Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 3381–3389
Marseille, 11–16 May 2020

The Competitiveness Analysis of the European Language Technology Market

Andrejs Vasiļjevs, Inguna Skadiņa, Indra Sāmīte, Kaspars Kauliņš, Ēriks Ajausks, Jūlija Meļņika and Aivars Bērziņš
Tilde
Vienibas gatve 75a, Riga, Latvia
{fistname.lastname}@tilde.lv

Abstract

This paper presents the key results of a study on the global competitiveness of the European Language Technology market for three areas – machine translation, speech technology, and cross-lingual search. EU competitiveness is analyzed in comparison to North America and Asia. The study focuses on seven dimensions (research, innovations, investments, market dominance, industry, infrastructure, and Open Data) that have been selected to characterize the language technology market. The study concludes that while Europe still has strong positions in Research and Innovation, it lags behind North America and Asia in scaling innovations and conquering market share.

Keywords: competitiveness analysis, language technology market, machine translation, speech technology, cross-lingual search

1. Introduction

This paper provides the key results of the competitiveness analysis of the European Language Technology (LT) market in comparison to North America (Unites States and Canada) and Asia (China, Japan, India, South Korea and Singapore). The study is part of broader research contracted by the European Commission (EC) to assess the European Language Technology market and identify potential actions that need to be initiated at the European Union level.

The study focused on three LT areas that are of the greatest interest for the EC – machine translation (MT), speech technology, and cross-lingual search. Seven dimensions were selected to assess the key factors for competitiveness in the global LT market: Research, Innovation, Investments, Market dominance, Industry, Infrastructure, and Open Data.

A number of criteria were used for each dimension to analyse the comparative position of the European LT market in respect to its main competitors – North America and Asia. A scale from 1 (weakest) to 3 (strongest) was used to rank markets within each dimension.

Taking into account limited time and resources allocated to the study its analysis is based on a desk research of secondary sources, data collected from previous studies, and overall economic indicators, such as studies and reports by Common Sense Advisory (Lommel et al., 2016), World Economic Forum (2017); TAUS (Massardo, 2016; Seligman, 2017; TAUS, 2017), CRACKER (SRIA, 2017) and META-NET (2015).

The full report of the findings from the study has been published by the European Commission (Vasiļjevs et al., 2019a). Besides an analysis of competitiveness, the report provides an assessment of the supply and demand sides of the European LT market, analysis of LT adoption by public services in the EU, and proposes a value proposition for the automated translation services provided by the European Commission.

In this paper we have summarized the key findings of the report in respect to the competitiveness of the EU market. The paper is structured by sections devoted to each of the analysed areas of LT market competitiveness. Speech and search technologies are covered in more detail, while only the essence is included for machine translation because it has already been discussed by Vasiljevs et al. (2019b).

2. Competitiveness of European 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 selected area
- research publications
- organizational infrastructure (e.g. associations, networks and research infrastructures).

When we analysed publicly available information about research centres in different countries, research institutions were not weighted for their size, since this information (e.g. number and qualification of researchers, research budget, number of projects) is not available in public sources.

In this study, we performed research on publications in the Scopus database[1]. The 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 and analytical tools that were needed for this study (e.g., number of publications per country, author, organization, etc.). 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 is the largest (22,800 titles from more than 5,000 international publishers) abstract and citation database of peer-reviewed literature (Scopus, 2017). To calculate regional distribution of publications, the methodology used by Scopus to count 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.

Research publications describe both academic and industrial research results. However, it could be that

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

industrial research is not entirely disclosed, since not all industrial research results are made public.

2.1 Research in Machine Translation

The main findings for Machine Translation have been summarized by Vasiljevs et al (2019b):

- Europe has the largest number of research centres, almost twice as many as North America (54 vs. 14 in Asia and 23 in North America).
- Number of publications in top conferences and journals is very similar for North America and Europe (Figure 1). However, it should be noted that the trend in the last two years is an increasing amount of research in Asia.
- When the 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.

