Competitiveness Analysis of the European Machine Translation Market

Andrejs Vasiļjevs, Inguna Skadiņa, Indra Sāmīte, Kaspars Kauliņš, Ēriks Ajausks, Jūlija Meļņika, Aivars Bērziņš
Tilde, Vienibas gatve 75a, Riga, LV 1004, Latvia
firstname.lastname@tilde.lv

Abstract

This paper presents the key results of a study on the global competitiveness of the European Machine Translation market in comparison to North America and Asia. The study focuses on seven dimensions that have been selected to characterize the machine translation market. The study concludes that while Europe still has strong positions in Research and Innovation, it lags behind North America and Asia in Industry and Investments, and is also weaker than North America in Infrastructure, Data availability, and Market visibility.

1 Introduction

The aim of this study was to analyze a competitiveness of the European machine translation (MT) market in comparison to North America (Unites States and Canada) and Asia (China, Japan, India, South Korea and Singapore).

This research is a part of a wider undertaking to identify possible shortcomings and opportunities for the European Language Technology (LT) market and identify potential actions that need to be addressed at the European Union level.

The analysis is based on an extensive desk research of various studies, policy papers, and online information sources. The quantitative foundation of the analysis is based on the surveys and interviews done by and analysed under the leadership of IDC in the framework of the SMART project[1]. It is also an aggregation and analysis of data collected from previous studies on MT and the broader localization and translation sector, and overall economic indicators (e.g., World Economic Forum, 2017; Common Sense Advisory (Lommel et al., 2016); TAUS (Massardo, 2016; Seligman, 2017; TAUS, 2017); CRACKER (2015; SRIA, 2017) and META-NET (2015)).

The study focuses on seven dimensions that can characterize the machine translation market as part of the broader language technology market: Research, Innovation, Investment, Market dominance, Industry, Infrastructure, and Open Data. These dimensions were analysed for global competitiveness, highlighting the most important achievements and gaps in the LT ecosystem between Europe and its largest global competitors – North America and Asia. To characterise each dimension, a number of criteria were analysed. Using these results, we have ranked the markets within each dimension on a scale from 1 (weakest) to 3 (strongest).

The full report of the findings from the study has been submitted to the European Commission. In this paper we have summarized the key findings of this report.

2 Competitiveness of European MT Research

The following criteria were used as quantitative indicators: number of research centres, number of research publications, organizational infrastructure (e.g. associations, networks and research infrastructures).

We analysed publicly available information about research centres in different countries. 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 their size.

[1] Study on service portfolio development and implementation of the "service desk" component of the CEF Automated Translation platform, SMART 2016/0103 Lot 1

2.1 Research Centres

The recent Wikipedia article "List of research laboratories for machine translation" lists 113 institutions, from which 91 are in scope of our study. This list includes academic, governmental, and corporate sites. This list confirms a strong research capacity in Europe, as it has 47 academic research centers compared to only 18 in America and 9 in Asia (see Table 1).

	ACADEMIC	GOVERN-MENTAL	CORPO-RATE	TOTAL
EUROPE	47	1	6	**54**
ASIA	9	4	1	**14**
AMERICA	18	1	4	**23**
TOTAL	**74**	**6**	**11**	**91**

Table 1. Number of research laboratories for machine translation in different regions

The higher number of European research centres compared to the number of North American research centres is also reflected in the membership of the European Association of Machine Translation (EAMT)[2] that lists 43 R&D groups and 16 corporate members. The American Association of Machine Translation (AMTA) lists 15 academic research organizations and 6 industrial research labs[3]. The Asia-Pacific Association for MT has 32 corporate members and over 66 individual members[4].

2.2 Publications

In this study, we researched publications in the Scopus database[5]. The research publications include both academic and industry researchers. However, it could be that industry research is underrepresented, since not all industry research results are made public. 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. Both Scopus and WoS are well established academic citation indexes that are widely used to assess the outcome and impact of scientific work. However, Scopus has better coverage for 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

We analysed the publications in the Scopus database retrieved by querying for "machine translation" in title, abstract, and keywords. Figure 1 shows the number of publications for the time period from 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.

