderate (we assume that most of the potential users have already adopted this technology). 7 out 15 express high interest with 3 having no or low interest. This is expressed on our scale with 1 (no interest) to 5 (very high).

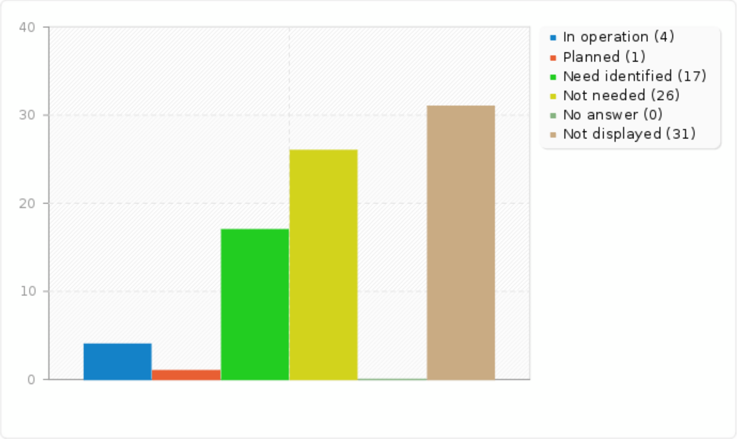
Figure 131 Interest to use translation workflow management in the future



6. Authoring tools

Authoring tools allow the production of documents and include technical writing tools or controlled language aided writing. Tools that help spell checking or grammar analysis were not excluded *per se*, but we did not expect them to be mentioned and this is what happened. They are used by all, and no one considered them as specific tools.

Figure 132 Status of use of authoring tools

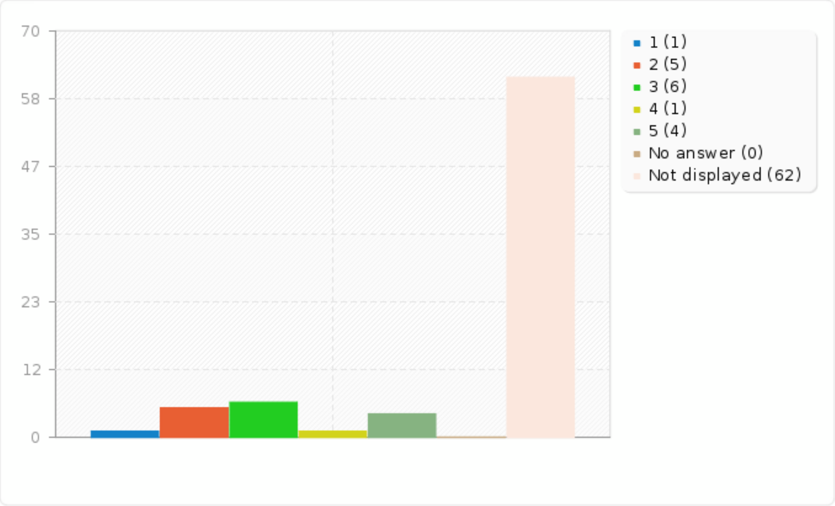


(N=48/79)

Interestingly, only 4 respondents out of 48 indicate that there are some tools in operation. It is probably because very few administration services are using the simplification writing which requires the use of authoring tools like those used in aerospace industry (use of controlled vocabularies, use of very strict and specific grammar rules, e.g. no passive mode, etc.). The one tool mentioned is FontoXML, an XML editor that helps to “create structured and intelligent content”.

The level of interest for future use of this technology is hard to interpret, 5 out of 17 respondents indicate a high interest.

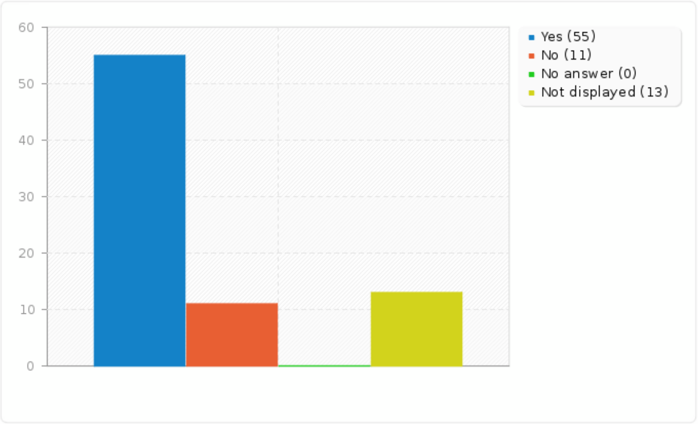
Figure 133 Interest to use authoring tools in the future



Terminology software

In this section, we collected information about both the terminology management and the terminology extraction tools. 66 participants responded to this question (out of the 79) and 55 indicated that they are using or interested to use terminology software.

Figure 134 Use or plans for the use of terminology software



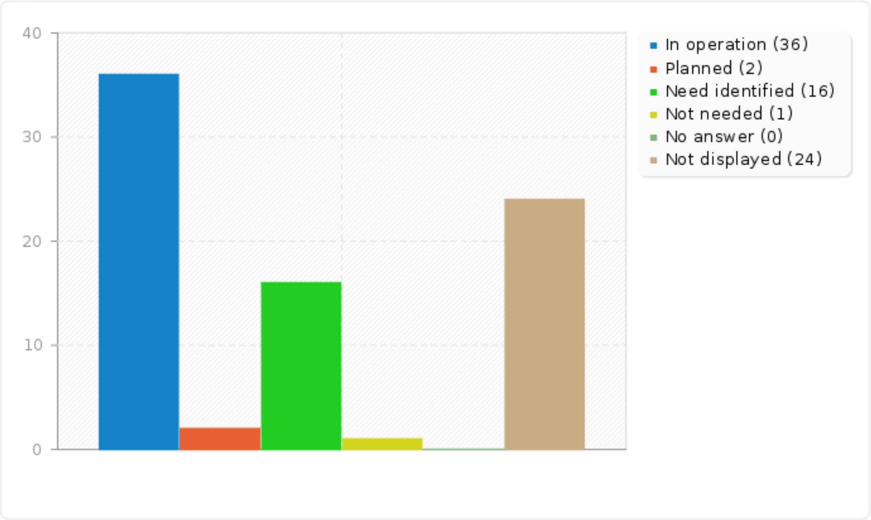
(N=66/79)

1. Terminology management systems

There are many services dealing with translation and writing activities which are helped by the use of terminology tools. Out of the 55 responses, we got 36 indicating that such tool is in operation, 2 that implementation is planned, 16 that the needs were identified and only 1 that such tool is not needed.

Again, on the list of suppliers, SDL figures prominently (multiple packages) but many other tools, including open source packages and/or in-house ones, are also listed. Examples are CrossLang and Terminotix.

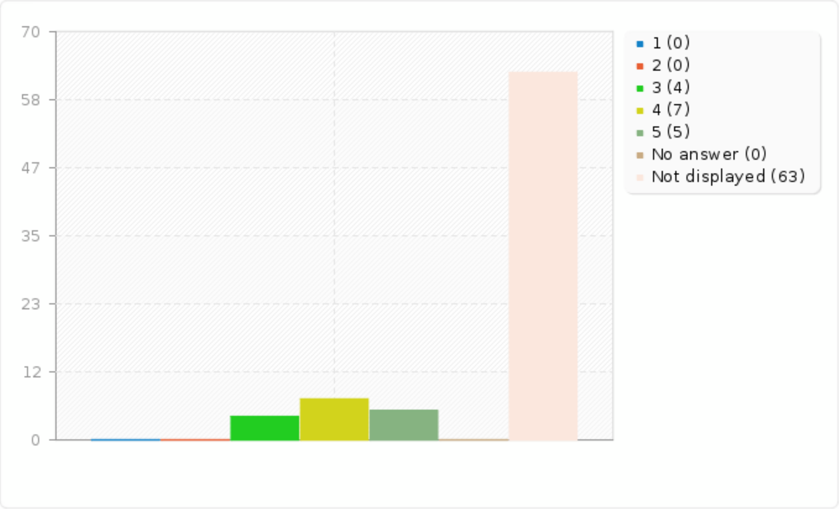
Figure 135 Status of use of terminology management systems



(N=55/79)

It is not surprising that the level of both use and interest is very high. 12 respondents out of 16 selected high or very high interest.

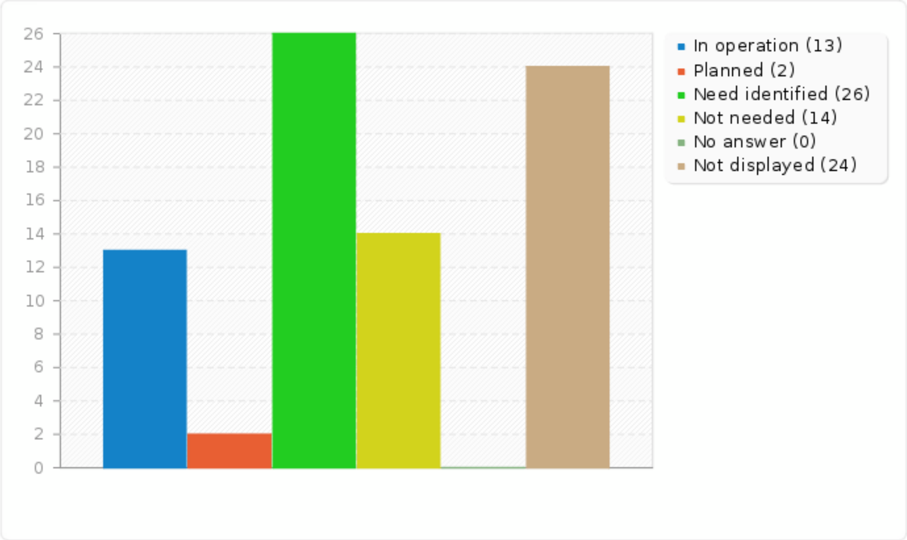
Figure 136 Interest to use terminology software in the future



2. Terminology extraction

The level of integration of terminology extraction is lower than the use of terminology database management software. We think this is because extraction requires more computational linguistics and domain expertise than management of existing databases. Nevertheless, 13 of the 55 responses indicate that such technology is in operation and 2 that it is planned, while 26 have identified that it is needed. 14 stated that it is not needed.

Figure 137 Status of use of terminology extraction

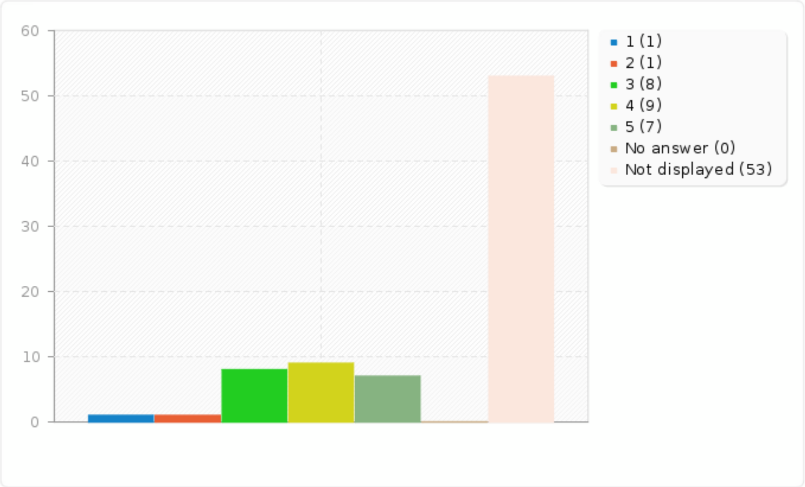


(N=55/79)

The suppliers listed by respondents include Trados, some small local suppliers such as Gridline and the Sketch Engine, and also academic packages such as TermoStat, developed at the university of Montreal, or SynchroTerm from Terminotix.

The level of interest for future use of this technology is very high, with 16 out of 26 scoring their interest at 4 or 5 and 8 at 3.

Figure 138 Interest to use terminology extraction in the future

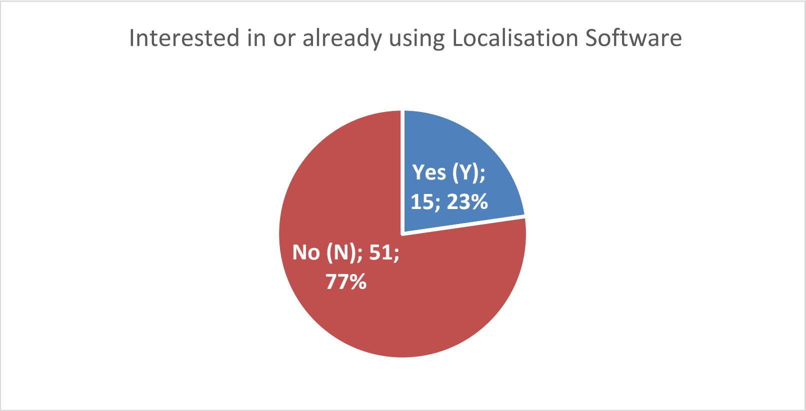


Localisation software

Localisation applications focus on a very narrow market in the case of public administrations. Very often localisation is most visibly showcased in the localisation of web sites, software and forms (in particular administrative forms) and localisation tools applied to subtitling/dubbing production. Only 15 of the 66 replies indicate current use of such technology.

13 of the 79 responses did not select the section on localisation aspects (included in the category "Not Displayed" in the figures below, as the questionnaire did not display the related sections in that case).

Figure 139 Use or plans to use localisation software

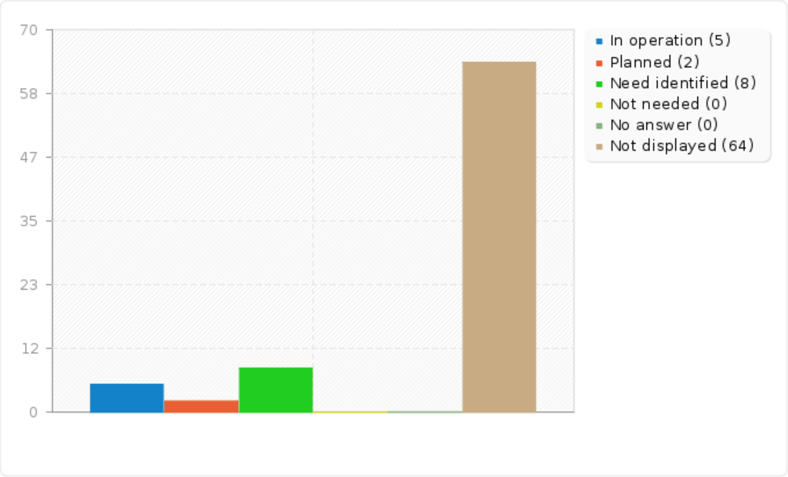


(N=66/79)

1. Localisation tools applied to websites

Although these tools are a critical part of the multilingual aspect of public services, only 5 respondents indicate that they have it in operation, 2 that it is planned and 8 that they have their needs identified. 64 participants did not respond to this question (55 participants skipped this section).

Figure 140 Status of use of localisation tools applied to websites

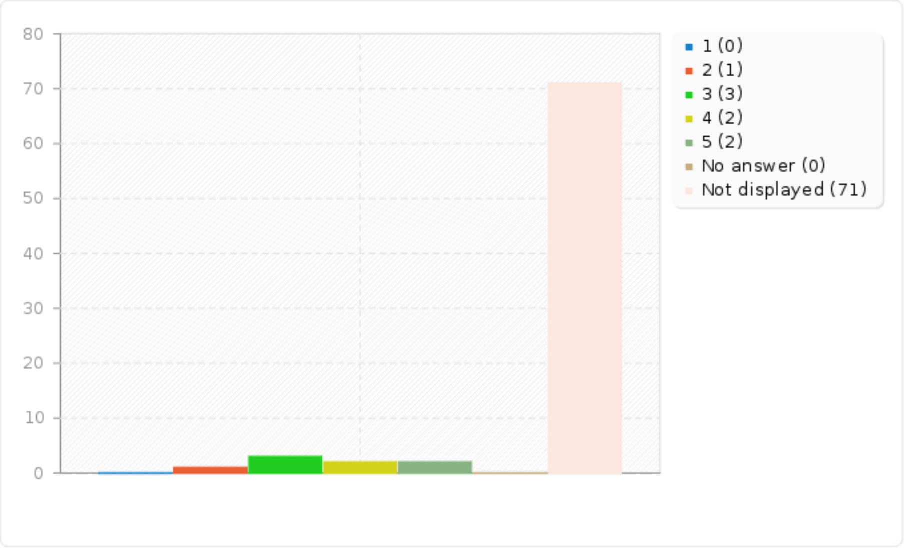


(N=15/79)

When asked about suppliers, the few responses indicated SDL (the Passolo package), Atril, and Google as the main tools.

Given the level of integration, we also expected a low level of interest in coming years. We received 8 answers, out of which 2 indicated high and 2 very high interest for future use.

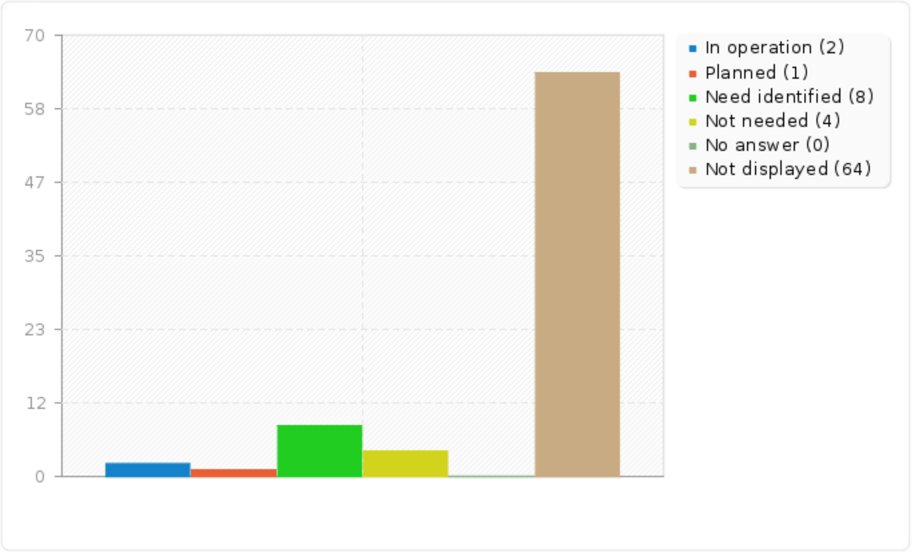
Figure 141 Interest to use localisation tools applied to websites in the future



2. Localisation tools applied to software

We did not anticipate a high level of response for this area, given that public bodies' interest in software localisation is generally low. Out of the 15 replies received, only 2 indicated this is in operation, the respondents are archiving houses which develop in-house applications for their own use. No suppliers are listed.

Figure 142 Status of use of localisation tools applied to software



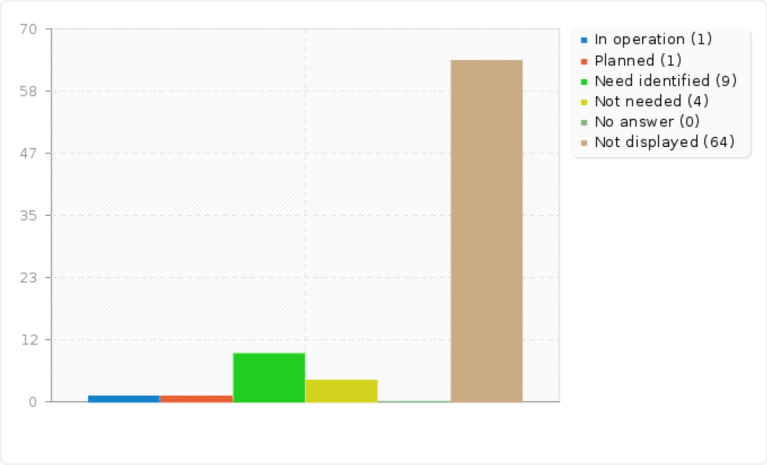
(N=15/79)

The level of interest is also very low (only 3 responses indicate high/very-high interest out of 8).

3. Localisation tools applied to forms

We estimated a large number of responses as many administrations operate through the use of web forms to collect structured information. However, only one response indicated that such tool is in operation.

Figure 143 Status of use of localisation tools applied to forms



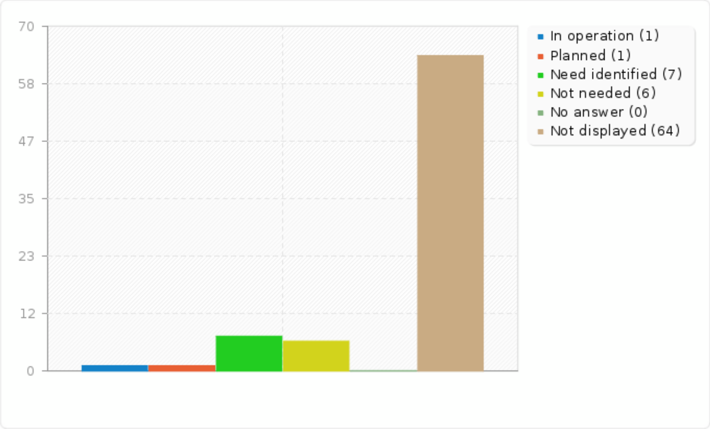
(N=15/79)

The level of interest is also low, 2 responses indicating high interest and 2 very high.

4. Localisation tools applied to subtitling/dubbing production

Here we had anticipated that there are some users of audio data who would also be interested in subtitling or dubbing their content. The only “in-operation” response came from an academic body which is using it in the teaching environment, and 7 respondents have already identified a need for such applications.

Figure 144 Status of use of localisation tools applied to subtitling/dubbing



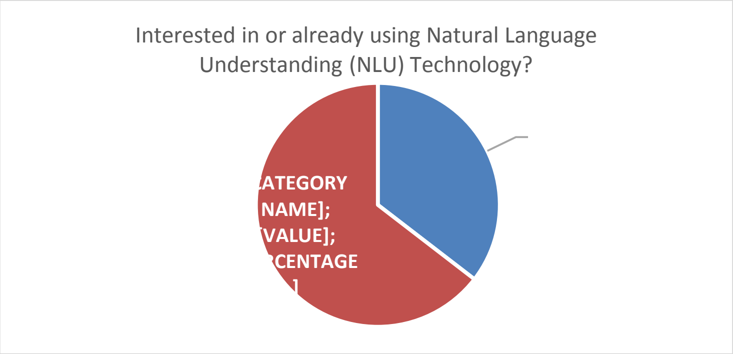
(N=15/79)

Similarly to the current implementation, the level of interest for future use is also very low, only 4 out of 7 responses indicate high or very high interest.

Natural Language Understanding (NLU) Technology

This section is related to the applications that exploit natural language understanding to implement human-machine interactions based on textual inputs.

Figure 145 Use or plans to use NLU



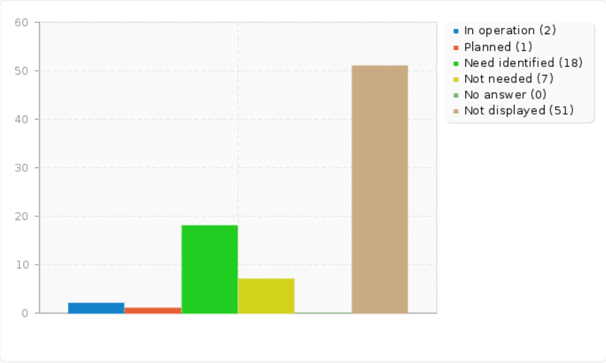
(N=79)

1. Chatbot/virtual assistant

Chatbots and virtual assistants are a hot topic in all areas of eCommerce and other Business to Consumer (B2C) services. We were looking forward to see how they are used or are of interest to the public sector. It is interesting to find that out of the 28 positive answers for NLU in general, only 2 indicated that such service is in operation (French National Railway Company and Finnish Patent and Registration Office).

The current implementation of these technologies is very low in our sample (2 in operation and 1 planned), although we see high interest with 18 respondents have identified their needs for this technology.

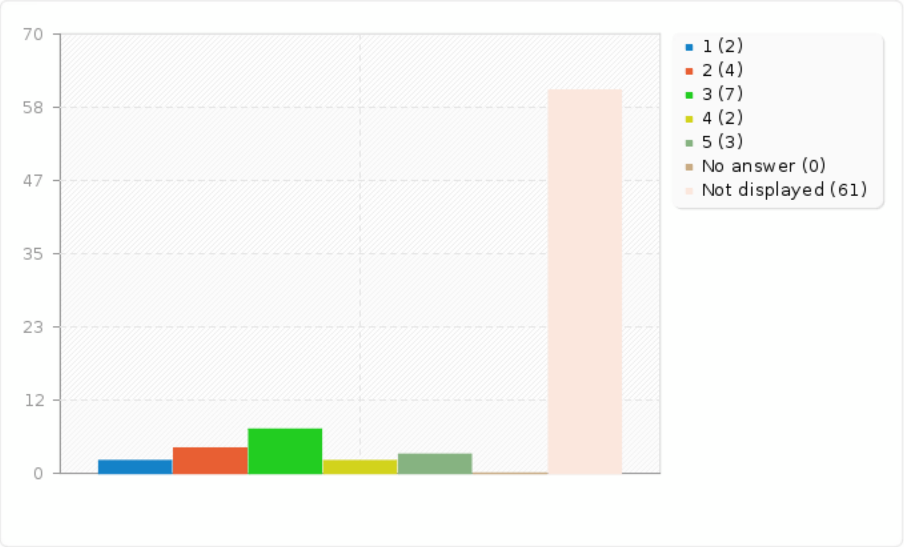
Figure 146 Status of use of chatbot/virtual assistant



(N=28/79)

Taking into account the 18 replies indicating an interest with identified needs, the level of interest is still low with having 5 high and very high interest.

Figure 147 Interest to use chatbot/virtual assistant in the future

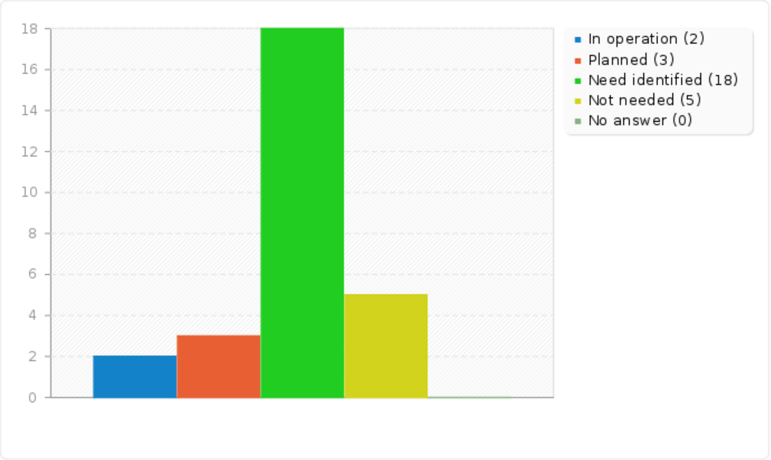


2. Keyword extractor

This component is rather technical but the technology suppliers surveyed indicated it as one of the heavily used technologies in their offer.

Interestingly, only 2 of the 28 participants indicated that such a tool is in operation in their administration and 3 that they plan to use such a tool. 18 have identified a need for it. The suppliers mentioned include Treetagger, TXM, Lexico3, Synomia, Viseo; or tailored solutions (provided by Tetracom Interactive solutions LTD). Keyword extractor modules are incorporated in many NLP systems. For example, TreeTagger is a part-of-speech tagger which assigns grammatical tags to words in a text.

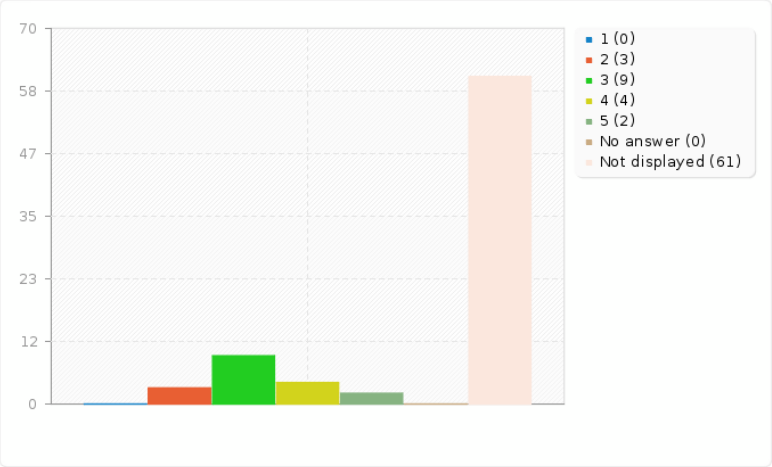
Figure 148 Status of use of keyword extraction tools



(N=28/79)

The level of interest is moderate on average (with 6 of 18 replies indicating high and very high interest and 9 moderate interest).

