Powering automatic translation with **TAUS Data Marketplace**

Milica Panić

Efforts to enable language data sharing and to get closer to language equality are not new or recent. They have, in fact, started some decades ago. However, to this date, no initiative has managed to gain enough traction to support the growing need for language data in the global language services industry where AI-enabled services are coming into play more and more. In this article, we introduce TAUS Data Marketplace, a platform for language data acquisition and monetization designed to serve as an enabler for the emerging Language Data for AI sector.

Early data sharing initiatives

Here are some of the well-known industry-wide collaborative actions:

- In 1992, the Linguistic Data Consortium (LDC) was founded in the USA as an open consortium of universities, libraries, corporations, and government research laboratories to address the data shortage faced by language technology research and development;
- 1995 saw the founding of the European Language Resources Association (ELRA), a non-profit organization with a mission to promote language resources and evaluation protocols and offering language data such as text and audio corpora, lexica, and terminology;
- In 2000, the Open Language Archives Community (OLAC) started as an international partnership of institutions and individuals interested in creating a worldwide virtual library of language resources;
- In 2014, the European Commission launched the European Language Resource Coordination (ELRC) program within the Connecting Europe Facility Automated Translation (CEF.AT) initiative.

Why language data matters

Almost three decades after the first initiative mentioned above, at the 4th ELRC Conference in Helsinki in 2019, it became clear that we were still far from language equality (digital or otherwise) in the increasingly multilingual world. The theme of the conference was

Milica Panić is a marketing professional with over 10 years in the field. As TAUS Head of Product Marketing she manages the positioning and commercialization of TAUS data services and products. Before joining TAUS in 2017, she worked in various roles at Booking.com, including localization management, project management, and content marketing. She holds two MAs in Dutch Language and Literature, from the University of Belgrade and Leiden University, and is passionate about inventing new ways to teach languages

Why Language Data Really Matters, stressing the importance of data sharing for the improvement of automated translation systems. Ironically, only a year later the whole world experienced firsthand the urgent need for timely access to multilingual information and rapid response in as many languages as possible, in the context of the worldwide COVID-19 crisis.

From the technological perspective, without sufficient language data we cannot speak of training or improving machine learning models, language expansion, or domain diversification. The biggest online machine translation (MT) service, Google Translate, currently supports only 108 languages, and even Google admits to facing difficulties when launching new languages due to meager amounts of available language data online. Commercial businesses wanting to reach users in emerging markets are constantly playing catch up when there are no available MT systems for new users' native languages. But there are also other, more philanthropic reasons to gather and document as much language data as we can. Think of all the research opportunities that performing large-scale cross-linguistic processing and studying the universal linguistics (as proposed by Steven Abney and Steven Bird) could offer, allowing for the preservation and fostering of languages, cultural heritage, and knowledge carried with them.

Bridging the gap with the Data Marketplace

Despite so many previous initiatives, public data collection efforts (e.g. Europarl, DGT Translation Memory, UN Parallel Corpus, Opus, etc.) and established data repository platforms (e.g. MyMemory, Linguee, Glosbe, TAUS Data Cloud), access to and sharing of language data remained limited.

History of TAUS Data

TAUS was an early mover in the language data space with the launch of TAUS Data Cloud in 2008. The platform consisted of a vast data repository supported by 45 founding members that grew to 73B words in 2,300+ language pairs by 2017. It used a reciprocal model, allowing users to upload data and earn credits to download other users' data. In the era of statistical MT, more data was always better data, and the Data Cloud served that purpose well. However, as the technology evolved into neural networks, and users became much more sophisticated, the old reciprocal model no longer sufficed. Users needed more domain-specific data, customization, and the ability to track the origin of the data. That is how the idea of an open market for data was born. What was still uncertain was how inclined the

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. Over time it has grown to offer the largest industry-shared repository of data, deep know-how in language engineering, and an immense network of language project professionals all over the world. Today it empowers global enterprises and their service and technology providers with data solutions that help them to communicate in all languages, faster, better and more efficiently.

data users and producers would be to adopt the marketplace model. In 2015, TAUS conducted a survey among data users and producers which showed that people are most likely to share their data if they receive an immediate benefit from it and if the usefulness of what they get in return outweighs the concerns around privacy, confidentiality, and data ownership (cf. TAUS Data Market White Paper, 2017). Similarly to how people are happy to share their location with Google Maps provided they get directions to their destination, translators tend to share their translations and edits if, as a result, they get immediate access to a good usable MT and if they know this engine will do yet better for their future projects. This insight confirmed our belief that establishing a Data Marketplace might be a more successful step towards creating a continuous supply of language data for all stakeholders in the global language and AI industries. In addition, the results of another survey from 2019 signaled that more and more players in the conventional language sector were acknowledging the fact that data-related services were becoming a new business stream. As more businesses join the Language Data for AI sector, the need for a platform where sellers and buyers can discover and reach each other directly became evident.

