LANGUAGE TECHNOLOGY LANDSCAPE CONFERENCE

CHARTING THE GLOBAL LANDSCAPE OF LANGUAGE TECHNOLOGY

June 18, 2024

A EUROPEAN COMMISSION INITIATIVE

BY NIMDZI INSIGHTS

Text-based LTs

(including OCR and alternative modalities)

Laszlo K. Varga, Nimdzi Insights
Jourik Ciesielski, Nimdzi Insights

LANGUAGE TECHNOLOGY LANDSCAPE CONFERENCE

Housekeeping

Q&A

In-session polls

Recording

Post-event feedback survey

LANGUAGE TECHNOLOGY LANDSCAPE CONFERENCE

Agenda

Time	Session
10:00-10:30	**Welcome & Keynotes** Hosts: Philippe Gelin, Laszlo K. Varga Keynotes: Renate Nikolay, Renato Beninatto
10:30-11:25	**The Landscape of Language Technology** Speakers: Laszlo K. Varga
11:30-12:00	**Multilingualism of European Websites and the Technology Solutions Supporting It** Speakers: Andrejs Vasiļjevs
12:00-13:00	**Lunch Break**
13:30-14:20	**Large Language Models and Foundational Language Technologies** Speakers: Laszlo K. Varga, Nadezda Jakubkova
14:25-15:15	**Text-based Language Technologies** Speakers: Laszlo K. Varga, Jourik Ciesielski
15:15:15:35	**Break**
15:35-16:25	**Speech-Based Language Technologies** Speakers: Laszlo K. Varga, Igor Szoke, Khalid Choudry
16:30-17:00	**Closing** Speakers: Philippe Gelin, Laszlo K. Varga

LANGUAGE TECHNOLOGY LANDSCAPE CONFERENCE

Agenda

Time	Session
10:00-10:30	**Welcome & Keynotes** Hosts: Philippe Gelin, Laszlo K. Varga Keynotes: Renate Nikolay, Renato Beninatto
10:30-11:25	**The Landscape of Language Technology** Speakers: Laszlo K. Varga
11:30-12:00	**Multilingualism of European Websites and the Technology Solutions Supporting It** Speakers: Andrejs Vasiļjevs
13:30-14:20	**Large Language Models and Foundational Language Technologies** Speakers: Laszlo K. Varga, Nadezda Jakubkova
14:25-15:15	**Text-based Language Technologies** Speakers: Laszlo K. Varga, Jourik Ciesielski
15:15:15:35	**Break**
15:35-16:25	**Speech-Based Language Technologies** Speakers: Laszlo K. Varga, Igor Szoke, Khalid Choudry
16:30-17:00	**Closing** Speakers: Philippe Gelin, Laszlo K. Varga

LANGUAGE TECHNOLOGY LANDSCAPE CONFERENCE

Text-based language technologies

Laszlo K. Varga

Lead Researcher and Analyst,

Nimdzi Insights

Jourik Ciesielski

Technology Consultant and Researcher

Nimdzi Insights

LANGUAGE TECHNOLOGY LANDSCAPE CONFERENCE

Text-based language technologies

Introduction.

Text LTs are a meta-category where written words are used as input and output to create value for intralingual and monolingual tasks alike.

Categories explained.

In this section, the following LT categories were grouped into text-based LTs for convenience:

- Machine translation
- Translation management systems
- QA tools
- Optical character recognition and handwriting-to-text
- Alternative input (Braille, Sign language)

LANGUAGE TECHNOLOGY LANDSCAPE CONFERENCE

Machine translation (MT) 1

Main LT category	Machine translation
Market size estimate (2023)	EUR 2-3 billion
Growth potential	Moderate
Investment interest	Low
Market character	Three levels: 1. Core tech is well-established, big-tech dominated, commoditised 2. Custom enterprise solution providers, professional services heavy 3. Base of composite LTs and services (TTS, MI/S2ST, NLP, Chatbots)
AI / ML adoption / disruption level	Already adopted, LLM adoption ongoing
Technology maturity level	Stable

Introduction. Demand.

Machine translation is the “obvious” language technology. Since Google Translate, it is widely accessible and used. Low resource languages are a gap in quality, as well as special domains.

Integrated into websites, mobile devices, chatbots, TMSs, and more.

Raw MT output is used in low-risk scenarios, otherwise MT used as a translation productivity tool, as well as an alternative to no-translation.