Figure 149 Interest to use keyword extraction tools in the future

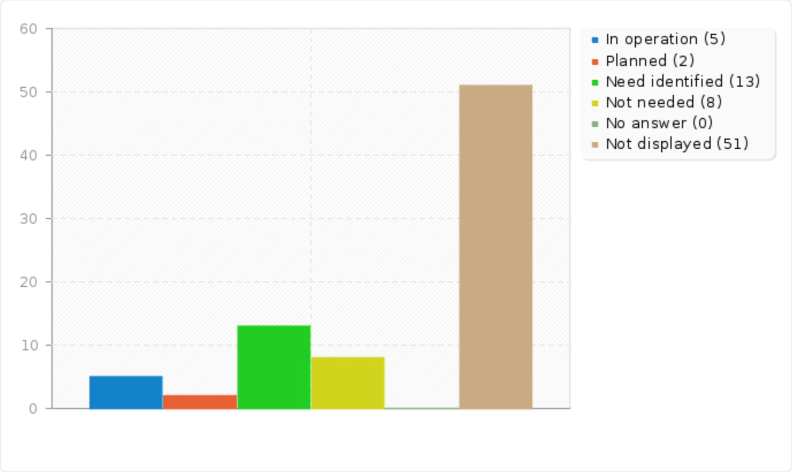


3. Topic modelling tools

This is also an important text analysis tool that allows modelling or extracting topics from textual material.

Although this is a technical module, 5 respondents point out it is already in use, for instance at the French Ministry for the Economy and Finances or at the National Library of Norway. 13 of the 28 responses indicate that needs have been identified. The few indications about suppliers list cooperation with academic partners who develop tailored solutions.

Figure 150 Status of use of topic modelling tools



(N=28/79)

Of the 10 replies on the level of interest for future use of these topic modelling tools, only 3 indicated high or very high interest. This reflects that this is a very specialised type of technology for general public administration use.

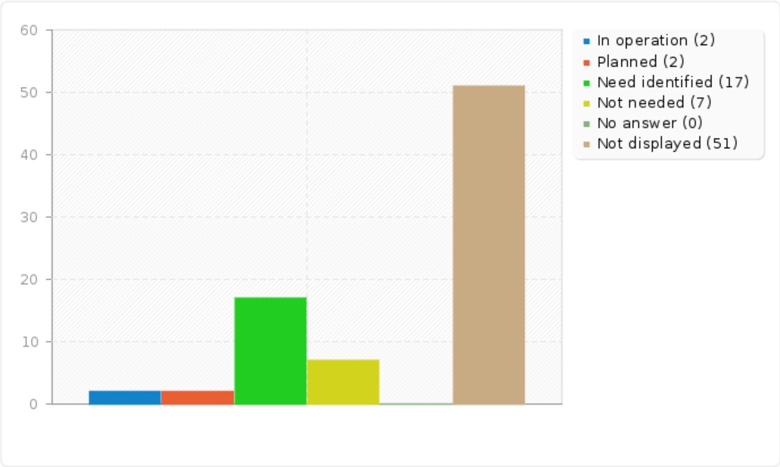
4. Automatic summarisation tools

Given the number of documents to analyse in public sector activities, we estimated this technology to be widely used, even though its maturity is still being debated.

Only two respondents (national language institutes) indicate the use of automatic summarisation tools and 2 indicated that it is planned. A very large number (17) replied that they identified it as needed while 7 that is not needed.

The list of suppliers include e.g. the Estonian language content collector *EstSum,* which also provides versions for Swedish etc.

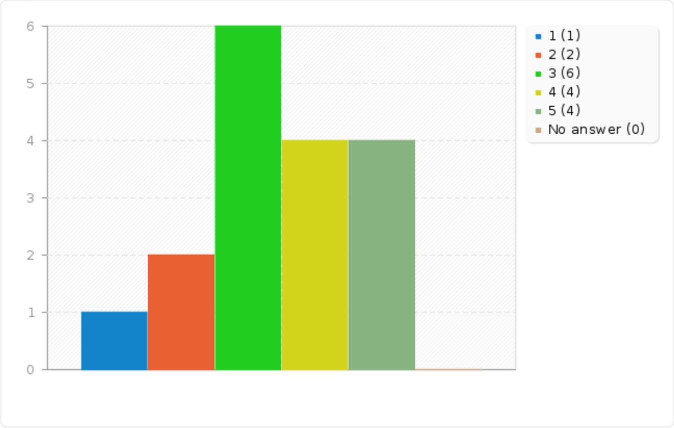
Figure 151 Status of use of automatic summarisation tools



(N=28/79)

As indicated above, many respondents reported about identified needs. When asked about their level of interest, 8 expressed high/very-high interest and 6 moderate. Only 3 pointed out low/very low interest.

Figure 152 Interest to use automatic summarisation in the future

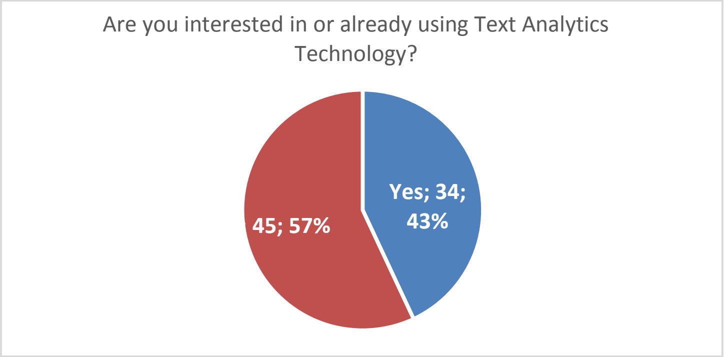


Analytics

Text analytics tools include text analysis and mining tools that covers text mining, sentiment analysis, text prediction, authorship attribution, etc.

Almost half of our respondents pointed out their interest in text analytics technology (34 out of 79).

Figure 153 Use or plans to use analytics



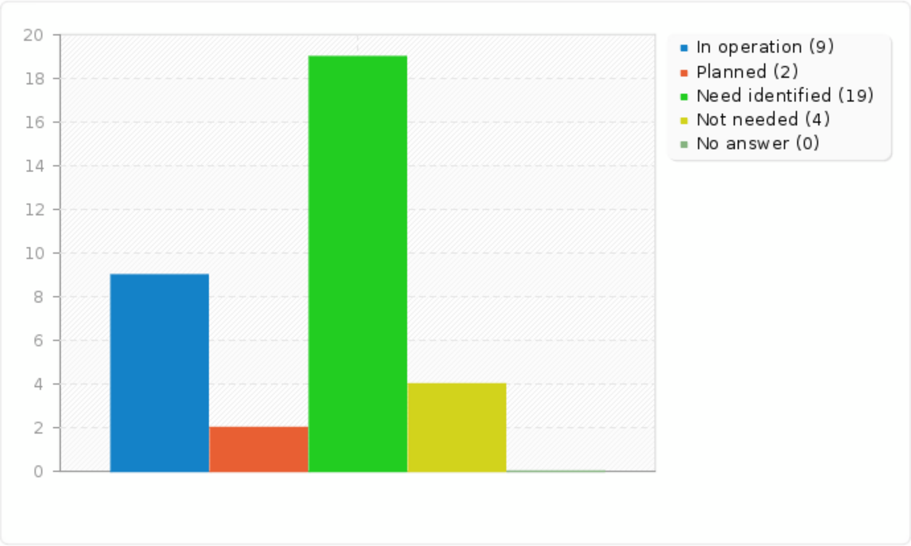
(N=79)

1. Text mining tools

Text mining tools allow extraction of information from textual material and we anticipated that this may be used by a number of ministries to identify specific texts and the associated relations.

The large number of positive responses on the use of analytics tools is in line with the responses about the level of incorporation of text mining tools in the respondents' activities. 9 indicated they are already in operation, these are mostly national archives and language institutes but also ministries. 2 indicated using these tools is planned and a very large number (19) pointed out that they identified their needs regarding text mining tools. Only 4 indicated that they have no need for them. A few suppliers are mentioned but most of the work is either internal or through service providers.

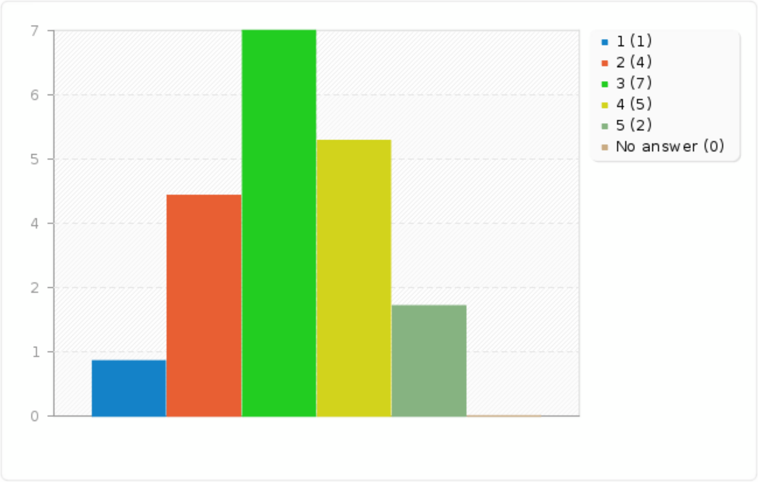
Figure 154 Status of use of text mining tools



(N=34/79)

Out of the 19 responses, 7 expressed high/very-high interest and 7 a moderate one, only 5 declared a very low interest.

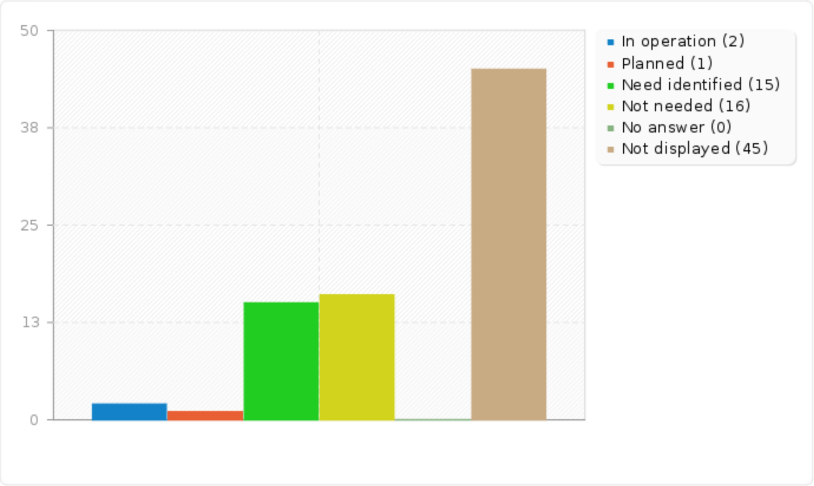
Figure 155 Interest to use text mining tools in the future



2. Sentiment analysis tools

Sentiment analysis tools identify sentiments, opinions in many streams of texts (e.g. Tweets) and other user-generated content (and are used in e.g. eReputation applications). The level of incorporation among the questionnaire's participants is rather low. 2 responses stated that an application is in use while 1 indicated it is planned. 15 indicated that the needs are identified and 16 that it is not needed. Only one supplier is mentioned (a software house that develops projects on demand).

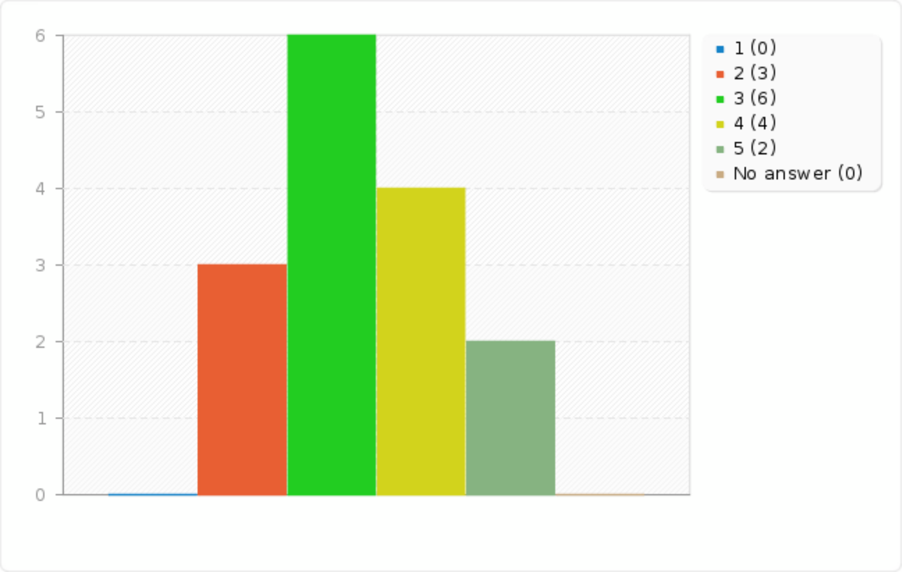
Figure 156 Status of use of sentiment analysis tools



(N=34/79)

The level of interest of those responding to this question is rather high, 6 indicate high/very-high interest and 6 moderate one, compared to 3 indicating low interest.

Figure 157 Interest to use sentiment analysis in the future



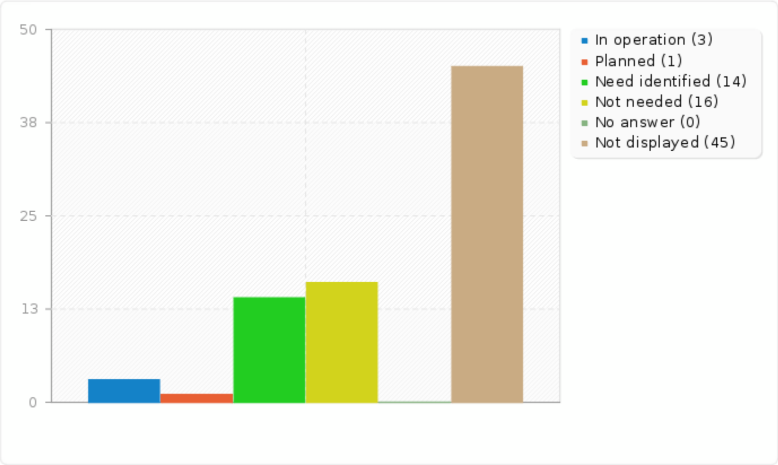
(N=15/79)

3. Text prediction tools

Text Prediction tools allow autocompletion and other efficient text input during text productions.

3 responses indicated that such application is in operation (and not only in language institutes) and 1 indicated that it is planned. 14 (out of the 34 responses) signaled that the needs are identified while another 16 indicate it is not needed. The indications on suppliers corroborated the information obtained through other questions (e.g. mention of the translation package MemSource).

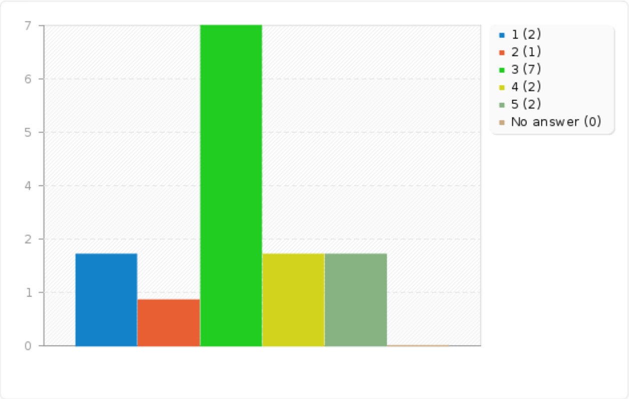
Figure 158 Status of use of text prediction tools



(N=34/79)

Despite the large number of respondents that have identified their needs, only very few expressed high/very-high interest (4 out 14). 7 signaled moderate interest and 3 low/very low interest.

Figure 159 Interest to use text prediction tools in the future



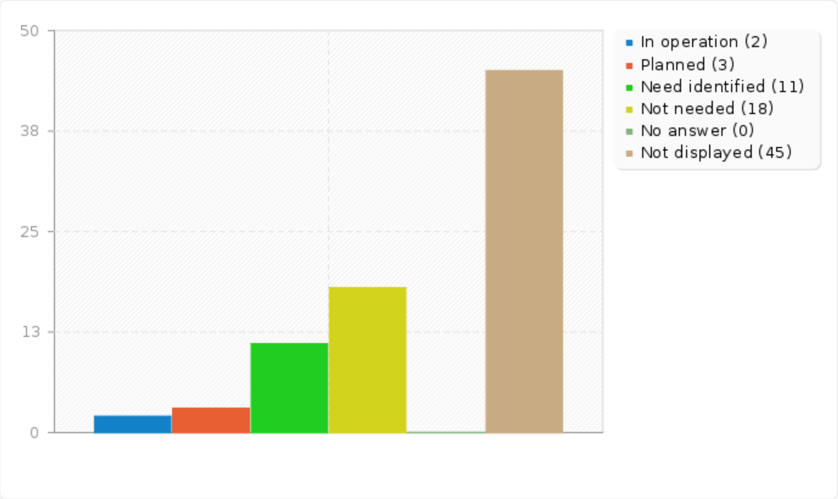
(N=14/79)

4. Authorship attribution tools

Authorship attribution tools identify the author of a given text and/or assign a text to the given author (e.g. it allows to detect plagiarism). They are usually applied by services handling publications (e.g. education institutions or national libraries).

Surprisingly this is not used as widely as expected, despite its certified maturity today. Only 2 respondents, both from language institutes, confirmed that these tools are operational at their end. 3 indicated that it is planned and they are also connected to national language organisations (libraries, archive, etc.). 11 signaled that the needs are identified but more than half respondent do not need it. The only supplier mentioned by the respondents is a private company offering a plagiarism detection system.

Figure 160 Status of use of authorship attribution tools



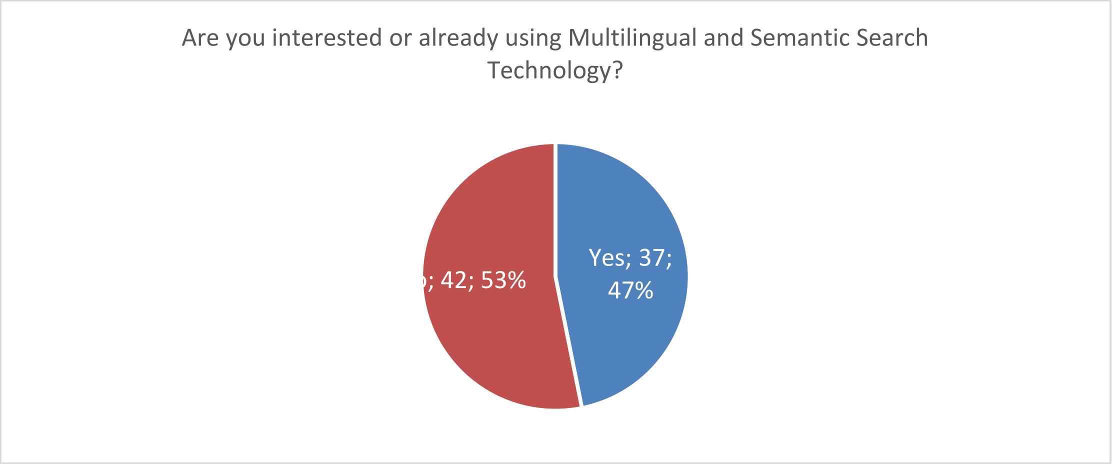
(N=34/79)

Only 11 expressed some kind of interest and out of these, 2 indicated high/very high, 4 moderate and 4 low/very low interest.

Multilingual and Semantic Search Technology

This is one of the most widely spread language technologies and we expected it to be used by a large number of services to power web sites and other information and knowledge databases. It includes the traditional search engine but also the question-answering systems. It was mentioned by almost half the respondents (37 out of the 79).

Figure 161 Use or plans to use multilingual and semantic search technology



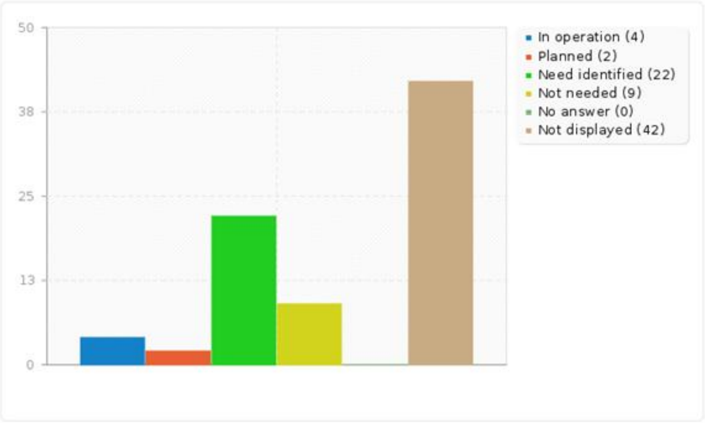
(N=79)

1. Question answering (QA) system

The QA systems go beyond the search and retrieval applications or the browsing of FAQs and use different NLP modules to build applications that automatically answer questions expressed by users in a natural language. These applications do not retrieve pages or sections from the internet but compile the information and generate short responses.

Very surprisingly, only 4 replies indicate it is in operation and 2 that it is planned. A very large number (22 out of 37) have identified the need for it. Only 9 indicate that it is not needed. No particular supplier is mentioned (use of internal tools, from the internet, or the SDL suite).

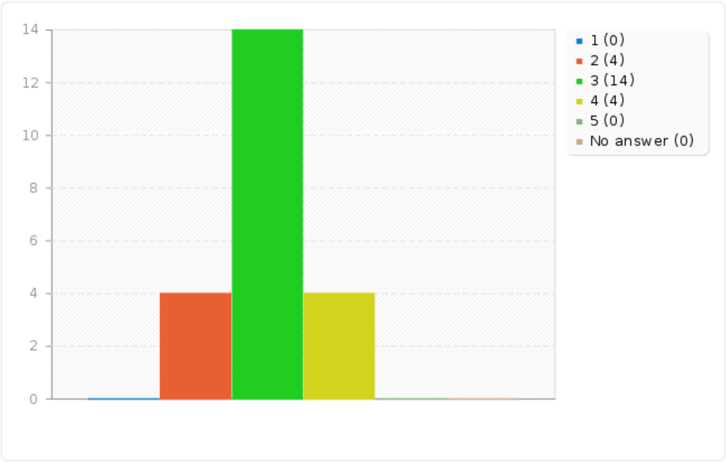
Figure 162 Status of use of question answering tools



(N=37/79)

Interestingly, the plans of coming implementations are very promising, according to the expression of interest. 4 indicated high level, 14 moderate and 4 low interest.

Figure 163 Interest to use question-answering technologies in the future



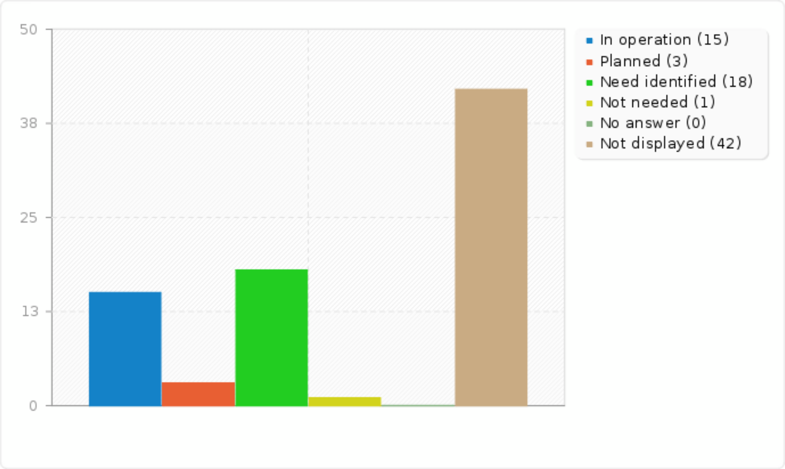
(N=22/79)

2. Search engine

Search engines are widely available for integration in web sites. They are often based on commercial products but also on open source packages provided by the research community and independent developers.

Out of the 37 responses, 15 indicated the use of a search engine and 3 the plans to use one. 18 have identified the needs and only 1 that is not needed. The list of suppliers includes the major players (Google Search, Microsoft Bing, Qwant, dtSearch, Wordbee, etc.), but also many open source packages (e.g. ElasticSearch) and internal tools tailored to the institution knowledge base (e.g. JocondeLab used at the Ministry of Culture in France).

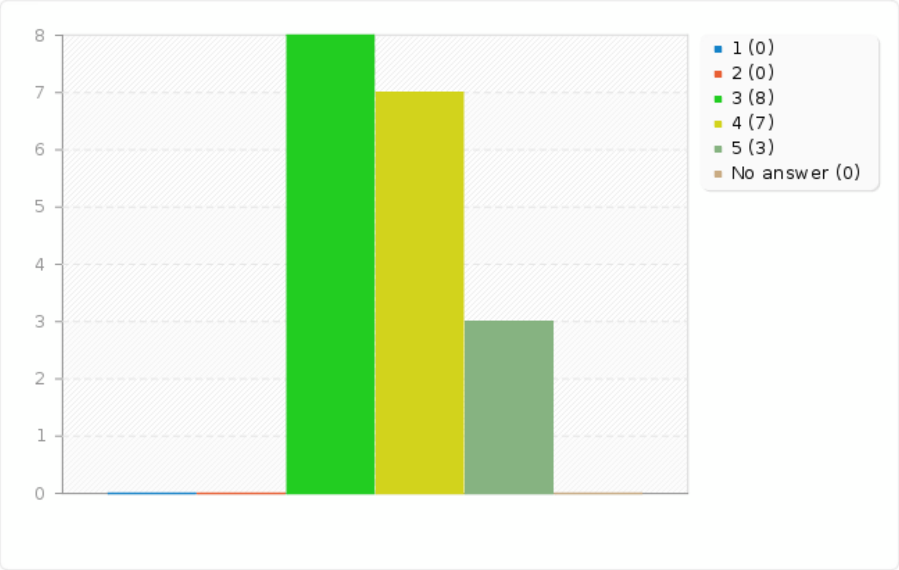
Figure 164 Status of use of search engines or tools



(N=37/79)

The level of interest is very high, as one may expect. 8 respondents selected level 3 (moderate) while 7 selected high and 3 very high.

Figure 165 Interest to use search engine technology in the future

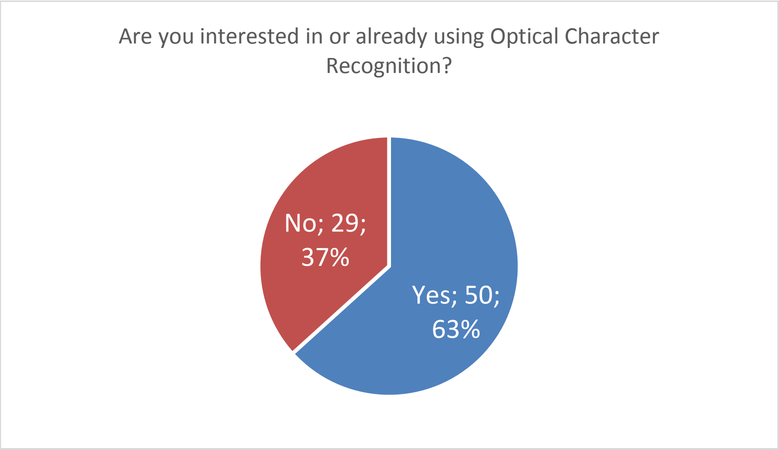


(N=18/79)

Optical character recognition

This technology is highly deployed by services for scanning, digitising and storing information as editable texts or even as forms and database entries.

