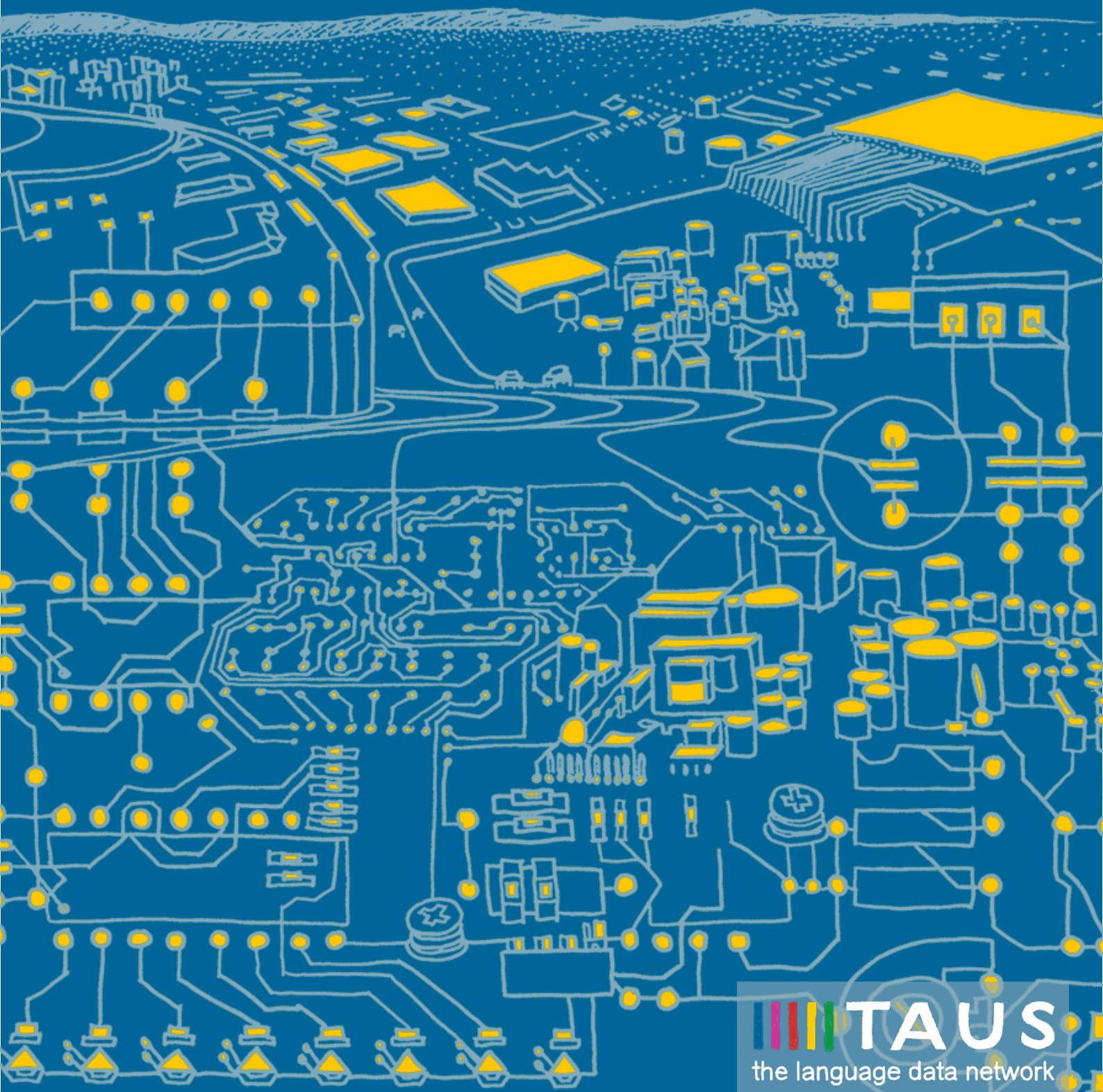
Language Data for AI

(LD4AI)

A TAUS Report, November 2020



TAUS
the language data network

Written and Reviewed by:

Andrew Joscelyne
Jaap van der Meer
Şölen Aslan

Published by TAUS Signature Editions, Danzigerkade 65A, 1013AP Amsterdam, The Netherlands
E-mail: memberservices@taus.net
www.taus.net

Design: Anne-Maj van der Meer

TABLE OF CONTENT

Executive Summary 4
1. Introduction 5
2. Birth of a New Industry Sector: Language Data for AI 6
3. Data and Language 7
    A History that Starts in the 1950s 7
    Corpus Linguistics as a Prelude to LD4AI 7
    The Unreasonable Effectiveness of Data 8
    Two Catalysts Driving LD4AI Forward 8
    Benefits of LD4AI for the Translation Industry 9
4. Translation Economics 10
    The Beginnings of the Translation Industry 10
    MT with 000-fold Productivity 10
    Sustainability of the Current Model 11
5. The ‘Data First’ Paradigm Shift 12
6. Language Data Landscape 16
    Fifty Percent of LSPs Active in LD4AI 16
    Types of Services 16
    Speech Data Tasks 17
    Data Preparation Tools 17
7. Profiling the New Cultural Professional 19
    Wide-Band Communications Talents 19
    New Role of the Professionals 19
    Adapting to New Ways of Digital Communication 19
    Emergence of Crowd-Based Platforms 20
8. Five Takeaways 21
9. TAUS Data Marketplace 23

EXECUTIVE SUMMARY

This report gives stakeholders in the language and AI industries insights into how language data is becoming a business on its own. The report is broken down as follows:

1. **Introduction**
   Previous TAUS reports and the rationale behind this new publication

2. **Birth of a New Industry Sector: Language Data for AI (LD4AI)**
   Converging trends in AI and language industries

3. **Data and Language**
   Noam Chomsky versus Randolph Quirk - the theoretical story behind the growing awareness of the unreasonable effectiveness of data.

