Revue des arts
et médiations humaines

Journal of arts
and Human Mediations

Journal *Hybrid*, no. 7

The Creative Web of Languages

Smart and/or diverse: the paradoxes of machine processed language

Claire Larsonneur

Claire Larsonneur is a senior lecturer at Paris 8 University within the TransCrit team. Her work focuses on digital literature, the economic, technological and social context of translation and digital humanities. She co-directed the Labex Arts-H2H project "Le sujet digital," the Cerisy symposium "Posthumans and digital subjectivities" and *Angles* edition on "Digital Subjectivities" and is part of the project team "Epistemologies and practices of digital humanities" Paris 8, UVSQ, OPERAS project). Her latest book, *When Translation Goes Digital*, co-edited with Dr Renée Desjardins (University of Saint-Boniface, Canada) and Dr Philippe Lacour (University of Brasilia, Brazil) will be published at the end of 2020 by Palgrave Macmillan.

Abstract

Machine language processing tools affect the global online landscape of languages in complex ways. Free NMT tools, such as Google Translate or DeepL, and smart assistants, such as Alexa or Siri, enable people from all over the world to access and produce content in many languages, furthering what I would call interface multilingualism in terms of quantity, visibility and accessibility. But upon closer examination this comes at a cost for linguistic diversity, as these tools tend to streamline and standardize expression, are based on a limited number of corpora and value user engagement metrics over the quality of content.

Keywords

natural language processing, diversity, multilingualism, artificial intelligence, AI, smart personal assistants

Published : 15 June 2021

Full text (PDF file)

According to the predictions of a team of AI experts for the Future of Humanity Institute, computers will outperform humans in translation by 2024 and in writing a *NY Times* bestseller by 2049.[1] These predictions rest on recent and remarkable advances in language processing technologies, due to improved AI techniques, increased computing power and huge datasets collected mostly on the net. Although a Booker Prize bot recipient is still science fiction, a growing proportion of texts in our daily lives are now produced industrially, from manuals to weather prediction. Machine translation is currently the best example of those language processing techniques becoming commonplace through such major platforms as Amazon, Twitter, Facebook or Google.

At the same time, the number of languages used online appears to be on a sharp increase. Multilingualism has indeed now become the desired standard for large corporations or public bodies, rather than the exception. One website, that of Jehovah's Witnesses, is currently available in 750 different languages, which earns it the title of "the world's most translated website." Wikipedia operates in 287 languages and Google in 149.[2] Facebook is available in more than a 100 different languages and boasts of enacting more than 6 billion translations per day involving more than 4,000 language pairs.[3] Globalization has boosted the need for translation all over the world, leading to a boom in the translation and localisation industry: the global language services market has doubled in size in ten years, reaching 46.9 billion dollars in 2019[4] and the global machine translation market is expected to grow by 19%

[1] Charlotte Edmond, "This is when a robot is going to take your job, according to Oxford University," 26 July 2017. [Online] https://www.weforum.org/agenda/2017/07/how-long-before-a-robot-takes-your-job-here-s-when-ai-experts-think-it-will-happen/ [accessed 18 September 2019].

[2] Mishel Shaji, "What is the most translated website in the world?," 17 March 2019. [Online] https://www.quora.com/What-are-the-top-three-most-translated-Websites-and-how-many-languages-do-they-each-accommodate [accessed 18 September 2019].

[3] Facebook does not specify whether the impressive figure of 6 billion translations refers to the number of characters, the number of words or the number of documents: Khari Johnson, "Facebook Messenger launches translations by intelligent assistant M," 1st May 2018. [Online] https://venturebeat.com/2018/05/01/facebook-messenger-launches-translations-by-intelligent-assistant-m/ [accessed 18 September 2019].

[4] Elena Mazareanu, "Market size of the global language services market," 9 August 2019. [Online] https://www.statista.com/statistics/257656/size-of-the-global-language-services-market/ [accessed 3 July 2020].

between 2020 and 2024.[5] This growing demand for translation fuels the demand for machine language processing technologies which can deliver instant and cheap solutions. In return, more efficient machine language processing enables the creation and circulation of more content in more languages. There appears to be a loop between these technologies and online multilingualism.

