

= <https://multilingual.com/issues/translation-economics-of-the-2020s/>

## Translation Economics of the 2020s

A journey into the future of the translation industry in eight episodes



Jaap van der Meer

Jaap van der Meer, founder and director of the Translation Automation User Society (TAUS), is a language industry pioneer who started his first translation company in the Netherlands in 1980. In 1985, his company INK published the first desktop term extraction and translation memory software.

The long-expected technical revolution is here. Automatic translation is no longer just a freebie on the internet. It is now entering the “real” economy of the translation sector, and it changes everything. The fundamental economic laws of intellectual property or of scarcity, for instance, do not apply in a world in which machines autonomously generate an abundance of translated material. Translation as a social good, not owned by anyone and free to the user, exists alongside translations owned and paid for by governments and corporations.

# 01










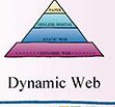
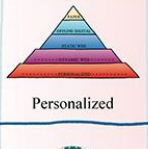
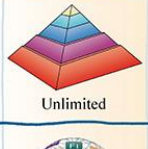











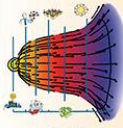
## A short history of the translation industry

Over a period of four decades, the translation sector has gone through regular shifts of adaptation, instigated by changes in the business and technological environment. The first shift happened in the late 80s when software conquered the world and personal computers started to occupy our desktops. Translation became localization. The art and craft of translation were now overshadowed by the challenges of computer literacy. Translation Memory (TM) and terminology management released localizers from tedious repetitive tasks, improved profitability, and ensured consistency.

The second shift in the evolution of translation happened in the nineties when it became more fashionable to refer to our business as globalization. Customers in the IT industry now became more sophisticated, deploying globalization management systems (GMS) and trying to centralize the benefits of translation leveraging. The business was flourishing with the expansion of markets and locales.

Yet localization remained a somewhat isolated activity, limited to products shipped to export markets. The attempt to open localization to the whole enterprise characterizes the integration phase in this evolution. Connecting the GMSs to the content management systems, the customer relationship management systems, and a slew of other enterprise applications allowed extremely large amounts of content to be translated. The integration phase of the past decade led quite naturally to the convergence phase that we find ourselves in right now. The difference lies in the web services that make the connection of various software apps a piece of cake. For a couple of years now you've been able to find your locations, payment apps, currency and time converters, everything you need, fully integrated into the tools you use. Our smartphones are the ultimate example of the convergence era: a single device that lets us take pictures, make appointments, do our shopping, make payments, navigate our journeys, make our phone calls, and now also translate everything into our own language (more or less).

However impressive the journey has been so far, nothing compares to what is still to come: the singularity. In this new phase, technology essentially takes over completely. The human translator is no longer needed in the process. Google and Microsoft alluded to this future state when they claimed that their machine translation (MT) engines translated as well as human professional translators. However, this has led to heated debates both in academic and professional circles about what this so-called human parity really means.

	1980 TRANSLATION	1990 LOCALIZATION	2000 GLOBALIZATION	2010 INTEGRATION	2020 CONVERGENCE	2030 SINGULARITY
INDUSTRY FOCUS	 Documents	 Software	 Simship	 Integration in enterprise systems	 Embedded in every app, on every screen	 Ubiquitous
TYPE OF CONTENT	 Paper	 Digital	 Static Web	 Dynamic Web	 Personalized	 Unlimited
LANGUAGES	 1 > 10	 1 > 25	 1 > 40	 6 <> 60	 60 <> 60	 150 <> 150
TECHNOLOGY	None	TM and terminology software	Workflow (GMS)	MT and Advanced Leveraging	Real-time customized MT	FAUT
DATA	Glossaries	Project TMs	Centralized TMs	Limited shared data	Web crawled	Plug and Play
COMMUNICATIONS	 B2B G2C B2C	 C2B	 C2C	 Social	 Internet of Things	 M2M

The evolution of translation since 1980. Source: TAUS.