4. **Translation Economics**
   Meanwhile in the real world: automatic translation became a thousand times more productive than human translation and laid a solid foundation for expansion in new services and business in the language sector.

5. **The Data-First Paradigm Shift**
   There is enough evidence, both from a technology and from an economic perspective, that the translation industry is in the midst of a paradigm shift. Computer-aided translation is making space for a data-first approach.

6. **The Language Data Landscape**
   How language service providers branch off into new services and change the translation landscape.

7. **Profiling the New Cultural Professional**
   The hunt for native organic data changes the profile of the workers needed in the global language business.

8. **Trends and Takeaways**
   The five main takeaways from our explorations in this new sub-sector.

9. **TAUS Data Marketplace**
   How the Data Marketplace serves as an enabler and accelerator in the Language Data for the AI sector.

INTRODUCTION

In this report, we describe *Language Data for AI*, a new industry sub-sector, and a result of converging trends in the AI and language industries. We dial back a few decades to analyze how language really became data and how that megatrend starts to overhaul the professional language industry. Our goal with this report is to offer context and perspective and assess the opportunities and challenges for both buyers and new providers entering this industry.

TAUS was an early mover in the language data sub-sector with the launch of the TAUS Data Cloud in 2008. The TAUS Data Cloud allowed users to upload data and earn credits to download other users' data. This reciprocal business model served many of the early users of statistical MT very well. The recipe was simple: *more data was always better data*. However, as the technology has evolved into neural networks, and users have become much more sophisticated, the old reciprocal model of the Data Cloud no longer suffices. Users need more domain-specific data, customization, even personalization. They want to track the origin of the data and access new languages with human-in-the-loop services.

The publication of this report coincides with the launch of the new TAUS Data Marketplace in November 2020. Our goal with this report is also to discover the market needs and opportunities for a central marketplace where language data producers and publishers can meet users and together create a vibrant platform for the advancement of global communications across all languages.

This report has been built on the earlier findings shared in two previous TAUS reports TAUS Data Market White Paper (June 2017) and Translation Data Landscape Report (Dec 2015). Moreover, we have carried out one-on-one interviews and questionnaires with data services providers including Lionbridge, Clickworker, Welocalize, Unbabel, Utopia Analytics, BasicAI, and PacteraEdge. To understand the tendencies within the traditional language industry, we have surveyed 205 LSPs, asking them their insights on whether the language industry is transforming into a data industry. And lastly, we have interviewed topic experts Paul Leahy (Oracle), Shahram Khadivi (eBay), Marcello Federico, and Suzanne Fogarty (Microsoft).

BIRTH OF A NEW INDUSTRY SECTOR: LANGUAGE DATA FOR AI

When a technology breakthrough disrupts an economy or heralds the arrival of new services and products, we typically see industries breaking down or branching off and new sectors emerging. The same is happening now with a sector that we refer to as *Language Data for AI*.

Artificial Intelligence (AI) is the biggest technology breakthrough of our times, and in its most accessible form as machine learning, is a disruptor for many sectors across the whole spectrum of manufacturing and services. AI changes fundamentally how we humans interact in our roles as consumers, citizens, patients, and passengers with businesses, devices, robots, vehicles, governments, health systems, networks, machines, and…. with each other.

Language Data for AI (or LD4AI) is a new industry sub-sector that attracts on the one hand well-established language service companies and on the other hand newcomers with innovative data science and Natural Language Processing strategies and roots in crowdsourcing.

The *LD4AI* industry is a branch of both the much larger AI industry and the global language services industry. In the AI industry, the researchers know that they can only do so much with algorithms and parameter settings. Again and again, they see that data outperform the models. They can't do without the humans in the loop interpreting, annotating, and validating the data. In the global language services industry, the entrepreneurs see business being taken over by machines. For them, data preparation is a welcome innovation and diversification.

According to Cognilytica, the market for AI and machine learning relevant data preparation solutions was worth more than *$1.5B in 2019 growing to $3.5B by the end of 2024*. What this business entails at a high level is adding human intelligence to massive amounts of data in the form of text, speech, images, videos, and sound, so that machines can be trained to understand the world better.

On the other hand...

On the basis of these Cognilytica estimates for the size of the AI data labeling and preparation market, we can't say that the market is 'huge' as some of the language service providers we spoke to want to believe. But it's certainly significant, and important enough to be taken seriously, especially in light of the phasing out of traditional, handcrafted translations.

DATA AND LANGUAGE

A History that Starts in the 1950s

The data story for language in both research and business in the digital age begins in the mid-1950s when two visions of language study came into conflict. This was some years before the modern translation industry had developed, and computing was still mostly a research project.

In the US, Noam Chomsky was revolutionizing linguistics by focusing on what he called *competence* – the underlying cognitive mechanisms that *explain* how humans generate and understand an infinity of language utterances. He argued that a child is not exposed to enough language data to justify stimulus-response learning, so there must be a sophisticated internal device - an inherited grammar in the brain - that drives the work of language learning. What people actually say or write is simply *performance*. In this influential theory, the job of linguistics was to focus on the mental machine, not the actual content it produced or recognized.