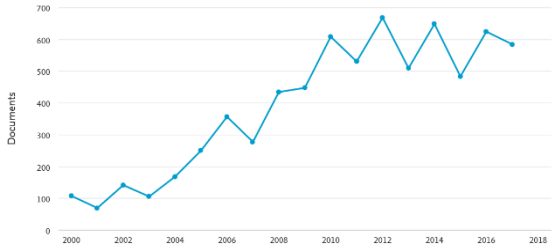

Figure 1 Number of publications for "machine translation" (2000-2017)

When querying for "machine translation" for the years 2010-2018, we found 4931 publications, 4723 of these publications are from the countries/regions addressed in this study (on July 10, 2018). Publications on CAT tools were not included and analysed in this study, because the number of publications on CAT tools alone[6] in the Scopus DB for 2010-2018 is very small (only 149 additional publications or about 3% were found).

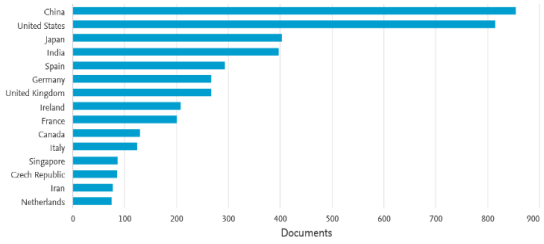

Figure 2. Number of MT related publications in Scopus database: top 15 countries (2010-June 2018)

Figure 2 shows the 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

[2] http://www.eamt.org/, retrieved on 12.07.2018
[3] https://amtaweb.org/resources, retrieved on 12.07.2018
[4] http://www.aamt.info/english/about/01.php, retrieved on 12.07.2018

[5] The Scopus database can be found in https://www.scopus.com/
[6] Publications that do not mention "machine translation" in title, abstract, or keywords

(208 publications), France (200 publications), Italy (124 publications), Czech Republic (85 publications), and the Netherlands (75 publications).

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 3).

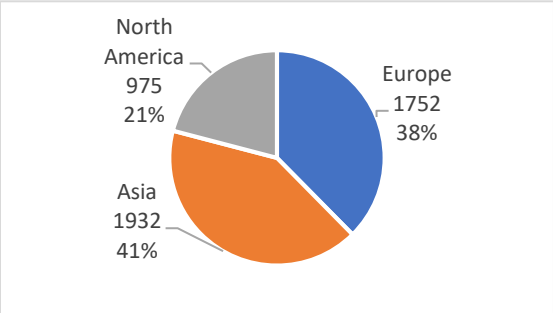

Figure 3. Distribution of publications between regions (2010-2018)

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 (Table 2).

AUTHOR NAME	NUMBER OF PUB.	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

Table 2. Authors publishing on MT (2010 - June 2018) with more than 25 publications (top 20) according to Scopus: distribution between countries and regions

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

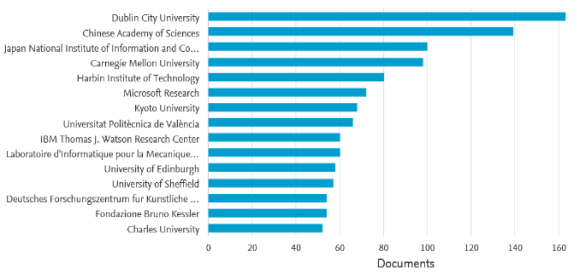

Figure 4. Top 15 organizations that published papers on machine translation (2010-June 2018) in Scopus

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

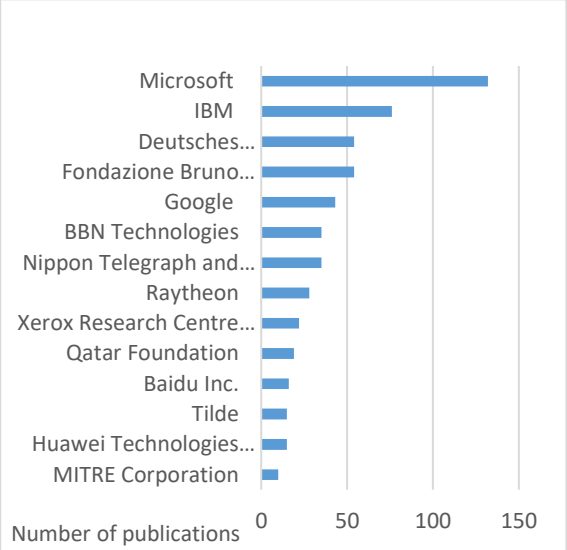

Figure 5. Industry and privately financed organisations that published on MT (2010-June 2018) in Scopus