But having more multilingual content online and offline does not mean that there are more multilinguals—defined as individuals who routinely use three or more languages.[6] On the contrary, easier access to content in one's first language may dampen the desire to learn other languages. And does multilingualism coincide with linguistic diversity? While the communication strategies of the most online platforms proudly emphasize the large number of languages they operate in, the European Union recently highlighted the danger of digital extinction facing more than 20 European languages, lesser spoken and insufficiently present on the web, in its resolution entitled "Language Equality in the Digital Age," passed on 11 September 2018. Paolillo argues that Internet favours large languages and the associated technical standards, while its impact on small and minority languages can either weaken or strengthen them.[7] Finally, one may wonder what representations and values are attached to machine language processing. Manohar Paluri, Director of AI for Facebook, famously stated in May 2019 that artificial intelligence "gives us a powerful tool to think about language problems in a language agnostic way," while Google Translate announced in July 2019 a shift towards massively multilingual neural machine processing. How can the same technology both be language agnostic and massively multilingual?

In the spirit of what Yves Citton calls the "archeology of media,"[8] I will investigate the specific technical cogs and wheels of language processing technologies to study their impact

[5] Technavio, "Globalization of business to boost growth." [Online] https://www.businesswire.com/news/home/20200529005034/en/Machine-Translation-Market-2020-2024-Globalization-Business-Boost, 29th May 2020 [accessed 3 July 2020].

[6] Charlotte Kemp, "Defining multilingualism," in Larissa Aronin and Britta Hufeigein (eds.), *The Exploration of Multilingualism*, Amsterdam, John Benjamins Publishing, 2009, p. 14.

[7] John Paolillo, "Language diversity on the Internet: examining linguistic bias," in UNESCO Institute for Statistics, *Measuring Linguistic Diversity on the Internet*, Paris, UNESCO, 2005, p. 55.

[8] Yves Citton, *Médiarchies*, Paris, Seuil, 2017, p. 194: "L'archéologie des media, c'est surtout une autre façon d'approcher dans le long terme la matérialité physique des modes de communication qui conditionnent tout à la fois nos organisations collectives et nos représentations mentales. C'est dans le fonctionnement matériel, généralement caché, des appareils qu'il convient d'aller chercher la raison des qualités occultes qui en émanent. L'hypothèse de départ est que nos appareils en savent davantage que nous-mêmes sur ce qu'ils font de nous quand nous croyons nous servir d'eux." [The point of media archeology is to apprehend differently, in the long

on multilingualism and linguistic diversity. Are we witnessing the advent of a more multilingual but less linguistically diverse world economy? In terms of machine language processing, could smart spell the end of diverse? I shall start by discussing the differences between multilingualism and linguistic diversity, before focusing on the case study of neural machine translation within the current ecosystem of human/machine linguistic cooperation. This will lead me to question the impact of machine language processing on linguistic diversity on several levels.

Multilingualism versus linguistic diversity

"With 10 languages accounting for nearly 90% of the top 10 million websites, a tiny fraction of the world's 7,097 languages dominate online life," notes Emma Charlton for the World Economic Forum.[9] Although the historically dominant share of English online is dwindling due to rapid growth of online content posted in Arabic, Russian, or Chinese, mapping languages online remains a complex issue. Internet World Statistics for instance focuses on the number of Internet users by language and Internet penetration within a given population.[10] It shows there can be stark discrepancies between the number of native speakers and internet users, as for Arabic where only 53% of Arabic speaking people are Internet users or French (only 35.2%), compared to German or Japanese, where the penetration rate is 93.8%. People may also use several languages online (for instance Korean and English) to varying degrees and in varying situations.

This linguistic complexity is reflected in the many and overlapping definitions of multilingualism, as shown by Charlotte Kemp. The word multilingual usually refers to individuals who speak three or more languages. She distinguishes this from polyglossia, "a term usually used in sociolinguistics to refer to communities where a number of languages or varieties are used by some or all individuals within a specified community where they have

run, the physical materialities of our modes of communication which frame both our collective organisations and our mental representations. Most devices exert an occult influence which is grounded in their operating mode, often hidden from view. Our premise is that those devices are built to gather intelligence about how they use us, unbeknowst to us while we think we use them. (My translation)]

[9] Emma Charlton, "The Internet has a language diversity problem," World Economic Forum, 13 December 2018. [Online] https://www.weforum.org/agenda/2018/12/chart-of-the-day-the-internet-has-a-language-diversity-problem/ [accessed 18 September 2019].

[10] Internet World Stats, [online] https://www.internetworldstats.com/stats7.htm [accessed 22 June 2020].

different roles."[11] However, one must note that both these definitions do not take into account the online production and circulation of languages: for instance individuals, even when they live offline in a predominantly monolingual space (France for instance), may belong to a bilingual online community such as the gaming community or the open source software community, which make extensive use of English. Viewing online communities as "translocal affinity spaces"[12] enables us to move away from the offline mapping of languages, which often corresponds to physical and political territories, and away from a purely quantitative assessment. Focusing instead on linguistic practices and their complexities may prove a more fruitful path of investigation.