Contrast that with a language project emerging across the Atlantic in the UK, where Randolph Quirk initiated an extensive *Survey of English Usage* in 1959. This involved recording and for the first time digitizing large amounts of spontaneous English language behavior (speech and text) on tape as raw data and collecting them in a database so that linguists could use the contents to *describe* (not explain) the grammar of English on the basis of usage statistics. Unlike Chomsky's theory-driven project, Quirk was interested in the details of what people speak and write, not simply that there was an infinite amount of it!

Corpus Linguistics as a Prelude to LD4AI

This effort helped launch the discipline of *corpus linguistics*, which used the added scale provided by computer databases to capture large volumes of written and transcribed language. As a "technology," the concept of corpus+search was to have a powerful influence on the emerging business of professional translation. Computer *termbases* such as the Canadian resource *Termium*, founded in the late 1970s, collected a wide corpus of FR-EN usage to provide online terminology services for human translators.

Meanwhile, a more Chomsky-inspired research project conceived in the 1960s began to produce the first machine translation (MT) engines. Using a "grammar+dictionary" or symbolic AI approach of *analyzing* the grammar of a source sentence and then *synthesizing* a target language output, the first systems were limited by the fact that machines were designed to mimic the human process of understanding and recreating sequences of language from scratch. This continued until the late 1980s in Japan and Europe. There were simply not enough existing data available in digital form to train machines to *learn* from existing translations.

Nevertheless, the use of practical automation to support translation by copying and updating terminology lists or quickly searching for words-in-context in a big document file began to

spread down from mainframes as applications for the first personal computers during the 1980s and early 1990s. In Europe, Russia, Japan, and the USA, individual translators were able to benefit from the first fruits of corpus data management as a CAT tool with programs such as Trados Multiterm.

This data-driven breakthrough using *translation corpora* (source + target bitexts) as a model made it possible to search for existing data strings in a new source text. Eventually, a statistically-driven *search and replace* operation itself could be automated, changing translation economics and models for good by the 2000s, by introducing a fully digital translation workflow, leveraging data to recycle existing content in new ways.

The Unreasonable Effectiveness of Data

By the time Google launched its free statistical Translate service in April 2006, parallel language data had become a critical, high-value resource, even though the range of languages used was limited, generating much poor quality output. TAUS had been founded the year before in recognition of the transformational power of language data to drive a new translation automation business. The need for a forum to solve the conceptual and commercial issues around data ownership, preparation, quality assessment, and sourcing was by then becoming self-evident.

In this same period, researchers began to realize that work based on complex theories similar to Chomsky's grammar but in other domains such as vision and robotics were proving to be less useful in building AI applications than access to lots of good data.

In a famous 2009 paper titled *The Unreasonable Effectiveness of Data*, Peter Norvig and colleagues argued that "for many tasks, words and word combinations provide all the representational machinery we need to learn from text." In other words, we didn't need to explain language as a rule system based on semantics to computers: they could "learn" all kinds of things simply by approaching texts as statistically significant distributions of words and phrases. Elaborate theories of grammar were irrelevant - Quirk's datafication rather than Chomsky's theorizing was coming to the aid of the translation industry.

The latest twist in this story, starting around 2015, has been the arrival of neural machine translation (NMT) engines, built using machines that produce constantly better output by learning from the right amounts of well-prepared parallel language data, or some variant of this approach. This has stimulated a new drive to find, collect, evaluate and prepare data for language tasks in general, as machine learning technology can be used to handle almost any linguistic task, from an NLP focus on text and speech analytics to text generation, summarization and question-answering systems.

Two Catalysts Driving LD4AI Forward

This story of language and data revolves around two developments that determine exactly what this report is about:

- *Language data annotation* as a new industry service is born. Language data annotation is crucial for encoding all communicative phenomena as sentiment, intent, argumentation or storytelling moves, gender, health, age, and more. This metadata initially added by humans,

then helps machines identify entities, emotional states, arguments, and intentions in all kinds of chatbot conversations, text understanding, translation automation, and similar language-led tasks.

- *The massive growth of User Generated Content (UGC)* comes just at the right time to feed and test the new AI systems. In both written text (e.g. social media and online comments of all kinds) and speech (virtual assistant dialog, transcribed podcasts, and virtual meeting output), UGC is vital for the brand platform and website owners to monitor what is being written/said across languages and evaluate the benefits and threats emerging from such "native" expression.

Benefits of LD4AI for the Translation Industry

In under a decade annotated language and speech data have evolved into a rich source of useful knowledge for smarter decision-making in sectors ranging from commerce, marketing, legal, and healthcare, to education, policing, security, and entertainment media. This, in turn, opens up new possibilities for the translation industry to:

- Lower costs, accelerate throughput, and improve language range and quality in its traditional delivery models;
- Build a more inclusive and innovative data collection tool and distribution service to harness the power of more of the world's languages;
- Offer multilingual speech and text data enrichment as an AI service around translation to a broader range of customer industries.

These, sequentially, depend on developing the right economic model for a data-driven translation industry, which we look at next.

On the other hand...

Don't forget that one of the most obvious disadvantages of LD4AI is the fact that most languages around the world do not (yet) have enough resources to join in this data preparation business. Neural translation models using low data requirements exist, and even enable massively-multilingual solutions for languages with almost no data. But there's a risk that the current paradigm will simply favor big-data languages.