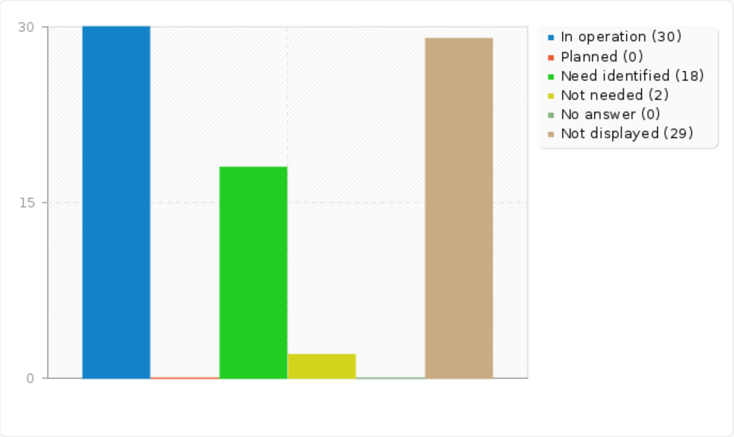
Figure 166 Use of or plan to use optical character recognition technologies



(N=79)

For this technology, the number of respondents indicating it is in use is very high (30 out of the 50 responses). 18 signaled that needs have been identified and only 2 stated they do not need it. For this technology, the market leaders are different from the traditional NLP players. The list of technology suppliers given by the respondents includes mostly commercial products from Abby, Nuance, Adobe, but also open source packages e.g. Tesseract or services like Jouve.

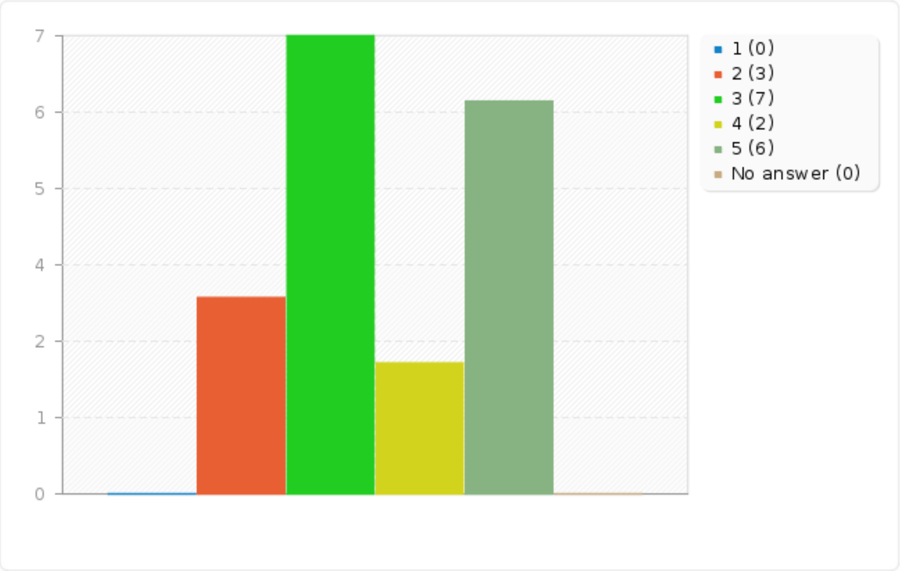
Figure 167 Status of use of optical character recognition technologies



(N=50/79)

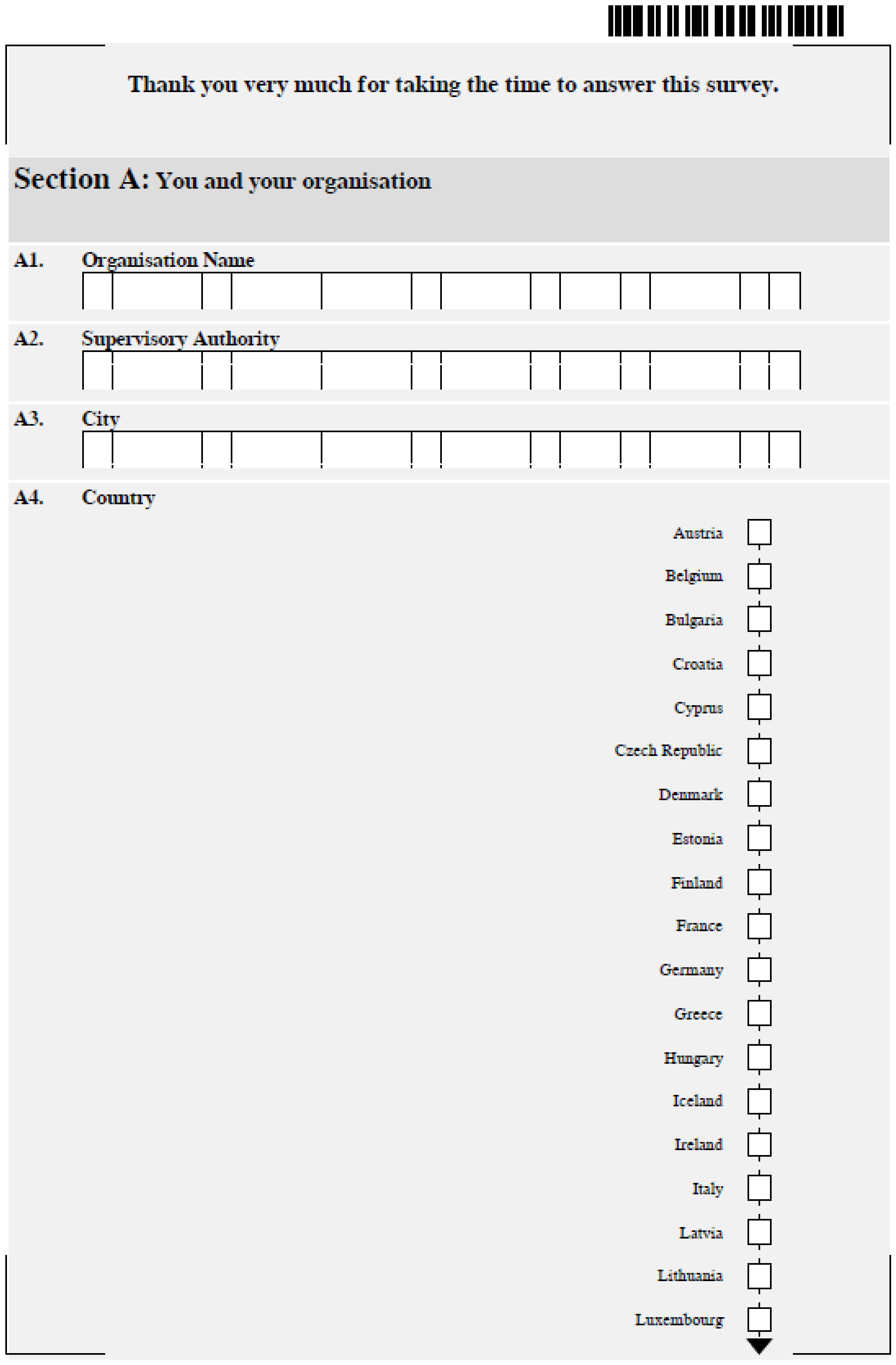
Out of the 18 respondents who expressed interest, 8 signaled high/very-high level and 7 a moderate one. Only 3 indicated low level of interest.

Figure 168 Interest to use optical character recognition technology in the future



O. Task 3: online questionnaire on LT adoption public services

The questionnaire is shown here as a number of pages.



Thank you very much for taking the time to answer this survey.

Section A: You and your organisation

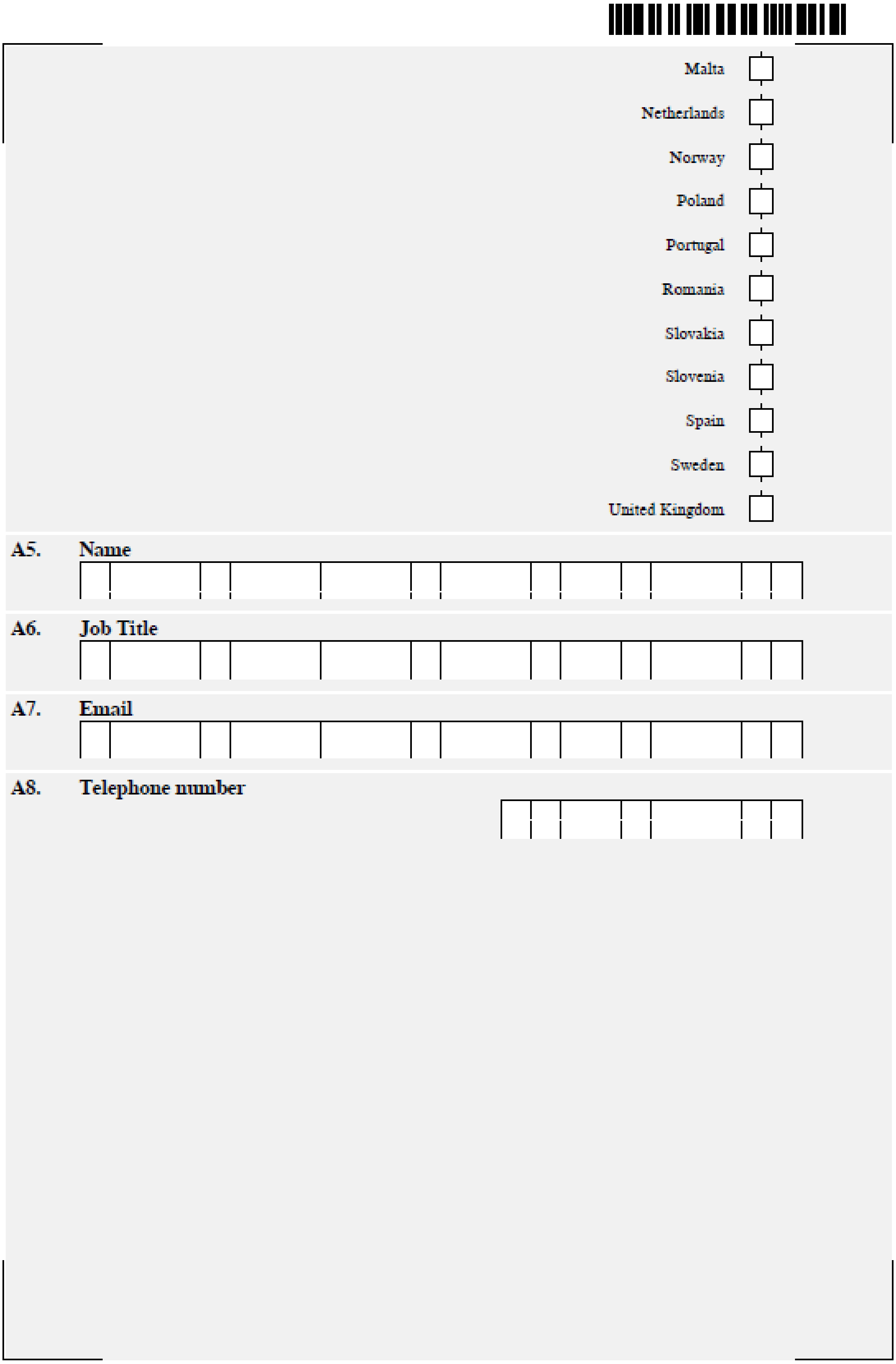
A1. Organisation Name

A2. Supervisory Authority

A3. City

A4. Country

- Austria □
- Belgium □
- Bulgaria □
- Croatia □
- Cyprus □
- Czech Republic □
- Denmark □
- Estonia □
- Finland □
- France □
- Germany □
- Greece □
- Hungary □
- Iceland □
- Ireland □
- Italy □
- Latvia □
- Lithuania □
- Luxembourg □



Malta □

Netherlands □

Norway □

Poland □

Portugal □

Romania □

Slovakia □

Slovenia □

Spain □

Sweden □

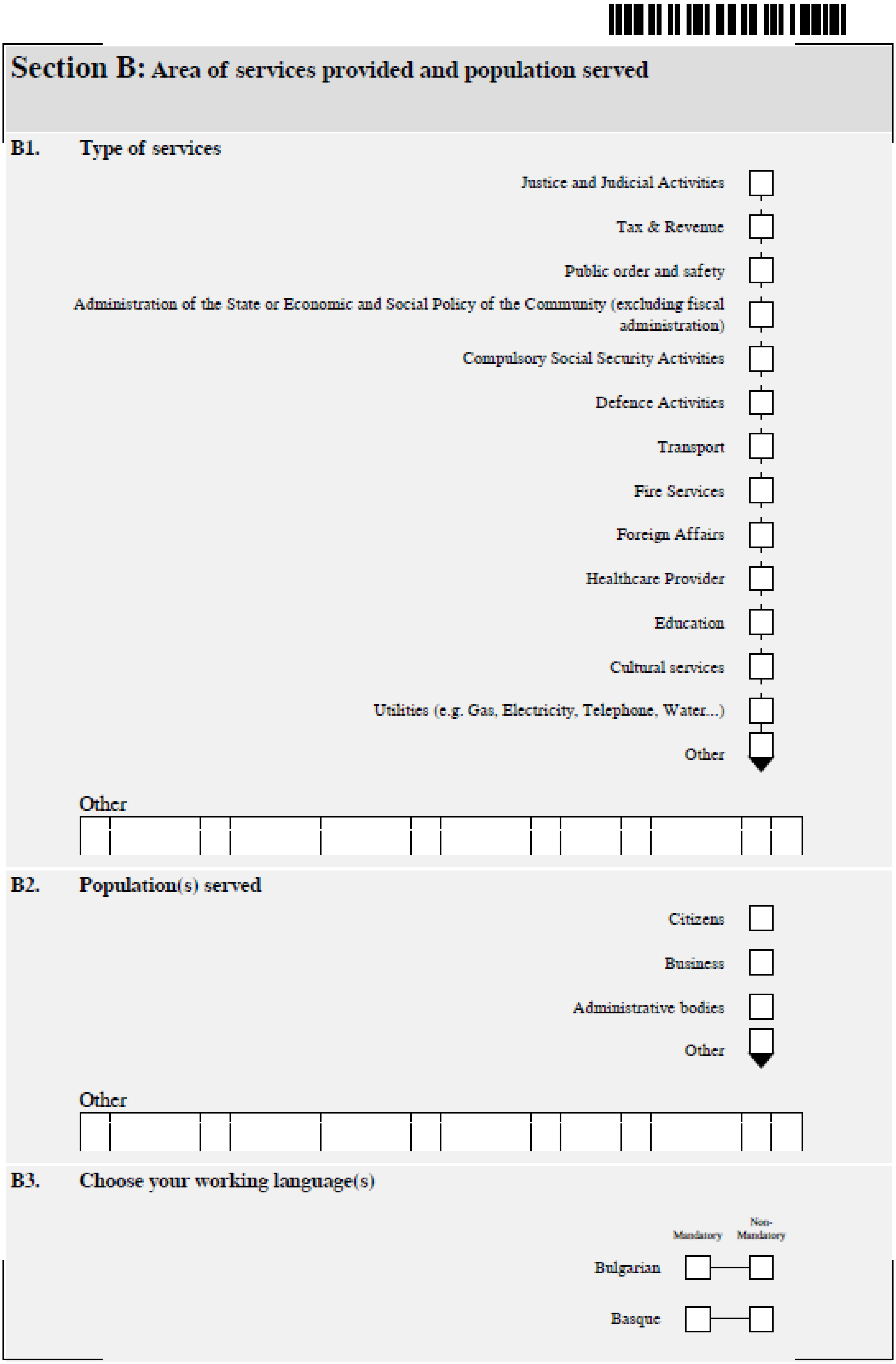
United Kingdom □

A5. **Name**

A6. **Job Title**

A7. **Email**

A8. **Telephone number**



Section B: Area of services provided and population served

B1. Type of services

Justice and Judicial Activities □

Tax & Revenue □

Public order and safety □

Administration of the State or Economic and Social Policy of the Community (excluding fiscal administration) □

Compulsory Social Security Activities □

Defence Activities □

Transport □

Fire Services □

Foreign Affairs □

Healthcare Provider □

Education □

Cultural services □

Utilities (e.g. Gas, Electricity, Telephone, Water...) □

Other □

Other

B2. Population(s) served

Citizens □

Business □

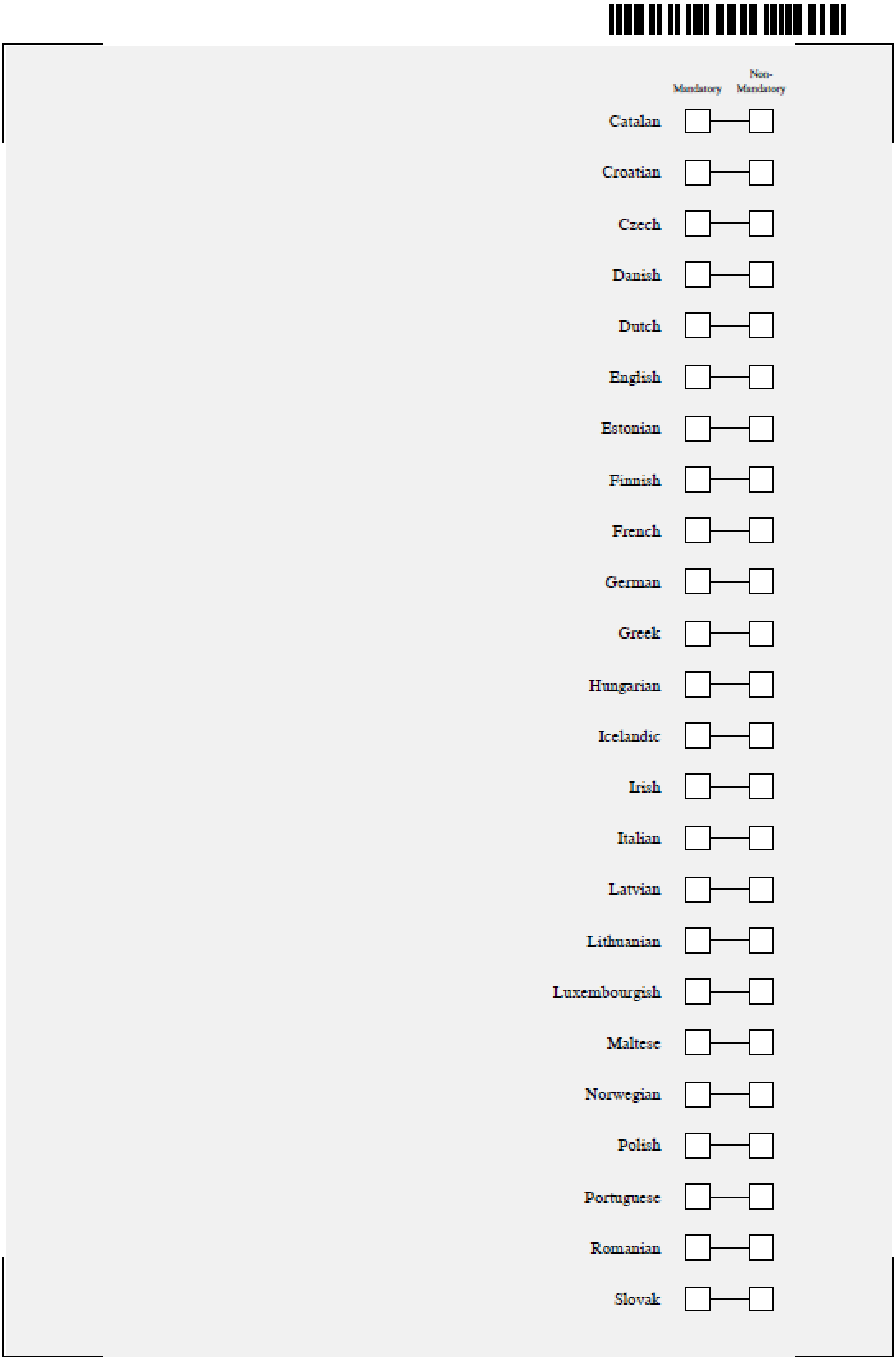
Administrative bodies □

Other □

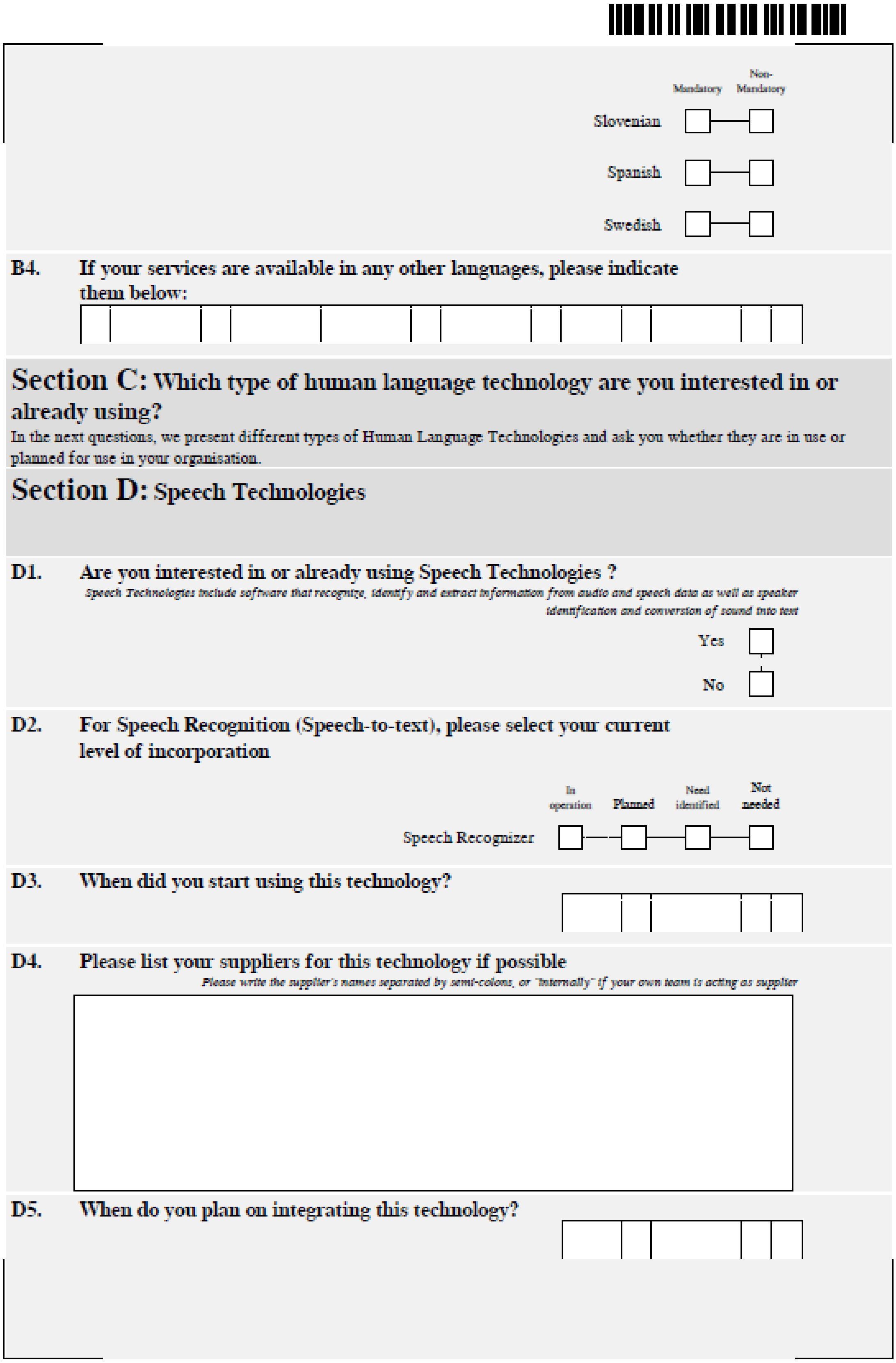
Other

B3. Choose your working language(s)

	Mandatory	Non-Mandatory
Bulgarian	□	□
Basque	□	□



	Mandatory	Non-Mandatory
Catalan	□	□
Croatian	□	□
Czech	□	□
Danish	□	□
Dutch	□	□
English	□	□
Estonian	□	□
Finnish	□	□
French	□	□
German	□	□
Greek	□	□
Hungarian	□	□
Icelandic	□	□
Irish	□	□
Italian	□	□
Latvian	□	□
Lithuanian	□	□
Luxembourgish	□	□
Maltese	□	□
Norwegian	□	□
Polish	□	□
Portuguese	□	□
Romanian	□	□
Slovak	□	□



	Mandatory	Non-Mandatory
Slovenian	□	□
Spanish	□	□
Swedish	□	□

B4. If your services are available in any other languages, please indicate them below:

Section C: Which type of human language technology are you interested in or already using?

In the next questions, we present different types of Human Language Technologies and ask you whether they are in use or planned for use in your organisation.

Section D: Speech Technologies

D1. Are you interested in or already using Speech Technologies ?

Speech Technologies include software that recognize, identify and extract information from audio and speech data as well as speaker identification and conversion of sound into text

Yes □

No □

D2. For Speech Recognition (Speech-to-text), please select your current level of incorporation

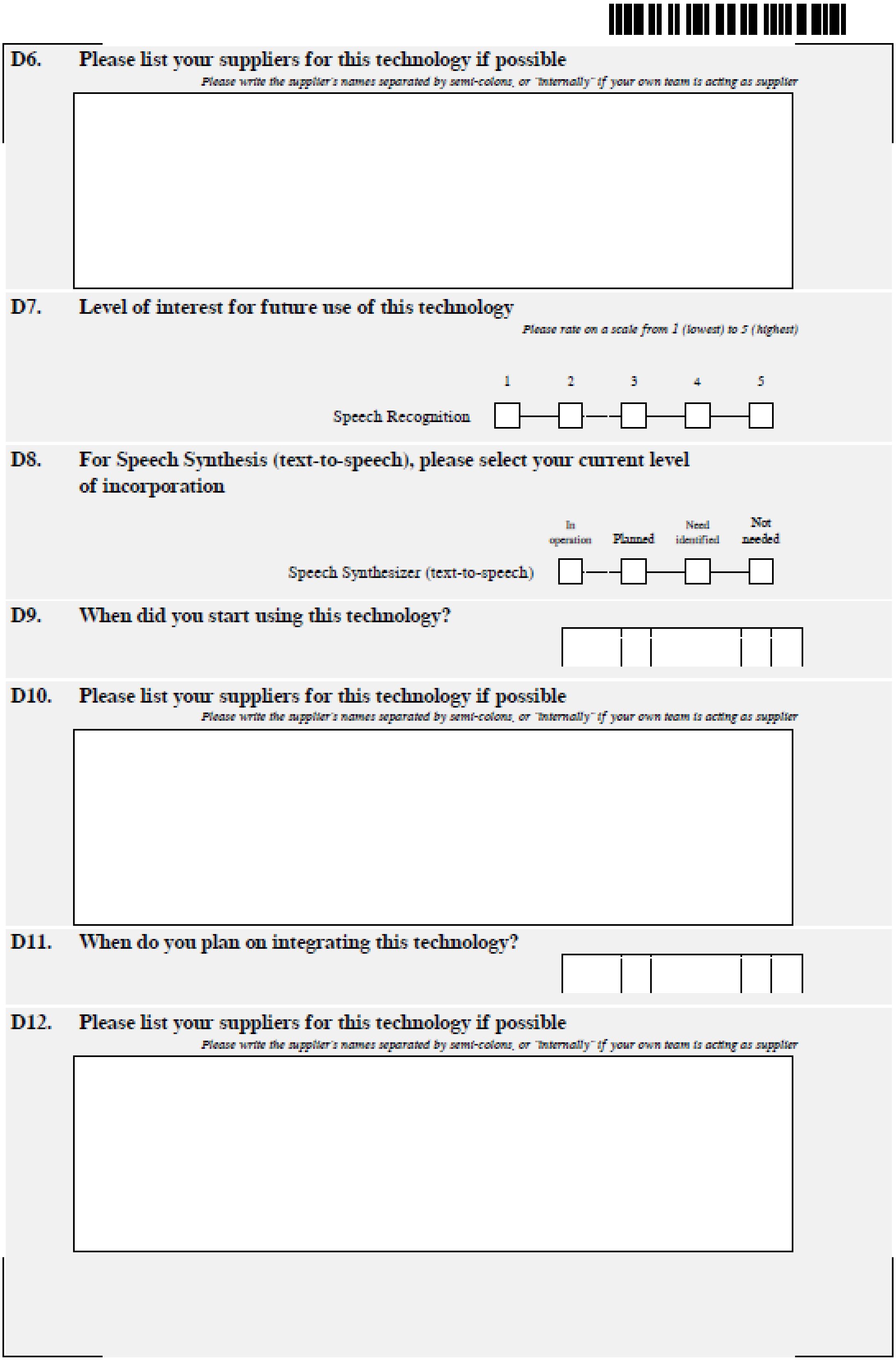
	In operation	Planned	Need identified	Not needed
Speech Recognizer	□	□	□	□

D3. When did you start using this technology?

