LANGUAGE TECHNOLOGY LANDSCAPE CONFERENCE

CHARTING THE GLOBAL LANDSCAPE OF LANGUAGE TECHNOLOGY

June 18, 2024

A EUROPEAN COMMISSION INITIATIVE

BY NIMDZI INSIGHTS

LANGUAGE TECHNOLOGY LANDSCAPE CONFERENCE

Your hosts

Philippe Gelin

Head of Sector "Multilingualism"

Directorate-General for Communications Networks, Content and Technology (DG CNECT), European Commission

Laszlo K. Varga

Lead Researcher and Analyst,

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

Keynote from the European Commission

Renate Nikolay

Deputy Director-General

Directorate-General for Communications Networks, Content and Technology (DG CNECT)
European Commission

LANGUAGE TECHNOLOGY LANDSCAPE CONFERENCE

Keynote from Nimdzi Insights

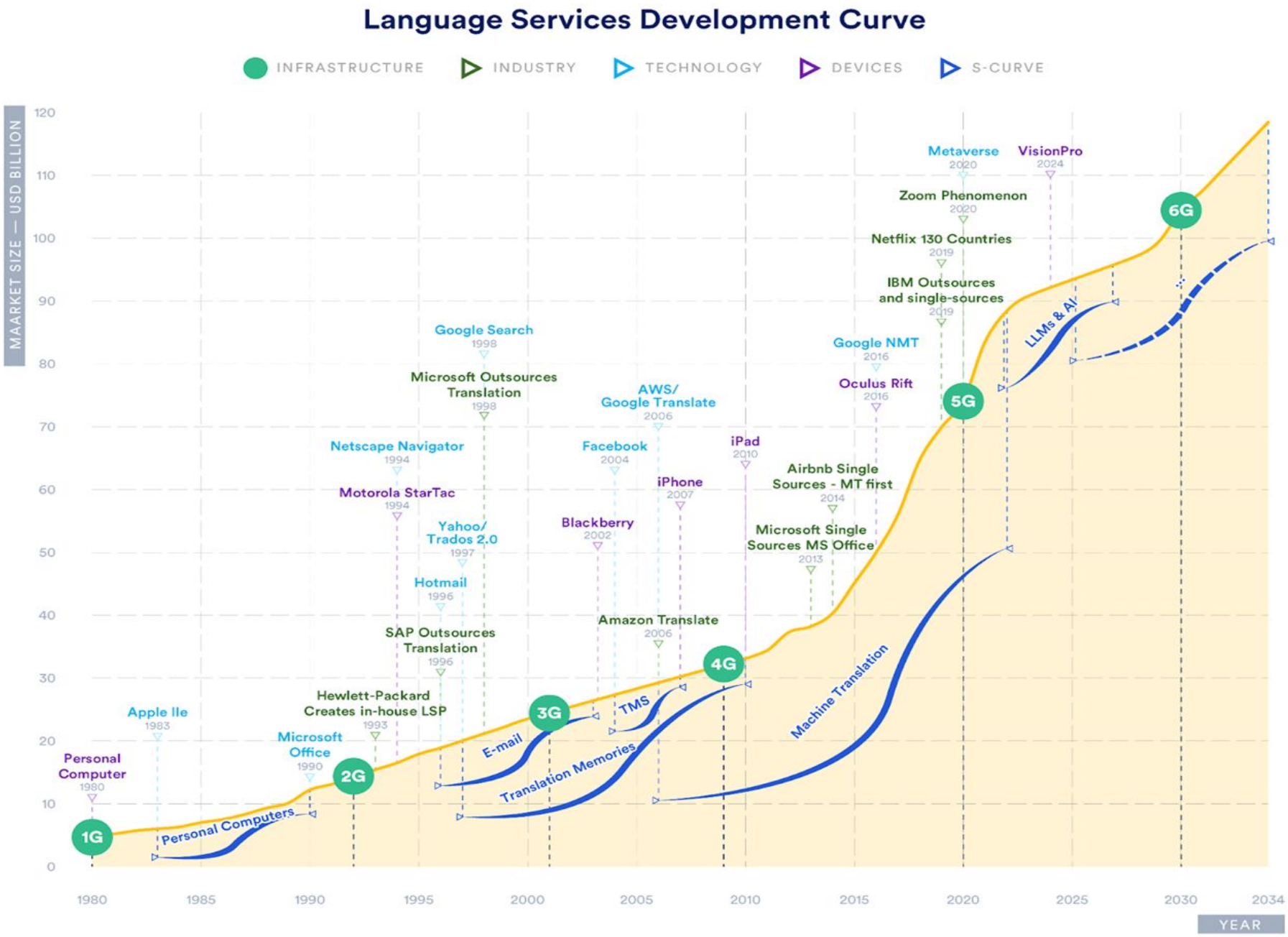
Renato Beninatto

Co-Founder, Advisor

Nimdzi Insights

LANGUAGE TECHNOLOGY LANDSCAPE CONFERENCE

Technology drives growth



LANGUAGE
TECHNOLOGY
LANDSCAPE
CONFERENCE

We estimate that the potential impact of language technologies on intra-EU trade is up to EUR 360 billion.

LANGUAGE TECHNOLOGY LANDSCAPE CONFERENCE

CHARTING THE GLOBAL LANDSCAPE OF LANGUAGE TECHNOLOGY

June 18, 2024

A EUROPEAN COMMISSION INITIATIVE

BY NIMDZI INSIGHTS

The Landscape of Language Technology

Laszlo K. Varga, Nimdzi Insights

LANGUAGE TECHNOLOGY LANDSCAPE CONFERENCE

Introduction

This presentation is based on the interim Market Study for LTs and adjacent market research projects.

This presentation has been created to disseminate the finding of first year interim delivery for CNECT/LUX/2022/OP/0030 - LANGUAGE TECHNOLOGY SOLUTIONS, LOT 3.

The Market Study on Language Technologies aims at mapping the current language technology (LT) landscape as well as providing an overview of the most relevant market trends.

The Study puts special emphasis on the European market in the context of the global LT market in order to understand how influential and competitive Europe is in this field.

The overall goal of the study is to raise awareness about LT and the potential it holds, and subsequently increase the use of LT in Europe and globally.

LANGUAGE TECHNOLOGY LANDSCAPE CONFERENCE

Untangling the web - executive summary 1

LTs are ubiquitous and beneficial.

Language technologies (LT) have multiple vital roles in the lives of individuals, communities, societies, small businesses, large enterprises, and public organisations, as well as in the local, regional, and global economy.