TRANSLATION ECONOMICS

The Beginnings of the Translation Industry

In a parallel track, we see, also in the 1950s, the birth of the professional translation services industry. Economies are growing. The Marshall Plan in Europe leads to stronger export industries. Job seekers and young families are migrating to the New World. Fluency in foreign languages forms a business opportunity for the first generation of translation entrepreneurs who establish agencies and advertise their services in the Yellow Pages. Their focus is financial and legal documents, user manuals for the expanding range of consumer products, diplomas, and certificates. The tools that the agencies used were typewriters and later in the seventies and eighties word processors. When personal computers entered our daily lives some smart guys invented translation memory as a productivity tool for translators.

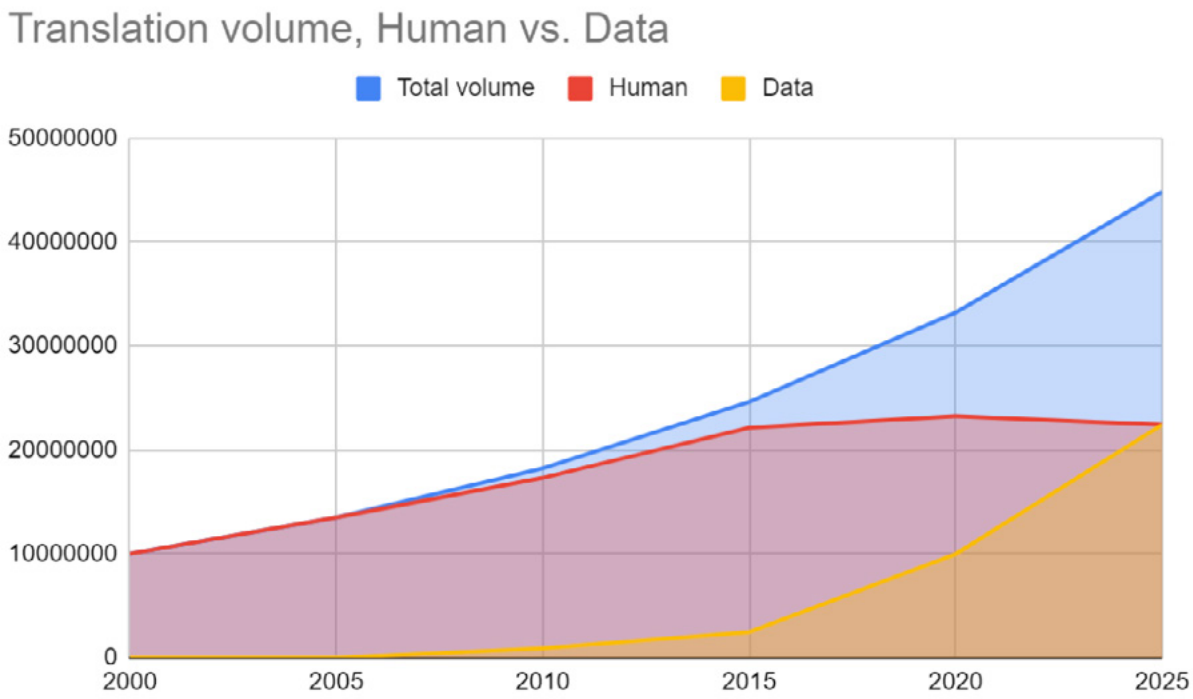
The need for translation and localization kept growing, leading to what is often described as a cottage industry with thousands of agencies. Even though the history of Machine Translation started around the same time as the beginning of the translation industry, AI researchers and translation practitioners managed to stay clear from each other most of the time. Apart from a few encounters which lead to interesting anecdotes, translation services remained craftsmanship and not something that could be scaled and automated very easily. That was the *communis opinio*. Until….

MT with 000-fold Productivity

Until data-driven MT systems started to show better results and Google launched their Translate feature in 2006. Soon more translation apps followed, from Microsoft, Yandex, Baidu. The audience of the translation apps started to become bigger and bigger. Today, the output from the free MT apps is already thousands of times bigger than the capacity of the 'old' industry. Google alone translated 300 trillion words in 2019 compared to an estimated 200 billion words total output from the entire professional translation industry.

After a slow start around 2010, we observe a rampant adoption of MT technology in 2015 and acceleration since the outbreak of COVID-19. As the chart on the next page illustrates translation volumes keep growing rapidly, but soon data becomes equally important as humans for the generation of output from the industry.

The global translation industry finds itself now in a 'mixed economy' condition: on one side a *vertical 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. In the old way of doing translation, every job is sent down the supply chain of project managers, linguists, reviewers, etcetera, all adding a further cost each and every time a new translation is needed.



*Based on estimates and meant to illustrate the trends.

Sustainability of the Current Model

This raises the following questions: is the mixed (hybrid) economic model that we have today sustainable? Can it scale up to reach the goals of world-readiness? How can the translation industry set itself free from the vertical, labor-based cost model? 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?

These fundamental economic questions lead to a reorientation in the translation space that becomes manifest through a spree of mergers and acquisitions, the rise of automated platforms with the credo of *human-in-the-loop*, and the shift to new outsourced activities: language data management.

The power of data becomes so strong that some companies consider creating basic datasets first when they are looking to expand into new locales. It is a subtle difference perhaps and the frontier becomes somewhat fluid, but the thinking goes towards putting data before translation. This is especially valid for what is called low-resource languages, where not so much language data is available and where the public is not 'spoiled' with high-quality translations or users are even excited to see content coming available in their own language.

So while the need for human translation starts to recede, the urge among language service providers to diversify their service portfolio is on the rise. The innovators in the translation industry will grasp these opportunities.