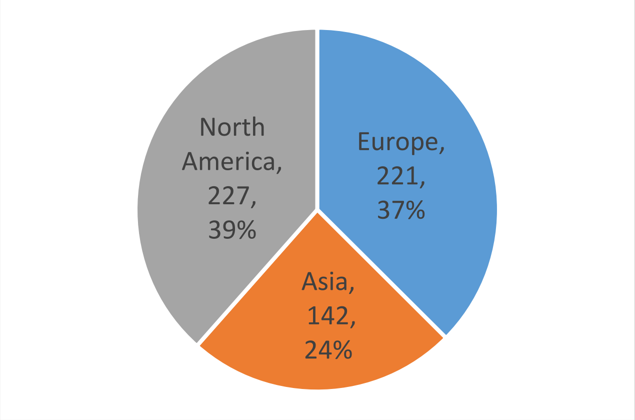

Figure 1: Distribution of publications on "machine translation" between regions in ACL, COLING, EACL, NAACL and NIPS proceedings (2010-2017).

2.2 Research in Speech Technologies

2.2.1 Research Centres

Research in speech technology occurs in companies and in academic research centres. EU multilingualism policy and language diversity of Europe is a reason for more research centres in Europe than in other regions. The International Speech Communication Association (ISCA) lists 176 speech laboratories from 37 countries around the world[2]. Europe is a leader with the largest number of laboratories (72 in total out of 176).

2.2.2 Publications

We have analysed publications found in the Scopus database by querying the database for "speech recognition" OR "text-to-speech" OR "speech synthesis" in the title, abstract and keywords. Figure 2 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 powerful and accurate deep learning methods. At the same time, although there is less interest in speech synthesis, this interest remains stable.

When regions are compared, the leader is Asia (41% of publications), followed by Europe with 11,596 or 34% of publications, while for 7,811 (25%) publications at least one author is from North America (Figure 3).

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 first place followed by North America. However, it should be noted that the trend in the last two years is an increasing number of research from Asia in comparison to Europe and North America.

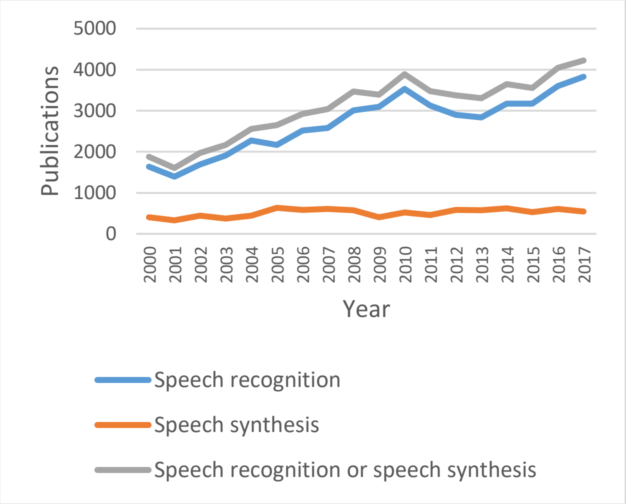

Figure 2 : Numbers of publications per year for "speech recognition" OR "text-to-speech" OR "speech synthesis" in the Scopus DB (2010-2017).

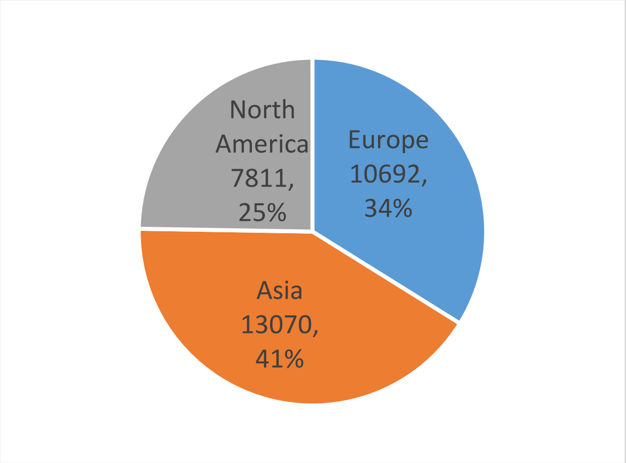

Figure 3 : Distribution of publications by region for "speech recognition" OR "text-to-speech" OR "speech synthesis" in the Scopus DB (2010-October 2018).