Demand will only keep increasing as content is exploding, and MT is a core component for compound LTs.

Market size and character.

The MT market is dominated by big-tech’s solutions: free options for consumer, low-cost for generic engines in the cloud, with customisation options. Domain-trained specific engines are important for niche markets.

Because of commoditisation, prices are low and and moderate growth is expected in market value.

LANGUAGE TECHNOLOGY LANDSCAPE CONFERENCE

Machine translation (MT) 2

Main actors.

Big-tech dominated, commoditized. Key big tech players include Google, Microsoft, Amazon, Baidu, Alibaba.

Many tech-enabled language service providers also develop their own proprietary MT systems, but most of them don't directly commercialise them. Exceptions include RWS's LanguageWeaver, Translated's ModernMT, and Unbabel.

Pure-tech challengers include European players such as DeepL, Globalese, Systran, Pangeanic.

Quality is a differentiating factor, and also the ability to adapt 'real-time' to edits made by humans.

Security and on-premise deployment is a sub-segment within machine translation.

"AI whitewashing" heavily applies.

Company	Country of origin	Languages supported	Estimate MT revenue (2023)	Investment / funding (till March, 2024)
Alibaba Group	China	210+	(not core)	(not core)
Amazon	USA	75	(not core)	(not core)
Apptek	USA	70+	EUR 20 million	undisclosed
Baidu	China	200+	(not core)	(not core)
DeepL	Germany	30+	EUR 100-150 million	>EUR 100M
Globalese	Hungary	30+	EUR 5-10 million	undisclosed
Google	USA	100+ (EU24)	(not core)	(not core)
Lengoo	Germany	50+	EUR 2 million	EUR 20 million
LILT	USA	55	EUR 25 million (MT share ND)	EUR 100 million
Microsoft	USA	110+	(not core)	(not core)
Mirai Translate	Japan	14	EUR 7.5 million	undisclosed
Translated	Italy	200+	EUR 55 million (MT share ND)	undisclosed
Omniscien	Thailand	60+	EUR 5-10 million	undisclosed
Pangeanic	Spain	9	EUR 2 million (MT share ND)	undisclosed
Promt	Russia	41	undisclosed	undisclosed
Rozetta	Japan	25	EUR 20 million	Publicly traded
RWS	UK	62	EUR 100-200 million	Publicly traded
Systran	France	50+	EUR 20 million	undisclosed
Tilde	Latvia	20+	EUR 10 million (MT share ND)	undisclosed
Tencent	China	19	(not core)	(not core)
Unbabel	USA / PT	30+	EUR 35 million (MT share ND)	EUR 100 million
XL8	USA	46	EUR 5-10 million	EUR 10 million
Yandex	Russia	100+	(not core)	undisclosed