THE 'DATA FIRST' PARADIGM SHIFT

We now have enough evidence, both from a technology and an economic perspective, to see that the translation industry is in the midst of a paradigm shift. Computer-aided translation is making space for a data-first approach. Front-loading the machine - adding value to data ahead of its modeling rather than editing language only after machine processing - becomes the new focus. So how can we best use humans in the loop to add value to these data? Smarter data due to annotation will help improve the training of the algorithms that drive the machines that translate (transcribe, summarize, rewrite, etc) the content.

In today's AI-assisted industry, we have seen that language data consist of any sequence of linguistic elements that can be used to solve a "language" automation problem. Strings of letters can be used as data to find patterns about words, strings of words about sentences, and so on. These sets of text and speech will usually be grounded in spontaneously-generated communicative content (documents, conversations, and voice recordings), although they can also be generated by machines themselves to provide "synthetic data."

Before the AI wave, we used the notion of data in the language community in two ways:

1. **The traditional translation approach** - data here referred to familiar information *about* human languages - their structures, lexicons, and other properties, described by science and the literate tradition. Facts about words, phrases, meanings, pronunciation indications, were all data that helped translators find useful information in reference works such as dictionaries and grammars. They were also used by teachers and others to describe the object of their study. In fact, these are a form of *metadata* - data about naturally occurring language, but which are not necessarily useful for machine processing.
2. **The new data-driven approach**. In the LD4AI paradigm, data has come to mean *language content* (natural or synthetic text and speech), as found in the digital corpora constructed by linguists such as Quirk (see chapter 3), as documents or their multiple language versions. Translation memories (TM) are a classic example of computer-friendly data collections that can be searched, compared, and used for NMT learning.

In today's pipelines, language data of many kinds can be used as *training data* to build a model – i.e. for the algorithm that is going to be used to actually translate (or summarize, generate or otherwise process) texts. This means that access to these data, plus an understanding of data provenance, relevance, and quality, are now critical to the forward march of NMT. An AI machine does not learn *meanings* but distributions of items that only make sense when read by humans. This represents a small revolution in the world of translation, and is transforming parts of the industry into a more data-driven rather than purely human decision-making business.

The crucial condition is that mission-critical data are needed to ensure a quality target text, especially in those increasing numbers of cases where NMT output is delivered straight to an end-user without human post-editing. This means that the process of selecting and preparing

training data at scale needs to be significantly improved. TAUS, and no doubt others too, is inventing solutions to ensure that appropriate domain-sensitive training data can be streamed straight to any translation pipeline.

Data preparation is relevant not only to translation training but to other kinds of language processing tasks as well, from NLP to speech technology. This means that designing appropriate workflows and tools to enable a growing work-force will be key to delivering these services at scale.

In the next section, we look at how the language services sector is adapting to the data-first paradigm shift and what kind of new data services are being developed and produced.

On the other hand...

There are some 11,000 LSPs in the world and this report mentions the reactions of about 200 of them. There is a possible risk that this emphasis on data services among LSPs only represents a minor thread in the industry's economic transition to the post-COVID world.

Categorization of
Language Data Services

Forms of Data

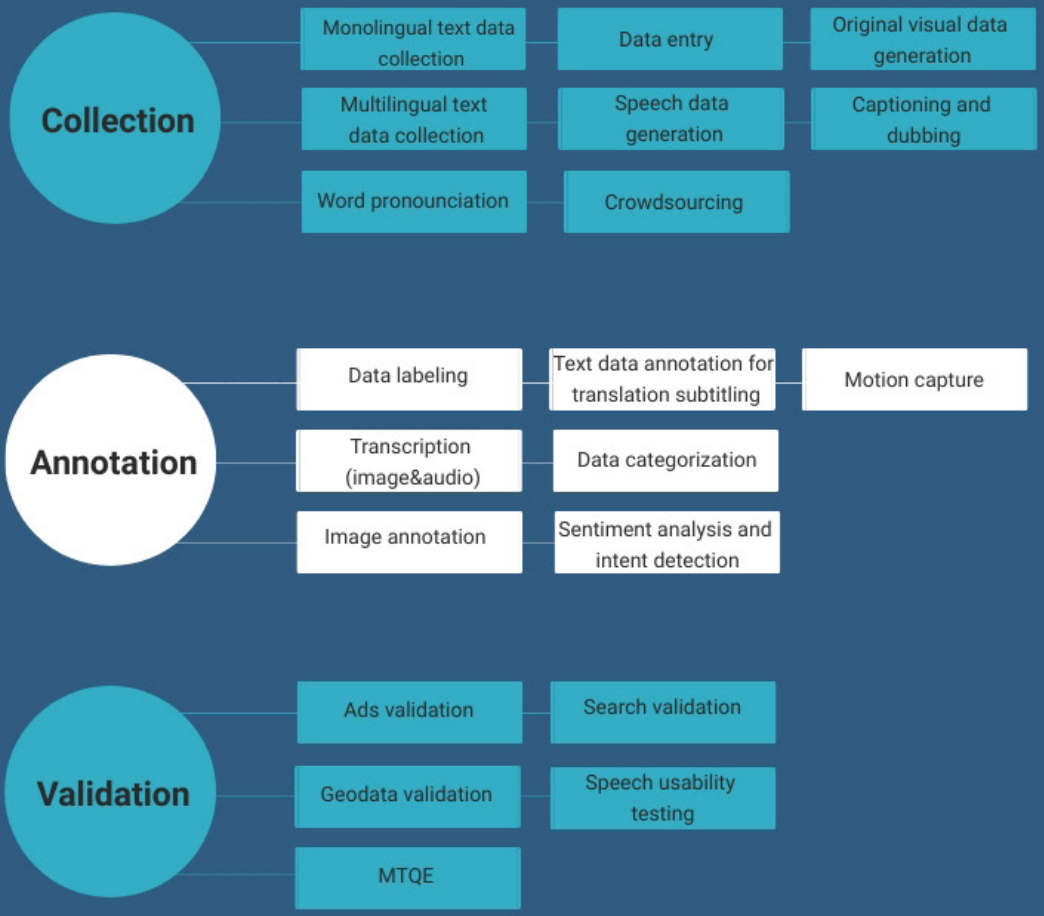
Text **Speech** **Visual (Image & Video)**