For the purpose of this research, I will propose to distinguish between online multilingualism, as defined by measures of quantity, visibility, and accessibility (such as the number of websites in a given language or the number of languages offered by a website), and linguistic diversity as characterized by practices and uses, involving levels of proficiency, degrees of mutual intelligibility and issues of identity.[13] It will here follow the line suggested by Peter Strevens:[14]

> A central problem of linguistic study is to reconcile a convenient and necessary fiction with a great mass of inconvenient facts. The fiction is the notion of a "language" – English, Chinese, Navajo, Kashmiri. The facts reside in the mass of diversity exhibited in the actual performance of individuals when they use a given language.

When analyzing the online linguistic policies of global corporations, this distinction between multilingualism and linguistic diversity can be useful. For instance, online corporate multilingualism often does not reflect the offline linguistic realities, either because of technical constraints or marketing choices. As of September 2019, the firm Zara is physically

[11] Charlotte Kemp, "Defining multilingualism," in Larissa Aronin and Britta Hufeigein (eds.), *The Exploration of Multilingualism*, Amsterdam, John Benjamins Publishing, 2009, p. 15.

[12] Sirpa Leppänen and Saija Peuronen, "Multilingualism on the Internet," in Marilyn Martin-Jones, Adrian Blackledge and Angela Creese (eds.), *The Routledge Handbook of Multilingualism*, London/New York, Routledge, 2015, p. 390.

[13] Charlotte Kemp, "Defining multilingualism," in Larissa Aronin and Britta Hufeigein (eds.), *The Exploration of Multilingualism*, Amsterdam, John Benjamins Publishing, 2009, p. 23.

[14] Peter Strevens (1982) quoted in Charlotte Kemp, "Defining multilingualism," in Larissa Aronin and Britta Hufeigein (eds.), *The Exploration of Multilingualism*, Amsterdam, John Benjamins Publishing, 2009, p. 16.

present in 96 countries, from Azerbaijan to Ecuador, and has online presence in 202 markets.[15] Its localisation strategy is precise: it covers all the languages of the EU since 2016 and the Swiss website is available in four languages.[16] Yet the main website for its home company Inditex is available only in English and Spanish, and a customer is automatically rerouted to the website for the country in which his or her computer's IP is listed, without being able to access content in other markets and languages. One could say that Zara's online linguistic presence is undeniably multilingual yet does not promote linguistic diversity for the internet user or customer.

Keeping in mind that we want to chart the impact of machine language processing on linguistic diversity, we can examine another online commercial feature: smart assistants like Siri for Apple, Alexa for Amazon, or Cortana for Microsoft. Their heterogeneous language settings offer insights into the diverse rationales of corporate linguistic policies.[17] As of June 2020, Cortana is available in seven languages. It does not distinguish between Australian, British, or American variants of English, nor between Mexican and Spanish variants, and the only bilingual region is Canada (English/French). Apple's Siri is more multilingual. As of September 2019, the settings on a Mac for Siri offer a choice of 41 languages, some localized: there are 9 variants of English, including Singaporean, 4 variants of Chinese, but only one in Arabic, when the ISO standards lists more than 30 varieties of Arabic within the 639-3 section. Discrepancies also characterise gender options in Siri: both male and female voices are available for the American, British, and Australian variants of English, but the Irish and the South African variants only come in the female option. Amazon's Alexa comes in seven languages but until recently offered only the female voice option.[18]

Such discrepancies in corporate linguistic policies appear to stem from a complex network of overlapping rationales: the identification of promising markets, the availability of data, the

[15] [Online] https://www.inditex.com/about-us/our-brands/zara [accessed 22 June 2020]

[16] *Internet retailing* (staff writer), "Zara. Communication without borders," 7 August 2017. [Online] https://internetretailing.net/research-articles/research-articles/zara-communication-without-borders [accessed 17 December 2019].

[17] Globalme.com, "Language support in voice assistants compared" (updated 2019), 28 November 2019. [Online] https://www.globalme.net/blog/language-support-voice-assistants-compared [accessed 17 December 2019].