LANGUAGE TECHNOLOGY LANDSCAPE CONFERENCE

Machine translation (MT) 3

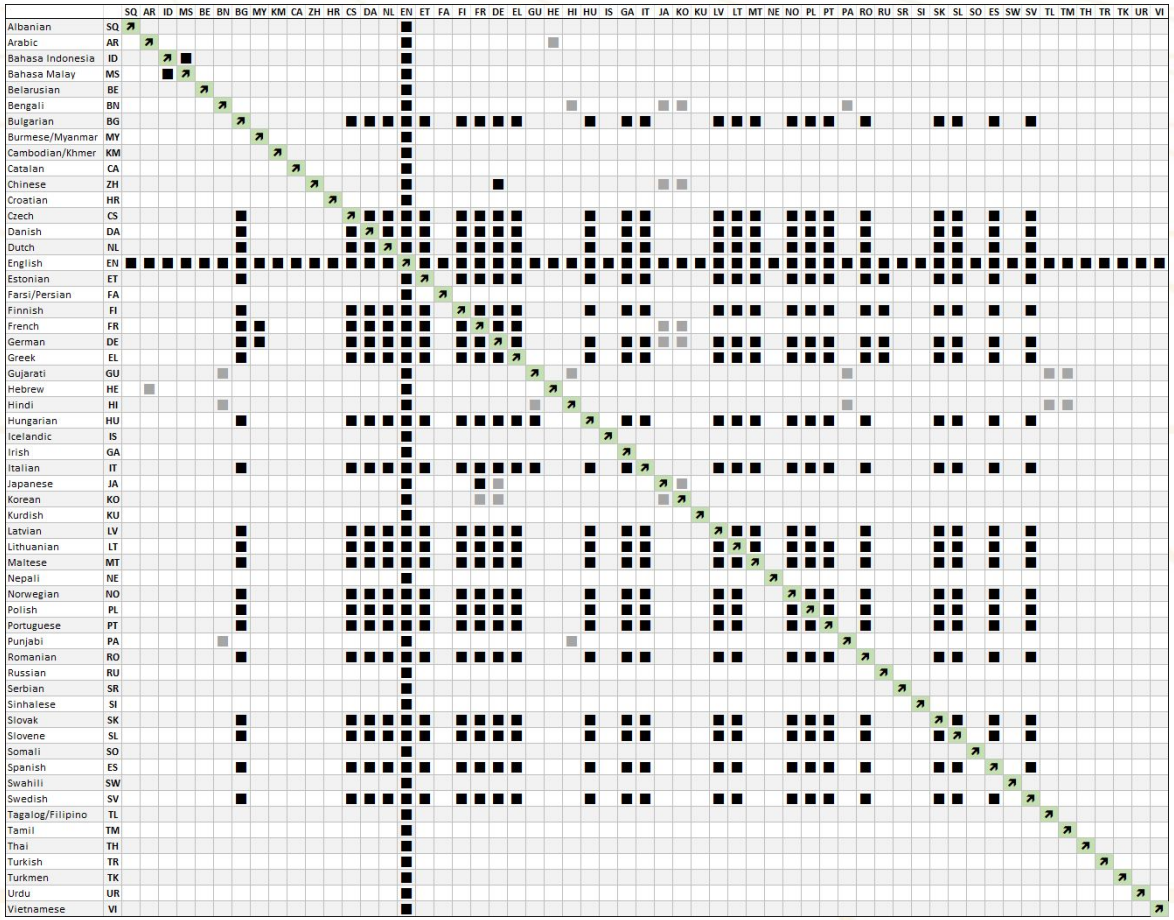
Quality.

High-resource languages are well-supported with diminishing quality returns on R&D, but **not all language combinations are equal.**

MT is widely used as a productivity tool for translation, with human post-editing is widely applied for better results.

Raw MT output is only used in low-risk, high-velocity scenarios.

Special domain coverage is a challenge solved by custom training.



Not all language combinations are created equally.
Source: Omniscien

LANGUAGE TECHNOLOGY LANDSCAPE CONFERENCE

Machine translation (MT) 3

Technology outlook.

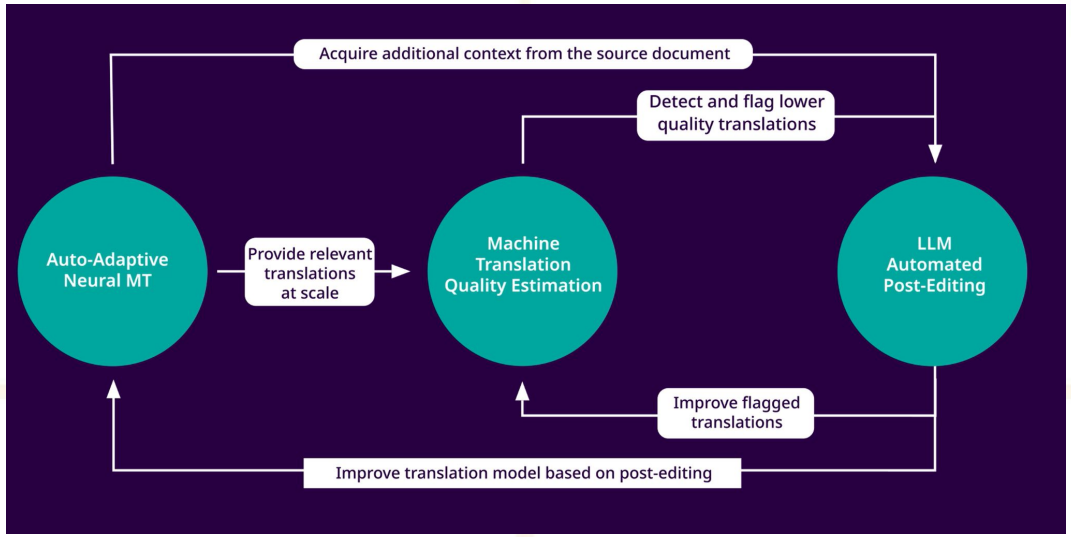
Stable technology core (transformers) since 2017 in all models. Generative **LLMs are an alternative to** the sequence-to-sequence **“traditional” neural MT solutions**.