2.3 Research in Search Technologies

2.3.1 Research Centres

We measure the number of research institutions working in information retrieval by comparing the number of organizations 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 organizations have been identified. 142 are from the countries included in our analysis. Most of the

[2] http://www.isca-students.org/?q=speechlabs

organizations (63 in total) are from Europe, while there are 47 institutions from North America and 32 from Asia.

2.3.2 Publications

Similarly, to the two other language technology 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. However, the analysis demonstrated a constant decrease of publications in 2010-2017, with less than 60 publications per year. 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. Our analysis of publications related to information retrieval for 2000-2017 also demonstrates more interest in this topic before year 2010. However, the number of publications after 2010 is rather stable (about 6K a year).

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 considered. 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 4).

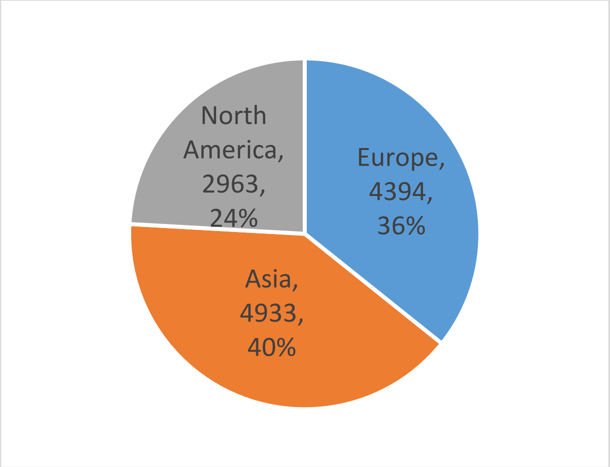

Figure 4 : Distribution of publications between regions when querying for "information retrieval" together with "text" or "word" (2010-November 2018).

3. Innovation

We defined innovation as "the implementation of a new or significantly improved product (good or service), or process, a new marketing method, or a new organizational method in business practices, workplace organization 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, the emergence of start-ups in the respective industry across regions and the known implementation of the latest technique in the respective area. The start-up companies were analysed using the AngelList database[3].

3.1 Innovation in Machine Translation

The analysis of comparative advances in MT by Vasiljevs et al. (2019b) concluded that Europe is a global leader in innovating translation technologies and services. They showed that Europe leads in development and implementation of translation automation tools, while North America comes in second. The same situation has been observed in the area of translation technology start-ups. However, global adoption of neural MT is led by global providers Google and Facebook but European companies and public services are quick to follow.

3.2 Innovation in Speech Technologies

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 a deterrent to market entry by local entrepreneurs and innovators. This conclusion has been corroborated by IDC data[4].

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, 2018 altogether there are 204 start-ups in the field of voice/speech recognition. The majority (113) are located in North America, while 51 are registered in Europe but 25 in Asia. 15 companies were registered in other regions (South America, Africa, Australia) or information regarding their location was not provided.

Developments in natural language processing and neural network technology have improved the speech and voice technology so much so that today it is reportedly on par with humans. However, although Google supports 119, Nuance over 86 languages and dialects, the speech recognition performance among languages is not equal.

It can be concluded that North America is the global leader in innovating speech technologies and services, with Europe coming in second.

3.3 Innovation in Search Technologies and Services

There is strong evidence that Google is a global leader in web search technologies covering 92% of 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% market share is Baidu while Google comes in at 0.61% and Bing at 0.37%.[5] Moreover, in Russia Yandex is aggressively expanding its ecosystem beyond its core search engine. As a result, Yandex leads the Russia market with 57.9%, leaving Google in second place with 43,3% of market share.[6]

[3] https://angel.co/

[4] IDC 2018 for SMART 2016-0103 Lot 1

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

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

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) The magazine *CIO Application* has collected information from enterprises and created their list of the most reputable tools[7]. (2) Analysts from G2 Crowd have done research on the 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 most efficient solution[8]. (3) At the same time the business review journal Business Online has created the list of top 20 companies that most fit enterprise needs.[9] (4) We also looked at the list which is purely based on reviews of open source tools.[10] Based on these findings we created a list that reflects the most popular search tools summarized in Table 1[11].