## 02

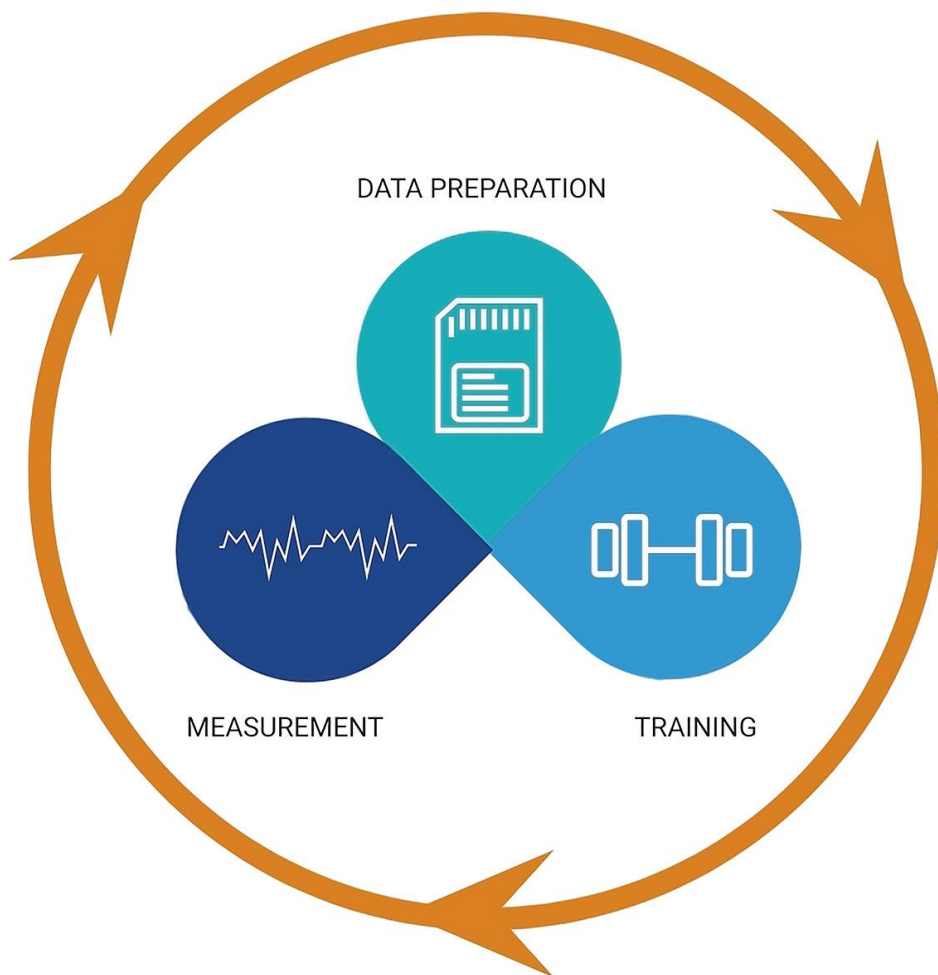
### The rise of zero-cost translation

Big tech companies are in severe competition with each other to develop machines that can generate an unlimited supply of best-in-class translations in as many languages as possible. And they offer this service at zero cost, to the end-users at least. In the translation service industry on the other hand, every job is sent down the supply chain of project managers, linguists, reviewers, etc., all adding a further cost each and every time a new translation is needed.

The global translation industry finds itself now in a “mixed economy” condition: on one side a vertically cascaded supply chain, and on the other, the new flat free machines model. The speed with which the machines are improving when fed with the right quality and volumes of data makes translation a near-zero marginal cost type of business (in the spirit of Jeremy Rifkin). This means that once the right infrastructure is in place, the production of a new translation costs nearly nothing and capacity becomes infinite.

Taking into account the fact that the ultimate measure of success in the globalization and localization business is world-readiness, meaning that we are able to speak with all customers and citizens on this planet in their own languages, we need to ask ourselves how sustainable this mixed economic model really is. Can the translation sector, as long as it is locked up in a vertical labor-based cost model, scale up to reach the goals of world-readiness? In other words: how realistic is it to think that we can just add more capacity and skills into our existing economic model to generate that global business impact?

For operators in the translation industry to follow the trend and transition to the new free machines model, they need to consider fundamental changes in the economics of their business. They need to be prepared to break down existing structures, adopt new behaviors of sharing and collaboration, eradicate the need for human tasks and work activities, and progress technology towards abundance in translation supply. Under the new economic model the vocabulary of linguistic quality, translation memories, and word rates will lose its meaning. Instead, we will talk about global business impact, data and models, and value-based pricing.



The diligent data pipeline.