- **Enable multilingual communication and international trade.**
- **Enhance productivity**, including information retrieval, question answering, summarisation, content drafting, creation and editing, software code drafting and creation, sentiment and intent analysis, and customer support.
- Ideally **help protect linguistic diversity** fostering language equality, multiculturalism and inclusion.

LANGUAGE TECHNOLOGY LANDSCAPE CONFERENCE

Untangling the web - executive summary 2

(Language) data is the new oil.

Modern machine learning techniques, including for language technologies, require vast amounts of data Collecting and curating data for LTs is no easy feat even for the largest US tech labs. After breaking a few glasses and facing legal action, language AI labs are turning towards more **legitimate data collection methods and data acquisition deals**.

In Europe, data sovereignty and governance is taken to the highest level, with the Language Data Space and various other EU- and member-level initiatives.

Generic language data provides generic results even with the best architecture LTs.

Data quality is a most often quoted aspect in LT development. The higher quality the data, the more efficient – smaller, cheaper, faster to train and run – the LT models can be.

Additional effort into language data yields superior results and thereby greater impact of language technologies.

LANGUAGE TECHNOLOGY LANDSCAPE CONFERENCE

Untangling the web - executive summary 3

About sovereignty.

Sovereignty is critical on corporate, country, and regional levels alike. There are stark differences between how major economic powers handle language technology and data.

- **The EU** strives for a single (digital) market, protects its sovereignty and aims to lead with safe, regulated AI and data governance. Fragmented languages, policies, and support initiatives hinder.
- **The US** is more language-unified, free-market principled, heavily relies on well-developed big-tech cloud companies' ability to innovate, and is fastest on technology adoption and dissemination.
- **China** goes its own way, supporting and controlling LT development even at the highest level of policy making.

Being in control of the underlying core infrastructures, technology architectures, and, very importantly, data are critical in achieving and maintaining this autonomy.