[18] They moved to a variety of accents and gender in October 2018 by integrating Amazon Polly, a text-reading application: B. J. Haberkorn, "Amazon Polly voices in Alexa skills now generally available," 23 October 2018. [Online] https://developer.amazon.com/fr/blogs/alexa/post/baee53c1-5b03-4580-b57a-ee9510413354/amazon-polly-voices-in-alexa-skills-now-generally-available [accessed 17 December 2019].

degree of linguistic savviness of decision-makers, costs incurred when adding a new language, etc. The language settings of the major software companies, most of which are situated in the USA, are logically conceived and implemented in US English, and only then localized in other languages, those corresponding to profitable markets. For Microsoft's Cortana, languages are distributed into thirteen "regions," an in-house category that corresponds to neither countries nor states, not even markets. Its services do not allow for linguistic switches: "If you change your region, you might not be able to shop at Microsoft Store or use things you've purchased, like memberships and subscriptions, games, movies, and TV."[19] Although Cortana is multilingual, here again one finds restrictions to linguistic diversity in terms of user experience. To the economic aspects of decision making listed above (cost, data etc.), one should add cultural representations. Luke Munn has shown how the initial restriction of Alexa's voice settings to a female only option was rooted in earlier representations of a phone operator as a soft-speaking servant.[20] Siri is on the contrary presented by Apple as a concierge or a butler, with a focus on its conversational skills, which may explain why the range of accents is wider and why it offers more male and female options.

The heterogeneity of the actual landscape of languages on websites and these interfaces of platforms thus still seems to favour dominant languages. To further investigate multilingualism and linguistic diversity in machine language processing, we need to study in more detail the design and technical logic of the AI technologies at play, starting with the case of neural machine translation.

The case of neural machine translation

AI powered neural machine translation can now provide convincing, fluent and continuously improving results. Mainstream freely accessible neural translation tools such as Google Translate or DeepL combine statistical machine translation technologies and artificial intelligence technologies. The statistical method relies on huge corpora of translated segments to match one source text with the most probable target text: the output corresponds the most probable match, a process that favours common occurrences over rarities. AI algorithms also need to be trained on huge bilingual or even multilingual corpora to predict the most probable

[19] [Online] https://support.microsoft.com/en-ie/help/4026948/cortanas-regions-and-languages last updated 27 May 2020 [accessed 22 June 2020].

[20] Luke Munn, "Alexa and the intersectional interface," *Angles. New Perspectives on the Anglophone World*, no. 7 "Digital Subjectivities," June 2018. [Online] https://angles.edel.univ-poitiers.fr:443/angles/index.php?id=1492 [accessed 22 June 2020].

follow-up to the beginning of a sentence. Current results are good enough to challenge human translation: domain-specific target texts produced by neural machine translation are now considered to be usually "fit for purpose."[21] The artificial intelligence technologies at hand in neural machine translation are very similar to those used in other fields where human expertise and agency was considered crucial: medical diagnoses, crop monitoring, self-driving cars, or autonomous drones[22]… Indeed, all these tools rely on massive datasets, pattern recognition, and the training of algorithms. They are also developed by the same teams as exemplified by the career path of Mr. Olszewski: the former head of the Amazon translate project joined Uber in 2019 to work on automated cars, using the same technology.[23] Those decision-making algorithms share one more feature: they have been designed to be embedded in various applications or even physical appliances, such as cars, fridges, smart clothes, or home assistants like Amazon's Echo or Google Home Hub. So AI technologies point to a new era of the digital in which computing technologies literally move beyond the screen to become integrated within our daily environment.

After this rapid overview of NMT technologies, we can examine their impact on the language landscape. Although the point of NMT is to enable fast translation between large numbers of language pairs, thereby increasing multilingualism globally, I contend that it does so at the expense of linguistic diversity. First of all, like other AI technologies, neural machine translation entails information loss, as argued by Pasquinelli, who backs his claim on three counts: the limitations of any training dataset (which are always samples), the use of patterns, categories and taxonomies, and the existence of biases. Such loss is particularly noticeable in the treatment of statistical anomalies such as metaphors, rare words, or coinages.

Drawing attention to the fact that AI is more a technique of information compression than a manifestation of non-human cognitive prowess, Pasquinelli further analyses the impact of market imperatives on the uses and fine-tuning of those tools:

> Since ancient times, algorithms have been procedures of an economic nature, designed to achieve a result in the shortest number of steps consuming the least amount of resources, such space, time, energy, etc. The current arms race between

[21] Joss Moorkens, Sheila Castilho, Federico Gaspari and Stephen Doherty (eds.), *Translation Quality Assessment. From Principles to Practice*, Cham, Springer International Publishing, 2018.

[22] Cedric Villani, *For a Meaningful Artificial Intelligence. Towards a French and European Strategy*, March 2018. [Online] https://www.aiforhumanity.fr/en/ [accessed 20 September 2019].

[23] Marion Marking, "Alon Lavie joins Unbabel. He's not the only exec who recently left Amazon's machine translation group," *Slator*, 18 April 2019. [Online] https://slator.com/people-moves/alon-lavie-joins-unbabel-hes-not-the-only-exec-who-recently-left-amazons-translation-group/ [accessed 22 June 2020).