D4. Please list your suppliers for this technology if possible

Please write the supplier's names separated by semi-colons, or "Internally" if your own team is acting as supplier

D5. When do you plan on integrating this technology?



D6. Please list your suppliers for this technology if possible

Please write the supplier's names separated by semi-colons, or "internally" if your own team is acting as supplier

D7. Level of interest for future use of this technology

Please rate on a scale from 1 (lowest) to 5 (highest)

	1	2	3	4	5
Speech Recognition	□	□	□	□	□

D8. For Speech Synthesis (text-to-speech), please select your current level of incorporation

	In operation	Planned	Need identified	Not needed
Speech Synthesizer (text-to-speech)	□	□	□	□

D9. When did you start using this technology?

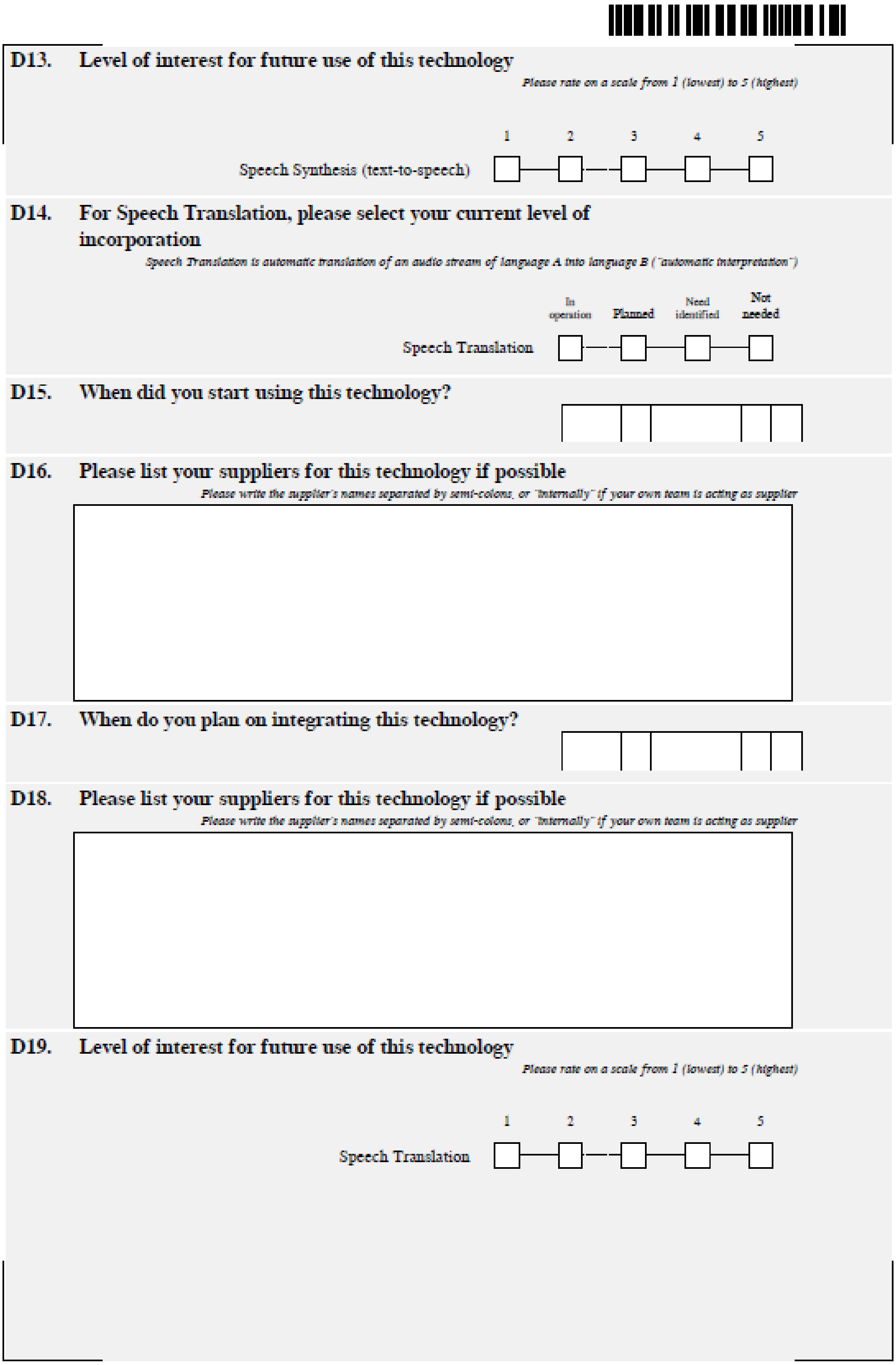
D10. Please list your suppliers for this technology if possible

Please write the supplier's names separated by semi-colons, or "internally" if your own team is acting as supplier

D11. When do you plan on integrating this technology?

D12. Please list your suppliers for this technology if possible

Please write the supplier's names separated by semi-colons, or "internally" if your own team is acting as supplier



D13. Level of interest for future use of this technology

Please rate on a scale from 1 (lowest) to 5 (highest)

	1	2	3	4	5
Speech Synthesis (text-to-speech)	□	□	□	□	□

D14. For Speech Translation, please select your current level of incorporation

Speech Translation is automatic translation of an audio stream of language A into language B ("automatic interpretation")

	In operation	Planned	Need identified	Not needed
Speech Translation	□	□	□	□

D15. When did you start using this technology?

□□ □□□□

D16. Please list your suppliers for this technology if possible

Please write the supplier's names separated by semi-colons, or "internally" if your own team is acting as supplier

D17. When do you plan on integrating this technology?

□□ □□□□

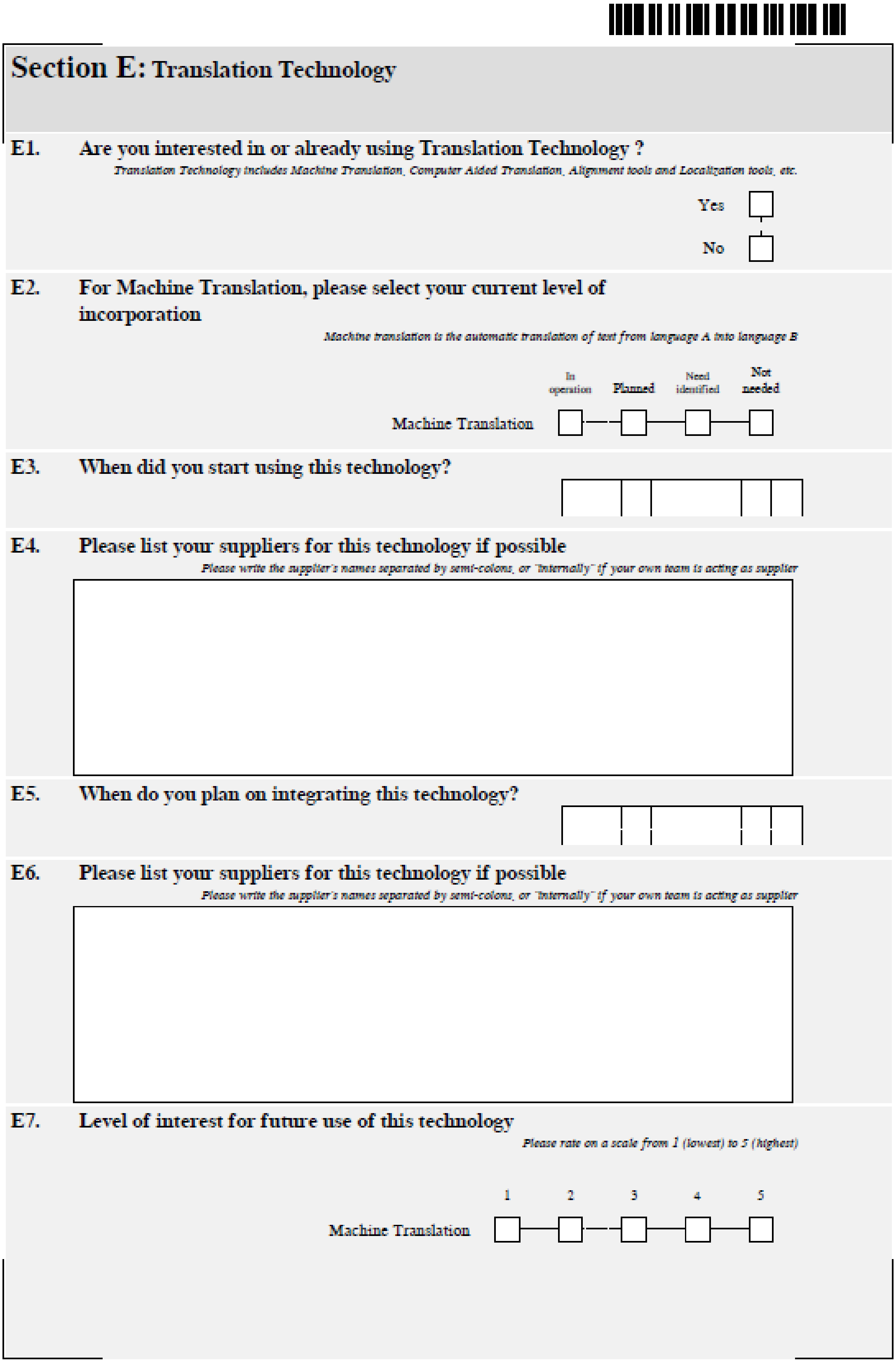
D18. Please list your suppliers for this technology if possible

Please write the supplier's names separated by semi-colons, or "internally" if your own team is acting as supplier

D19. Level of interest for future use of this technology

Please rate on a scale from 1 (lowest) to 5 (highest)

	1	2	3	4	5
Speech Translation	□	□	□	□	□



Section E: Translation Technology

E1. **Are you interested in or already using Translation Technology ?**

Translation Technology includes Machine Translation, Computer Aided Translation, Alignment tools and Localization tools, etc.

Yes □

No □

E2. **For Machine Translation, please select your current level of incorporation**

Machine translation is the automatic translation of text from language A into language B

	In operation	Planned	Need identified	Not needed
Machine Translation	□	□	□	□

E3. **When did you start using this technology?** □□/□□

E4. **Please list your suppliers for this technology if possible**

Please write the supplier's names separated by semi-colons, or "Internally" if your own team is acting as supplier

□

E5. **When do you plan on integrating this technology?** □□/□□

E6. **Please list your suppliers for this technology if possible**

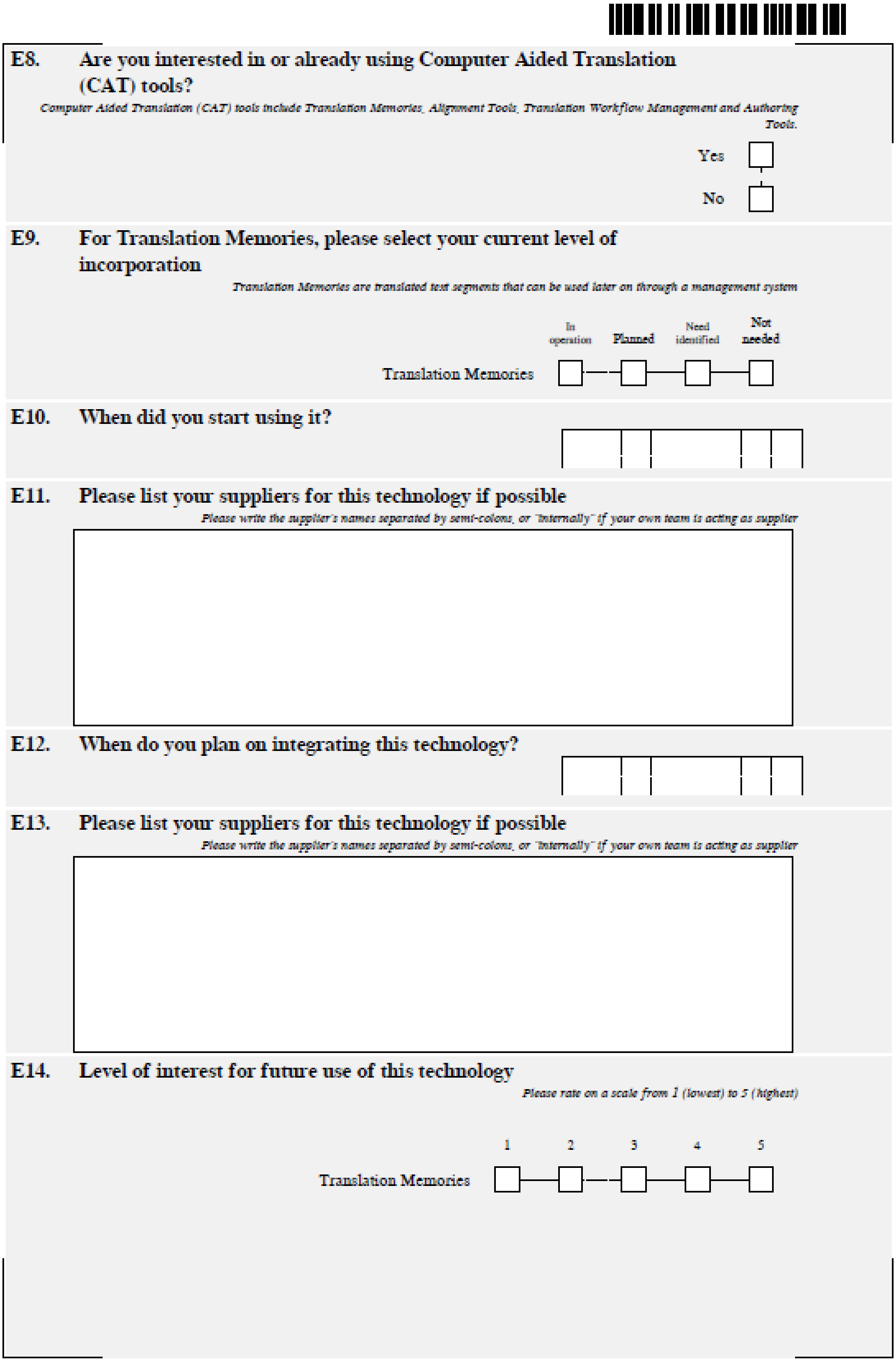
Please write the supplier's names separated by semi-colons, or "Internally" if your own team is acting as supplier

□

E7. **Level of interest for future use of this technology**

Please rate on a scale from 1 (lowest) to 5 (highest)

	1	2	3	4	5
Machine Translation	□	□	□	□	□



E8. Are you interested in or already using Computer Aided Translation (CAT) tools?

Computer Aided Translation (CAT) tools include Translation Memories, Alignment Tools, Translation Workflow Management and Authoring Tools.

Yes □

No □

E9. For Translation Memories, please select your current level of incorporation

Translation Memories are translated text segments that can be used later on through a management system

	In operation	Planned	Need identified	Not needed
Translation Memories	□	□	□	□

E10. When did you start using it?

E11. Please list your suppliers for this technology if possible

Please write the supplier's names separated by semi-colons, or "internally" if your own team is acting as supplier

E12. When do you plan on integrating this technology?

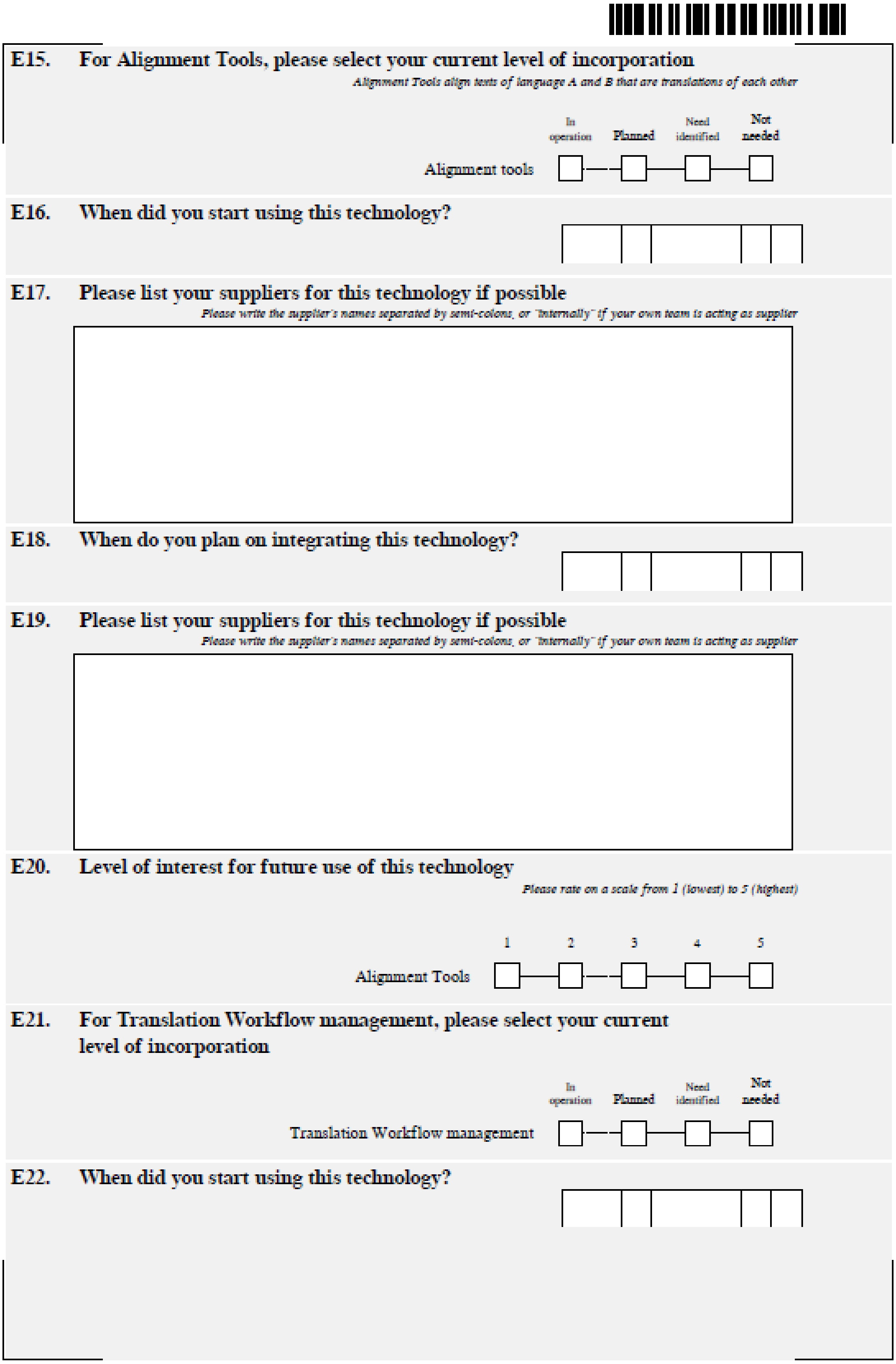
E13. Please list your suppliers for this technology if possible

Please write the supplier's names separated by semi-colons, or "internally" if your own team is acting as supplier

E14. Level of interest for future use of this technology

Please rate on a scale from 1 (lowest) to 5 (highest)

	1	2	3	4	5
Translation Memories	□	□	□	□	□



E15. For Alignment Tools, please select your current level of incorporation

Alignment Tools align texts of language A and B that are translations of each other

	In operation	Planned	Need identified	Not needed
Alignment tools	□	□	□	□

E16. When did you start using this technology?

E17. Please list your suppliers for this technology if possible

Please write the supplier's names separated by semi-colons, or "internally" if your own team is acting as supplier

E18. When do you plan on integrating this technology?

E19. Please list your suppliers for this technology if possible

Please write the supplier's names separated by semi-colons, or "internally" if your own team is acting as supplier

E20. Level of interest for future use of this technology

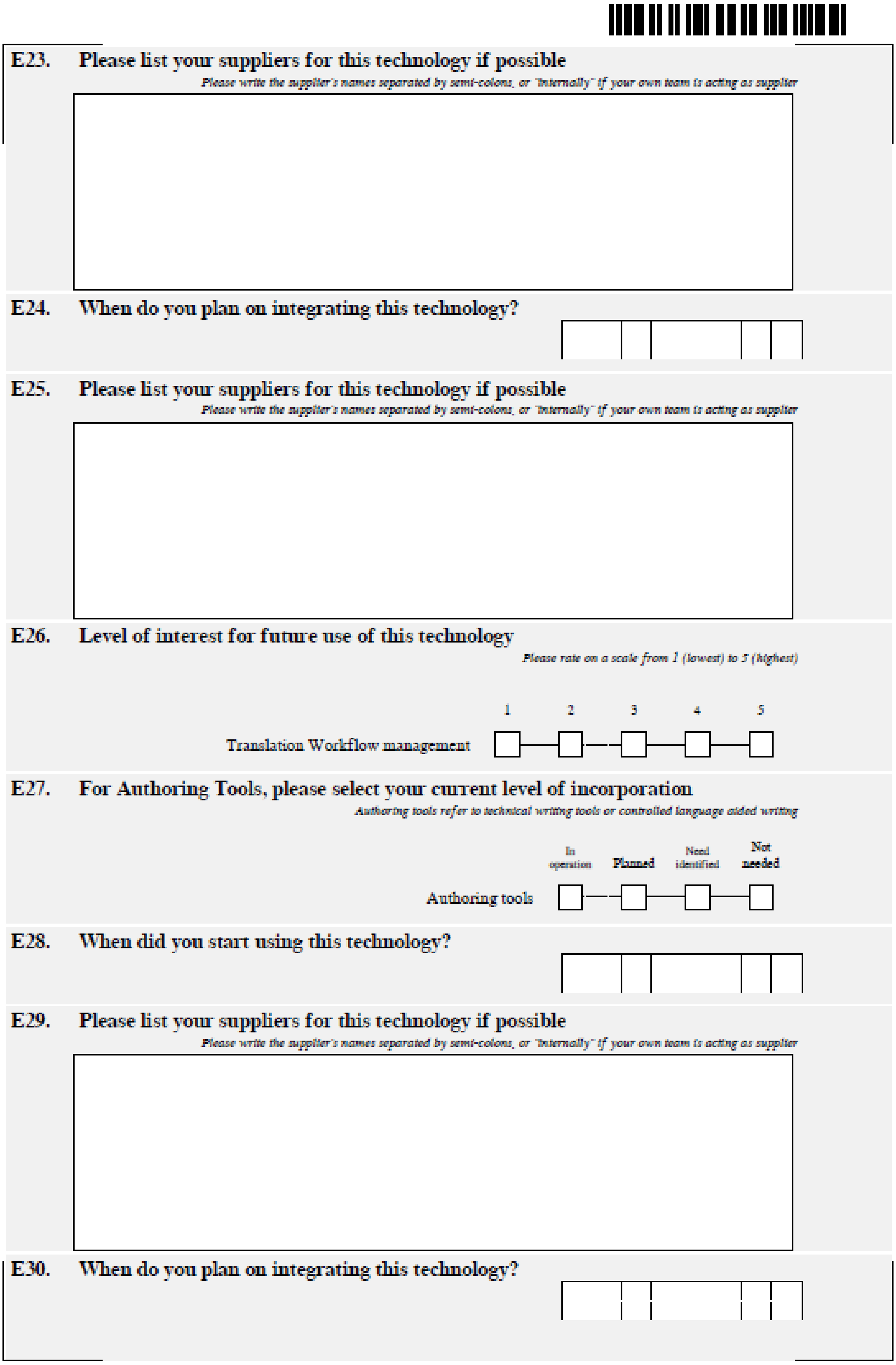
Please rate on a scale from 1 (lowest) to 5 (highest)

	1	2	3	4	5
Alignment Tools	□	□	□	□	□

E21. For Translation Workflow management, please select your current level of incorporation

	In operation	Planned	Need identified	Not needed
Translation Workflow management	□	□	□	□

E22. When did you start using this technology?



E23. **Please list your suppliers for this technology if possible**

Please write the supplier's names separated by semi-colons, or "Internally" if your own team is acting as supplier

E24. **When do you plan on integrating this technology?**

E25. **Please list your suppliers for this technology if possible**

Please write the supplier's names separated by semi-colons, or "Internally" if your own team is acting as supplier

E26. **Level of interest for future use of this technology**

Please rate on a scale from 1 (lowest) to 5 (highest)

	1	2	3	4	5
Translation Workflow management	□	□	□	□	□

E27. **For Authoring Tools, please select your current level of incorporation**

Authoring tools refer to technical writing tools or controlled language aided writing

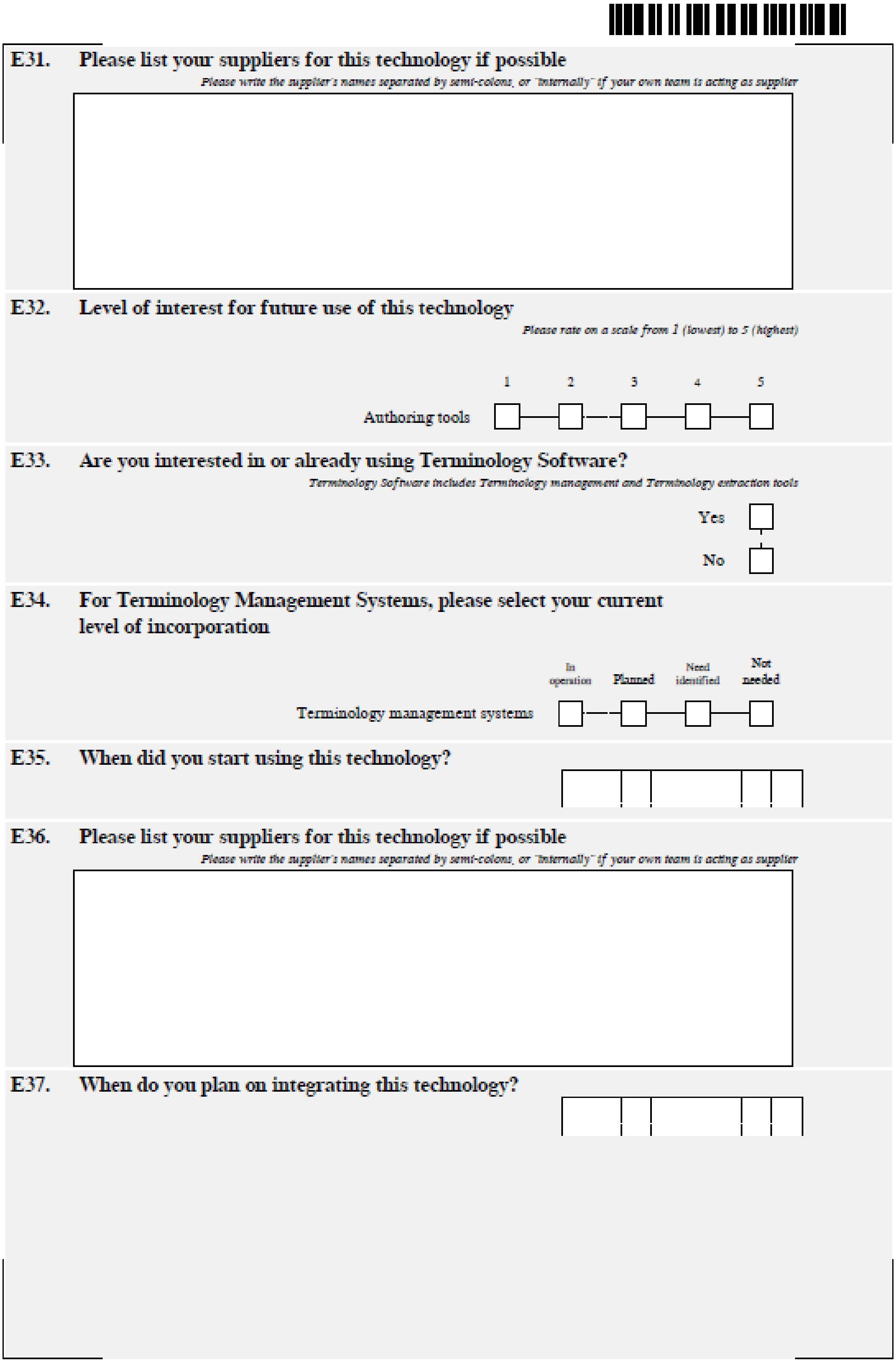
	In operation	Planned	Need identified	Not needed
Authoring tools	□	□	□	□

E28. **When did you start using this technology?**

E29. **Please list your suppliers for this technology if possible**

Please write the supplier's names separated by semi-colons, or "Internally" if your own team is acting as supplier

E30. **When do you plan on integrating this technology?**



E31. Please list your suppliers for this technology if possible

Please write the supplier's names separated by semi-colons, or "internally" if your own team is acting as supplier

E32. Level of interest for future use of this technology

Please rate on a scale from 1 (lowest) to 5 (highest)

	1	2	3	4	5
Authoring tools	□	□	□	□	□

E33. Are you interested in or already using Terminology Software?