Served sectors:

- Augmented Reality (AR) & Virtual Reality (VR)
- Virtual Assistants
- Automotive (autonomous vehicles)
- Facial recognition
- Machine Translation
- Social Media
- Text-to-speech technology
- Robotics
- Medical
- Security

Main Services

Collection **Annotation** **Validation**



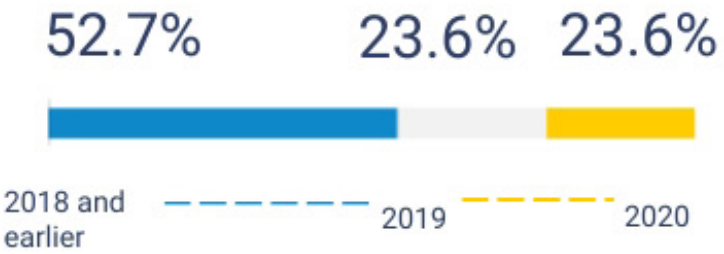
*Categorization based on the questionnaire answers submitted by data services providers

LSP perspective on whether translation industry is transforming into a language data industry

Based on 205 submissions

Recently received data-related project requests

When did the rise of data-related projects start?



Kinds of data-related tasks performed or requested:

- QA on term and info accuracy
- Speech data annotation
- Capturing speech as text
- Text data annotation
- Transcription
- Data creation for MT engine training
- Data tagging/labeling
- Producing original photo content
- Speech data collection for voice assistant training
- Post-editing MT output
- Language quality rating
- Gender bias detection and gender specific pronunciation
- Text detection/transcription in pictures
- Data classification
- Chatbot training
- Pronunciation data for visual assistant

Already have or planning to organize a "crowd-driven" business unit

*Results based on the survey submission by 205 language service providers

LANGUAGE DATA LANDSCAPE

Fifty Percent of LSPs Active in LD4AI

As the infographic on the prevous page illustrates, a significant proportion of LSPs (50% of the 200 or so surveyed for this report) say they have carried out LD4AI tasks for clients on some aspect of language training data. This ranges from the largest companies in the industry, some of which have significant plans to expand their LD4AI service offering into a business unit, down to much smaller companies that may be addressed by a client due to a specific language geography skill. In general, larger LSPs identified this market over five years ago, having understood that machine learning depends crucially on smarter data supplied by a mix of engineering and human intelligence.

So far, there are no reliable estimates available for the size of the specific LD4AI market. It is emerging as a sub-sector of the overall AI training data market, which could reach USD 4.8 Billion in value by 2027.

Although LD4AI as an industry is still embryonic, AI services have clearly gone mainstream as a new revenue source for many larger players in the localization industry. Various LSPs may also have embraced AI services to compensate for a downturn due to the impact of COVID-19 on their usual clientele, especially in the hospitality sector. We therefore note here the variety of AI activities which are contingent on LSP and translation expertise, and attempt to characterize their significance for the industry as a whole going forward.

Types of Services

Text versus speech has not been a natural terminological opposition in the translation industry; language has always meant *text* or writing, in contrast to interpretation for language as speech. Now that content is multimedia, text references written activity as part of a larger semiotic whole (image, speech, touch, signing, etc.). In the AI industry, text data is typically processed by "text analytics," with a global market worth around USD 5.3 billion in 2019, and expected to grow at 17% to reach a value of around USD 16 billion by 2025. However, text analytics is not necessarily associated with the kind of work performed on multilingual text by operators in the language data sector, so the size of the "language data" segment in this sector would be much smaller.

Services covered by LSPs and similar suppliers in this segment, as well as by larger data pure-players such as Appen and DefinedCrowd, include:

- Named entity recognition – helping machines distinguish brands and specific proper noun references from common word occurrences ("Apple" from "apple"). Note how many names of new startups tend not to use common vocab words or names to avoid this problem;
- Anonymization – to ensure privacy to obey the law;

- Semantic analysis (e.g. is a sentence sentiment positive or negative? Errors in linguistic description tests by potential Alzheimer patients can predict the disease);
- Data validation (e.g. in ads, search responses, for specific geographies, etc.);
- Document labeling - e.g. an email about X is different from one about Y, enabling documents to be routed automatically;
- Ambiguities - spotting structures and word usage that change the underlying 'syntax tree' and therefore the meaning;
- Cleaning – e.g. solving anything from punctuation issues to proper name spelling, and other text editing tasks, for example for OCR output, across all languages;
- Personally Identifiable Information (PII) - these data may need to be erased/anonymized;
- Bias suppression - identifying and repairing bias and ideological non-acceptability (gender, race, etc.) in lexical items and linguistic structures.