Search engines	***Language[12],[13]***	***Region (HQ)***	***Market Share***
Google	Multiling.	N. America	92.31
bing	Multiling.[14]	N.America	2.27
Yahoo! (powered by bing)	Multiling.[15]	N. America	2.51
Baidu	Chinese	Asia	0.85
YANDEX RU	Multiling.[16]	Other (Russia)	0.61
Shenma	Chinese	Asia	0.18
YANDEX[17]	Multiling.[18]	Other (Russia)	0.31
DuckDuckGo	Multiling.	N. America	0.33
Naver	Korean	Asia	0.18
Haosou	Chinese	Asia	0.08
Sogou (runs CLIR platform 'Sogou English')	Chinese/ English[19]	Asia	0.1
MSN (powered by Bing)	Multiling.	N. America	0.08
Daum	Korean	Asia	0.02
Mail.ru		Other (Russia)	0.04
Seznam	Czech	Europe	0.04
Ask Jeeves/ Ask.com	Multiling.	N. America	0.01
CocCoc (powered by Google)	Vietnamese/ English[20]	Other (Vietnam)	0.02
Other			0.06

Table 1 : Market of origin of most popular search tools.

Another indicator of innovation is the emergence of start-up companies. The analysis of the regional distribution of search technology start-ups from the AngelList database shows that North America is the leader in the number of emerging start-up's followed by Europe, and Asia in a distant third position.

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 to seek solutions.

4. Investments

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.

4.1 Investments in Machine Translation

Although Europe may have a global lead in research, as noted above, it lags in investment capacity. North America has a dominant presence in machine translation developed by the U.S.-based technology giants (such as Facebook, Google, Amazon, and Microsoft). In addition, North America also dominates the translation sector and by association also the machine translation component (more details in Vasiljevs et al., 2019b).

4.2 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 can be seen from the extensive list of Speech technologies recent venture capital and start-up financing transactions[21] and summary in Table *2*, 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.

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

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

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

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

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

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

[13] www.searchengineshowdown.com/language/limits.shtml

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

[15] https://angel.co/
https://developer.yahoo.com/search/languages.html

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

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

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

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

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

[21] https://index.co/market/speech-recognition/investments

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

Table 2 : Funding by region.

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 larger.

4.3 Investments in Search Technologies

Based on information gathered by Index.co,[22] Asia dominates the market in terms of attracted investment by search companies. From 2012 to 2018 Asian search companies have attracted more than 9 billion USD of funding, while North American companies have attracted nearly 5 billion and the EU – 1.1 billion (Table *3*).

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

Table 3 : Investments in search technology companies by region.

5. Market Dominance

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 dedicated web domains of the largest providers of the respective Language Technology services.

We selected two web traffic analysis tools for collecting and analysing data. Semrush[23] is used to analyse market dominance in all categories and subcategories, except Web search service providers, where we selected specialised analysis tool Statcounter[24] that specifically provides web search tool traffic analytics.

5.1 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 their services. With their relatively few, 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 overall.

5.2 Market Dominance in Speech Technologies

In this study we 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.

The speech and voice recognition market is almost completely dominated by the large US based global corporations that are using speech recognition technology as part of their product or service functionality to enhance closer communication with the end user. Our analysis (Figure 5) clearly demonstrates the extensive market dominance by the North American players followed by a tiny fraction of the web traffic to the three Asia based company (Brianasoft, IFlytek, Auraya Systems) websites. The web traffic of the leading Speech recognition service provider Nuance has a tenfold advantage over the closest follower Google, and Google has an eightfold advantage over the next in the row. There are no European companies among the 15 largest speech/ voice recognition service providers.