## 03

### **Kill your business before it kills you**

In the heyday of the dot.com revolution Jack Welch, a former chief executive officer (CEO) of General Electric, called on his executive team to subject each of their business divisions to a destroy-your-own-business (DYOB) scrutiny. This was at the time when companies that had grown up as brick-and-mortar enterprises came under threat from new online businesses. The message was clear: kill your business before it kills you. The process of creative destruction lies at the heart of small and big revolutions in companies and markets. A good example is how Sony managed to get a share of the digital camera market, or on the negative side how Kodak completely missed this transformation. Think also about the disappearance of the video sector as a result of streaming businesses such as Netflix, and the threats to the taxi industry as a result of online rideshare services such as Uber. Payment platforms such as Stripe and PayPal have already taken over a core function of the old banking sector. A fundamental characteristic of the capitalist economic system (according to the economist Joseph Schumpeter) is that it constantly destroys and reconfigures previous economic orders until it, eventually, may lead to its own demise as a system altogether.

In 2019, Google alone translated 300 trillion words compared to an estimated 200 billion words translated by the professional translation industry. Google Translate was launched in 2006, Microsoft Bing Translator three years later, followed by Yandex MT in 2011, Alibaba in 2012, with Tencent and a couple of other Asia-based tech companies joining around 2017. Back in the US, we see Amazon joining the MT space in 2017 and Apple in 2020, being the latest addition to the premier league of MT providers. The total output from MT engines is probably already tens of thousands times bigger than the overall production capacity of all the professional translators on our planet.

Until only two or three years ago, after the new neural MT (NMT) success stories started to trickle in, human professional translation and MT existed in two parallel worlds. They didn't seem to affect each other that much. Even inside Google and Microsoft, the product localization divisions didn't make use of their company's own MT engines. But that has changed now. MT is integrated into almost every translation tool and workflow.

Creative destruction in the translation space has begun in different ways. Lionbridge, one of the largest translation companies in the world, spun off a large part of its business as a dedicated artificial intelligence (AI) company, thereby providing a tremendous capital gain to its investors. New AI-enabled translation start-ups such as Lengo, Lilt, and Unbabel have managed to attract \$155 million in growth capital.

The question for operators in the translation industry, therefore, is whether the two processes — human and machine translation — continue to co-exist, or whether MT will completely wash away the old business. A good part of the established cohort of language service providers, is already feeling the increasing pressure and fleeing into data services or transcreation services. Alternatively, they are jumping on the bandwagon and starting to build their own MT systems and services.

The outcome of all this turbulence will most likely be yet another rebranding and labeling of our sector. Gartner, in their recent research report on the sector, talks about AI-enabled translation. Unbabel's CEO Vasco Pedro coined the new term language operations (or LangOps), as a more holistic approach to managing global business communications. Whatever the most fashionable label for the sector will be, it's clear that under the hood nothing will be the same anymore. Gartner reckons that by 2025, enterprises will see 75% of the work of translators shift from creating translations to reviewing and editing machine translation output.

The takeaway for LSPs is that, if they don't want to be the Eastman Kodak of the translation industry, ignoring AI and MT is not an option. To grow the business they need to get out of their localization niche and use data and technology to scale up and expand into new services.

“The takeaway for LSPs is that if they don't want to be the Eastman Kodak of the translation industry, ignoring AI and MT is not an option. To grow the business they need to get out of their localization niche and use data and technology to scale up and expand into new services.”