Speech Data Tasks

To contextualize the speech dimension, the global intelligent spoken virtual assistant market was valued by various analysts at around USD 3.5 billion in 2019 and is thought to be growing massively at a CAGR of 34 - 37%, giving it an estimated value of some USD 45 billion by 2027. The global captioning and subtitling solutions market meanwhile is expected to reach around USD 440 million by the end of 2026 growing by 7 to 8%.

The need for *multilingual speech data verification and management* is therefore growing relatively fast, especially for chatbot or virtual assistant technology, which is becoming a major new engagement channel for businesses of all kinds. Parts of this market obviously rely on human-in-the-loop data correction and validation to maintain quality.

Machine learning has radically improved speech recognition and transcription solutions in the last decade, and voice provides a number of rich markers for collecting further customer data to feed business intelligence. The data needs in this medium are fundamental to verify the identity and accuracy of the sentiment or pronunciation signals expressed by human voices subscribing to services. They cover the following datasets:

- Monologue speech - single-speaker scripted, guided, or spontaneous speech (broadband or narrowband, from speaker identification and pronunciation correction to transcription and annotating information points).
- Dialog speech - e.g. phone and chatbot interactions in both guided and spontaneous speech (aiding transcription tools to identify language elements, identify intent etc.)
- Speech-to-Text Transcription (STT) - collecting and correcting STT data to constantly improve transcription systems for any end uses including eventual translation (e.g. transcripts of law courts, business, and public sector meetings, as well as a broad range of phone calls and other recordings for security reasons, videos, etc.)

It would naturally be useful to pinpoint exactly how these tasks break down in terms of volume, and language spreads. Changes in their distribution would then provide useful input about the skill sets and manpower needs for different cases.

Data Preparation Tools

The nascent data preparation tools market is obviously poised to expand. One estimate for the size of the global data preparation tools market by 2025 is nearly USD8.5 billion.

There is no specific market for LD4AI tools as yet, but it is fairly predictable that the training data sector will begin to employ AI engineers to develop technology fixes to automate more tasks in language data verification and repair and thus speed them up, lower the cost, and broaden the business base.

Critically, the end-to-end automation of language validation issues is unlikely to occur any time soon, due to the complexity of the problem and the business risk of using technology that overlooks bias or error in data. However, the inevitable evolution of AI as a service (AIaaS) which can be provided without a human in the loop will produce solutions that attempt to deal with specific data problems automatically. **Utopia Analytics** (FI) is one such outlier in this market. It uses an innovative, language-independent analysis solution that can potentially adapt to any language or dialect by using a small data footprint to transcribe and process UGC text and speech data where ethical moderation is an issue.

On the other hand...

It would be useful to have some real numbers - i.e. data! - about the growing role of USG in demands for translation. There could also be limits to the amount of voice/speech work that smaller LSPs could bid for in the current state of the market. This extension of services to new skill sets will more likely attract new entrants who may not share the same commitment to the "multilingual quality" values that have inspired the industry until now.

PROFILING THE NEW CULTURAL PROFESSIONAL

Wide-Band Communications Talents

As language data in an AI framework grows in importance, it seems inevitable that the role of human intervention will change. On the one hand, humans will become more closely involved in checking, evaluating, and augmenting language content in AI workflows, as we have seen in the previous chapters. On the other, the profile of these human activities in the language industry will significantly evolve. The tradition has been to privilege well-trained language experts as translators; the future sketched out above is likely to require different skills more closely suited to supporting machine intelligence.

From our one-on-one interviews with big data providers, it is clear that machines need to be fed a wide variety of data to properly process sensitive language content. For instance, to properly recognize voice commands, an AI system needs data that includes male/female and young/old voices across a broad variety of accents for all kinds of languages.

To deliver all these language services, data providers typically create crowdsourcing platforms and apps where all types of data can be generated and evaluated to deliver appropriate training output. These platforms, in turn, are creating new communities of human resources – or 'crowds' – that play a new role within the translation industry economy which we can describe as "cultural professional". The skill set of this new role extends beyond language alone into the entire media universe found across cultures. By contributing knowledgeable annotations and evaluations to data training tasks, whether operating on voices, faces, images, truth judgments, or language, these contributors become "wide-band" cultural professionals for the part of the world or community they know best.

New Role of the Professionals

As this phenomenon becomes more widespread, what impact could it have on more "narrow-band" professional linguists, translators, and language experts that have traditionally supplied the industry? As AI technology assumes greater responsibility for many of the large-scale mechanical aspects of translation, translators will be working more on high-quality production, contextual text design, and transcreating content tailored to target cultures, as well as on all those language pairs where there is currently a lack of digital data. All in the knowledge that their human output will be recycled as new data, and aided by cultural professionals to train the AI.

Adapting to New Ways of Digital Communication

One obvious property of human language is the fact that it is constantly changing under the impact of communication dynamics. Pronunciations and intonations, vocabulary choices, pet phrases, memes, syntactic forms, borrowings from other languages and more are all subject to almost random shifts. Apart from legal and similar public regulatory documents, discourse is typically being influenced by the accelerating pace of change on two-way digital media networks.

In the new environment of “organic” (UGC) data collection and annotation, language specialists will be expected to adapt to this new style of rapid-change communication in line with what one topic expert interviewed called “the TikTok age of communication” in which punctuation, spelling, and usage are becoming more free-form than has been the case in traditional written language. They will also be able to play a role in testing sonic branding.