We also analysed conference proceedings from ACL, COLING, EACL, NAACL and NIPS[7] - five important computational linguistics conferences by querying for "machine translation". We found more recent (2015-2017) papers from United States (68) and China (42), but fewer from Germany (34), United Kingdom (27), Ireland (21) and other European countries. While US authors have more publications as authors from each single EU or Asian country, European countries are still leaders, when the regional distribution of publications are compared.

3 Innovation

As proxies for innovation by region, we analysed the market of origin of the most popular tools,

[7] ACL (2010-2017), COLING (2010, 2012, 2013, 2014, 2016), EACL (2010, 2012, 2013, 2014, 2017), NAACL (2010, 2012, 2013, 2015, 2016), NIPS (2010-2017) proceedings were indexed in Scopus by the time of this study.

emergence of start-ups in the respective industry across regions, and known implementation of the latest technique in each respective area.

3.1 Market of origin of the translation automation tools

Parallel to MT technologies, we have witnessed dynamic innovation in computer assisted translation (CAT) tools that play a major role in automating professional translation. Despite a huge improvement in the quality of machine translation thanks to the advances in neural MT (e.g. Bojar et al., 2018), recent research has shown that MT systems are still not able to produce translations of sufficient quality at the sentence level and even more so on document level. Often machine translation output still requires post-editing by a human to correct errors and improve the quality of the translation (Läubli et al., 2018; Hassan et al., 2018; Toral et. al., 2018).

CAT incorporates this manual editing stage into translation software, making translation an interactive process between human and computer. 11 out of the 24 recognized CAT tools that are used by the majority of translation companies have been developed in Europe.

3.2 Translation technology start-ups

Another indicator of innovation is the emergence of start-up companies that introduce new technologies, innovative ways of addressing business needs, and novel business models. For this analysis we collected a list of translation technology start-ups from AngelList[8] – a U.S. website for startups and angel investors – and assigned their regional attribution based on the location of their headquarters. Europe is the leader in the number of emerging start-ups (54) closely followed by North America (51), leaving Asia in a distant third position (28).

3.3 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 NMT is led by Google (Wu et al., 2016) and Facebook but European companies and public sector have been quick to follow. In a few months from the first release of the Chinese-English NMT by Google there were numerous NMT systems launched by European companies Tilde (Pinnis et al, 2017), KantanMT, SDL, and DeepL. The European Commission also is on the fast track to adopt NMT by replacing the MT@EU statistical MT systems with the NMT systems on the eTranslation platform[9].

4 Investments

Based on data from translation industry research by Common Sense Advisory (2017) and the Slator 2018 Language Service Provider Index (2018), Table 3 lists the top 20 global translation companies by turnover. Nearly all the top 20 are investing in MT by either developing their own or buying existing MT service providers. Many have the latest NMT technologies illustrating how very important cutting-edge technologies are in the language services sector.

COMPANY	COUNTRY	ACTIVITIES & ACQUISITIONS	TURNOVER[10]
Lionbridge	US	Bought CLS Communication (2014) 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
LanguageLine Solutions	US	Sold to Teleperforma (FR) for $1.5 b (2016)	$451m
SDL	GB	Aquired Language Weaver for $42.5 (2010) In-house NMT	$388.5m $56 m LT turnover
RWS Group	GB	Uses SDL MT	$221.5m
Welocalize	US	Uses 3rd party MT (Microsoft, Iconic MT etc.)	$200 m
STAR Group	CH	In-house MT	$166.2m
Amplexor	LU	Aquired Sajan for $28.5 (2017)	$175.6m
Moravia	CZ	In-house MT Acquired by RWS (2015)	$100m
Hogarth Worldwide	GB	No info	$177m
CyraCom International, Inc.	US	Interpreting, looking for early stage investment	$161m
RR Donnelley Language Solutions	US	In spin-off mode	$93m
Semantix	SE	No info	$107m
Honyaku Center Inc.	JP	Acquired Media Research Inc for $4.8 (2017)	$26m
Pactera Technology International Ltd	CN	Sold for $675m to HNA EcoTech (2016)	$85.2m
Ubiqus	FR	Interpretation, no known MT	$82.6
Keywords Studios	GB	Games, audio	$180.1m
United Language Group (ULG)	US	ULG purchased Lucy MT for an undisclosed amount (2017)	$79m
Logos Group	IT	No information on MT available	No info