The first version of TAUS Data Marketplace was released in October 2020. The platform development is a collaborative project between TAUS, Translated, and FBK Trento, co-financed by the European Union under the Connecting European Facility Program (CEFR). By October 2021, all scoped functionalities will be available to all users.

Platform value proposition and users

TAUS Data Marketplace business model is intended to disintermediate the language data supply chain by connecting data producers and consumers directly. To incentivize trading, the platform offers advantages to both end-users and providers through high-value data processing services. While data sellers benefit from free cleaning and anonymization during data upload and the monetary reward when their data is purchased, the data buyers get to use the advanced clustering technique (known as Matching Data), to find the most useful quality language data that is fit for their purpose. As a result, the Data Marketplace establishes a more equal-level playing field for all those stakeholders wanting to invest in language-based AI, by making access to language data easier, more affordable and less of a privilege reserved to the big-tech companies that can afford doing data collection, crawling and curation themselves.

So far, sellers on the Data Marketplace come from different

TAUS Data Services include end-to-end language data solutions and off-the-shelf custom datasets for training ML and AI systems. They enable organizations to scale their business and reach their users in high-growth markets. TAUS has been a provider of data since 2008 and has participated in numerous EU-funded projects. Today, the TAUS in-house NLP team includes researchers, data scientists, and data engineers providing tailored, fit-for-your-use-case solutions, while its Human Language Project (HLP) communities of native speakers build on-demand datasets in a variety of low-resource domains and language pairs. Data cleaning, anonymization, and clustering are available through the TAUS Data Marketplace, while data creation, annotation, transcription and content licensing are a part of the HLP platform.

backgrounds such as publishers, data companies, language service providers, translators and linguists who look for ways to monetize the language data they have collected or generated over time. Data buyers, on the other hand, are any users or developers of MT and AI looking to expand into new languages, new domains, and new applications quickly.

Addressing data ownership and copyright concerns

One of the first steps in the process of establishing the Data Marketplace was diving deeper into the evident uncertainties around language data ownership, copyright, control, and liabilities. No matter how useful the platform could be, it was our responsibility to keep the potential users informed about all questions. Intellectual property and data protection laws vary from country to country and are still to catch up with the AI technologies. Moreover, the translation ecosystem is complex, making it hard to draw conclusions about the responsibilities and legitimacy of specific use cases. Therefore, TAUS partnered with the acclaimed Baker McKenzie law firm to produce the Who Owns My Language Data white paper as a blueprint for the industry.

Additionally, TAUS Data Marketplace is established on a strong legal framework that complies with privacy policies in Europe and North America, with great emphasis on the importance of transparency on the origin and lawful usage of language data.

How it works

Anyone is welcome to upload or explore language data on TAUS Data Marketplace.

As a seller, one should be aware of the rights sellers have when it comes to sharing and selling translation memories or any other type of language data. On the Data Marketplace, they have the opportunity to prepare their data with cleaning and anonymization as part of the publishing process, ensuring that published data is unique and of good quality. The data price is defined by the seller or based on smart suggestions coming from the Price Index Table. In the last step of the publication process, they can add metadata such as domain and content type and a description to make their dataset stand out. Once data is published it is visible to all potential buyers in the Search area and under the Sellers tab.

For those looking for data to buy, the process is even simpler. They can either search by the language pair, available domain, or content type or upload a sample to search for data that corresponds to the domain represented by their sample data. The search will return a sample to review and decide if it suits their needs. When ready to buy, they will be directed to a secure payment environment, and after performing the transaction a link will appear to download the data and to rate and review it, if they like. After a two-week cancellation period, payment is transferred to the sellers and the trading is complete.

In the first six months since launching the Data Marketplace, the platform sparked considerable interest, especially among data sellers. More than 30 sellers have joined the platform, sharing over 500 datasets including low-resource languages like Kurdish, Pashto, Sinhala, Turkmen, and Yoruba. Success stories are available here. Seeing how well the model is received, we are increasingly confident that TAUS Data Marketplace is a great platform to drive the industry agenda forward in support of the advancement of global communication across all languages.