LANGUAGE TECHNOLOGY LANDSCAPE CONFERENCE

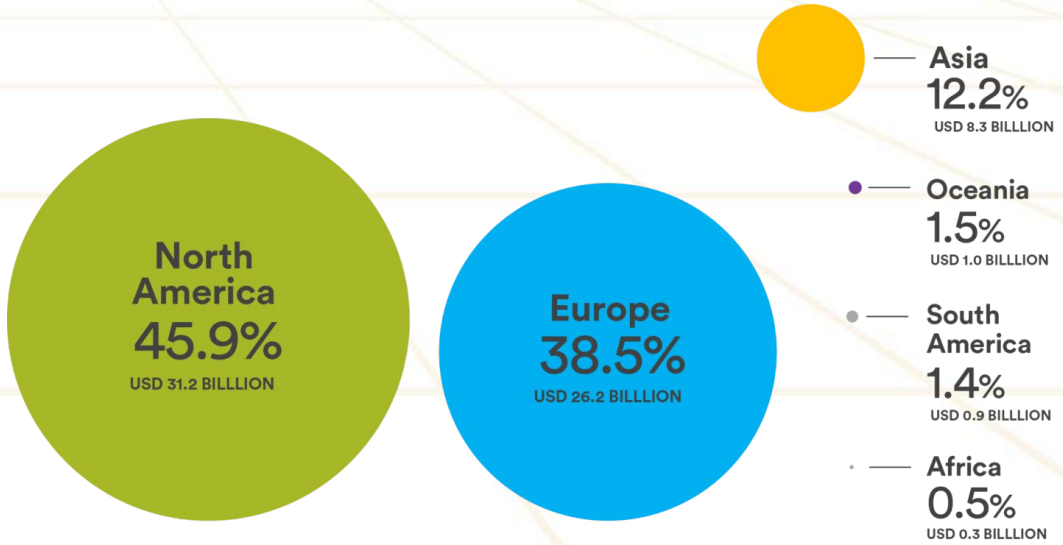
Untangling the web - executive summary 4a

LTs impact trade.

According to a 2023 Nimdzi research, **7 out of 10 users will always select their native language over English in ecommerce**. Businesses extensively use technology-enabled localisation to reach international markets.

Localisation enables international trade with LTs such as machine translation at the forefront both as efficiency tools and as enablers, alternatives to "no translation == no trade".

The European examples of Booking.com, Zalando, IKEA, or SAP all heavily rely on language technologies to reach their customers.



US enterprise demand represents about 46% of the language market, while European companies constitute about 39% of language services and technology spends.
(Source: the Nimdzi 2024 100)

LTs as costs in localisation are insignificant when compared to the doors they help open for opportunities with new customers in different countries.

LANGUAGE TECHNOLOGY LANDSCAPE CONFERENCE

Untangling the web - executive summary 4b

LTs impact intra-EU trade.

99% of EU's companies are SMEs, adding 53% of value created; for them, localisation cost is a high barrier to trade.

Language technologies as enablers can boost EU SME's participation in cross-border trade, which has a direct economic impact on the EU.

Based on previous studies, LTs such as machine translation have the potential to add up to EUR 300 billion value to intra-EU trade.

In addition, LTs open the doors to new businesses that previously did not consider cross-border trade, by EUR 60 billion.

We estimate that the potential impact of language technologies on intra-EU trade can be in the order of EUR 360 billion,
via reducing or removing the language barrier.

LTs as costs in localisation are insignificant when compared to the doors they help open for opportunities with new customers in different countries.

LANGUAGE TECHNOLOGY LANDSCAPE CONFERENCE

Untangling the web - executive summary 5

LTs impact productivity.

GenAI – according to McKinsey – might have an EUR 5.6-7.2 trillion global economic impact.

This indicates that, at the current (and rapidly evolving) level of LLMs,
the EU has a EUR 700 billion – 1.4 trillion economic opportunity with large language models
including newly added value and labour productivity gains.

The big **underlying assumption is that LLMs support all languages at the same level** for productivity and other economic gains – which, unfortunately, does not reflect reality, but is skewed towards English-speaking use-cases and businesses.

Nevertheless, even a 10% productivity gain (such as in program coding or knowledge-based work) is a major competitive advantage locally, globally, on business, national, and regional level.

Focus on language equality, data, and accessible language data markets are crucial in the larger economic landscape.

LANGUAGE TECHNOLOGY LANDSCAPE CONFERENCE

Untangling the web - executive summary 6

LT is (US) big-tech dominated.

The language technology market is centralised and dominated by big-tech servicing large enterprise and generic demand.

This is followed by a **long tail of specialised language solution providers**, providing for locale, industry vertical, and function-specific needs. They often rely on big-tech's ability to drive the expensive core innovations.