> AI companies is still about finding the fastest algorithms to compute statistical models. Information compression, therefore, measures the ratio of profit in these companies, but also the ratio of information loss – and that loss often means a loss of the world cultural diversity.[24]

Second, the very logic of neural machine translation is to wrestle free from the constraints of natural languages by encoding them into mathematical representations. Using massive corpuses of bilingual and monolingual phrases, scientists produce semantic maps of relations. In supervised bilingual machine learning, natural language expressions in the source texts are encoded into semantic vectors within that map and then decoded into the target text. The algorithms select the most probable pairings, based on statistical recurrence within the corpus.[25] NMT is thus based on a double departure from the logic of natural languages: through the encryption and decryption process and through statistical selection.

This trend has been furthered by the latest development in NMT research: unsupervised massively multilingual machine translations. Instead of working on language pairs, the research teams of Google and Facebook have built a representation space shared among multiple languages. Facebook trains its multilingual model in 93 languages, 30 language families, and 22 different scripts.[26] Google's model handles 103 languages trained on 25 billion examples, retrieved from the Web by crawling and extracting parallel sentences. In 2019 the Facebook AI teams thus won a translation contest for machine translation between English and Burmese, using unsupervised retro-translation techniques.[27]

However these technologies entail a transfer/interference trade-off in which performance is improved on low-resource languages while it is degraded for high-resource languages due to

[24] Matteo Pasquinelli, "How a Machine Learns and Fails — A Grammar of Error for Artificial Intelligence," *Spheres*, no. 5, "Spectres of AI," November 2019. [Online] http://spheres-journal.org/how-a-machine-learns-and-fails-a-grammar-of-error-for-artificial-intelligence/ [accessed 13 December 2019].

[25] Dorothy Kenny, "Machine translation," in Piers Rawling and Philip Wilson (eds.), *The Routledge Handbook of Translation and Philosophy* (p. 428-445), Milton Park, Routledge, 2018.

[26] Gino Diño, "Facebook says it now has a powerful tool to think about language problems in a language agnostic way," *Slator*, no. 6, May 2019. [Online] https://slator.com/technology/at-f8-facebook-says-it-now-has-a-powerful-tool-to-think-about-language-problems-in-a-language-agnostic-way [accessed 18 September 2019].

[27] Philippe Pajot, "La traduction automatique s'attaque aux langues rares," *La Recherche*, no. 554, December 2019, p. 20 *sq*.

constrained capacity and interference.[28] But these technologies are improving fast, fuelled by massive investment from the Google, Facebook, Apple, Amazon, and their Asian equivalents.[29] One should note that these investments in neural machine translation branch out into fields that go beyond text processing, the aim being for the machines to navigate smoothly between representation learning, content understanding, dialogue systems, information extraction, sentiment analysis, summarization, data collection and cleaning, and voice/text/image navigation.[30] Neural machine translation is part of a range of language processing technologies that includes writing help (Microsoft Ideas), assessment and annotation tools for education (Criterion), smart assistants (Siri, Alexa, Cortana), conversational bots, etc. Though they undeniably contribute to an extension of online multilingualism, these tools do so at a cost for linguistic diversity and towards aims that have little to do with quality language interactions. In the following section, we will extend the inquiry beyond translation and draw upon all these interrelated technologies.

Diversity issues in machine language processing

To further understand in what ways linguistic diversity in enhanced or downgraded by machine language processing, one should examine the different dimensions of the technical and economic ecosystem in which they operate. This means one should look at technics used, metrics involved to assess results, human-machine interaction, recent evolutions in this market and governance issues.

Since language processing technologies rely on vast corpora of monolingual and bilingual phrases, the content they generate reflects the scope and quality of the corpus on which they are trained. In the case of Alexa, the aim is to fuel plausible and engaging conversations with the users, which implies topic detection models, sentence completion technologies, and entity linking. And so the IT teams focused on datasets from newspapers such as the Washington

[28] Esther Bond, “What’s so massive about Google’s massively multilingual neural machine translation?,” *Slator*, no. 18, July 2019. [Online] https://slator.com/technology/whats-so-massive-about-googles-massively-multilingual-neural-machine-translation/ [accessed 18 September 2019].

[29] Claire Larsonneur, “The disruptions of neural machine translation,” *Spheres*, no. 5, “Spectres of AI,” November 2019. [Online] http://spheres-journal.org/the-disruptions-of-neural-machine-translation/ [accessed 17 December 2019].