[8] https://angel.co/

[9] https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/eTranslation

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

Capita Translation and Interpreting	GB	Acquired through merger SmartMate MT	$178m

Table 3. Top 20 global translation companies: Activities and acquisitions

5 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. We analysed the market dominance in all three regions by comparing total web traffic (e.g. number of times a unique IP address has opened the webpage of the said company) received by the dedicated web domains of the largest providers of MT services. Based on this analysis, North America clearly dominates the market in terms of attracting customers to their services. With their relatively few companies, 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.

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 their market power and dominance. North American MT providers also have strong market position in Asia and Europe. In Asian markets they face strong local competition from Baidu, Tencent, Sogou and others.

The global MT market has a very high degree of concentration – 20% of the market players[11] account for more than 80% of the revenue. A majority of companies earn on MT less than a million euros annually, indicating that MT market is underdeveloped 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 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 a majority of European based companies face challenges in gaining market visibility and increasing their brand awareness.

Free online MT as a service, e.g. Google Translate, freetranslation.com (powered by Microsoft), Reverso, has a major impact on the MT market. In terms of the perceived value – MT services have been commoditized, even devalued, with a concurrent strong impact on the perceived quality expectations by both individual consumers as well as businesses. "Large players such as Google, Microsoft and Apple have some positive effects, as they strongly contribute to create or increase market awareness. On the other hand, they are tough competitors as they offer mass market free software which is difficult to compete with, especially for SMEs."[12]

6 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 is the market capitalization and estimates of market revenues of the companies that can be identified as being engaged in language services and specifically in MT development and implementation (Table 4).

COMPANY	COUNTRY	INDUSTRY	MARKET CAP 2018 ($B)	IN-HOUSE MT
Apple	US	Tech	851	MT
Alphabet	US	Tech	719	MT
Microsoft	US	Tech	703	MT
Amazon	US	Consumer Services	701	MT
Tencent	China	Tech	496	MT
Berkshire Hathaway	US	Financials	492	
Alibaba	China	Consumer Services	470	MT
Facebook	US	Tech	464	MT

Table 4. Top Global Companies by Market Capitalization and their activities in MT, as of March 31, 2018

[11] "mix of big Internet, pure-play MT and Large LSP/MLV companies such as Google, Systran, Microsoft, SDL" (TAUS, 2017)

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

Table 4 shows the impact of MT on the global economy, by highlighting that 7 of the largest 8 companies by market cap have a notable presence in this technology sector. In addition, comparing independent estimations, we can assume that the global MT market in 2017 was worth $300m – $350m with an annual growth rate close to 20%.

According to the IDC study[13], the estimated European market for translation technologies is EUR 67m ($78.3m). This would lead to an estimation of the share for European MT market in a range of 22%-26% or about a quarter of the global market.

7 Infrastructure

Europe is lagging behind other global economic powers in providing computing power for computing intensive applications such as MT. Although Europe consumes 29% of global HPC resources it supplies less than 5% of them.

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.[14] Although this estimate includes investments in fibber-optics networks, 5G networks and other ICT infrastructure, a substantial part of these investments is needed to meet European demand for high performance computing power.

8 Data for Machine Translation

Availability of data is crucial as almost all contemporary machine translation systems are based on data-driven techniques.

As indicators for data availability, we analysed the availability of open data, access to proprietary data resources, and legal regulations of data usage. 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. The European Open Data Portal[15] provides access to diverse language resources. It also contains a dedicated repository of public sector language resources for MT created and populated by the European Language Resource Coordination Action[16], funded by the EU Connecting Europe Facility programme (Lösch et al., 2018).