## 04

### **Unbundling machine translation**

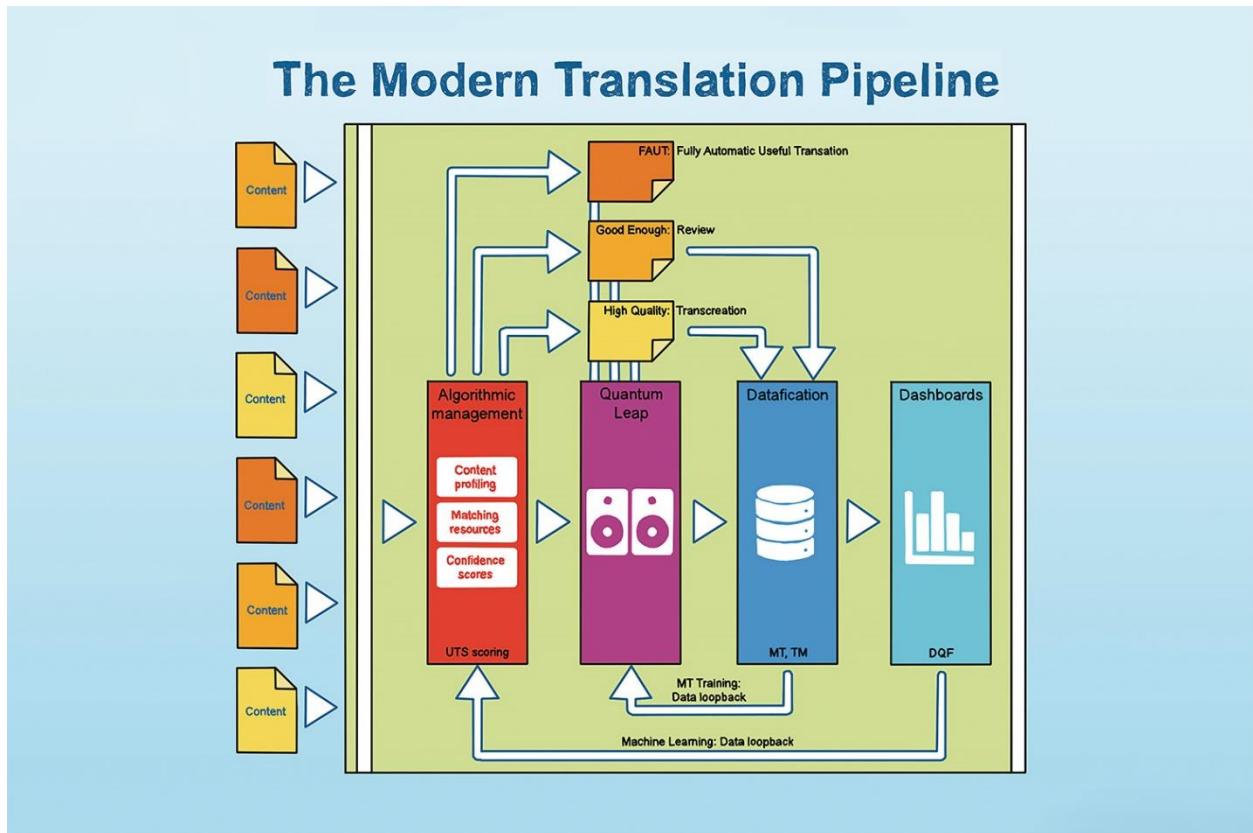
Putting aside all the technical talk about neural network encoders, vectors and encoder-decoder architectures, attention-based models, etc., one could say that MT is a simple sum of algorithms and data. The algorithms are developed in close cooperation between tech companies and universities. Breakthroughs on this front are shared among the gurus in the world of MT through academic papers, and the code is usually shared under open-source licenses. This is a great benefit, of course, because it helps speed up the development and resolution of problems.

The core competency driving innovation in MT is natural language processing (NLP). Translation is one of the tasks that AI is trying to solve through NLP models. Other tasks text summarization, sentiment analysis, automatic speech recognition, and question-answering. A testament to the value and importance of NLP is the new round of \$40 million investments that the start-up Hugging Face received in March 2021, bringing the total amount of investments to \$55 million. Hugging Face builds and tests models that help automate all these language-related tasks. The size of these models is growing exponentially and the parameters that can be used to tweak the models is increasing to hundreds of millions now in the largest models. The online tutorials by Clement Delangue (founder of Hugging Face) about the future of NLP are interesting in this context. He talks about the thousands of choices the developers have in changing the parameters hoping to get better outcomes (read: “better quality translation”), and then sighs, and says “more data outperforms a smarter model.” Time and again, the MT developers come to that same bitter conclusion: no matter how much they try to optimize the models, more and better data is always more effective in generating better results.

That brings us home to the core competency of the translation sector. For decades linguists, translators, and quality managers have already worked on polishing their translations to satisfy their customers. In other words, the translation industry sits on a wealth of data that is crucial to solving core AI problems. Putting this in a wider economic context of world-readiness or



readiness for the future, we are starting to get a glimpse of where new business opportunities may be opening up for the translation sector. Whether in the financial services industry, consumer technology, healthcare, or any other industry, business processes everywhere can be reduced in their most basic form to text or speech input triggering a process or a transaction; translation is simply one variant of that process.



The modern translation pipeline.