Market dynamics currently indicate that the free market-driven approach of the US is prevailing in LT innovation and adoption.

However, **big-tech often obscure the data problem in training and usage**. With rising data and privacy concerns (also in the US, potentially influenced by the EU's AI Act) **there is a potential that the EU will get back into the race as it takes control of their data**, especially to meet the demand of the sensitive European markets.

LANGUAGE TECHNOLOGY LANDSCAPE CONFERENCE

Untangling the web - executive summary 7

Innovation is undirected.

Innovation can come from anywhere – academia, startups, big tech.

Protection of R&D results is a competitive edge more than ever, less disclosure from the top actors.

ChatGPT put LTs into the limelight, resulting in:

- A **new wave of LT investment** and startup scene;
- **Attention from executive managers' to LTs** as a a 'new' way of productivity gains, 'doing language work', and a new competitive factor.

B2C success is transferable.

Google Translate's or OpenAI ChatGPT's **success with consumers can be leveraged for B2B** also due to the level of visibility and exposure.

European LT actors such as DeepL or Translated followed this recipe to success, challenging the 'commodity' nature of machine translation. Others starting in the B2B space, such as Aleph Alpha, Mistral, or Silo, have a harder time.

Quality is a tough conversation to dislodge B2B incumbents, which – especially in times of technology paradigm shift – indicates an **early mover advantage for providers**.

LANGUAGE TECHNOLOGY LANDSCAPE CONFERENCE

Market size estimations

The global market for language technology is estimated

between EUR 19 and 33 billion in 2023,

to grow to EUR 29-52 billion in 2024 with a strong growth outlook.

This estimate from the interim research findings does not directly include the professional services attached to custom LT development, which could double the market size for the applicable LT categories.

In general, **the closer an LT is to an actual business use-case, the higher the share of professional services in their development and deployment.**

LANGUAGE TECHNOLOGY LANDSCAPE CONFERENCE

Market size estimations

Main LT category	Market size estimate (2023)	Growth potential (2024)	Investment interest	AI / ML adoption / disruption level	Technology maturity level
Generative AI - language models (LLM)	EUR 1-2 billion	Very high; EUR 5-10+ billion	Very high	Fundamental	Emerging foundational technology Fast growing with high impact on other LTs
Multilingual content generation (MLCG)	EUR 1-2 billion	Low / moderate	Moderate	Fundamental	Emerging / evolving
Machine translation (MT)	EUR 2-3 billion	Moderate	Low	Already adopted, LLM adoption ongoing	Stable
Translation management systems (TMS)	EUR 0.3-0.5 billion	Low	Moderate	Low	Stable
Quality assurance and review tools	<EUR 100 million	Low	Low	High	Evolving
Machine interpreting (MI)	<EUR 1 billion	High; EUR 1-2 billion	Moderate	Fundamental	Emerging
Speech-to-speech translation (S2ST)	<EUR 1 billion	High; EUR 1-2 billion	Moderate	Fundamental	Emerging / evolving
Speech-based technologies (ASR, STT)	EUR 3-5 billion	High, EUR 5-10 billion	Moderate	Fundamental	Evolving
Text-to-speech (TTS) incl. AI dubbing	EUR 2-5 billion	Very high; EUR 5-10 billion	High	Fundamental	Mature (core TTS), Evolving (AI dubbing)
Chatbots and virtual assistants	EUR 5-10 billion	Moderate	High	High	Maturing / evolving
Optical character recognition (OCR) incl. handwriting-to-text (HTT)	<EUR 1 billion	Low / none	Low	Ongoing, via VLMs and MLLMs	Stable
Braille technologies	<EUR 200 million	Low / none	Low	Low disruption	Stable
Sign language technologies	<EUR 100 million	Low / moderate	Low	Ongoing	Immature
Language learning apps	EUR 2 billion	Moderate	Low	Moderate, ongoing	Evolving

LANGUAGE TECHNOLOGY LANDSCAPE CONFERENCE

Key takeaways 1/8

Language technology is ubiquitous and ever-developing

- **LTs solve human communication challenges, but…**
  - The volume of text and speech processed by technologies is increasing exponentially. Humans alone can't process this flux.
  - This places an ever-increasing demand on both humans and machines, separately, augmented, or connected, in the language technology and services space.
- **LTs can help achieve new levels of productivity, but…**
  - The need for growth and the competitive nature of markets mean there is no such thing as "enough" gain.
- **LTs could create a foundation for language equality and equity, but…**
  - There is a strong asymmetry in the performance of language technologies across languages.