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, Japan government open data portal) provide only few if any language resource.

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 MT.

This is also true for Chinese firms like Alibaba and Tencent, which have become similarly dominant in their home market (Giles, 2018).

European copyright regulation is much more restrictive for data usage comparing to the United States. Lack of the fair use principle makes huge volumes of copyright protected data unavailable for use by European researchers and machine translation developers (Hugenholtz, 2013; Von Lohmann, 2017). At the same time US businesses and research institutions reap an advantage by applying the fair use exception and using this data.

9 Summary

Figure 6 summarizes the global position of the European MT market using a simple 3 point score representational graph.

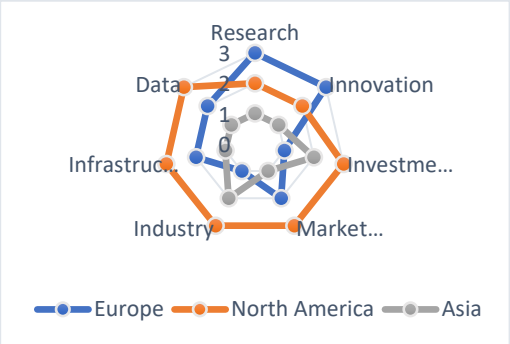

Figure 6. Comparative position of European machine translation market versus North America and Asia regions (1 – weakest, 3 - strongest).

[13] SMART 2016/0103 Lot 1

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

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

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

References

Bojar, O., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Koehn, P., Monz, C. 2018. Findings of the 2018 Conference on Machine Translation (WMT18). *Proceedings of the Third Conference on Machine Translation (WMT),* Volume 2: Shared Task Papers, 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*. http://www.cracking-the-language-barrier.eu/wp-content/uploads/SRIA-V1.0-final.pdf

Crego, J., Kim, J., Klein, G., Rebollo, A., Yang, K., Senellart, J., … & Enoue, S. 2016. SYSTRAN's Pure Neural Machine Translation Systems. *arXiv preprint arXiv:1610.05540*

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., ... & Liu, S. 2018. Achieving Human Parity on Automatic Chinese to English News Translation. *arXiv preprint arXiv:1803.05567.*

Hugenholtz, P. B. 2013. Fair use in Europe. *Communications of the ACM*, 56(5), 26-28.

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.

Läubli S., Sennrich, R., Volk, M. 2018. Has Machine Translation Achieved Human Parity? A Case for Document-level Evaluation. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing,* 4791–4796.

Lösch, A., Mapelli, V., Piperidis, S., Vasiļjevs, A., Smal, L., Declerck, T., Schnur, E., Choukri, K. and Van Genabith, J. 2018. European Language Resource Coordination: Collecting Language Resources for Public Sector Multilingual Information Management. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018),* 1339-1343.

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. http://www.meta-net.eu/projects/cracker/multimedia/mdsm-sria-draft.pdf.

Pinnis, M., Krišlauks, R., Miks, T., Deksne, D. and Šics, V. 2017. Tilde's Machine Translation Systems for WMT 2017. *Proceedings of the Second Conference on Machine Translation*, Volume 2: Shared Task Papers, 374-381.

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

Slator. 2018. The Slator 2018 Language Service Provider Index: Slator.

Strategic Research and Innovation Agenda. 2017. *Language Technologies for Multilingual Europe: Towards a Human Language Project*. Retrieved from: http://cracker-project.eu/wp-content/uploads/SRIA-V1.0-final.pdf

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

Toral, A., Castilho, S., Hu, K., & Way, A. 2018. Attaining the Unattainable? Reassessing Claims of Human Parity in Neural Machine Translation. *arXiv preprint arXiv:1808.10432*.

Von Lohmann, F. 2017. Fair use as innovation policy. *Copyright Law* (pp. 169-205). Routledge.

World Economic Forum. 2017. Schwab, K. (Ed.), *The Global Competitiveness Report 2017-2018*. Geneva: World Economic Forum.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., … & Klingner, J. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*