Adaptive and contextual MT is the new frontier.

The language industry is now adopting
LLMs for augmentative MT tasks such as:

- Source optimisation
- Quality estimation
- Automatic post-editing
- Automatic (L)QA

Automated translation bench	
Neural MT	**LLM**
Purpose-built	General purpose
Faster	Slower
High resources for training and fine-tuning	Even higher resources
Low inference compute requirements	High inference compute requirements
Hallucinate mostly under out-of-domain inputs	Hallucinate in general
More accurate, less fluent	Good accuracy, more fluent
Once trained, ready for direct use	Require carefully devised and evaluated prompt techniques for initiation and context
Hard to insert context and language assets	Injection of in-context learning (prompts)
ROI established	ROI still under scrutiny



Source: RWS Evolve

LANGUAGE TECHNOLOGY LANDSCAPE CONFERENCE

Translation management systems (TMS) 1

Main LT category	Translation management systems
Market size estimate (2023)	EUR 0.3-0.5 billion
Growth potential	Low
Investment interest	Moderate
Market character	Well-established
AI / ML adoption / disruption level	Low
Technology maturity level	Stable

Introduction. Demand.

TMS platforms allow language service providers (LSP) and buyers (LSB) to manage the translation process of multiple content streams effectively. TMSs grew out from CAT tools with integrated features such as for project and vendor management.

LSPs have widely adopted TMSs, the major players having developed their own systems. Enterprise LSBs, on the other hand, typically rely on 3rd party solutions, while SMEs would use their LSPs' TMS.

TMSs come both in cloud and (some) as on-premise installations, often integrated with buyer-side content and product management systems and additional LTs such as MT, LLMs, OCR, and QA tools.

Market size and character.

The TMS market has both bundled and standalone solutions. Some major LSPs don't resell their TMS outside of their clientele (bundling) – exceptions include RWS Trados, while pure-tech standalone solutions are popular with buyers and LSPs due to their independence. Big tech is not present in this space with commercial solutions.

Market size is estimated to be between EUR 300-500 million, with moderate growth outlook, as content localization with MT proliferation may circumvent TMS in some cases; however, enterprise use is predicted to be stable.

LANGUAGE TECHNOLOGY LANDSCAPE CONFERENCE

Translation management systems (TMS) 2

Main actors.

The main TMS actors come from across the globe, European and US products are all available, generalists and niche providers alike.

Most actors have been on the market for more than 5-10 years.

AI-enablement is at the forefront of marketing communications.

Quality and technology outlook.

A main differentiator of TMSs in terms of functionality is the level of integration and number of connections with 3rd party systems and the coverage of use cases – from focused niche providers of SW localisation TMS to generalists for all use cases.

A new disruptor is – of course – GenAI, being adopted for MT and pre- and post-processing ot MT. Orchestration of localisation tasks in a user-friendly way is a new buzzword. Another new frontier for major TMS providers is incorporating modalities other than text, possibly via multimodal language models.

Company	Country of origin	Founded	Product name	Estimated TMS revenue (2023)
Across Systems	DE	2005	Across	EUR 10 million
Bureau Works	US	2004	BWX.io	EUR 3 million
Tarjama	AE	2019	CleverSo	EUR 0.5 million
Crowdin	EE	2009	Crowdin	EUR 2 million
TransPerfect	US	1999	GlobalLink	EUR 100 million
Lokalise	US	2017	Lokalise	EUR 7 million
memoQ	HU	2004	memoQ	EUR 13 million
Phrase	CZ	2010	Phrase TMS	EUR 35 million
Smartcat	US	2016	Smartcat	EUR 25 million
Smartling	US	2009	Smartling	EUR 35 million
STAR Group	CH	1986	STAR Transit	EUR 4 million
RWS	UK	~1984	Trados Enterprise	EUR 100-150 million
Transifex	GR/US	2009	Transifex	EUR 12 million
MittagQI	DE	2009	Translate5	EUR 1 million
Translated	IT	1999	TranslationOS	EUR 6 million
Weglot	FR	2016	Weglot	EUR 1 million
TransPerfect	US	2008	Wordbee	EUR 2.5 million
XTM International	UK	2002	XTM Cloud	EUR 30 million

LANGUAGE TECHNOLOGY LANDSCAPE CONFERENCE

QA and review tools

Main LT category	Quality assurance and review tools
Market size estimate (2023)	<EUR 100 million
Growth potential	Moderate
Investment interest	Low
Market character	Well-established
AI / ML adoption / disruption level	High
Technology maturity level	Mature / evolving

Introduction. Demand.