[30] Facebook has created a consortium funding numerous research projects and Amazon has invested US$ 200 million for voice recognition technologies: Kyle Wiggers, “Facebook founds AI Language Research Consortium to solve challenges in natural language processing,” 28 August 2019, Venturebeat.com. [Online] https://venturebeat.com/2019/08/28/facebook-founds-ai-language-research-consortium-to-solve-challenges-in-natural-language-processing/ [accessed 17 December 2019].

Post, from social media such as Twitter or Reddit, and knowledge bases like Amazon's Evi, Freebase, Wikidata (Wikidata, 2017), Microsoft Concept Graph, Google Knowledge Graph, and IMDb.[31] These are all American-based sources of information, and they only reflect a portion of what humans actually produce as language interaction, moreover in specific contexts. And so they do not reflect the full diversity of anglophone interactions worldwide. Similarly the focus on standard American English in Criterion, an automated writing evaluation software (AWE), means the programme was unable to accept effective rhetorical and stylistic uses of language from alternative traditions derived from class or race.[32] Even within the boundaries of one language, in this case English, these machine language processing technologies appear to restrict linguistic diversity.

Assessment metrics are key: not only do they help promote language processing technologies for prospective users, but they also have an influence on research. Until recently, most neural machine translation outputs were assessed by comparison to human translation through BLEU.[33] But as of 2019, some investors in neural machine translation research have moved on to another metric: user engagement. The focus has thus switched from trying to render the meaning of the source texts as accurately as possible into the target text, to maximising the number of clicks or the amount of time a given user spends on the page or interface. Facebook has patented a new user engagement metrics,[34] and Amazon includes in its research targets the performance of conversations with Alexa, which is "measured using duration, turns and ratings obtained from engagement evaluators."[35] However, user engagement metrics may also contribute to a weakening of linguistic diversity as humans tend to engage more frequently

[31] Ashwin Ram *et al*., "Conversational AI: The science behind the Alexa Prize," *First Proceedings of the Alexa Prize*, 2017, p. 9.

[32] Nickly Hockly, "Automated writing evaluation," *ELT Journal*, vol. 73, no. 1, January 2019, p. 85. [Online] https://doi.org/10.1093/elt/ccy044 [accessed 22 June 2020]

[33] Joss Moorkens, Sheila Castilho, Federico Gaspari and Stephen Stephen (eds.), *Translation Quality Assessment: from Principles to Practice*, Cham, Springer International Publishing, 2018, p. 9 *sqq*.

[34] Esther Bond, "Facebook patents alternative to 'expensive' BLEU," *Slator*, 22 August 2019. [Online] https://slator.com/technology/facebook-patents-alternative-to-expensive-bleu/ [accessed 21 November 2019].

[35] Ashwin Ram *et al*., "Conversational AI: The science behind the Alexa Prize," *First Proceedings of the Alexa Prize*, 2017, p. 12.

and more intensely with specific contents, often news that are controversial, kinky, bizarre, those that are most popular and those that feed into their own vision of the world.[36]

Another dimension of those programmes we need to investigate is the design of human/machine interaction. Most of those tools are designed to enable interaction between several agents, some technical (a smart fridge, a retail platform) and some human. To work smoothly, these tools need to pare down the number of variants for the expression of a given idea. In other words, query input from humans needs to be streamlined. "Requests can either be predefined *formal prompts* that users must know to trigger a desired action, *natural language* requests or accumulations of *sensor data* which, from a user perspective, is often collected unconsciously."[37] Siri's help button for instance displays a number of prompts, such as "launch photos," "set up a meeting at 9," or "FaceTime John." Google translate and other automated interpreting applications such as VoiceTra, developed in Japan for the Olympics, also require the user to simplify and streamline his sentences so that the machine can correctly process them. Processing queries thereby fuels standardisation of language.[38]

In addition to the streamlining of language production by machine language processing, one should pay attention to the recent evolutions of the market for language technologies. One of the most concerning issues is the arrival on the market of a handful of very large actors, namely Google, Amazon, Facebook, Apple, Microsoft, and their Asian equivalents. They dominate the market on several accounts: through their sheer size, through the massive investments they pour into machine language processing, and through the number of users of their services. "Recent forecasts predict the worldwide user count for SPAs such as Amazon Alexa, Apple's Siri or Microsoft Cortana to increase from 390 million in 2015 to 1.8 billion in 2021, which results in 2.3 billion USD average sales growth per year."[39] The market for language processing technologies may soon turn into a fringed oligopoly, dominated by

[36] Heather O'Brien, "Exploring user engagement in online news interaction," *Proceedings of The American Society for Information Science and Technology*, 11 January 2012. [Online] https://doi.org/10.1002/meet.2011.14504801088 [accessed 3 July 2020].