"Basic" language technologies are (almost) free - such as machine translation.

We predict that many other LTs will follow suit to become generally and freely available. (GPT-4o, anyone?)

LANGUAGE TECHNOLOGY LANDSCAPE CONFERENCE

Key takeaways 2/8

Why is there a language technology industry and market?

The quality argument:
Mission- and business-critical use cases need more than what generic solutions offer.
They are not upfront scalable and suitable for custom, specialised, confidential, top-quality, and – most importantly – revenue-, profit-, or brand image-impacting language tasks.

The niche argument:
While generic LT solutions cover generic use cases, businesses have specific needs in their domains, functions, and use cases. These can be catered to by 3rd party LT providers focusing on the specific niche that is not covered by big-tech.

But language is typically not the core function of most organisations.
Outsourcing the technology and related services is common practice.

LANGUAGE TECHNOLOGY LANDSCAPE CONFERENCE

Key takeaways 2/8

Why is there a language technology industry and market?

Businesses are always looking for new customers, international expansion, customer retention and user attention, revenue growth, and cost savings.

Language technologies can enable all of these, although – at their current level of development – don’t solve the overarching language barrier problem out-of-the-box.

Custom tool creation, domain adaptation, connectors and integrations, bespoke deployment, are all part of the value proposition of the language technology market.

The key driver of LT revenues (and market size) is reliable quality at scale.

In a B2B setting, 80% or 95% doesn’t get you there.

And there’s (almost) always an element of human-in-the-loop service.