## 05

### Buying the best MT engine

DeepL often gets rave reviews for its astonishingly good quality translations. And it's common practice among MT developers to use Google Translate as a benchmark for their own engines. The question is asked many times: which MT engine is the very best? Or perhaps you can tell us which MT engine is best for this language or in that domain. The answer to all these questions is that there is no magic MT box that does a perfect job on every document or piece of content. DeepL performs well because it started off as Linguee, an online concordance and dictionary tool, and has therefore collected more valuable language data than most other MT engines. Google Translate is used as a benchmark because it has been out on the market the longest and has therefore gathered more language data in more language pairs than any other MT system developer, including the multilingual content crawled from the World Wide Web.

MT developers use the same frameworks or models at the core, like Marian, BERT, or OpenNMT, which are shared under open-source licenses on GitHub. The “best” MT engine out-of-the-box, therefore, does not exist. MT is not static: the models are constantly being improved of course, but even more important, the output of the machine is totally dependent on the data that is fed into the models to train and customize. Buying and using MT is like constantly tuning a process and measuring the results.

The race among the translation memory system (TMS) providers to have the widest selection of MT engines connected through automatic programming interfaces (APIs) in their workflows is not very helpful to language service companies. For them, it is much more important to have an easy way to customize the MT engines with their own high-quality language data. How easy that can be is demonstrated by some of the disruptive innovators in the translation industry that have implemented a real-time or dynamic adaptive MT process, which means that the engine learns almost immediately from the edits made by humans. We could call this “predictive translation,” similar to the predictive writing feature that Google offers.

Real-time, adaptive MT is the state-of-the-art in the professional translation space at the moment, especially for customers who do not want to worry too much about implementation and customization. It’s only available in a closed software-as-a-service offering, which is understandable because of the immediate data-feedback loop that is required to optimize the speed and success of the learning process.

Language companies that need more flexibility and control over the technology they use need to build and customize their own end-to-end solution. Their main challenge is to set up a diligent pipeline for data preparation, training, and measurement. Their language operations then become a data-driven solution.

In the meantime, the NLP gurus are pushing the frontier on the technology front. The models they are developing will become much better at “understanding” context in a document. Context-aware MT is therefore just around the corner. At the same time, Google and Microsoft are working on making MT much more economic with massively multilingual MT: a single model that can tackle any language pair in the world. Another idea that isn’t science fiction is to skip translation altogether and have the engines generate text directly in other languages — a machine version of transcreation if you like.

## 06

### **No data, no future**

One thing is perfectly clear: data is the new oil, and without data, the translation industry has no future. The modern translation pipeline is designed around two distinctive data streams: the language data and the translation data. The language data is the text in the source and target languages, stored as bilingual corpora. The translation data is the data about the translation, also referred to as metadata. Examples of translation data are: resources used for a given segment (the MT engine, the editor/reviewer), throughput time, edit-distance score, and quality score. These are all data points collected through the dynamic quality framework (DQF). In the most



sophisticated workflow, we can even track the data point of how many eyes look at each translated segment, which allows us to establish the most accurate return on investment calculation for translation. Now, link this data point with your marketing automation tools and we are getting pretty close to measuring the economic impact of our localization efforts.

In the past years, the translation industry has accumulated a massive amount of texts in source and target languages that are stored in databases referred to as translation memories. It's understandable that operators in the translation industry fall back on their archives of translation memories when they are asked to provide data for the customization of their MT engines. They then find out that legacy translation memories do not always make the best training data. Their translation memories may be too specific and too repetitive, they contain names and attributes that can mess up the MT engines, and they are often not maintained very well over time.

To optimize quality output from MT, therefore, the language data needs to be of the highest possible quality. Data cleaning and corpus preparation should include steps like deduplication, tokenization, anonymization, alignment checks, named entity tagging, and more. To ensure that the language data that is used for customization of MT engines is on topic, more advanced routines can be used to select and cluster data to match the domain. Legacy translation memories may be used as part of a larger pool of basic data, but most of the time they are not productive enough. Those translation memories, carefully built up over many years, are unfortunately not the kind of linguistic assets that we wanted them to be.

Today, switching to the new paradigm of LangOps or AI-enabled translation requires new skills, new talents, and a new organizational structure. Operators in the translation industry would do well to create a chief data officer function and identify their NLP talents, even if they may decide to outsource most of the data-related activities.