[37] Robin Knote, Andreas Janson, Matthias Söllner and Jan Marco Leimester, "Classifying smart personal assistants: An empirical cluster analysis," *Proceedings of the 52d Hawaii International Conference on System Sciences*, 2019, p. 2027.

[38] Rory Smith, "The Google Translate World Cup," *New York Times*, 13 July 2018. [Online] https://www.nytimes.com/2018/07/13/sports/world-cup/google-translate-app.html [accessed 22 June 2020].

[39] Robin Knote, Andreas Janson, Matthias Söllner and Jan Marco Leimester, "Classifying smart personal assistants: An empirical cluster analysis," *Proceedings of the 52d Hawaii International Conference on System Sciences*, 2019. p. 2024.

American and Chinese mega-corporations, which do not primarily aim at providing quality linguistic solutions but rather at extending their monitoring of human interactions.

This leads us to the last issue in terms of linguistic diversity: the type of governance implemented by the major providers of language processing technologies. They are all concerned with the nature of content published and distributed through their platforms and work to embed content monitoring technologies in their tools. Amazon thus advertises its proactive policy to ensure positive customer experience what would otherwise be classified as a form of censorship:

> While our initial content monitoring system was based on a blacklist mechanism, we also made initial strides toward building a more sensitive and contextually aware classifier to identify 1) profane content, 2) sexual content, 3) racially inflammatory content, 4) other hate speech, and 5) violent content. This system will be integrated into future Alexa Prize competitions from the outset as they are key to ensuring a positive customer experience for end users.[40]

Conclusion

The influence of machine language processing tools on the global landscape of languages online appears thus both contrasted and complex. Free neural machine translation tools and smart assistants enable people from all over the world to access and produce content in many languages, furthering what I would call interface multilingualism in terms of quantity, visibility and accessibility. But upon closer examination this comes at a cost for linguistic diversity. Levels of proficiency, both in foreign languages and in one's mother tongue, may be lowered by the ubiquitous presence of language technologies that streamline and standardize expression. Mutual intelligibility is of course enhanced by machine language processing, yet it fails to carry the untranslatables, all the quirks and specificities of a given language community. Finally, the monitoring of much of our online linguistic interactions by the major platforms of retail and social media according to their own and situated conception of governance could have an impact on the conception and enactment of cultural identities. If one agrees that languages are more than a vehicle for communication but the cornerstones of cultural practices and identities, it is concerning to see machine language processing expanding at a fast pace under the rule of a handful of corporations. Like AI technologies in other fields, these programmes and their providers should come under scrutiny and be held accountable. We would need a fully-fledged debate on the role and consequences of machine

[40] Ashwin Ram *et al*., "Conversational AI: The science behind the Alexa Prize," *First Proceedings of the Alexa Prize*, 2017, p. 6.

language processing in society, gathering providers, institutions (educational, cultural, legal), and representatives from user communities.

Bibliography

Bond, Esther, "What's so massive about Google's massively multilingual neural machine translation?," *Slator*, 18 July 2019. [Online] https://slator.com/technology/whats-so-massive-about-googles-massively-multilingual-neural-machine-translation/ [accessed 18 September 2019].

Bond, Esther, "Facebook patents alternative to 'expensive' BLEU," *Slator*, 22 August 2019. [Online] https://slator.com/technology/facebook-patents-alternative-to-expensive-bleu/ [accessed 21 November 2019].

Charlton, Emma, "The Internet has a language diversity problem," World Economic Forum, 13 December 2018. [Online] https://www.weforum.org/agenda/2018/12/chart-of-the-day-the-internet-has-a-language-diversity-problem/ [accessed 18 September 2019].

Citton, Yves, *Médiarchie*, Paris, Seuil, 2017.

Diño, Gino, "Facebook says it now has a powerful tool to think about language problems in a language agnostic way," *Slator*, 6 May 2019. [Online] https://slator.com/technology/at-f8-facebook-says-it-now-has-a-powerful-tool-to-think-about-language-problems-in-a-language-agnostic-way [accessed 18 September 2019].

Edmond, Charlotte, "This is when a robot is going to take your job, according to Oxford University," 26 July 2017. [Online] https://www.weforum.org/agenda/2017/07/how-long-before-a-robot-takes-your-job-here-s-when-ai-experts-think-it-will-happen/ [accessed 18 September 2019].

Haberkorn, B. J., "Amazon Polly voices in Alexa skills now generally available," 23 October 2018. [Online] https://developer.amazon.com/fr/blogs/alexa/post/baee53c1-5b03-4580-b57a-ee9510413354/amazon-polly-voices-in-alexa-skills-now-generally-available [accessed 17 December 2019].