Terminology Software includes Terminology management and Terminology extraction tools

Yes □

No □

E34. For Terminology Management Systems, please select your current level of incorporation

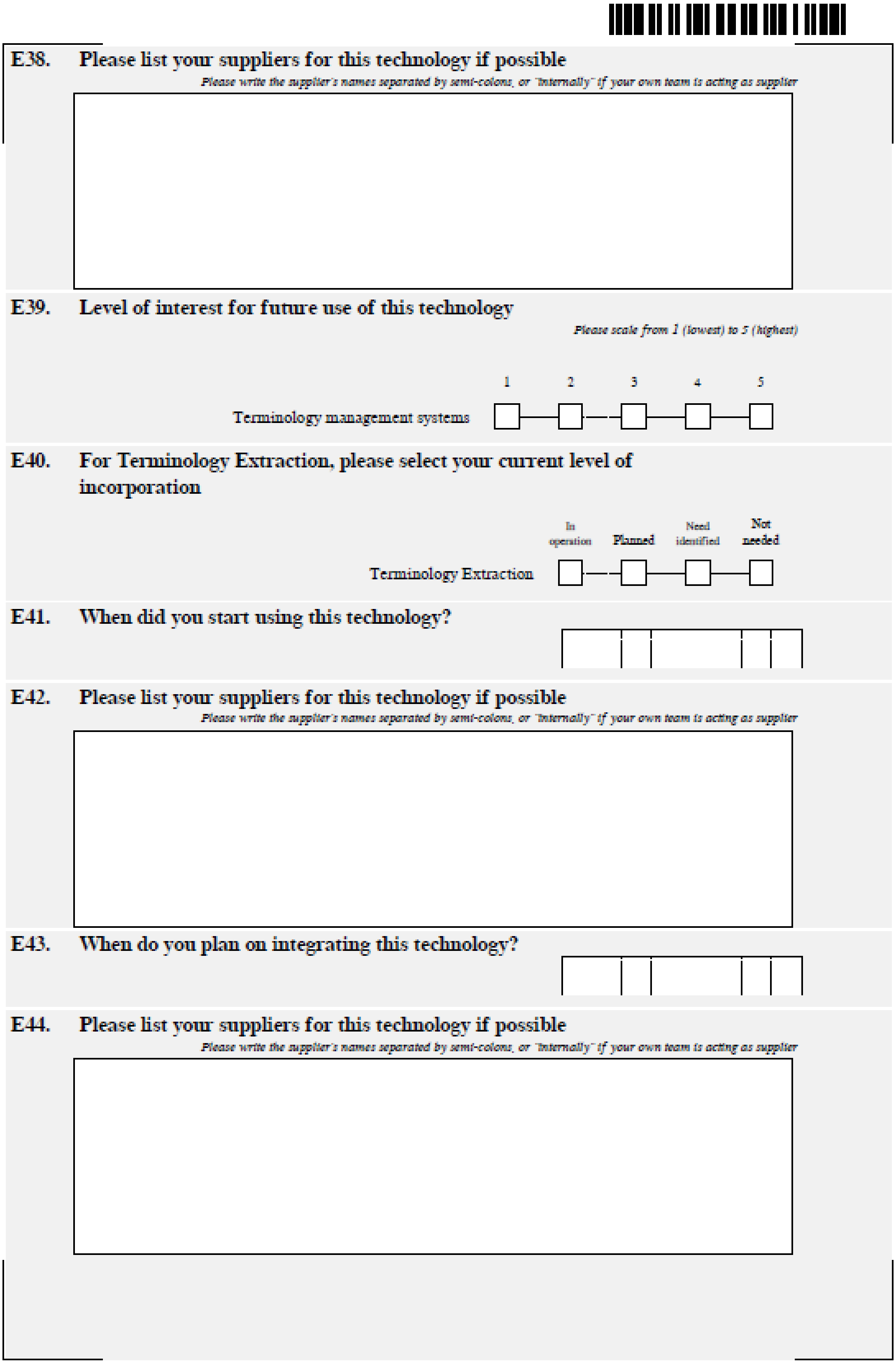
	In operation	Planned	Need identified	Not needed
Terminology management systems	□	□	□	□

E35. When did you start using this technology?

E36. Please list your suppliers for this technology if possible

Please write the supplier's names separated by semi-colons, or "internally" if your own team is acting as supplier

E37. When do you plan on integrating this technology?



E38. Please list your suppliers for this technology if possible

Please write the supplier's names separated by semi-colons, or "Internally" if your own team is acting as supplier

E39. Level of interest for future use of this technology

Please scale from 1 (lowest) to 5 (highest)

	1	2	3	4	5
Terminology management systems	□	□	□	□	□

E40. For Terminology Extraction, please select your current level of incorporation

	In operation	Planned	Need identified	Not needed
Terminology Extraction	□	□	□	□

E41. When did you start using this technology?

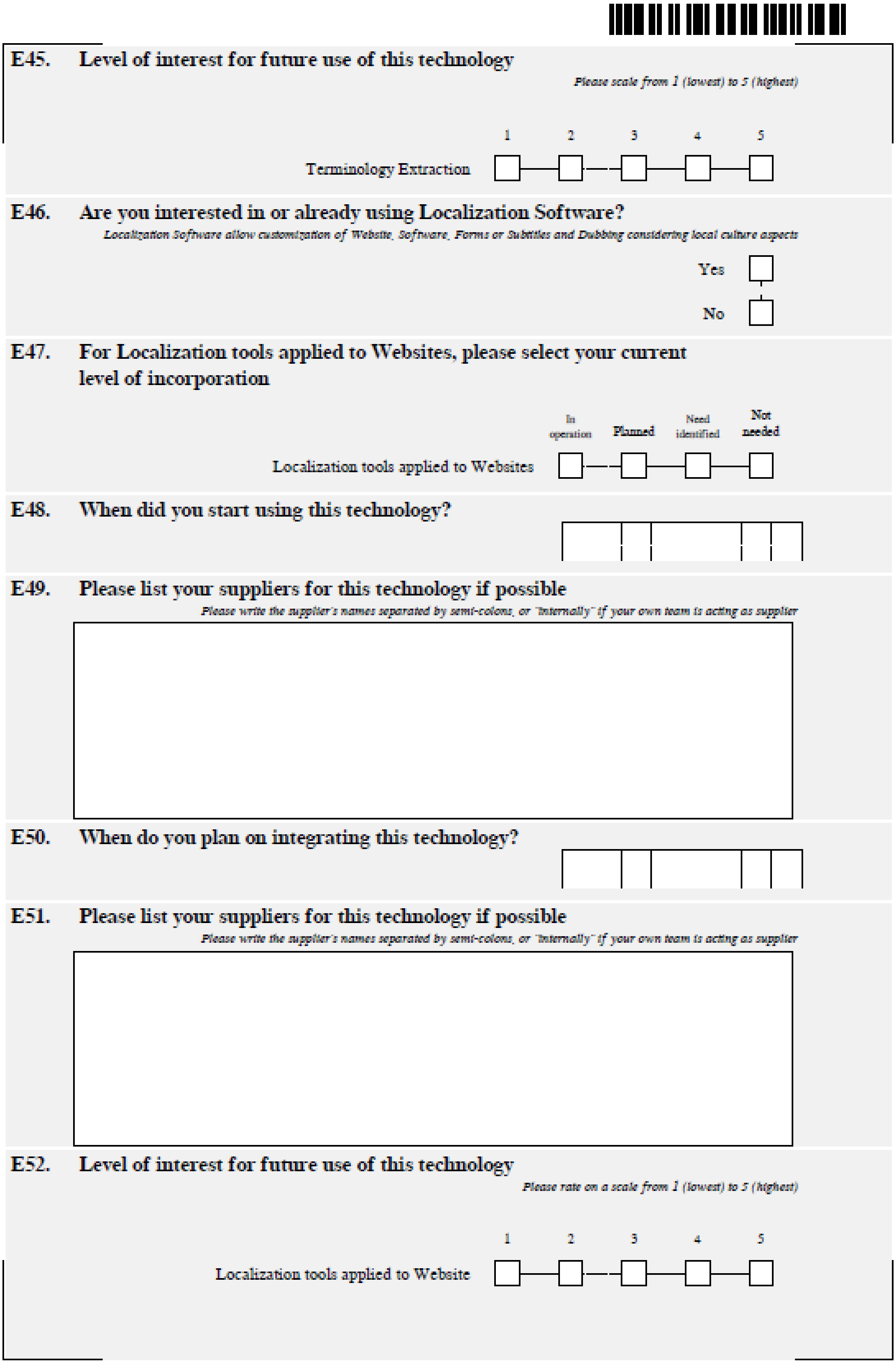
E42. Please list your suppliers for this technology if possible

Please write the supplier's names separated by semi-colons, or "Internally" if your own team is acting as supplier

E43. When do you plan on integrating this technology?

E44. Please list your suppliers for this technology if possible

Please write the supplier's names separated by semi-colons, or "Internally" if your own team is acting as supplier



E45. Level of interest for future use of this technology

Please scale from 1 (lowest) to 5 (highest)

	1	2	3	4	5
Terminology Extraction	□	□	□	□	□

E46. Are you interested in or already using Localization Software?

Localization Software allow customization of Website, Software, Forms or Subtitles and Dubbing considering local culture aspects

Yes □

No □

E47. For Localization tools applied to Websites, please select your current level of incorporation

	In operation	Planned	Need identified	Not needed
Localization tools applied to Websites	□	□	□	□

E48. When did you start using this technology?

E49. Please list your suppliers for this technology if possible

Please write the supplier's names separated by semi-colons, or "internally" if your own team is acting as supplier

E50. When do you plan on integrating this technology?

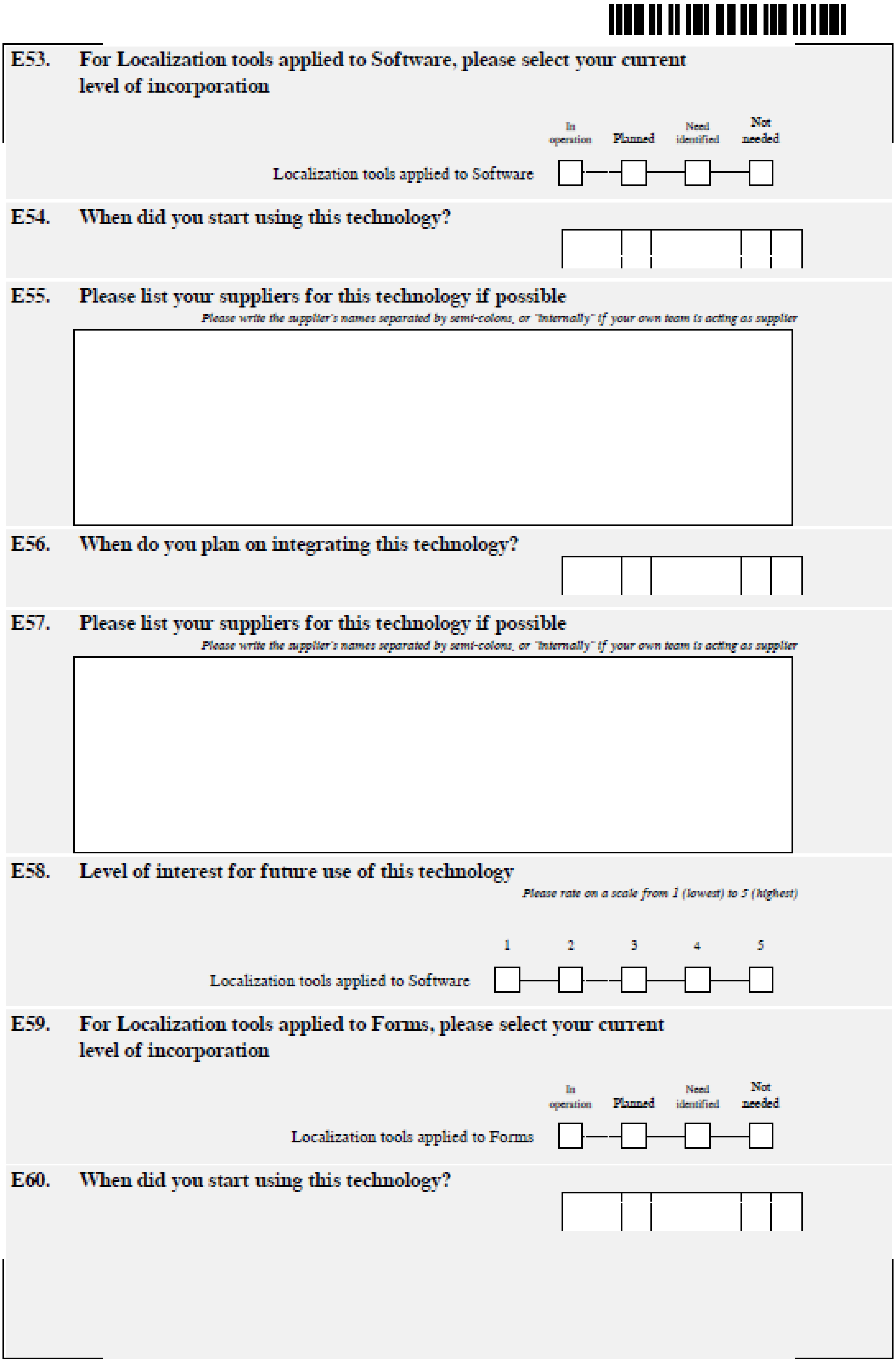
E51. Please list your suppliers for this technology if possible

Please write the supplier's names separated by semi-colons, or "internally" if your own team is acting as supplier

E52. Level of interest for future use of this technology

Please rate on a scale from 1 (lowest) to 5 (highest)

	1	2	3	4	5
Localization tools applied to Website	□	□	□	□	□



E53. For Localization tools applied to Software, please select your current level of incorporation

	In operation	Planned	Need identified	Not needed
Localization tools applied to Software	□	□	□	□

E54. When did you start using this technology?

E55. Please list your suppliers for this technology if possible

Please write the supplier's names separated by semi-colons, or "internally" if your own team is acting as supplier

E56. When do you plan on integrating this technology?

E57. Please list your suppliers for this technology if possible

Please write the supplier's names separated by semi-colons, or "internally" if your own team is acting as supplier

E58. Level of interest for future use of this technology

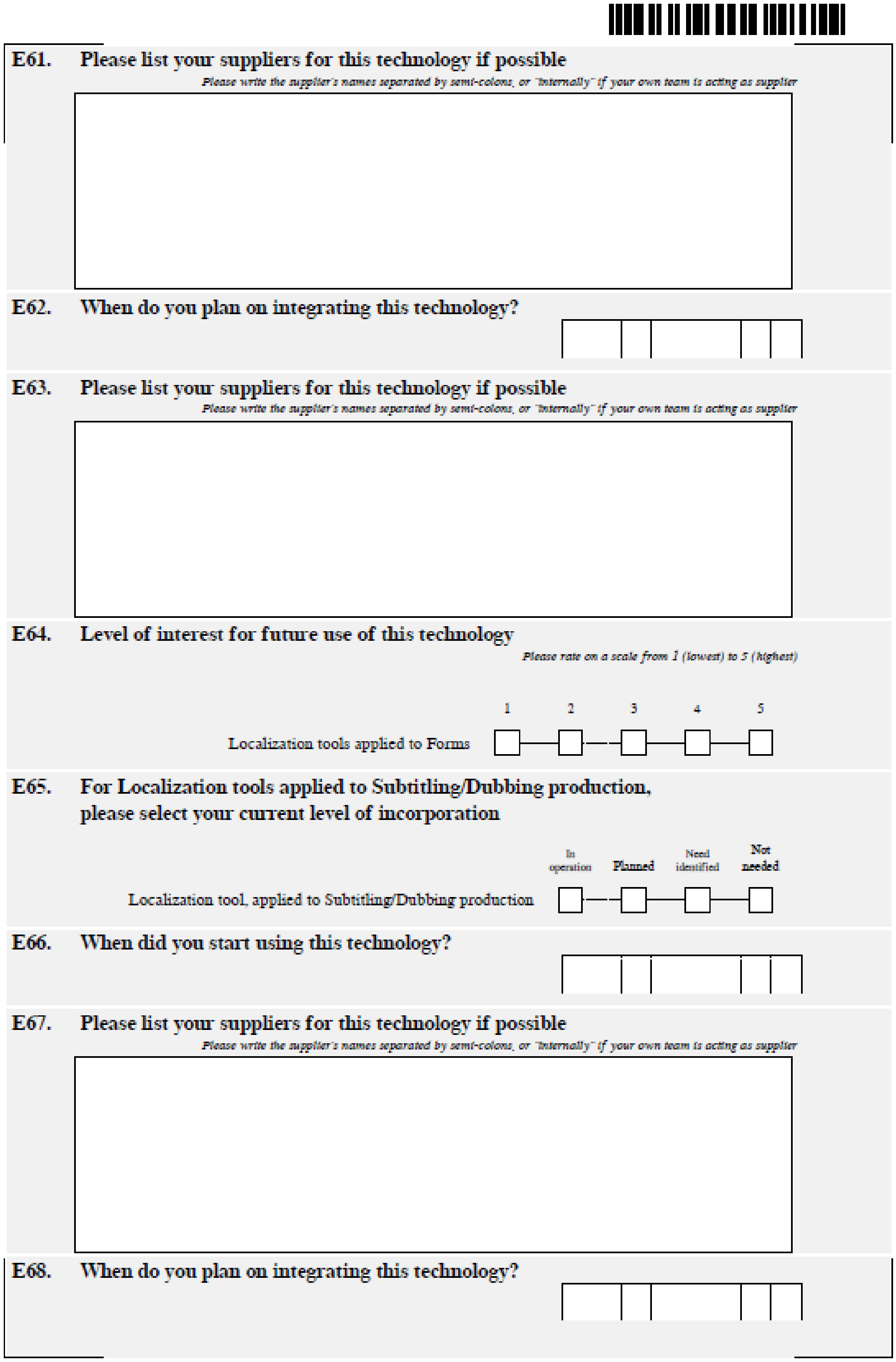
Please rate on a scale from 1 (lowest) to 5 (highest)

	1	2	3	4	5
Localization tools applied to Software	□	□	□	□	□

E59. For Localization tools applied to Forms, please select your current level of incorporation

	In operation	Planned	Need identified	Not needed
Localization tools applied to Forms	□	□	□	□

E60. When did you start using this technology?



E61. Please list your suppliers for this technology if possible

Please write the supplier's names separated by semi-colons, or "internally" if your own team is acting as supplier

□

E62. When do you plan on integrating this technology?

□□ / □□

E63. Please list your suppliers for this technology if possible

Please write the supplier's names separated by semi-colons, or "internally" if your own team is acting as supplier

□

E64. Level of interest for future use of this technology

Please rate on a scale from 1 (lowest) to 5 (highest)

	1	2	3	4	5
Localization tools applied to Forms	□	□	□	□	□

E65. For Localization tools applied to Subtitling/Dubbing production, please select your current level of incorporation

	In operation	Planned	Need identified	Not needed
Localization tool, applied to Subtitling/Dubbing production	□	□	□	□

E66. When did you start using this technology?

□□ / □□

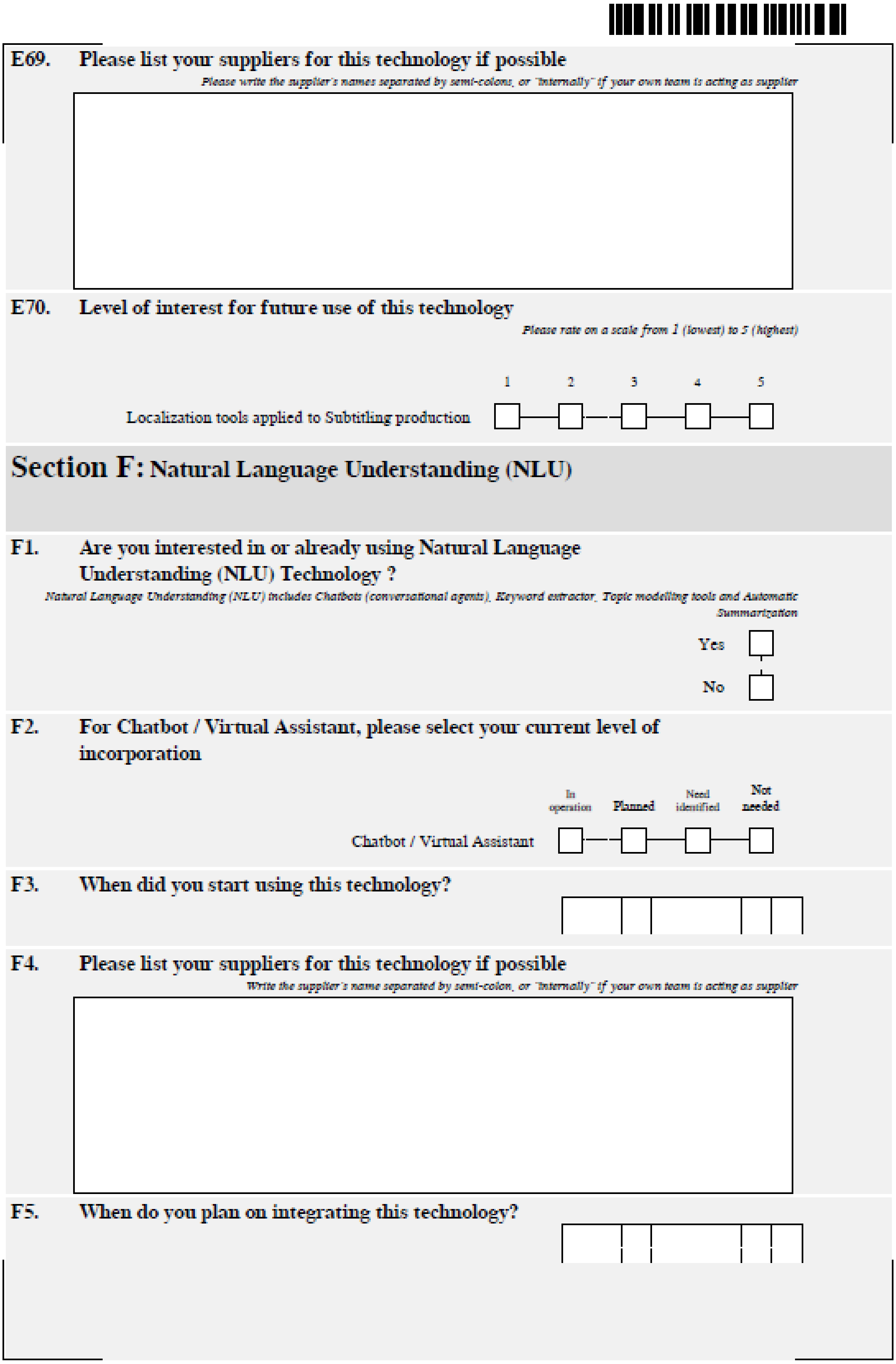
E67. Please list your suppliers for this technology if possible

Please write the supplier's names separated by semi-colons, or "internally" if your own team is acting as supplier

□

E68. When do you plan on integrating this technology?

□□ / □□



E69. **Please list your suppliers for this technology if possible**

Please write the supplier's names separated by semi-colons, or "Internally" if your own team is acting as supplier

E70. **Level of interest for future use of this technology**

Please rate on a scale from 1 (lowest) to 5 (highest)

	1	2	3	4	5
Localization tools applied to Subtitling production	□	□	□	□	□

Section F: Natural Language Understanding (NLU)

F1. **Are you interested in or already using Natural Language Understanding (NLU) Technology ?**

Natural Language Understanding (NLU) includes Chatbots (conversational agents), Keyword extractor, Topic modelling tools and Automatic Summarization

Yes □

No □

F2. **For Chatbot / Virtual Assistant, please select your current level of incorporation**

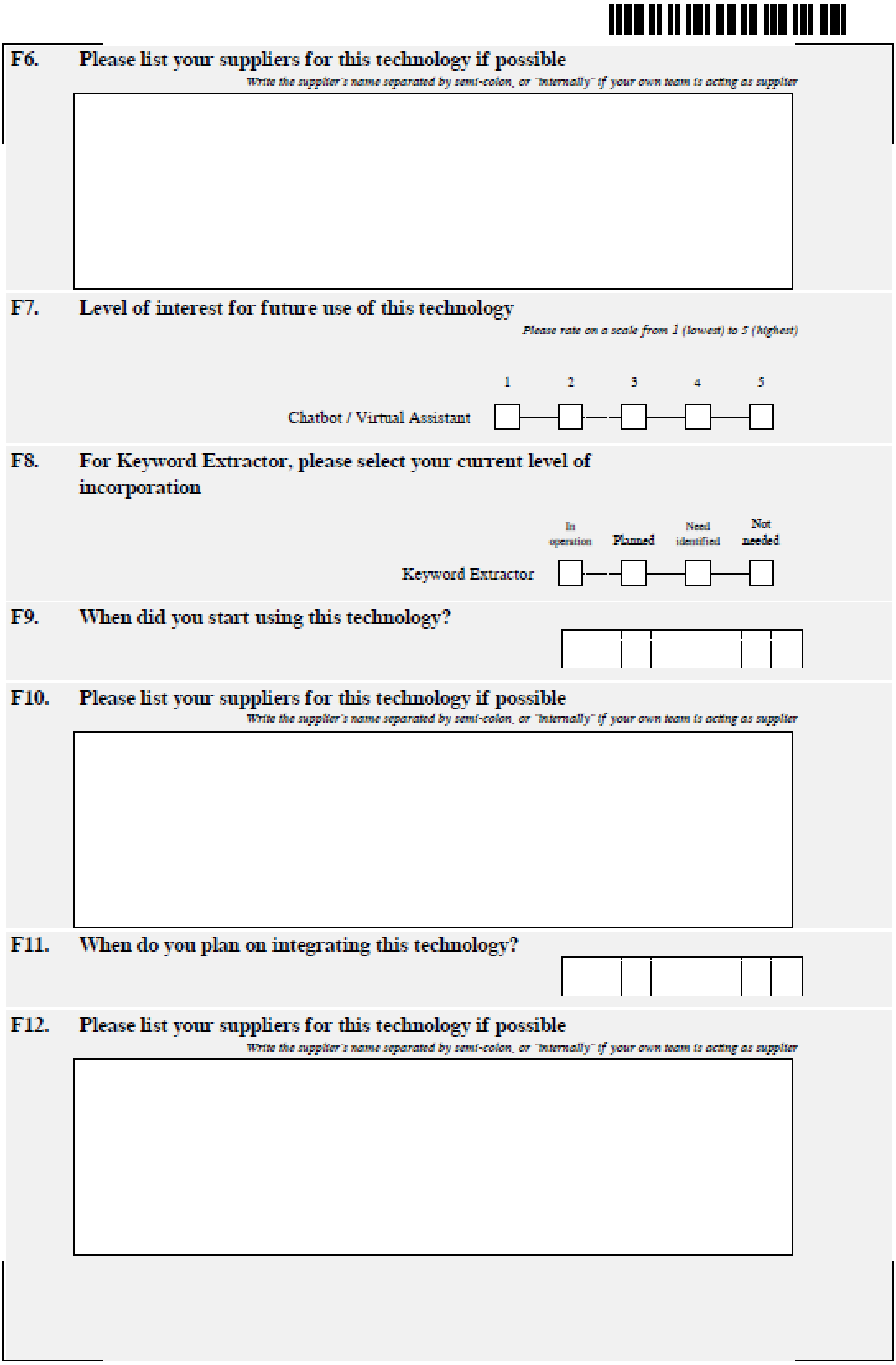
	In operation	Planned	Need identified	Not needed
Chatbot / Virtual Assistant	□	□	□	□

F3. **When did you start using this technology?**

F4. **Please list your suppliers for this technology if possible**

Write the supplier's name separated by semi-colon, or "Internally" if your own team is acting as supplier

F5. **When do you plan on integrating this technology?**



F6. Please list your suppliers for this technology if possible

Write the supplier's name separated by semi-colon, or "Internally" if your own team is acting as supplier

F7. Level of interest for future use of this technology

Please rate on a scale from 1 (lowest) to 5 (highest)

	1	2	3	4	5
Chatbot / Virtual Assistant	□	□	□	□	□

F8. For Keyword Extractor, please select your current level of incorporation

	In operation	Planned	Need identified	Not needed
Keyword Extractor	□	□	□	□

F9. When did you start using this technology?