QA (AutoQA) tools are commonly used is localisation workflows by LSPs and LSBs alike. Large LSPs and buyers may also develop their own in-house solutions.

With the emergence of LLMs, automated review (AutoLQA) and evaluation / estimation (MTQE) tools are being developed to save human reviewer time for localisation efficiencies.

Main actors.

AutoQA: ContentQuo, Kaleidoscope's GlobalReview, Yamagata's QA Distiller, Xbench (ApSIC), and Verifika (Palex).

AutoLQA / MTQE as the next frontier: TAUS, ModelFront, Phrase, RWS (Evolve), and Translated are commercial solutions, while major tech-enabled LSPs are all experimenting with such solutions.

Market size and character.

Demand for QA and review tools is limited to localisation programs and companies, and is estimated in the vicinity of EUR 50-100 million. With the disruption of LLMs opening new uses cases and horizons, additional growth is possible, although this will be shared between internally developed tools (that are invisible on the market) and commercial solutions.

LANGUAGE TECHNOLOGY LANDSCAPE CONFERENCE

OCR and handwriting (HTT) 1

Main LT category	Optical character recognition (OCR) and handwriting-to-text (HTT)
Market size estimate (2023)	<EUR 1 billion (OCR) <EUR 1 billion (HTT)
Growth potential	Low/none
Investment interest	Low
Market character	Embedded Well-established
AI / ML adoption / disruption level	Ongoing, via VLMs and MMLMs
Technology maturity level	Stable

Introduction. Demand.

Optical character recognition (OCR) is a data extraction technology that turns scanned documents (camera images or non editable PDFs) into editable digital files. It is one of the Achilles' heels of downstream tasks including localization and digital document management.

Because of a non-0 error rate prevalent in most machine learning systems, most often, OCR/ICR output is further corrected by human editors. Time-saver applications.

SImilar considerations are true for handwriting, which, in some sense, are a subcategory of OCR.

.

Market size and character.

OCR is a widely used piece of technology. Healthcare, transportation, retail, legal, insurance and banking, government and public organisations, and in general, any business or institution that has large support departments – such as accounting (invoice automation), customer service, legal, finance, or HR – in need of digitisation of printed or written forms and other documents.

In addition, OCR/ICR is also used in applications such as image translation (like on mobile phones).

The most popular use case of HTT is converting handwritten notes to digital text from touchscreen devices, and, as such, is a widely used tool in consumer settings as well.

LANGUAGE TECHNOLOGY LANDSCAPE CONFERENCE

OCR and handwriting (HTT) 2

OCR main actors.

OCR tools are fundamental features within intelligent document management (IDM) systems, and rather auxiliary features for big-tech's cloud platforms - such as in MS Office. IDMs tend to call OCR as ICR for Intelligent, which may learn from continuous usage instead of from pre-set rules.

Many OCR tools are also freely available, while corporate document management solutions are a vast industry with players such as Adobe or OpenText.

Quality and technology outlook.

Visual (VLM) and multimodal (MMLM) language models are disruptors in this space, also dubbed "OCR-free document understanding" tools. While all IDMs are working or experimenting with this, solutions such as Rossum's (CZ) Aurora or JPMorgan's DocLLM also provide glimpse into the future of GenAI disruptions in the field.