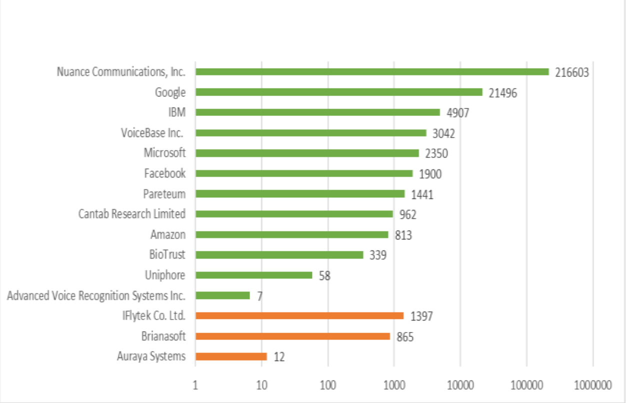

Figure 5 : Monthly speech recognition dedicated website visits, averaged from 03.-09.2018, green – North America, orange – Asia. Logarithmic graph.

In the speech synthesis market (Figure 6) the US based tech giants outperform the top companies focussed on speech synthesis alone (exc. Hoya) multiple times with Google being the 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.

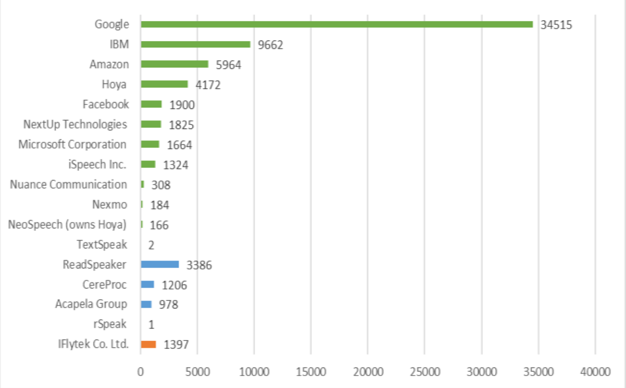

Figure 6: Monthly speech synthesis dedicated website visits, averaged from 03.-09.2018, green – North America, blue – Europe, orange – Asia.

[22] https://index.co/market/search

[23] https://www.semrush.com/company/

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

5.3 Market Dominance in Search Technologies

In order to conduct 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 by Google Search in the web search market of the respective regions, whereas the enterprise search tool market is led by the European company Elasticsearch.

6. Industry

Industry in the context of this study is defined as the commercial language technology developers and service providers. The criteria for measuring the Industry dimension is the market capitalization and estimates of market revenues of the companies that can be identified as being engaged in language services.

6.1 Machine Translation Industry

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 capitalization in comparison to the EU based companies. The EU lags behind as European companies have a lesser global presence.

6.2 Speech Technology Industry

Within the speech technology industry there are two distinct segments. The first segment is large developers for whom speech technologies are a competitive advantage technology for enhancing, popularizing, and marketing other products and services such as Alexa by Amazon. This could be considered a B2C (business to consumer) segment. The second segment is developers for whom the technology itself is the product, and who supply speech technologies as a service such as Nuance which supplies speech recognition software for use by Daimler in automobiles. This could be considered a B2B (business to business) segment. As can been seen from Table 4, leading market players in voice recognition software development are located in North America (predominantly the US).

In both segments speech technology has overwhelmingly been developed by North America based companies, followed by their Asian counterparts for whom speech technologies are 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.

6.3 Search Technology and Service Industry

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

All three segments are 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 the information retrieval and analysis potential. While the giant North America based search companies have the European and Arabic language markets wrapped up, Asian companies are fighting it out in their home markets for dominance. There are 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.

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

Table 4 : Leading market players in voice recognition (listed in alphabetical order).[25]

7. Infrastructure

We define infrastructure as the technical (computing) infrastructure needed for developing, running and utilizing

[25] 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

computationally intensive services. While organizational infrastructure (associations and other networking structures) could be evaluated as rather similar for all three regions, the computational infrastructure is more developed by North American headquartered global players (e.g., Google, Microsoft, and Amazon). Europe lacks high performance computational resources which could be an obstacle and result in slower R&D of computationally heavy LTs in Europe.