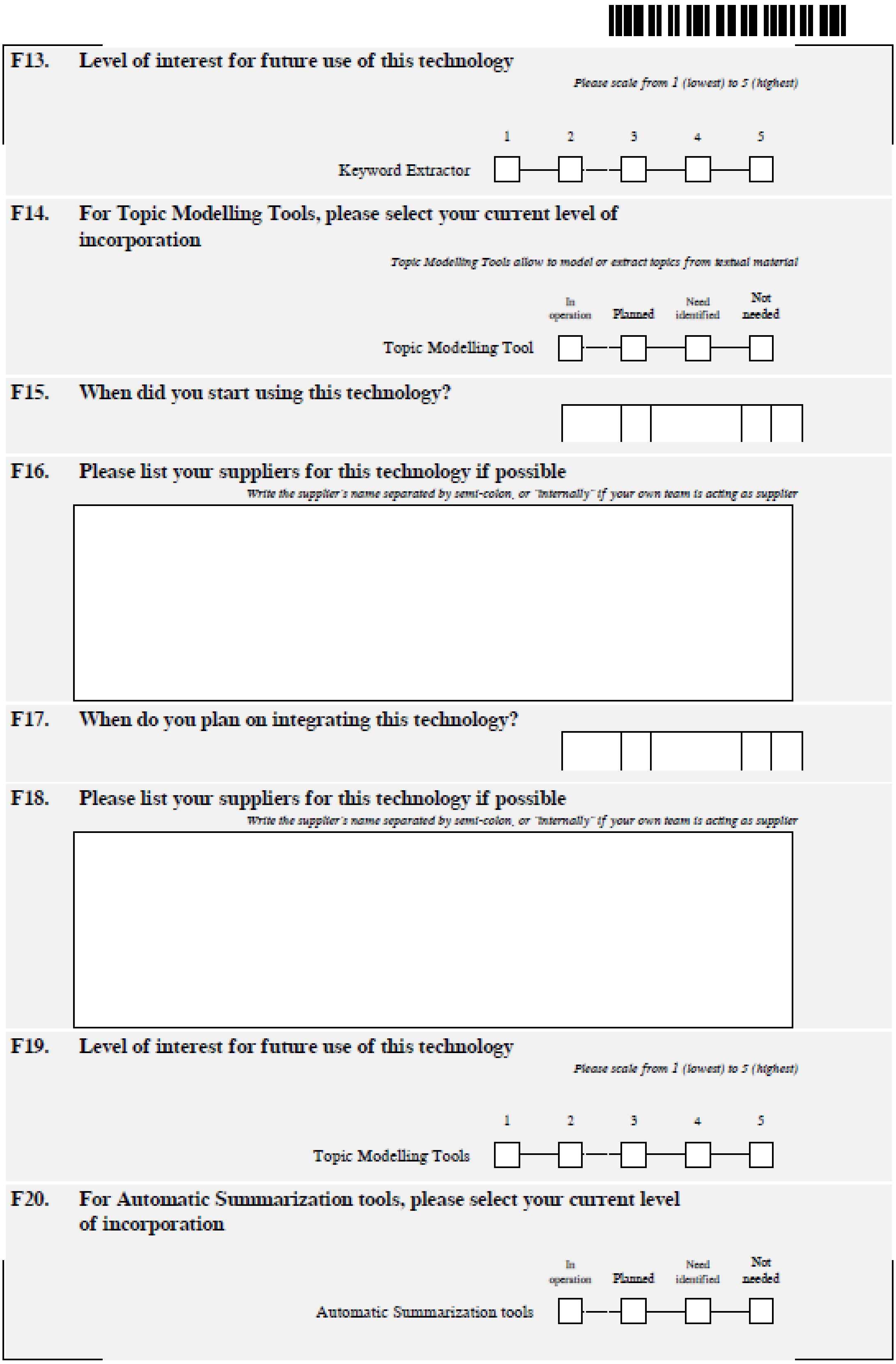
F10. Please list your suppliers for this technology if possible

Write the supplier's name separated by semi-colon, or "Internally" if your own team is acting as supplier

F11. When do you plan on integrating this technology?

F12. Please list your suppliers for this technology if possible

Write the supplier's name separated by semi-colon, or "Internally" if your own team is acting as supplier



F13. **Level of interest for future use of this technology**

Please scale from 1 (lowest) to 5 (highest)

	1	2	3	4	5
Keyword Extractor	□	□	□	□	□

F14. **For Topic Modelling Tools, please select your current level of incorporation**

Topic Modelling Tools allow to model or extract topics from textual material

	In operation	Planned	Need identified	Not needed
Topic Modelling Tool	□	□	□	□

F15. **When did you start using this technology?**

□□□□□

F16. **Please list your suppliers for this technology if possible**

Write the supplier's name separated by semi-colon, or "internally" if your own team is acting as supplier

F17. **When do you plan on integrating this technology?**

□□□□□

F18. **Please list your suppliers for this technology if possible**

Write the supplier's name separated by semi-colon, or "internally" if your own team is acting as supplier

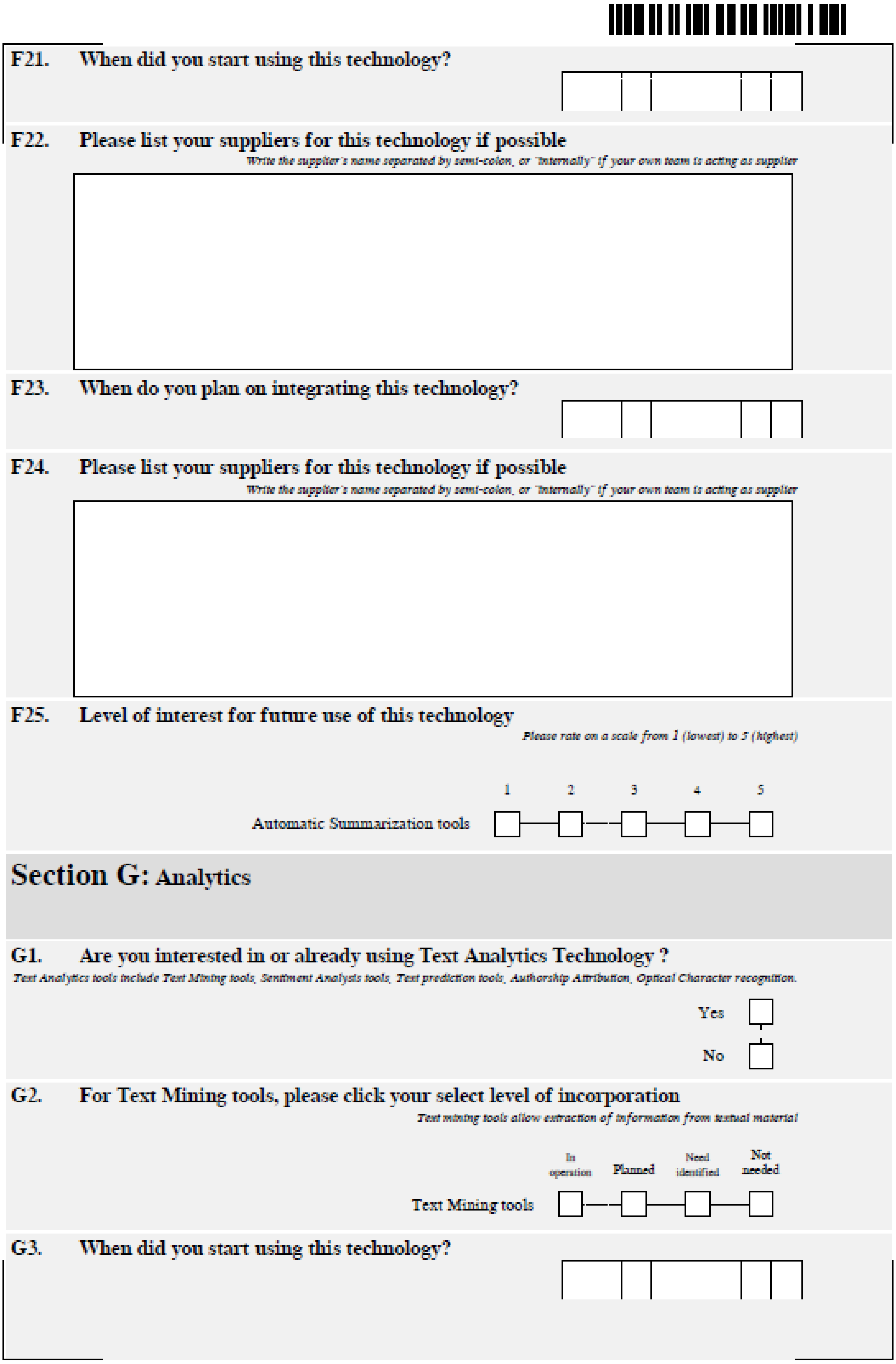
F19. **Level of interest for future use of this technology**

Please scale from 1 (lowest) to 5 (highest)

	1	2	3	4	5
Topic Modelling Tools	□	□	□	□	□

F20. **For Automatic Summarization tools, please select your current level of incorporation**

	In operation	Planned	Need identified	Not needed
Automatic Summarization tools	□	□	□	□



F21. When did you start using this technology?

□ □ □ □ □

F22. Please list your suppliers for this technology if possible

Write the supplier's name separated by semi-colon, or "internally" if your own team is acting as supplier

□

F23. When do you plan on integrating this technology?

□ □ □ □ □

F24. Please list your suppliers for this technology if possible

Write the supplier's name separated by semi-colon, or "internally" if your own team is acting as supplier

□

F25. Level of interest for future use of this technology

Please rate on a scale from 1 (lowest) to 5 (highest)

	1	2	3	4	5
Automatic Summarization tools	□	□	□	□	□

Section G: Analytics

G1. Are you interested in or already using Text Analytics Technology ?

Text Analytics tools include Text Mining tools, Sentiment Analysis tools, Text prediction tools, Authorship Attribution, Optical Character recognition.

Yes □

No □

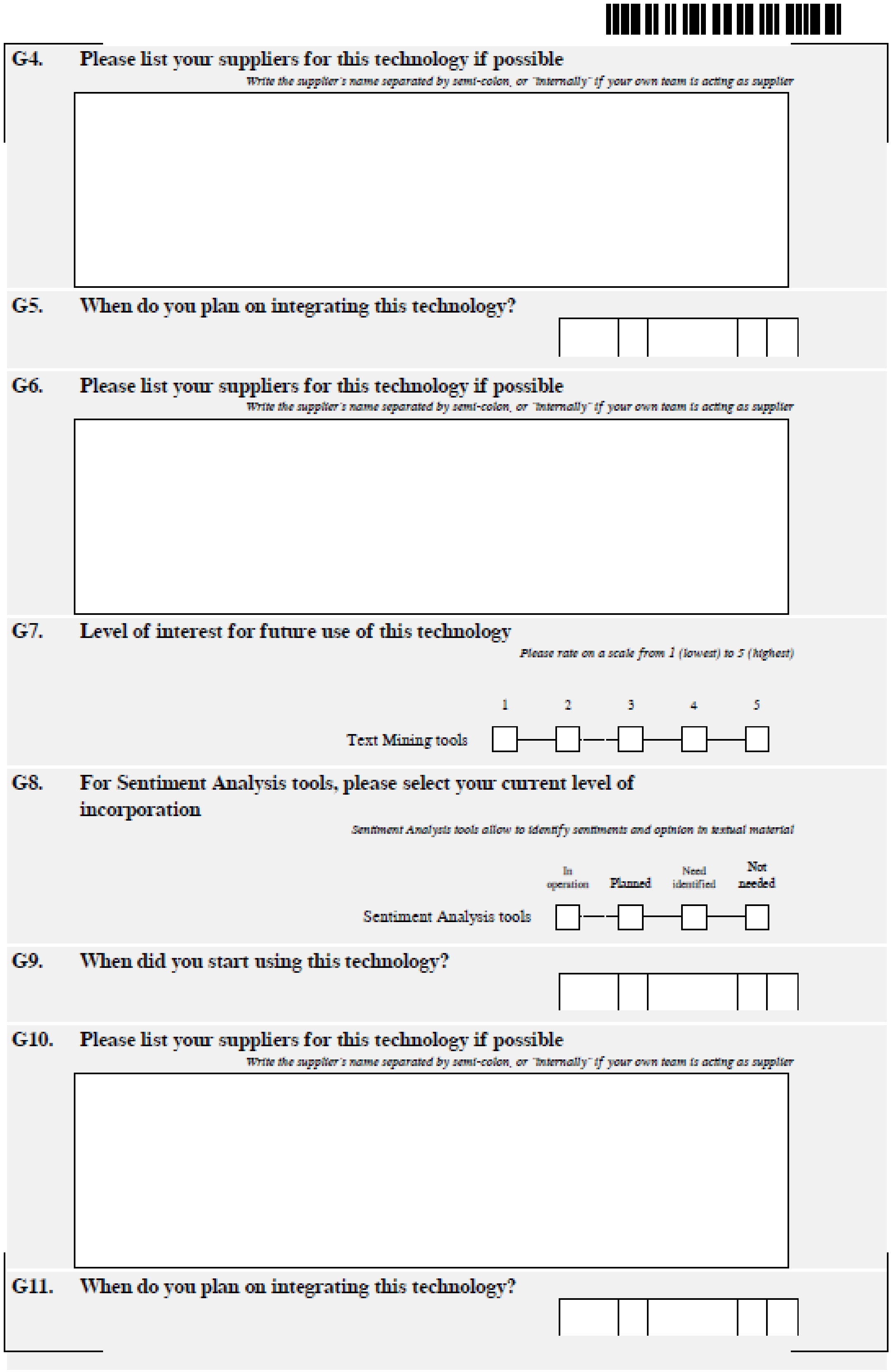
G2. For Text Mining tools, please click your select level of incorporation

Text mining tools allow extraction of information from textual material

	In operation	Planned	Need identified	Not needed
Text Mining tools	□	□	□	□

G3. When did you start using this technology?

□ □ □ □ □



G4. Please list your suppliers for this technology if possible

Write the supplier's name separated by semi-colon, or "internally" if your own team is acting as supplier

G5. When do you plan on integrating this technology?

□ □ □ □ □

G6. Please list your suppliers for this technology if possible

Write the supplier's name separated by semi-colon, or "internally" if your own team is acting as supplier

G7. Level of interest for future use of this technology

Please rate on a scale from 1 (lowest) to 5 (highest)

	1	2	3	4	5
Text Mining tools	□	□	□	□	□

G8. For Sentiment Analysis tools, please select your current level of incorporation

Sentiment Analysis tools allow to identify sentiments and opinion in textual material

	In operation	Planned	Need identified	Not needed
Sentiment Analysis tools	□	□	□	□

G9. When did you start using this technology?

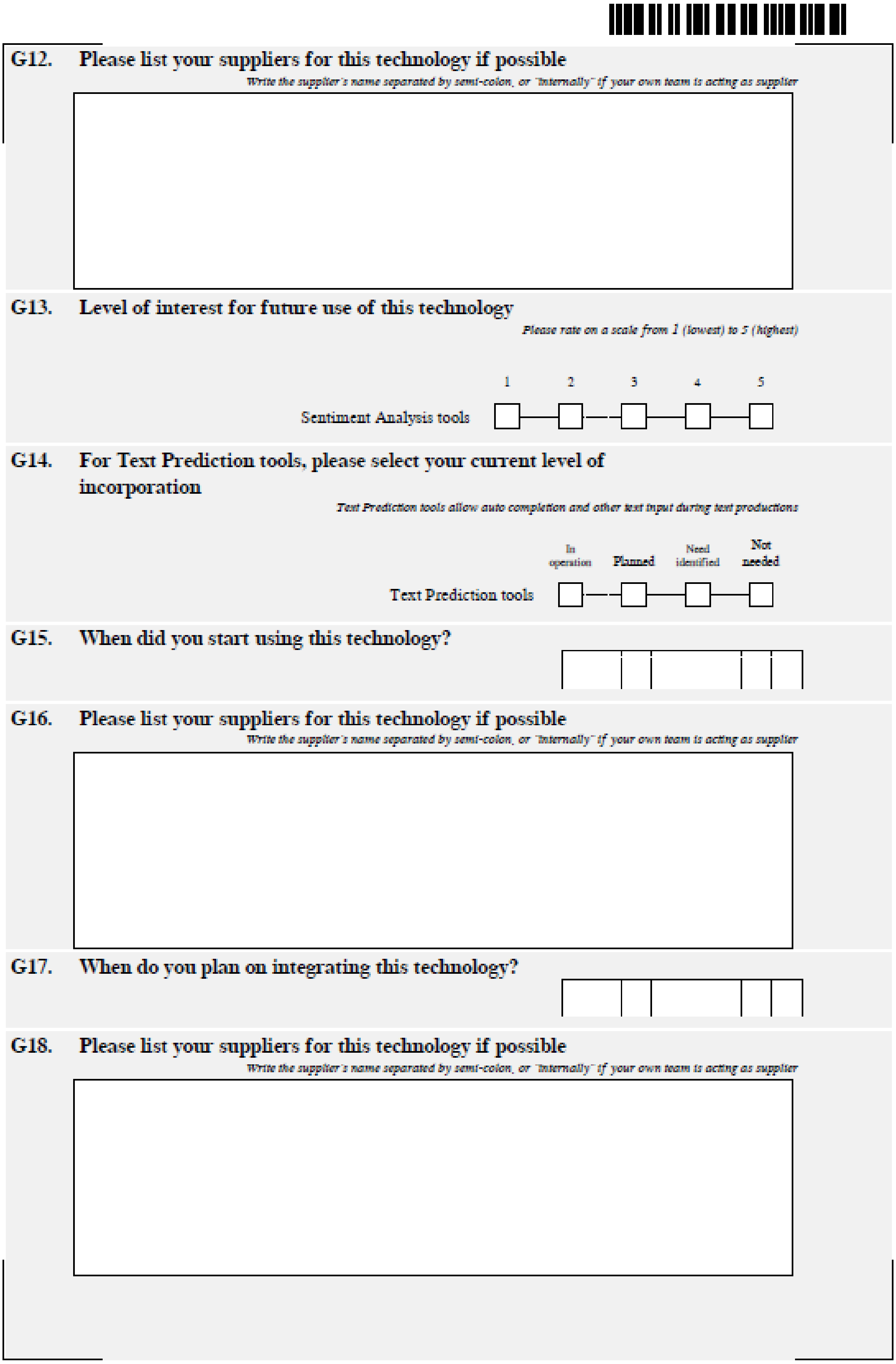
□ □ □ □ □

G10. Please list your suppliers for this technology if possible

Write the supplier's name separated by semi-colon, or "internally" if your own team is acting as supplier

G11. When do you plan on integrating this technology?

□ □ □ □ □



G12. **Please list your suppliers for this technology if possible**

Write the supplier's name separated by semi-colon, or "Internally" if your own team is acting as supplier

G13. **Level of interest for future use of this technology**

Please rate on a scale from 1 (lowest) to 5 (highest)

	1	2	3	4	5
Sentiment Analysis tools	□	□	□	□	□

G14. **For Text Prediction tools, please select your current level of incorporation**

Text Prediction tools allow auto completion and other text input during text productions

	In operation	Planned	Need identified	Not needed
Text Prediction tools	□	□	□	□

G15. **When did you start using this technology?**

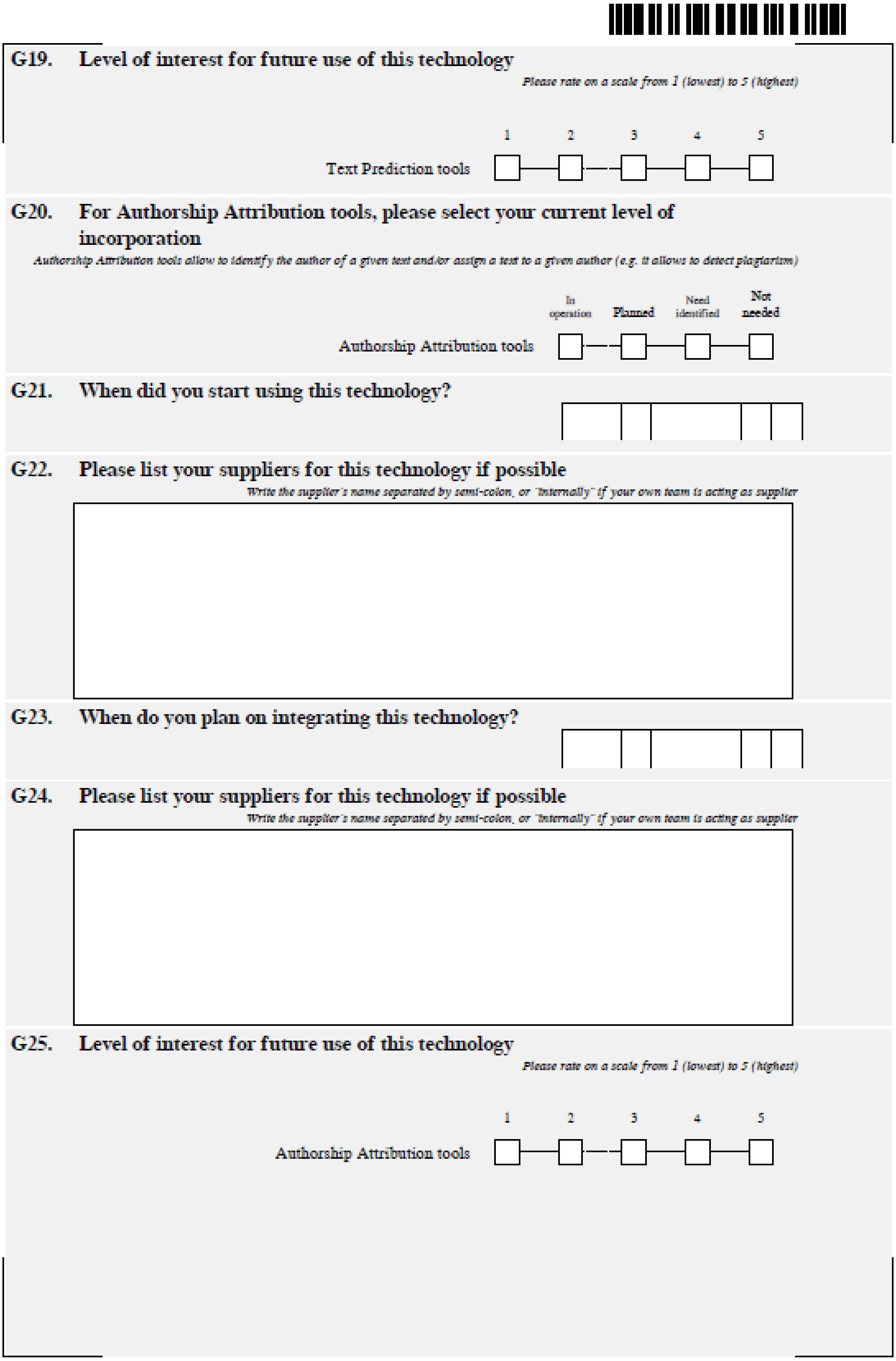
G16. **Please list your suppliers for this technology if possible**

Write the supplier's name separated by semi-colon, or "Internally" if your own team is acting as supplier

G17. **When do you plan on integrating this technology?**

G18. **Please list your suppliers for this technology if possible**

Write the supplier's name separated by semi-colon, or "Internally" if your own team is acting as supplier



G19. Level of interest for future use of this technology

Please rate on a scale from 1 (lowest) to 5 (highest)

	1	2	3	4	5
Text Prediction tools	□	□	□	□	□

G20. For Authorship Attribution tools, please select your current level of incorporation

Authorship Attribution tools allow to identify the author of a given text and/or assign a text to a given author (e.g. it allows to detect plagiarism)

	In operation	Planned	Need identified	Not needed
Authorship Attribution tools	□	□	□	□

G21. When did you start using this technology?

G22. Please list your suppliers for this technology if possible

Write the supplier's name separated by semi-colon, or "Internally" if your own team is acting as supplier

G23. When do you plan on integrating this technology?

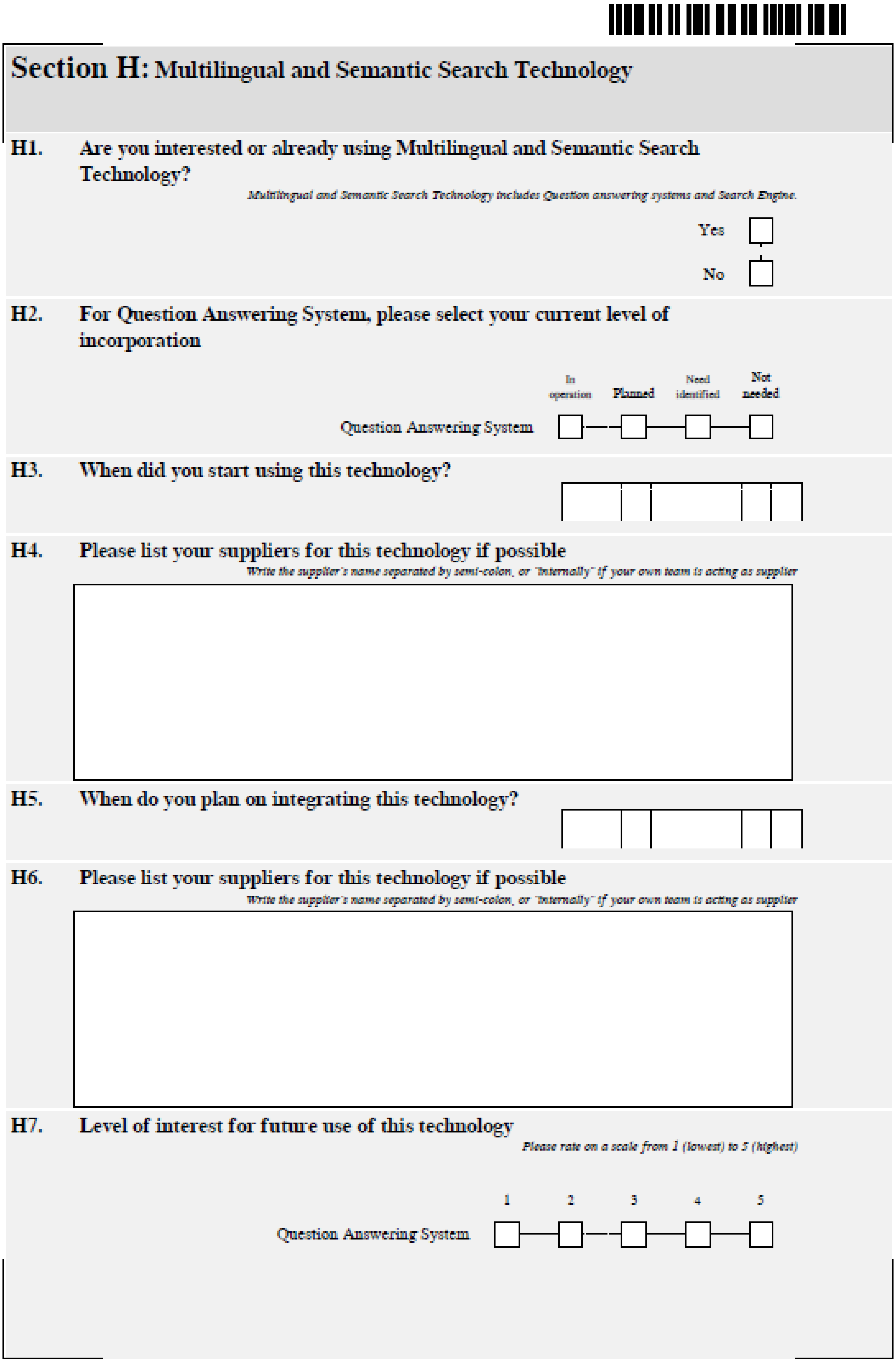
G24. Please list your suppliers for this technology if possible

Write the supplier's name separated by semi-colon, or "Internally" if your own team is acting as supplier

G25. Level of interest for future use of this technology

Please rate on a scale from 1 (lowest) to 5 (highest)

	1	2	3	4	5
Authorship Attribution tools	□	□	□	□	□



Section H: Multilingual and Semantic Search Technology

H1. **Are you interested or already using Multilingual and Semantic Search Technology?**

Multilingual and Semantic Search Technology includes Question answering systems and Search Engine.