Company	Country of origin	Product name	Languages supported	Estimated revenue (2023) (not OCR only)	Investment / funding (till March, 2024)
ABBYY	US	FineReader, OCR Container	>200, all EU24	EUR 200 million total	"significant" funding
OpenText	CA	OpenText Intelligent Capture	>120 (SAP: >60; web client: 11)	EUR 4 billion total	EUR 1 billion
Tungsten Automation	US	OmniPage	~100, all EU24	~EUR 600 million total	undisclosed
Adobe	US	Acrobat	34, non-EU24	EUR 4.5 billion	Publicly traded
Konica Minolta	JP	Document Navigator	n/a	EUR 7.8 billion	Publicly traded
IRIS	BE		137, all EU24	EUR 5 million	undisclosed
Rossum	CZ	Aurora	20, non-EU24	EUR 25 million total	EUR 90 million
UiPath	RO		~100, all EU24	EUR 370 million total	EUR 2 billion
Nanonets	US		>40, non-EU24	EUR 3 million total	EUR 26 million funding
Accusoft	US	SmartZone OCR	~100, non-EU24	EUR 20 million total	EUR 2 million
Foxit	CN/US		16, non-EU24	EUR 80 million total	EUR 380 million
Infrrd	US		>21, non-EU24	EUR 13 million	EUR 0.7 million
Docsumo	SG		60, non-EU24	EUR 5-10 million	EUR 3.5 million
IBM	US		>20, non-EU24	(not core)	(not core)
Google	US	Google Document AI, Google Cloud Vision	Wide range, all EU24	(not core)	(not core)
Microsoft	US	Azure AI Vision	Wide range, all EU24	(not core)	(not core)
Amazon	US	AWS Textract	6	(not core)	(not core)

LANGUAGE TECHNOLOGY LANDSCAPE CONFERENCE

OCR and handwriting (HTT) 3

HTT main actors.

While all major IDMs (as described in the OCR section) provide handwriting recognition support, there are specific solutions worth mentioning as standalone tools. For example, Mysrcipt's (FR) solution was integrated into the reMarkable 2 paper tablet, or Transkribus, a wide consortium for HTT technologies used for historic documents.

In addition, big tech companies also offer various HTT tools in the platforms such as mobile devices or office productivity tools via touchscreen. It is unknown which HTT tools are used my most HW makers (such as Apple, Samsung, or Lenovo.

Quality and technology outlook.

While HTT tools have always used machine learning algorithms, the new vision language models may be a step change for casual use.

In addition, handwriting generation is also made possible by new developments, leading to the possibility of handwriting cloning with already impressive (and alarmingly good) results.

Company	Country of origin	Products	Languages supported	Estimated HTT revenue (2023)	Investment / funding (till March, 2024)
Nanonets	US	Back-office automation and OCR	>40 (non-EU24)	EUR 2-5 million	EUR 40 million
Myscript	FR	Note taking Myscript Nebo. Integrated into reMarkable 2.	66 (non-EU24)	EUR 10-20 million	undisclosed
Transkribus	AT (20 country consortium)	HTT solution especially for historic documents. Horizon 2020 origin.	n/a (custom training possible)	undisclosed	undisclosed
Amazon	US	AWS Textract	6	(not core)	(not core)
Google	US	Gboard, Google Document AI, Google Cloud Vision	71	(not core)	(not core)
Microsoft	US	Ink to text Azure AI Computer Vision	58, non-EU24 5	(not core)	(not core)

LANGUAGE TECHNOLOGY LANDSCAPE CONFERENCE

Braille technologies

Main LT category	Braille technologies
Market size estimate (2023)	<EUR 200 million
Growth potential	Low / none
Investment interest	Low
Market character	Consumer and DEI / public sector driven Demand is compartmental
AI / ML adoption / disruption level	Low disruption
Technology maturity level	Stable

Introduction. Demand.

The Braille alphabet is used by 6 million people worldwide, our of 250 million blind and low vision individuals. Diversity, equality, and inclusion (DEI) principles drive the technology landscape for Braille users.

The market includes visually-impaired people, teachers, educators, parents or legal guardians, as well as public institutions and national-level organisations of public use (such as transport companies and emergency services).

Main actors.

Duxbury (UK), Humanware (US), Eurobraille (FR), Help Tech (DE), Freedom Scientific (US), Don Inc (KO) round up the most important actors in the Braille technologies space, including Braille translators, tactile input and smart devices.

Language support for Braille devices varies, typically includes only major languages, between 5-15, whereas Duxbury's translator solution supports a wide range of languages including all official EU24.

Market size and character.

The market for Braille devices (which may cost EUR 1-5 thousand) is estimated at around EUR 50 million, which, together with other solutions, limits the market size to less than EUR 200 million.

Demand for Braille niche providers' technologies is comparmentalised due to the nature of customer base.

LANGUAGE TECHNOLOGY LANDSCAPE CONFERENCE

Sign-language technologies

Main LT category	Sign-language technologies
Market size estimate (2023)	<EUR 100 million
Growth potential	Low / moderate
Investment interest	Low
Market character	Consumer and DEI / public sector driven Demand is compartmental
AI / ML adoption / disruption level	Ongoing
Technology maturity level	Immature / emerging

Introduction. Demand.