Availability and access to computing infrastructure is key to developing competitive high-performance LT 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 LT services. Rapid development of cloud computing democratizes access to the high performance computing needed for developing state-of-the-art machine translation systems. Running machine translation services on the cloud also dramatically extends the reach of machine translation.

According to estimations by the European Commission, Europe needs to invest close to $800bn in its digital infrastructure to catch up with the United States and China.[26] Although this is a total estimation 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 (HPC) power.

Europe is lagging behind other global economic powers in providing computing power for computing intensive applications. Although Europe consumes 29% of global HPC resources it supplies less than 5% of them (Figure 7).[27]

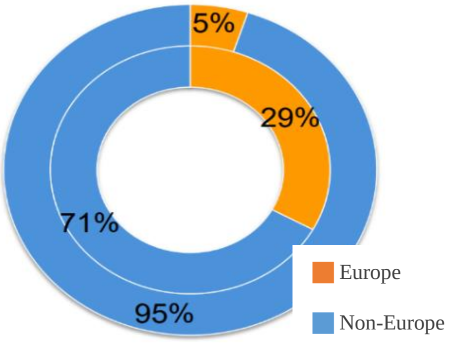

Figure 7 : Europe's consumption of the global HPC resources (29%) versus HPC resources supplied in Europe (5%).

8. Data

As indicators for data availability by region, we analysed the (1) availability of open data, (2) access to proprietary data resources and (3) legal regulations of the data usage.

8.1 Data for Machine Translation

In our study we came 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. In regard to proprietary data and user generated content, global online US and Asia companies have a strong advantage versus European players. Finally, European copyright regulation is much more restrictive for data usage compared to the United States. Lack of the fair use principle makes huge volumes of copyright protected data inaccessible to European researchers and machine translation developers. At the same time US businesses and research institutions reap an advantage by using this data based on the fair use exception.

8.2 Data for Speech Technologies

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. Most speech data are available for English and Mandarin, some data are available for German, French, Italian and Spanish. A lack of open 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 key reasons for the speech technology quality gap between languages.

8.3 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 top10 most popular web search sites and (2) usage of language in internet.

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 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. InternetWorldStats estimates the number of English language Internet users is25.4%, while Chinese is used by 19.30% of Internet users.[28] Although Chinese is the second largest language in terms of number of users, the total number of European Internet users exceeds it. Therefore, “top ten language” as criteria used to identify advantages for data must be carefully evaluated.

9. Conclusions and Recommendations

Graphical summaries of the comparative ranking below (Figure 8, Figure 9, Figure 10) provide a visual overview of the relative positions (based on a score from one to three)

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

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

[28] https://www.internetworldstats.com/stats7.htm

of the major economic regions (markets) within the dimensions we have selected to juxtapose.

The study demonstrates that Europe is traditionally strong in research and innovation but has problems in scaling innovations and conquering market share. A fragmented market is one of the issues strongly influencing the development of European language technology. 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 datasets for all languages. Technology providers, potential customers, and research should have a place to cooperate.

Language technology is a powerful enabler allowing small and large European businesses to reach 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. Thus, it is extremely important for Europe to develop its own language technologies in order to avoid dependence on US/Asian providers.

Public intervention is needed to address market failures. Public procurement is an efficient approach to drive public demand for 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, 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, encourage their faster and more efficient production, and 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 in basic and applied research. However, an increase in research efficiency is necessary, and the next scientific breakthrough is very much awaited. There is a need for an holistic approach on the European, national, and regional levels for coordination of actions and policies and to line up research activities and projects. Politics, business, research and society should all participate in the initiative. 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.

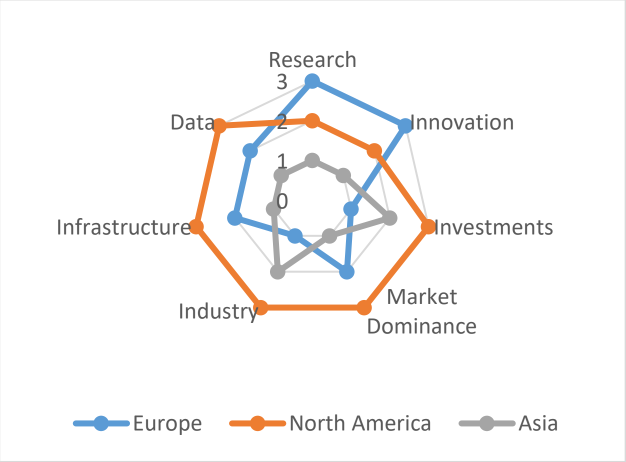

Figure 8 : Comparative position of European machine translation market versus North America and Asia regions.