Yes □

No □

H2. **For Question Answering System, please select your current level of incorporation**

	In operation	Planned	Need identified	Not needed
Question Answering System	□	□	□	□

H3. **When did you start using this technology?** □□□□□□

H4. **Please list your suppliers for this technology if possible**

Write the supplier's name separated by semi-colon, or "Internally" if your own team is acting as supplier

□

H5. **When do you plan on integrating this technology?** □□□□□□

H6. **Please list your suppliers for this technology if possible**

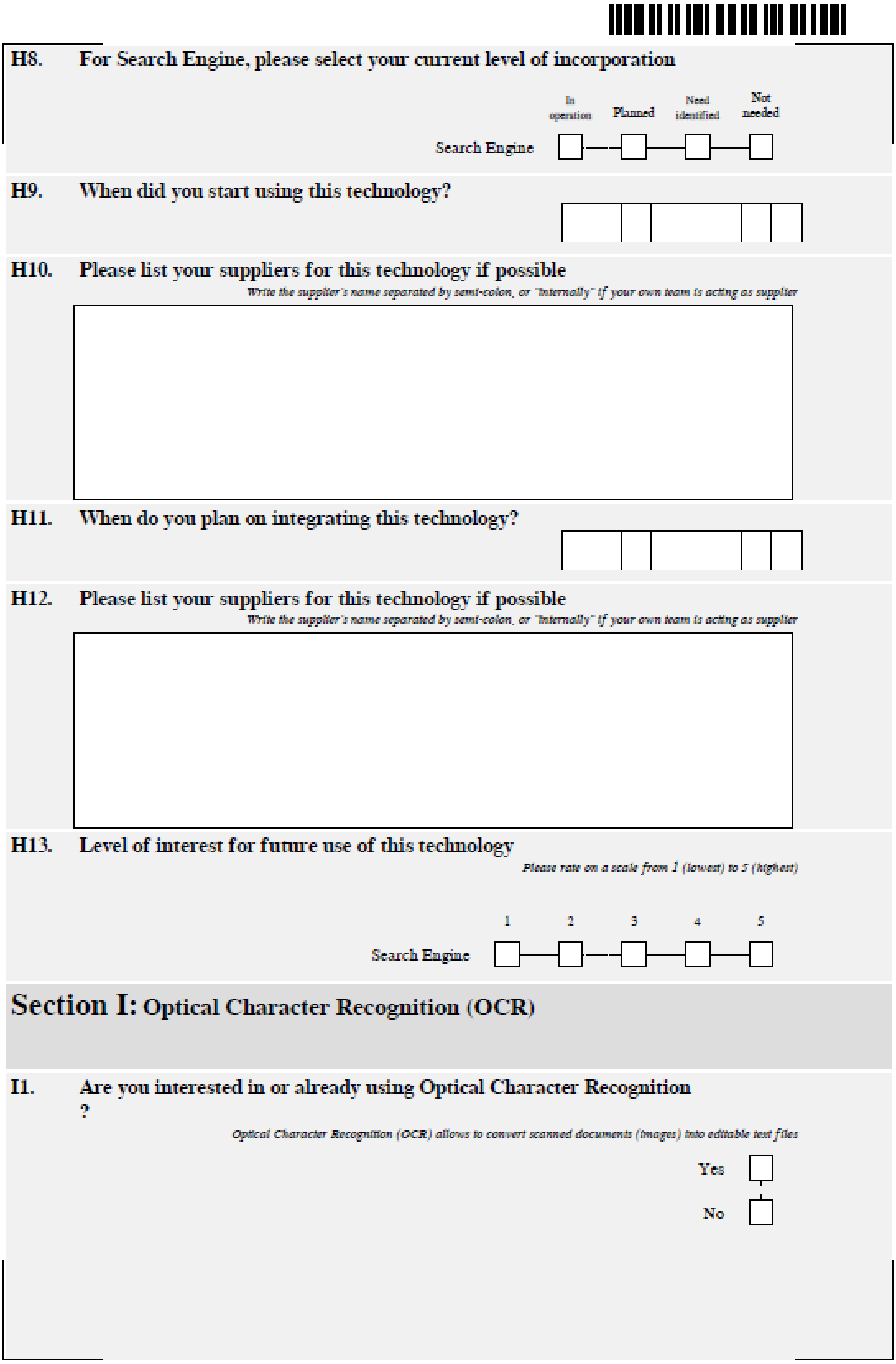
Write the supplier's name separated by semi-colon, or "Internally" if your own team is acting as supplier

□

H7. **Level of interest for future use of this technology**

Please rate on a scale from 1 (lowest) to 5 (highest)

	1	2	3	4	5
Question Answering System	□	□	□	□	□



H8. For Search Engine, please select your current level of incorporation

	In operation	Planned	Need identified	Not needed
Search Engine	□	□	□	□

H9. When did you start using this technology?

H10. Please list your suppliers for this technology if possible

Write the supplier's name separated by semi-colon, or "Internally" if your own team is acting as supplier

H11. When do you plan on integrating this technology?

H12. Please list your suppliers for this technology if possible

Write the supplier's name separated by semi-colon, or "Internally" if your own team is acting as supplier

H13. Level of interest for future use of this technology

Please rate on a scale from 1 (lowest) to 5 (highest)

	1	2	3	4	5
Search Engine	□	□	□	□	□

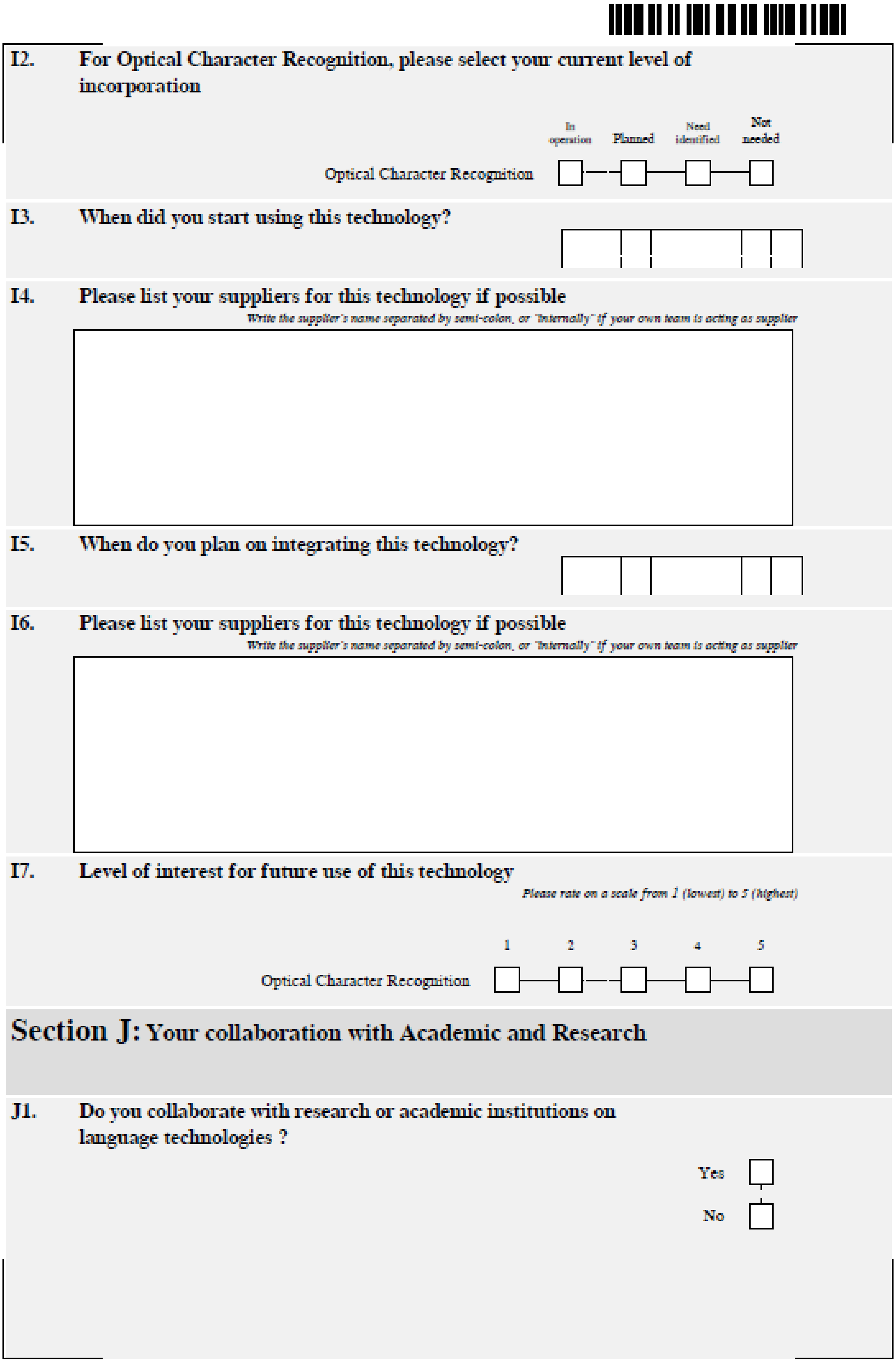
Section I: Optical Character Recognition (OCR)

I1. Are you interested in or already using Optical Character Recognition ?

Optical Character Recognition (OCR) allows to convert scanned documents (images) into editable text files

Yes □

No □



I2. For Optical Character Recognition, please select your current level of incorporation

	In operation	Planned	Need identified	Not needed
Optical Character Recognition	□	□	□	□

I3. When did you start using this technology?

I4. Please list your suppliers for this technology if possible

Write the supplier's name separated by semi-colon, or "internally" if your own team is acting as supplier

I5. When do you plan on integrating this technology?

I6. Please list your suppliers for this technology if possible

Write the supplier's name separated by semi-colon, or "internally" if your own team is acting as supplier

I7. Level of interest for future use of this technology

Please rate on a scale from 1 (lowest) to 5 (highest)

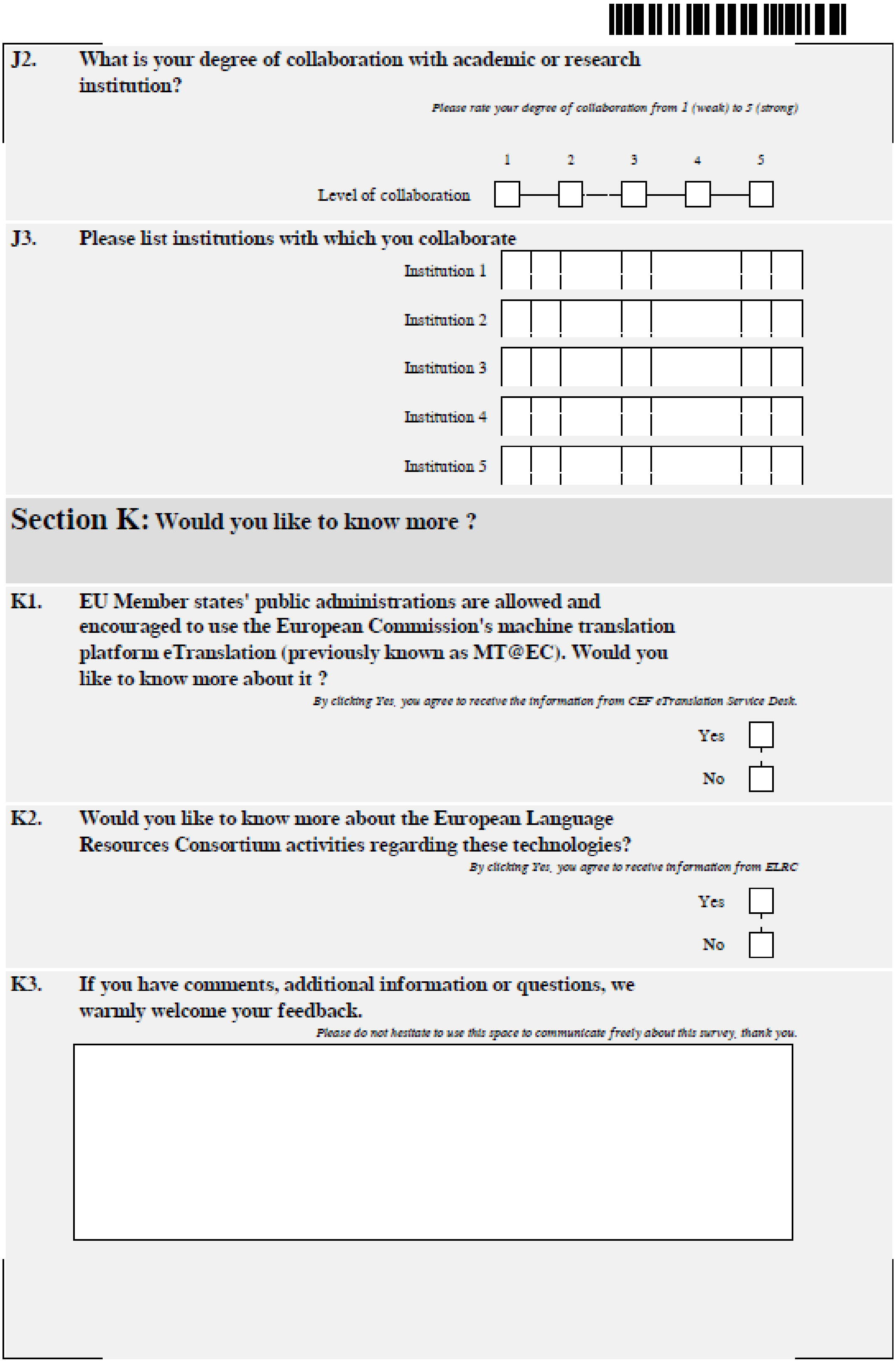
	1	2	3	4	5
Optical Character Recognition	□	□	□	□	□

Section J: Your collaboration with Academic and Research

J1. Do you collaborate with research or academic institutions on language technologies ?

Yes □

No □



J2. What is your degree of collaboration with academic or research institution?

Please rate your degree of collaboration from 1 (weak) to 5 (strong)

	1	2	3	4	5
Level of collaboration	□	□	□	□	□

J3. Please list institutions with which you collaborate

Institution 1

Institution 2

Institution 3

Institution 4

Institution 5

Section K: Would you like to know more ?

K1. EU Member states' public administrations are allowed and encouraged to use the European Commission's machine translation platform eTranslation (previously known as MT@EC). Would you like to know more about it ?

By clicking Yes, you agree to receive the information from CEF eTranslation Service Desk.

Yes □

No □

K2. Would you like to know more about the European Language Resources Consortium activities regarding these technologies?

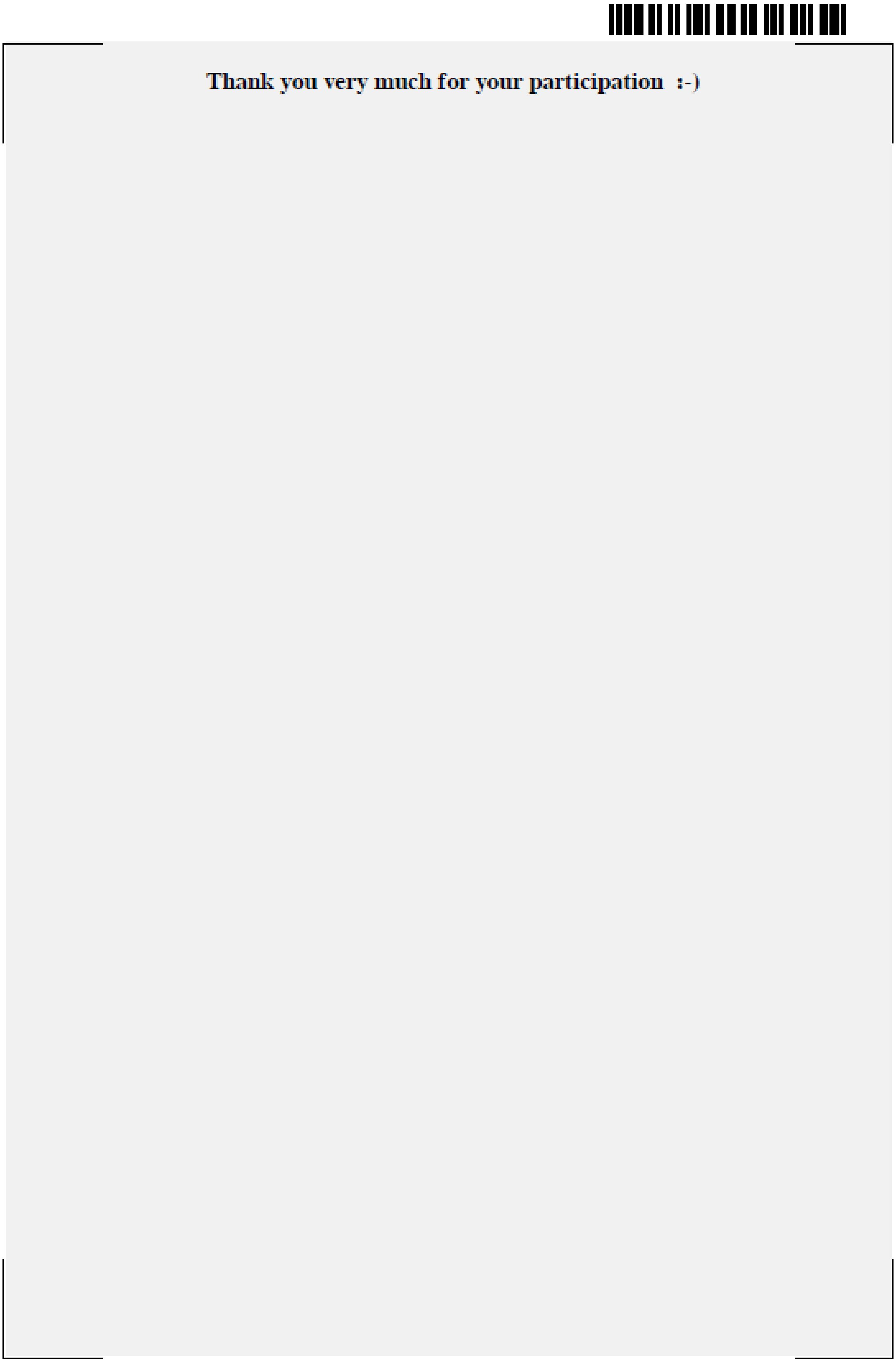
By clicking Yes, you agree to receive information from ELRC

Yes □

No □

K3. If you have comments, additional information or questions, we warmly welcome your feedback.

Please do not hesitate to use this space to communicate freely about this survey, thank you.



Thank you very much for your participation :-)

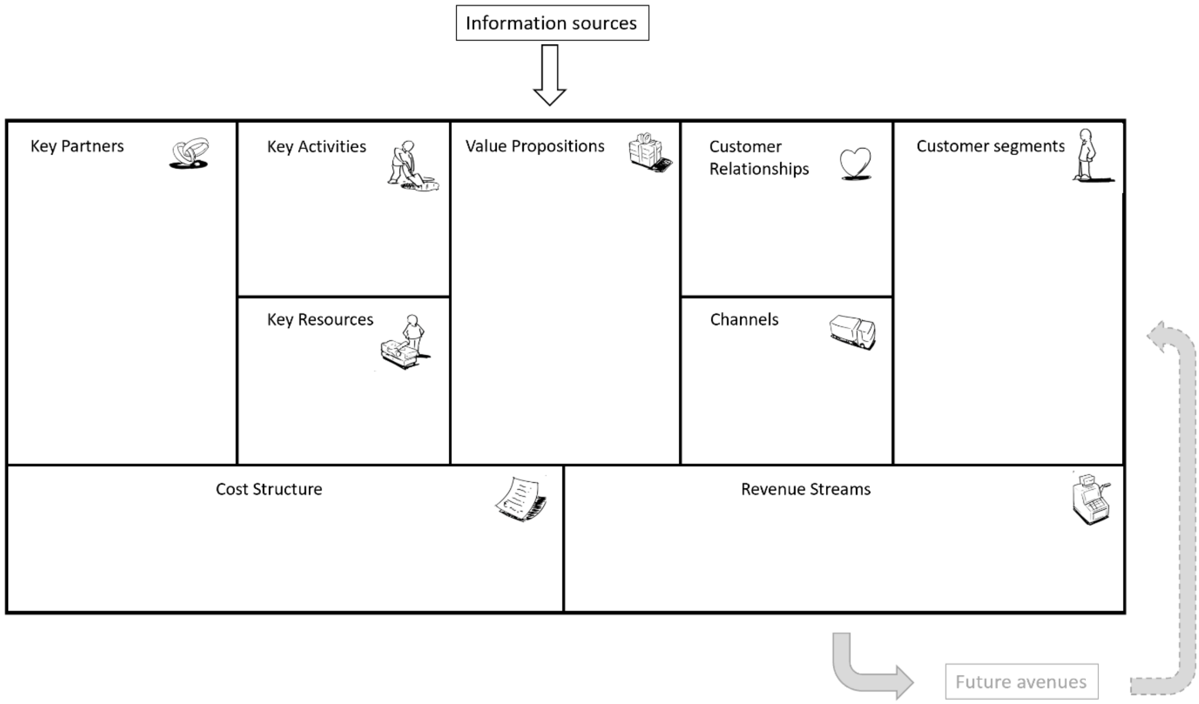
P. Task 4: methodology

Business Model Canvas

A business model is generally described as the rationale of how an organisation creates, delivers and captures value. It is typically used in a commercial context (for companies), but experience shows that it can also be applied to other types of organisations, like public administrations. In the latter case, the business model is constrained. For instance, in the case of a public sector department, constraints hold because of policy reasons, available budgets etc. A business model helps a public administration to view itself as a service-oriented business implementing private sector principles.

The business model canvas methodology describes a model through nine basic blocks that cover the four main areas of a business: customers, offer, infrastructure and financial viability. These model blocks are shown in Figure 169. Each model block is described by answering a number of questions, which are listed below. The answers to these questions are provided by information sources. For instance, in case of CEF AT, they include the results from other Lot 1 tasks. Based on a business model, potential future avenues can be identified, which may then lead to an adaptation of the business model.

Figure 169 Business model blocks



Model block questions

This section describes the nine model blocks and lists the questions associated to them.[229]

1. Customer Segments

This model block defines the different groups of people or organisations an enterprise aims to reach and serve. Segmentation into groups depends on needs, channels within which people or organisations are reached, willingness to pay for certain aspects of the offer, etc. An organisation must make a conscious decision about which groups to serve and which ones to ignore.

Questions associated to this model block are the following:

For whom are we creating value?

Who are our most important customers?

2. Value Proposition

This model block describes the bundle of products and services that create value for specific customer segments. It is an aggregation of benefits offered to customers. Some value propositions may be innovative and represent a new or disruptive offer. Others may be similar to existing market offers, but with added features and attributes.

Questions associated to this model block are the following:

What value do we deliver to customers?

Which one of our customer's problems are we helping to solve?

Which customer's needs are we satisfying?

What bundles of products and services are we offering to each customer segment?

3. Channels

This model block describes how an organisation communicates with its customer segments and reaches them in order to deliver a value proposition. An organisation interfaces with its customers through communication, distribution, and sales channels. Channels are customer touch points that play an important role in the customer experience. They serve several functions, such as raising awareness about products and services, providing the possibility to purchase specific products and services, providing post-purchase support etc.

[229] These were taken from Osterwalder and Pigneur (2010).

Questions associated to this model block are the following:

Through which channels do our customer segments want to be reached?

How are we reaching them now?

How are our channels integrated?

Which ones work best?

Which ones are most cost-efficient?

How are we integrating them with customer routines?

4. Customer Relationships

This model block describes the types of relationships an organisation establishes with specific customer segments. Relationships may be driven by the following motivations: customer acquisition, customer retention, or boosting sales. The customer relationships called for by an organisation's business model deeply influence the overall customer experience.

Questions associated to this model block are the following:

What type of relationship does each of our customer segments expect us to establish and maintain with them?

Which ones have we established?

How costly are they?

How are they integrated with the rest of our business model?

5. Revenue Streams

This model block represents the revenue an organisation generates for each customer segment (costs must be subtracted from revenues to create earnings). While customers comprise the heart of a business model, revenue streams are its arteries. Each stream may have different price mechanisms, such as fixed prices, volume-dependent prices, etc.

Questions associated to this model block are the following:

For what value are our customers really willing to pay?

For what do they currently pay?

How would they prefer to pay?

6. Key Resources

This model block describes the most important assets required to make a business model work, i.e. to create and offer a value proposition, reach markets, maintain relationships with customer segments, and earn revenues. Different key resources are needed depending on the type of business model. Some models are more focused on capital-intensive production facilities while others focus more on human resources. Resources can be owned, leased, or acquired from key partners. Key resources can be categorised in different ways. Physical resources consist of manufacturing facilities, buildings, systems, etc. Intellectual resources consist of software, databases, proprietary knowledge, patents, copyrights, etc.; they are difficult to develop but may offer substantial value. Human resources are crucial in knowledge-intensive and creative industries. Financial resources consist for instance of lines of credit.

Questions associated to this model block are the following:

What key resources do our value propositions require (physical, intellectual, …)?

Our distribution channels?

7. Key Activities

This model block describes the most important things a company must do to make its business model work. These are the most important actions a company must take to operate successfully. Like key resources, key activities are required to create and offer a value proposition, reach markets, maintain customer relationships, and earn revenues. And like key resources, they differ depending on the business model type. They can be categorised in different ways. Production activities relate to designing, making and delivering products. Problem solving activities relate to coming up with new solutions to individual customer problems. Platform/network activities relate to software platforms, online transaction platforms, etc.

The question associated to this model block is the following:

What key activities do our value propositions require (production, problem solving …)?

8. Key Partnerships

This model block describes the network of partners that make the business model work. Organisations forge partnerships for many reasons and partnerships are becoming a cornerstone of many business models. Alliances are created for various reasons. They allow optimising the allocation of resources and activities, which takes place usually for cost reduction and through outsourcing or infrastructure sharing. They allow reducing risks in a competitive environment characterised by uncertainty. Finally, they allow for acquiring particular resources and activities.

Questions associated to this model block are the following:

Who are our key partners?

Which key resources are we acquiring from partners?

Do partners perform key activities?

9. Cost Structure

This model block describes all costs incurred to operate a business model. Creating and delivering value, maintaining customer relationships, and generating revenue all incur costs. While costs should be minimised in every business model, low-cost structures are more important to some business models than to others. On the one extreme, there are cost-driven models, on the other value-driven ones.

Questions associated to this model block are the following:

What are the most important costs inherent in our business model?

Which key resources are most expensive?

Which key activities are most expensive?

Q. Presentation of study during 1st CEF eTranslation Conference

The results of the study presented in this report were presented during the 1st CEF eTranslation Conference organized in the framework of Smart 2016/0103 Lot 2 (Service Desk) on 29 and 30 November 2018 in Brussels. The presentation was followed by a panel discussion. The slides and a description of the panel discussion are included in the present annex.