Around 430 million people - including 34 million children - are affected by a disabling hearing loss. Diversity, equality, and inclusion (DEI) principles drive the technology landscape for sign-language users. Unlike Braille, there is no single sign-language: there are 300 variants across the world, including 30 in the EU itself.

The market includes deaf and hearing-impaired people (D/HH), teachers, educators, parents or legal guardians, as well as public institutions and national-level organisations of public use (such as transport companies and emergency services).

Main actors.

Sign language solutions include translators and (mocap) AI avatar. Main actors include Kara Technologies (NZ) and Signapse.ai (UK), as well as Signer.ai (IN), Slait.ai (DE), SignForDeaf (TR), PopSign (US), Hand Talk (BR), and Lingvano (AT).

Sign-language support of solutions vary greatly; some focus on local needs, but American and British (ASL and BSL) are the most common.

Market size and character.

The market for sign-language technologies is estimated at to be under EUR 100 million. As demand is often DEI / principle driven, moderate growth of these technologies can be expected.

At the same time, sign-language support of these technologies is only expected to increase with deeper public and government embracement and support.

LANGUAGE TECHNOLOGY LANDSCAPE CONFERENCE

Language education technologies 1

Main LT category	Language education technologies
Market size estimate (2023)	EUR 2 billion
Growth potential	Moderate
Investment interest	Low
Market character	Maturing, mainly consumer-focused
AI / ML adoption / disruption level	Moderate, ongoing
Technology maturity level	Evolving

Introduction. Demand.

Language education technologies fit primarily into the EdTech sector, but because of their language nature, they are also researched in the Study..

From book-based methods and video-conferencing remote tutoring, new LangEdTech applications are immersive, interactive, and even personalized.

The introduction of adaptive AI tutors instead of pre-defined learning paths is the new innovation in the space, primarily targeting consumer buyers, while also enabling businesses for team licences.

Market size and character.

Language education technologies are often freemium applications; basic functionalities are offered for free, and additional packages or features (such as personalisation or ad-free apps) are available on purchase. The main money-making model is subscription-based revenue.

Estimating from the main players' revenues, the market size currently is approximately EUR 2 billion.

LANGUAGE TECHNOLOGY LANDSCAPE CONFERENCE

Language education technologies 2

Main actors.

Prominent LangEdTech companies include the market leader Duolingo (US), Babbel (DE), Busuu (UK), Memrise (UK), and Rosetta Stone (US).

The language supported by the actors vary, but are typically in the 14-50 range, primarily to cover the main demand (low resource languages excluded). Differentiators also include audience segmentation (children, adults, etc)

Quality and technology outlook.

LangEdTech is heading towards (even) more immersive and interactive modes, often heavily gamified in alignment with the “fight for attention” trend of social networks.

The level of personalisation available, and the technologies used (audio, VR, and chatbot-like solutions with LLMs) is increasing to attract customers from a slowly growing pie.

At the same time, LLMs via their chat interface can also create language learning experiences, which is why LangEdTech companies are rushing forward with features, LLM implementations, and community building, to keep differentiating their offering.

Company	Country of origin	Languages supported	Estimated revenue (2023)	Investment / funding (till March, 2024)
Duolingo	US	39 from EN 100+ total combinations	EUR 500 million	EUR 160 million
Babbel	Germany	14	EUR 250 million	EUR 30 million
Busuu	UK	14	EUR 50 million	n/a (part of Chegg)
Memrise	UK	23	EUR 20 million	EUR 20 million
Qlango	Slovenia	45	EUR 5 million	undisclosed
Mondly	Romania	41	EUR <5 million	undisclosed
LingoDeer	Singapore	16 (Asia focus)	EUR <5 million	undisclosed
Drops (by Kahoot!)	Norway	50	n/a (part of Kahoot! platform)	n/a (part of Kahoot! platform)
Rosetta Stone	US	24	EUR 200 million (part of IXL Learning)	n/a (part of IXL Learning)

LANGUAGE TECHNOLOGY LANDSCAPE CONFERENCE

Q&A

Thank you.

Have more questions later?
Find me on LinkedIn or reach out to the project team at ltsurvey@nimdzi.com.

Feedback survey