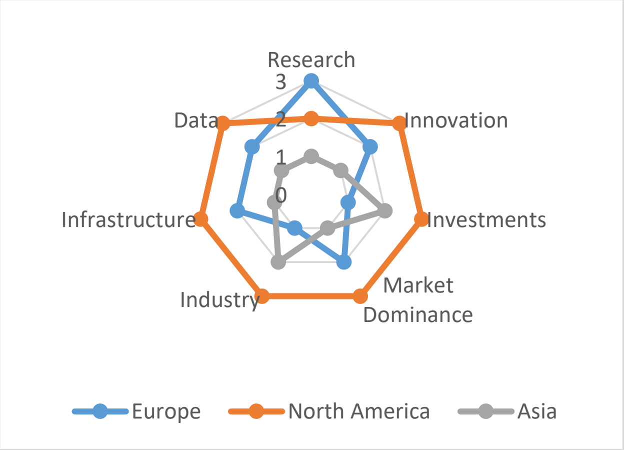

Figure 9 : Comparative position of European speech technology market versus North America and Asia regions.

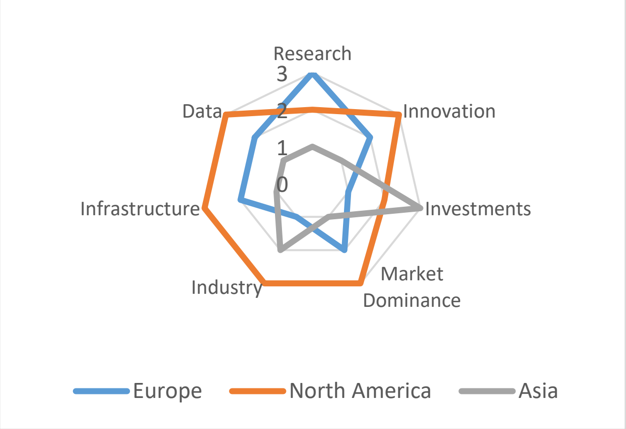

Figure 10 : Comparative position of European search technology market versus North America and Asia regions

9. Acknowledgements

This study has been funded by the CEF Automated Translation programme project SMART 2016/0103.

10. Bibliographical References

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.

Lommel, A. R., and DePalma, D. A. 2016. Europe's Leading Role in Machine Translation: How Europe is Driving the Shift to MT. Cambridge, Massachusetts: Common Sense Advisory.

Massardo, I., van der Meer, J., and Khalilov, M. 2016. TAUS Translation Technology Report. TAUS.

META-NET. 2015. Strategic Research Agenda for the Multilingual Digital Single Market. Technologies for Overcoming Language Barriers towards a truly integrated European Online Market.

Scopus. 2017. Scopus.Content Coverage Guide. https://www.elsevier.com/?a=6945

Seligman, M., Waibel, A., and Joscelyne, A. 2017. TAUS Speech-to-Speech Translation Technology Report. TAUS.

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

TAUS. Joscelyne, A. (Ed.), 2017. TAUS Machine Translation Market Report. TAUS.

Vasiljevs, A., Choukri, K., Meertens, L., Aguzzi, S. (2019a). Final study report on CEF Automated Translation value proposition in the context of the European LT market/ecosystem. Publications Office of the European Union.

Vasiļjevs, A., Skadiņa, I., Sāmīte, I., Kauliņš, K., Ajausks, Ē., Meļņika, J. and Bērziņš, A. (2019b). Competitiveness Analysis of the European Machine Translation Market. In Proceedings of MT Summit XVII, volume 2, pages 1-7.