Hockly, Nicky, "Automated writing evaluation," *ELT Journal*, vol. 73, no. 1, January 2019, p. 82-88. [Online] https://doi.org/10.1093/elt/ccy044 [accessed 18 April 2021].

Internet retailing (staff writer), “Zara. Communication without borders,” 7 August 2017. [Online] https://internetretailing.net/research-articles/research-articles/zara-communication-without-borders [accessed 17 December 2019].

Johnson, Khari, “Facebook Messenger launches translations by intelligent assistant M,” 1st May 2018. [Online] https://venturebeat.com/2018/05/01/facebook-messenger-launches-translations-by-intelligent-assistant-m/ [accessed 18 September 2019].

Kemp, Charlotte, “Defining multilingualism,” in Aronin, Larissa and Hufeigein, Britta (eds.), *The Exploration of Multilingualism*, Amsterdam, John Benjamins Publishing, 2009.

Kenny, Dorothy, “Machine translation,” in Rawling, Piers and Wilson, Philip (eds.), *The Routledge Handbook of Translation and Philosophy*, Milton Park, Routledge, p. 428-445, 2018.

Knote, Robin, Janson, Andreas, Söllner, Matthias and Leimester, Jan Marco, “Classifying smart personal assistants: An empirical cluster analysis,” *Proceedings of the 52d Hawaii International Conference on System Sciences*, 2019.

Larsonneur, Claire, “The disruptions of neural machine translation,” *Spheres*, no. 5, “Spectres of AI,” November 2019. [Online] http://spheres-journal.org/the-disruptions-of-neural-machine-translation/ [accessed 17 December 2019].

Leppänen, Sirpa and Peuronen, Saija, “Multilingualism on the Internet,” in Martin-Jones, Marilyn, Blackledge, Adrian and Creese, Angela (eds.), *The Routledge Handbook of Multilingualism,* London/New York, Routledge, 2015, p. 384-403.

Mazareanu, Elena, “Market size of the global language services market,” 9 August 2019. [Online] https://www.statista.com/statistics/257656/size-of-the-global-language-services-market/ [accessed 3 July 2020].

Moorkens, Joss, Castilho, Sheila, Gaspari, Federico and Doherty, Stephen (eds.), *Translation Quality Assessment. From Principles to Practice*, Cham, Springer International Publishing, 2018.

Pajot, Philippe, “La traduction automatique s’attaque aux langues rares,” *La Recherche*, no. 554, December 2019, p. 20 *sq*.

O’Brien, Heather, “Exploring user engagement in online news interaction,” *Proceedings of The American Society for Information Science and Technology*, 11 January 2012. [Online] https://doi.org/10.1002/meet.2011.14504801088 [accessed 3 July 2020].

Paolillo, John, “Language diversity on the Internet: Examining linguistic bias,” in UNESCO Institute for Statistics, *Measuring Linguistic Diversity on the Internet*, Paris, UNESCO, 2005.

Pasquinelli, Matteo, “How a Machine Learns and Fails — A Grammar of Error for Artificial Intelligence,” *Spheres*, no. 5, “Spectres of AI,” November 2019. [Online] http://spheres-journal.org/how-a-machine-learns-and-fails-a-grammar-of-error-for-artificial-intelligence/ [accessed 17 December 2019].

Ram, Ashwin *et al.*, “Conversational AI: The science behind the Alexa Prize,” *First Proceedings of the Alexa Prize*, 2017.

Shaji, Mishel, “What is the most translated website in the world?,” 17 March 2019. [Online] https://www.quora.com/What-are-the-top-three-most-translated-Websites-and-how-many-languages-do-they-each-accommodate [accessed 18 September 2019].

Smith, Rory, “The Google Translate World Cup,” *New York Times*, 13 July 2018. [Online] https://www.nytimes.com/2018/07/13/sports/world-cup/google-translate-app.html [accessed 18 September 2019].

Technavio, “Globalization of business to boost growth,” 29 May 2020. [Online] https://www.businesswire.com/news/home/20200529005034/en/Machine-Translation-Market-2020-2024-Globalization-Business-Boost [accessed 3 July 2020].

Villani, Cedric, *For a Meaningful Artificial Intelligence. Towards a French and European Strategy.* March 2018. [Online] https://www.aiforhumanity.fr/en/ [accessed 20 September 2019].

Wiggers, Kyle, “Facebook founds AI Language Research Consortium to solve challenges in natural language processing,” 28 August 2019. [Online] https://venturebeat.com/2019/08/28/facebook-founds-ai-language-research-consortium-to-solve-challenges-in-natural-language-processing/ [accessed 17 December 2019].