“To optimize quality output from machine translation, therefore, the language data needs to be of the highest possible quality. Data cleaning and corpus preparation should include steps like deduplication, tokenization, anonymization, alignment checks, named entity tagging and more.”

## 07

### **Who owns my language data?**

Convinced as they may be that the future lies in taking control of the data, many owners of agencies, as well as translation buyers, still hesitate to move forward because they are in doubt about their legal rights to the data. There is a strong feeling across the translation industry that translations are copyright-protected and can never be used to train systems. Is that true? If uncertainty over ownership of data is a factor that slows down innovation, it is time to get more clarity in this matter.

In practice, the industry operates with a mixed form of ownership. Taking a translation memory file as an example, the author holds the copyright of the text in the source language, and the translator, together with the author, holds the copyright of the text in the target language. In addition, and only in Europe, the party that creates the translation memory holds the so-called *sui*

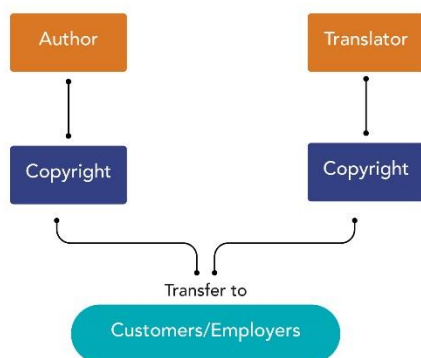
*generis* rights to the database. Often, but not always, translators and authors transfer their copyrights to their customers or, if they are employed, the copyrights reside with their employers. These copyrights are intended to apply to complete works or parts of works (documents, sections, products, features) more so than to individual segments. Individual segments can hardly be protected by copyright law, unless the sentence on its own has a distinctive creative value, like the lyrics of a song or a line of text from a poem. Case law from the European Court of Justice has confirmed that a sequence of at least 11 words can be a copyrighted “work,” provided that one can still distinguish “the hand of the author” in that segment.

Since MT developers normally use datasets that consist of randomly collected segments for the training of their engines, the chances of a copyright clash are minimal.

For US-based developers of MT, the risks of infringing copyright laws are further reduced thanks to the fair use ruling, which makes an exception for the use of data for research purposes. The fair use ruling has for many years put US-based MT developers at an advantage, but since 2019 the EU has introduced a similar exception for text and data mining (TDM) in the Digital Single Market strategy.

Technically speaking, copyright on language data is complex and involves multiple stakeholders and many exceptions. But at the end of the day, practitioners in the translation industry have to make up their minds. Customers expect them to use the best tools and resources available, and today that means using MT and data to customize the engines. To our knowledge, there is no precedent of a lawsuit over the use of translation memories for training of MT engines, and the risk of being penalized is negligible. But in case of a doubt, why not consult with your stakeholders? After all, we are all in this together.

Finally, a recommendation: for safety’s sake it is a good practice to filter out named entities (personally identifiable information) from the segments before sharing the data for training purposes, especially as that will most likely have a positive impact on the output quality as well.



Language data ownership.

## Breaking the data monopolies

We are all in this together. If data is the new oil and there is no future in translation without access to data, then it is in the interest of the thousands of LSPs and their customers, and the hundreds of thousands of translators in the world, to break the monopolies on language data. Right now, a handful of big tech companies and a few dozen large LSPs have taken control of the most precious resource of the new AI-driven translation economy. At the same time the call for change is getting louder: a more circular, sharing, and cooperative economic model would fit better into our modern way of working.

One solution is to unbundle the offerings tied up in the AI-driven translation solution and to recognize the value of all the different contributors. Hosting the powerful, scalable infrastructure that can support the ever-growing AI systems is a task that can only be managed by the largest companies.

Note that customizing the models for specific domains and languages is a specialized service that may best be left to service companies that have expertise in these fields and are capable of adding value through their offering. And, since the best quality training data is vital for everyone, why not let translators and linguistic reviewers who produce this data take full responsibility and earn money from their data every time an engine is trained with their data?

The process of creative destruction is now in full swing and may lead to a redesign of our entire ecosystem. The first marketplaces to enable this new dispensation are now out there. SYSTRAN launched a marketplace that allows service providers to train and trade translation models. TAUS launched a data marketplace that allows stakeholders in the translation industry to monetize their language data. These first steps should lead to healthy debate across the industry as we feel the shock waves of an industry reconfiguration driven by radical digitization, human reskilling, and exponential data intelligence.

=