Emergence of Crowd-Based Platforms

In response to this, with 29% of LSPs already claiming a crowd-based platform and another 15% planning to establish one to cater to the increasing number of data-related requests, it appears likely that the role of the cultural professional will gradually involve more consistency checking of text and speech than content creation from scratch, especially in the case of social media translation and analytics. We, therefore, predict that there will be a reconfiguration of staffing needs within the translation industry within the next five years. A subset of translation companies will begin to employ more AI experts and data engineers to benefit maximally from data training opportunities; others will build new communities of “cultural professionals” to handle data tasks involving new forms of local expression.

On the other hand...

If this scenario goes real, there will be two problems to solve: acceptable working conditions and pay for data workers, and rigorous training protocols to maintain quality standards as the range of casual staff skills required reaches beyond core multilingual capabilities. These skills will affect both in-company management, and hiring modalities involving the “crowd”. Existing freelance translators could also protect their interests by creating new alliances to provide the best talent for these new tasks.

FIVE TAKEAWAYS

Synthesizing all the information and data we have gathered while working on this report, we arrive at the conclusion that the emergence of the LD4AI sector is a 'small revolution' indeed that affects the global language industry and will change the way it operates fundamentally in the coming years.

Here are the five main takeaways that we offer to the technologists and the entrepreneurs in the language industries and their customers to consider when they do their strategic planning for 2021 and beyond:

1. **Language is Core to AI**. Language, and therefore language data, is at the core of the AI revolution because it is the gateway to *augmenting* the key human intelligence skills of speaking and understanding. Machines using LD4AI bring scale to the intimacies of discourse, providing a universal processing opportunity for all languages.

   And so, it happens that the global language industry now potentially gets a much bigger and strategic role in the grander scheme of all business functions.

2. **Data-First Paradigm Shift**. When starting a new translation project, the focus changes from hiring translators to do the job to collecting relevant language data and preloading the models to do the job: we call that a *transformation* or a *paradigm shift*. The success formula in translation today (as in every other AI endeavor) is to create a virtuous loop between humans and data - or human in the loop.

   And so we expect drastic changes in technologies and workflows in the global language sector to support the data-first approach.

3. **Acceleration of Change**. Machines are thousands of times more productive than humans in translation and language tasks. The current economic model in the translation sector is not sustainable. The COVID-19 global crisis only makes this more visible. Language service providers diversify quickly into data annotation services. Freelancers and smaller agencies offer to sell their data rather than translations.

   And so, watch out for an acceleration of change in the economics of the language sector: business models, roles, pricing, everything.

4. **Rise of the New Cultural Professional**. The pursuit of perfection in linguistic quality that localization professionals are accustomed to is no longer the criterion for every job and task. This top-down focus is being replaced more and more by the urge to find and collect native organic data. The established cascaded supply chains of global vendors contracting with local vendors who in turn hire the freelance translators make place for, or exist alongside, human intelligence task platforms where natural talents (hundreds

of thousands of them) are invited to perform small tasks where the main criterion for success is the roots and orientation in the local culture.

And so the shift from pure competence to the need to speak the language of the people leads to the rise of the new cultural professional.

5. **New Markets Move Faster**. When everything is basically up for change, as we see in the four takeaways above, those markets with the least to replace in terms of legacy systems and established roles and businesses will typically move faster. This is the law of the handicap of a head start. We observe a passionate embrace of new working methods and an eagerness to create data in India, South-East Asia, the Middle East and Africa so as to promote their languages to the 'premier league'.

So expect globalization to shift and new markets to accelerate in the language industries as well.

TAUS DATA MARKETPLACE

There is an urgent need to establish a more equal level playing field for access to language data. Language data are too often locked up in silos and access is restricted to the big tech companies. This is a barrier to the healthy and prosperous development of the language and AI industry in general and the LD4AI sector in particular.

TAUS is grateful for the funding from the European Union under the Connecting European Facility Program to develop the Data Marketplace. Together with partners Translated and FBK Trento, TAUS is rolling out this new platform where all producers and users of language data can meet and trade their data. See datamarketplace.taus.net.

The TAUS Data Marketplace comes just at the right time to support the grand shifts in the translation and AI industries and to realize the best results out of the investments in automation and innovation. The rationale behind the launch of the TAUS Data Marketplace can be summarized as follows:

- The Data Marketplace establishes a more equal level playing field for the thousands of companies in the world that want to optimize automatic translation and invest in language-based AI. Today, access to sufficient language data is precluded to a few big-tech companies who can afford to invest. The TAUS Data Marketplace makes access to language data universal and affordable.

- The Data Marketplace creates an open, fair, and transparent market for data, allowing producers and owners of data to monetize their hard work. Language workers from parts of the world that have a low representation on the internet, for instance, will be able to create data in their languages and sell them directly to the end-buyers. Buyers and sellers of data from all language markets and all domains can connect and trade data directly.

- The Data Marketplace is the best platform to drive the industry agenda forward in support of language expansion and domain diversification.

- The Data Marketplace hosts already at the launch date the largest collection of language data (more than 35B words in 600+ language pairs) and it can help users of MT and AI to expand into new languages, domains, and applications very quickly.

- The Data Marketplace provides economies of scale in linguistic features, such as cleaning, anonymization, and clustering of data and access to the most advanced NLP services.

- The Data Marketplace has a solid legal framework and provides full transparency on the origin and usage of the language data.

TAUS
the language data network

TAUS was founded in 2005 as a think tank with a mission to automate and innovate translation. *Ideas transformed into actions.*

TAUS became *the language data network* offering the largest industry-shared repository of data, deep know-how in language engineering and a network of Human Language Project workers around the globe.

Our mission today is to empower global enterprises and their service and technology providers with data solutions that help them to communicate in *all languages*, *faster, better and more efficiently*.