Presentation

Study on service portfolio development and implementation of the “service desk” component of the CEF Automated Translation platform

SMART 2016/0103

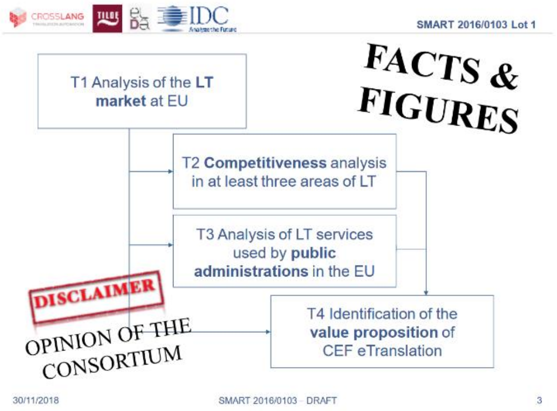
CEF eTranslation Conference
30 November 2018

SMART 2016/0103

Language Technology Market: State-of-the-art, Trends and Value Proposition

CEF eTranslation Conference
30 November 2018

SMART 2016/0103



SMART 2016/0103 Lot 1

FACTS & FIGURES

T1 Analysis of the **LT market** at EU

T2 **Competitiveness** analysis in at least three areas of LT

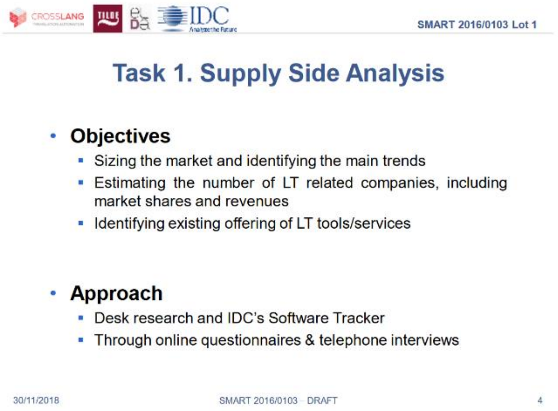
T3 Analysis of LT services used by **public administrations** in the EU

T4 Identification of the **value proposition** of CEF eTranslation

DISCLAIMER

OPINION OF THE CONSORTIUM

30/11/2018 SMART 2016/0103 – DRAFT 3

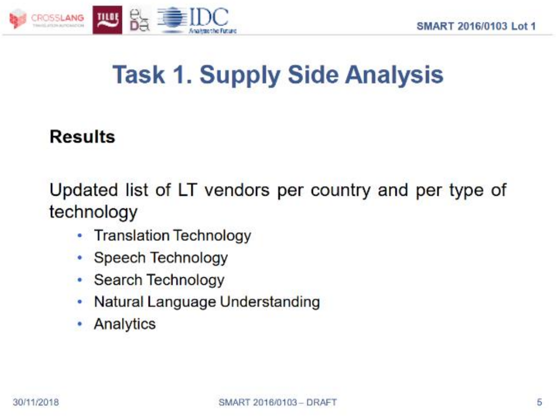


SMART 2016/0103 Lot 1

Task 1. Supply Side Analysis

- **Objectives**
  - Sizing the market and identifying the main trends
  - Estimating the number of LT related companies, including market shares and revenues
  - Identifying existing offering of LT tools/services
- **Approach**
  - Desk research and IDC's Software Tracker
  - Through online questionnaires & telephone interviews

30/11/2018 SMART 2016/0103 – DRAFT 4



SMART 2016/0103 Lot 1

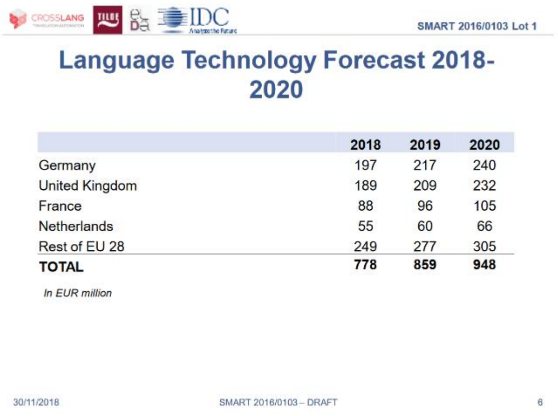
Task 1. Supply Side Analysis

Results

Updated list of LT vendors per country and per type of technology

- Translation Technology
- Speech Technology
- Search Technology
- Natural Language Understanding
- Analytics

30/11/2018 SMART 2016/0103 – DRAFT 5



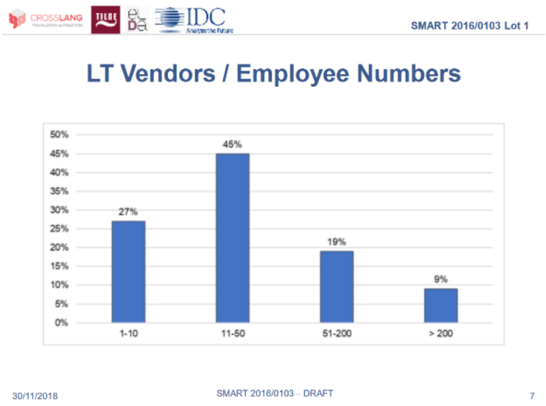
SMART 2016/0103 Lot 1

Language Technology Forecast 2018-2020

	2018	2019	2020
Germany	197	217	240
United Kingdom	189	209	232
France	88	96	105
Netherlands	55	60	66
Rest of EU 28	249	277	305
TOTAL	**778**	**859**	**948**

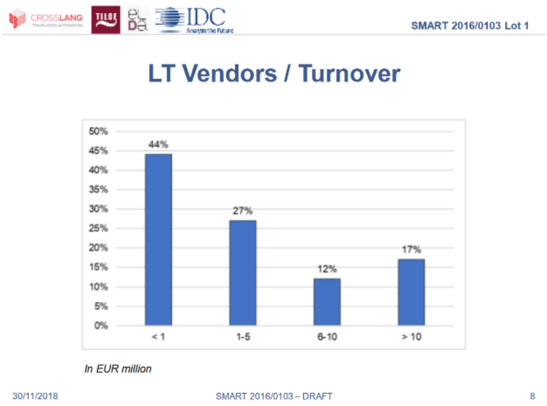
In EUR million

30/11/2018 SMART 2016/0103 – DRAFT 6


SMART 2016/0103 Lot 1

LT Vendors / Employee Numbers

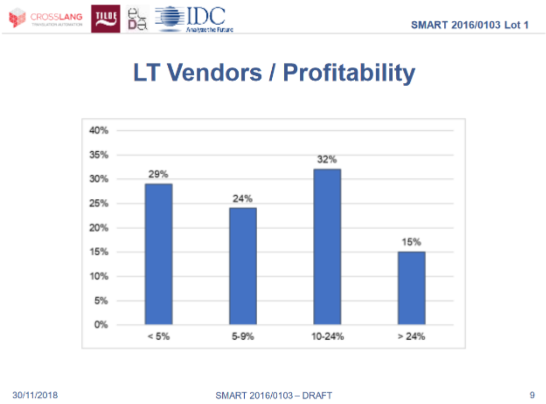
30/11/2018 SMART 2016/0103 – DRAFT 7


SMART 2016/0103 Lot 1

LT Vendors / Turnover

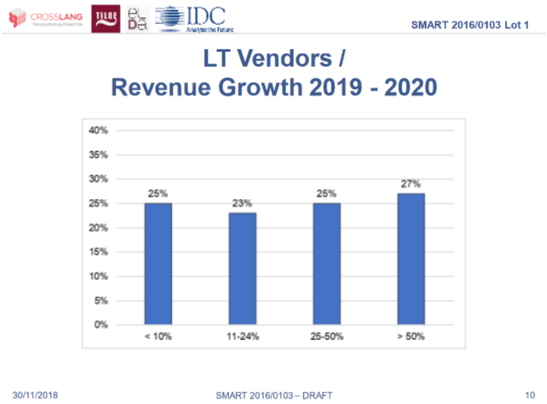
In EUR million

30/11/2018 SMART 2016/0103 – DRAFT 8


SMART 2016/0103 Lot 1

LT Vendors / Profitability

30/11/2018 SMART 2016/0103 – DRAFT 9


SMART 2016/0103 Lot 1

LT Vendors / Revenue Growth 2019 - 2020

30/11/2018 SMART 2016/0103 – DRAFT 10

SMART 2016/0103 Lot 1

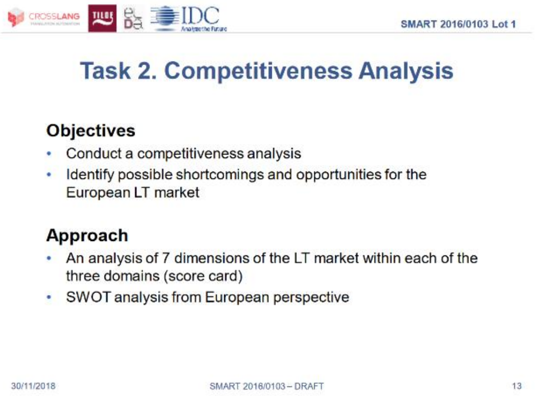
https://www.youtube.com/watch?v=yv_8dx7g-WA

30/11/2018 SMART 2016/0103 – DRAFT 11

SMART 2016/0103 Lot 1

Google I/O 2018

30/11/2018 SMART 2016/0103 – DRAFT 12



SMART 2016/0103 Lot 1

Task 2. Competitiveness Analysis

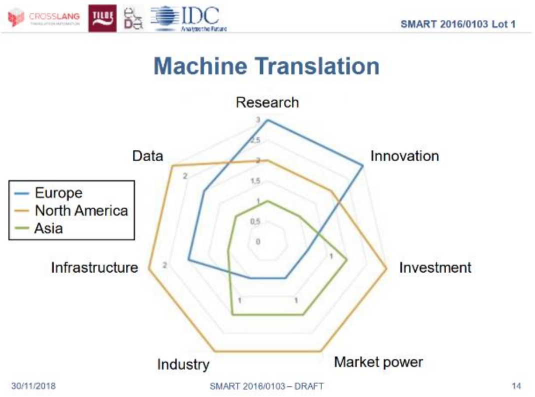
Objectives

- Conduct a competitiveness analysis
- Identify possible shortcomings and opportunities for the European LT market

Approach

- An analysis of 7 dimensions of the LT market within each of the three domains (score card)
- SWOT analysis from European perspective

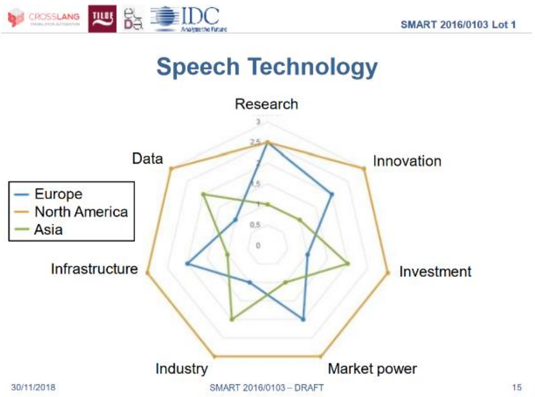
30/11/2018 SMART 2016/0103 – DRAFT 13



SMART 2016/0103 Lot 1

Machine Translation

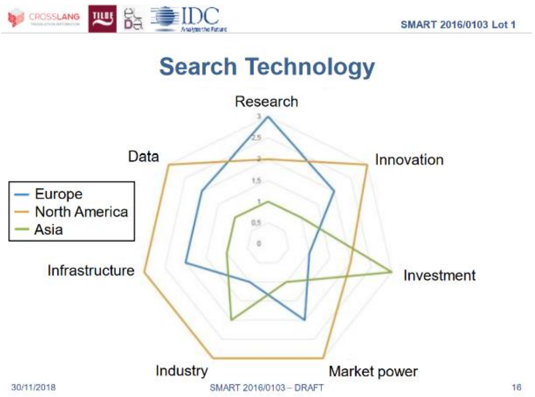
30/11/2018 SMART 2016/0103 – DRAFT 14



SMART 2016/0103 Lot 1

Speech Technology

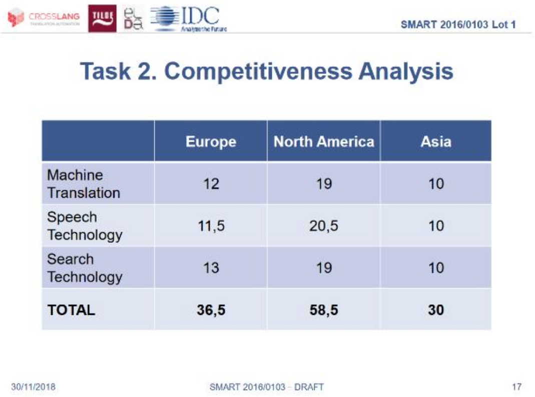
30/11/2018 SMART 2016/0103 – DRAFT 15



SMART 2016/0103 Lot 1

Search Technology

30/11/2018 SMART 2016/0103 – DRAFT 16

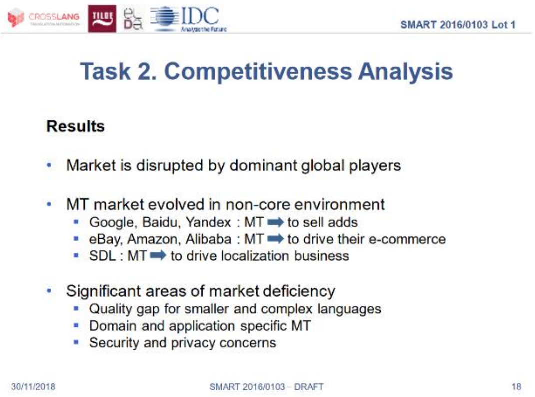


SMART 2016/0103 Lot 1

Task 2. Competitiveness Analysis

	Europe	North America	Asia
Machine Translation	12	19	10
Speech Technology	11,5	20,5	10
Search Technology	13	19	10
TOTAL	**36,5**	**58,5**	**30**

30/11/2018 SMART 2016/0103 – DRAFT 17



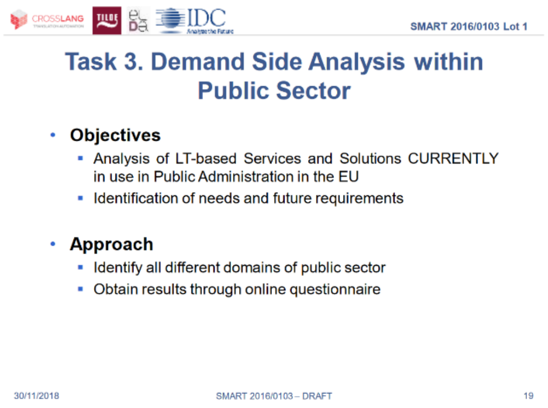
SMART 2016/0103 Lot 1

Task 2. Competitiveness Analysis

Results

- Market is disrupted by dominant global players
- MT market evolved in non-core environment
  - Google, Baidu, Yandex : MT ➡ to sell adds
  - eBay, Amazon, Alibaba : MT ➡ to drive their e-commerce
  - SDL : MT ➡ to drive localization business
- Significant areas of market deficiency
  - Quality gap for smaller and complex languages
  - Domain and application specific MT
  - Security and privacy concerns

30/11/2018 SMART 2016/0103 – DRAFT 18

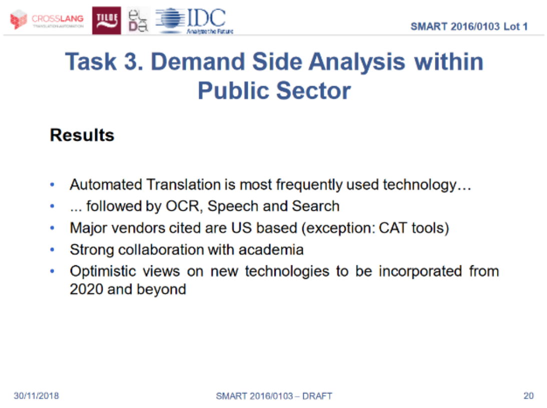


SMART 2016/0103 Lot 1

Task 3. Demand Side Analysis within Public Sector

- **Objectives**
  - Analysis of LT-based Services and Solutions CURRENTLY in use in Public Administration in the EU
  - Identification of needs and future requirements
- **Approach**
  - Identify all different domains of public sector
  - Obtain results through online questionnaire

30/11/2018 SMART 2016/0103 – DRAFT 19



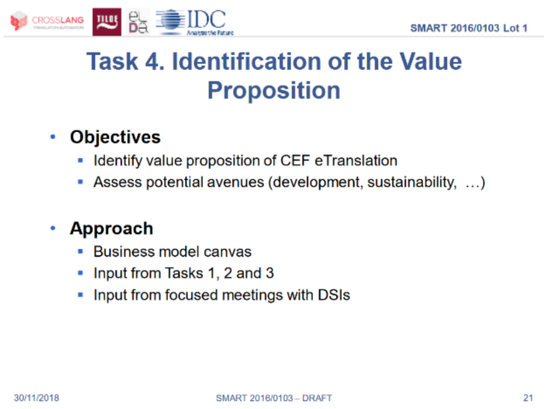
SMART 2016/0103 Lot 1

Task 3. Demand Side Analysis within Public Sector

Results

- Automated Translation is most frequently used technology…
- ... followed by OCR, Speech and Search
- Major vendors cited are US based (exception: CAT tools)
- Strong collaboration with academia
- Optimistic views on new technologies to be incorporated from 2020 and beyond

30/11/2018 SMART 2016/0103 – DRAFT 20

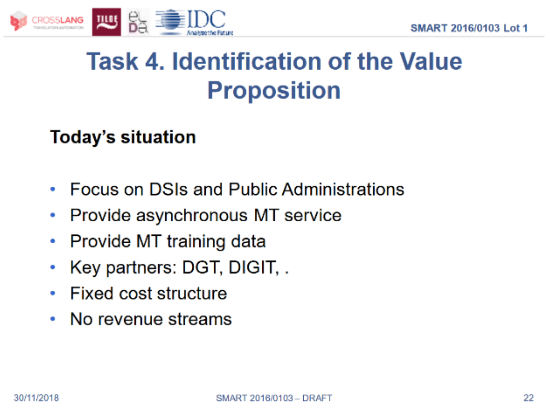


SMART 2016/0103 Lot 1

Task 4. Identification of the Value Proposition

- **Objectives**
  - Identify value proposition of CEF eTranslation
  - Assess potential avenues (development, sustainability, …)
- **Approach**
  - Business model canvas
  - Input from Tasks 1, 2 and 3
  - Input from focused meetings with DSIs

30/11/2018 SMART 2016/0103 – DRAFT 21



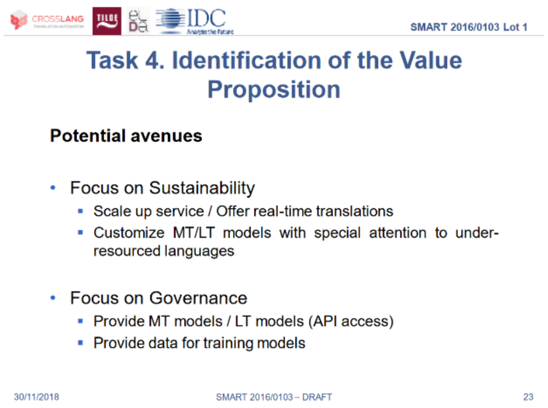
SMART 2016/0103 Lot 1

Task 4. Identification of the Value Proposition

Today's situation

- Focus on DSIs and Public Administrations
- Provide asynchronous MT service
- Provide MT training data
- Key partners: DGT, DIGIT, .
- Fixed cost structure
- No revenue streams

30/11/2018 SMART 2016/0103 – DRAFT 22



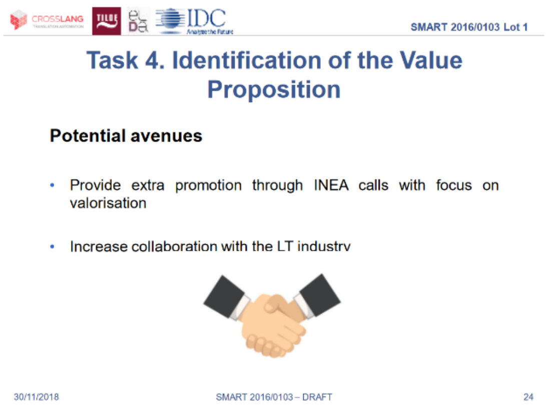
SMART 2016/0103 Lot 1

Task 4. Identification of the Value Proposition

Potential avenues

- Focus on Sustainability
  - Scale up service / Offer real-time translations
  - Customize MT/LT models with special attention to under-resourced languages
- Focus on Governance
  - Provide MT models / LT models (API access)
  - Provide data for training models

30/11/2018 SMART 2016/0103 – DRAFT 23



SMART 2016/0103 Lot 1

Task 4. Identification of the Value Proposition

Potential avenues

- Provide extra promotion through INEA calls with focus on valorisation
- Increase collaboration with the LT industrv

30/11/2018 SMART 2016/0103 – DRAFT 24

Panel discussion

The discussion was moderated by Luc Meertens. Panellists were Khalid Choukri (ELDA), Philippe Gelin (EC), Tom Vanallemeersch (CrossLang) and Andrejs Vasiljevs (Tilde). Key questions included the approach taken by the study, the focus on the public sector, and the expected trends in the EU technology market. Yet, the first and introductory question was actually raised by the audience, i.e. why there were not more investments in the EU language technology (see joining ETSI vocabularies, responding to reportings on AI etc.). It was proposed to have more of a public-private partnership collaboration. Given that the Commission has built a translation tool, which needs development and investment to be able to compete with the others, the money should not be given to the public sector to then invest it in Google. In response to this issue being raised, Philippe Gelin referred to his presentation on the first day of the conference, which actually detailed the EC's spending in LT as part of the new MFF. One of the topics there is in fact AI.

Andrejs Vasiljevs pointed out that the Latvian example of Hugo.lv and the EU Council Presidency Translator actually was a good example of collaboration on a national and European level, of collaboration between the public sector and private companies. The platform which was showcased on MT technologies is run by a European company (Tilde). The platform collects data to improve technologies in the country, but it also helps develop European infrastructure by sharing data with ELRC and eTranslation and by helping to develop Latvian LT that can serve other European services. By focusing on the needs of smaller European languages and specific domains, this example shows that we can provide better quality and solutions than global players like Google.

The discussion then returned to the panel, and the first question elaborated was about the approach used within the study. It was explained by Tom Vanallemeersch that there was actually a combination of desk research, questionnaires and interviews. About 200 questionnaires were sent to a group of companies, around 60 responses were received. There were also about 10 companies who were prepared for a telephone interview. The investigators asked about their opinion on how the market evolved in the last years and what they would expect for the next few years.

An interesting question concerned the fact that the focus of the study was on the public sector. The answer was very simple as this sector is the main user (and intended user) of CEF eTranslation. The goal was to find out who is using which tools and who is familiar with which technologies, trying to cover the whole panorama of technologies from MT to speech technologies, taking also into account the supply side. To some extent, public administrations are even earlier adopters of these technologies. As such, the survey also aimed to boost the use of the technologies by (i) raising awareness and informing the public sector about CEF eTranslation and translation services and by (ii) promoting the use of these technologies and sharing of language resources.

Philippe Gelin was then asked how much of the study results would actually be made public. He confirmed that in fact most of the study will be made public. Only the last section providing suggestions for the future development of CEF eTranslation, however, which is still under decision-making process and which was meant as internal information for the EC, would not be made available.

The next couple of questions focused on the trends of the EU technology market and on the future of CEF eTranslation. With regard to the future of CEF eTranslation, Philippe Gelin pointed out that lowering the language barriers is the overall key concept. This, however, cannot be done alone, so collaboration with private and public sectors will be required. It is important for the system uptake and improvement to raise awareness. When one leaves the circle of LT experts and translators, people do not always think of the potential this technology has achieved. In the future, it should and it will be possible to provide more resources and more tools and services.

As regards the expected trends in the technology market, Andrejs Vasiljevs explained that Europe is still quite strong in research and in finding new ways, methods, approaches, opportunities and breakthroughs. He pointed out that many startups are emerging, but when it comes to scaling up and building a business, many of these startups and successful companies are being bought by US global players. On the question whether he thinks that Europe lacks the commercial flair to scale research results into businesses, Andrejs Vasiljevs explained that there is an even greater issue than the successful transfer of research results, and that is the difference between vast market and global applications. For that, investments are critical. We have to be aware of that and we have to invest more in scaling up to reaching the market with successful solutions.

Since the study actually showed that there is a great deal of optimism with regard to LT, panellists were asked whether they shared this feeling and what would be their preferred technology. Khalid Choukri explained that he in fact had received very optimistic responses in the demand-side survey. The big difference between the European scene and the rest of the world is that in Europe it is not expected that investors go for such technologies. Therefore, public services have to act as early adopters to show that it is operational and that the results meet their expectations. Awareness must be raised.

Tom Vanallemeersch pointed out that when it comes to optimism, it is often about chat-bots and NLU, which are becoming more and more important. NLU is a very important aspect in Europe, but the major (i.e. non-European) players are also very strong in these areas. In Europe, because of the multilingualism, it is very important to support languages which are not necessarily supported by the large players. However, in order to support them, the systems need data. The more often the systems are used and the more data are collected, the better the systems will become. In Europe, there is a huge demand for data, so it will be important that this movement will happen early enough if Europe wants to support a number of languages which are not supported elsewhere. It is about investing timely enough to make this movement happen in Europe. Unfortunately, the investment attitude in Europe is not the same as in the US or Asia.

Andrejs Vasiljevs explained that the AI hype is actually helping to look at the language challenge from a different perspective. Solutions should not only be text-based, but also speech-based, raising awareness of the general importance of LT technology. This creates optimism of LT companies and providers and opens up new business opportunities. In the public sector, there is a great interest in chat-bot technologies (e.g. Latvia, Sweden, Finland etc.). As such, AI indeed is a good vehicle to make a public service more efficient and to decrease costs.

Philippe Gelin pointed out that from a European Commission perspective, the main target is to lower the language barriers. Within the last 3 years, there has been a drastic improvement on research level and there are a lot of opportunities. He explained that it is now important to carefully consider and exactly understand what will be needed in the future.

The evaluation of the future of LT and corresponding market trends triggered several questions from the audience. One participant noted that people often have no trouble understanding French and German, but the challenge and their bigger need is translating other languages like Swahili, Arabic etc. He asked whether these languages are also considered in surveys for the further development of CEF eTranslation. Khalid Choukri explained that out of the responses received in the study, the team has indeed received the feedback from many organisations that they are also looking for support in non-EU languages, so there is definitely a need for these languages. Philippe Gelin further explained that non-EU languages are a matter of interest, but providing all services in all languages is not feasible at the moment.

Another question raised by the audience concerned the LT market development. In Luc Meertens' presentation it was really surprising to see the evaluation of Asian research. The participant claimed that, in reality, the Asian research would be much stronger and easily on a path with European and North American research. Andrejs Vasiljevs explained that in order to determine the level of research in the US, Europe and Asia, the team had relied on the number of citations and publications. They showed that Asia is catching up with Europe and the US since the last few years, but it is still a bit behind. Khalid Choukri further explained that for the Olympics 2020 in Japan, for instance, a huge effort was made to make all Olympic games multilingual, both for journalists and for participants. This, however, would also impact on the industrial part and not only on the research part. Philippe Gelin too hinted that looking at the latest developments, e.g. WMT competitions, it becomes evident that Asia is taking over.

Another participant pointed out that there is a lot of potential for eTranslation in the private sector. His question was why the EC keeps the eTranslation Building Block mostly restricted for use in public administrations and why they do not extend it to for use in private sectors. Philippe Gelin pointed out that there are existing solutions on the market, which are not too expensive. The EC has limited resources and has to be careful with how taxpayers' money is spent. Moreover, there are legal constraints in spending such public money in particular with regard to deterring competition. Other MT companies should not be threatened by CEF eTranslation.

Another participant insisted that it would be important to support a market drift and to have a big project which relates to the needs of European companies and individual citizens to make translation services available to them on their desktop. The public sector should also have a leading role in developing the market. Furthermore, there should be more investments in making the digital publications multilingual on the EC websites: all information produced by the EU should be made available (at least) in all EU languages. This would also promote multilingualism. Philippe Gelin agreed with the latter and confirmed that multilingual access indeed is needed.

European Commission

Final study report on CEF Automated Translation value proposition in the context of the European LT market/ecosystem

Luxembourg, Publications Office of the European Union

2019 – 352

ISBN 978-92-76-00783-8

doi: 10.2759/142151

KK-03-19-154-EN-N

doi: 10.2759/142151 ISBN 978-92-76-00783-8

Digital
Single
Market