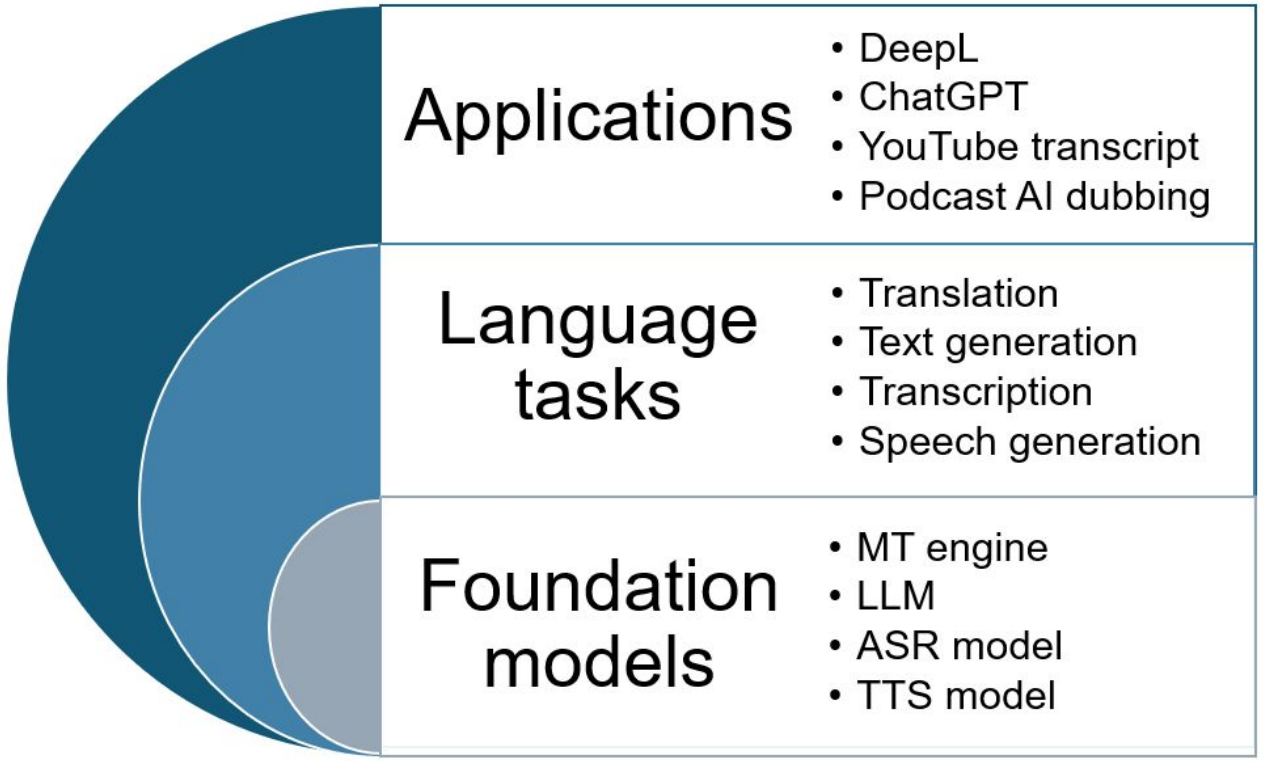
LANGUAGE TECHNOLOGY LANDSCAPE CONFERENCE

Key takeaways 3/8

Language technologies are interconnected

Foundation models are core and expensive.
Only a LT few applications outside of big-tech companies use proprietary foundation models.
Possibly even fewer in the future, as models get more complex.

LT applications and products are the most visible on the market.
These drive brand recognition, customer engagement, and market share.
We see and expect an increase in the number of LT applications and products.
As long as foundation model actors don't cover their niches and domains.



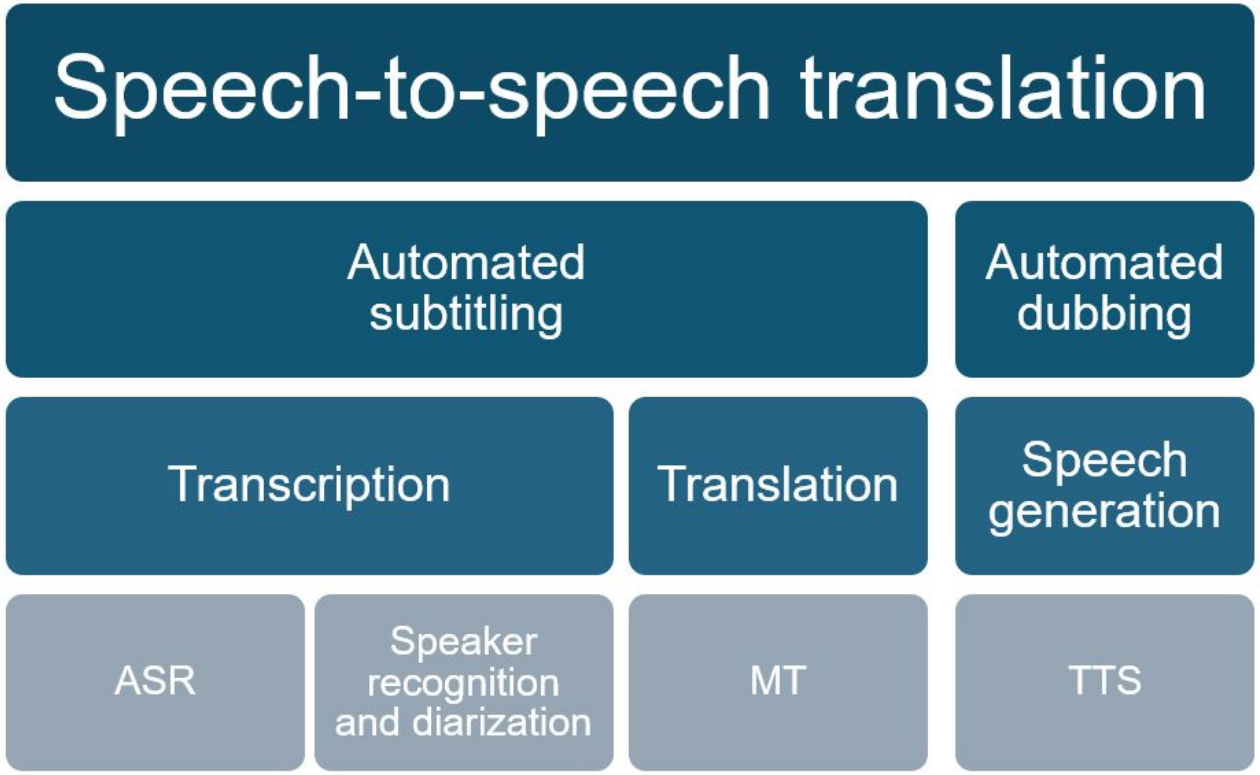
LANGUAGE TECHNOLOGY LANDSCAPE CONFERENCE

Key takeaways 3/8

Language technologies are interconnected

Many LTs are also building blocks into compound (language) technologies.

The iPhone effect.
The most attractive applications of language technology don't need to detail what combination of LTs are used and how, as long as they perform the actions and produce the outputs needed from them.

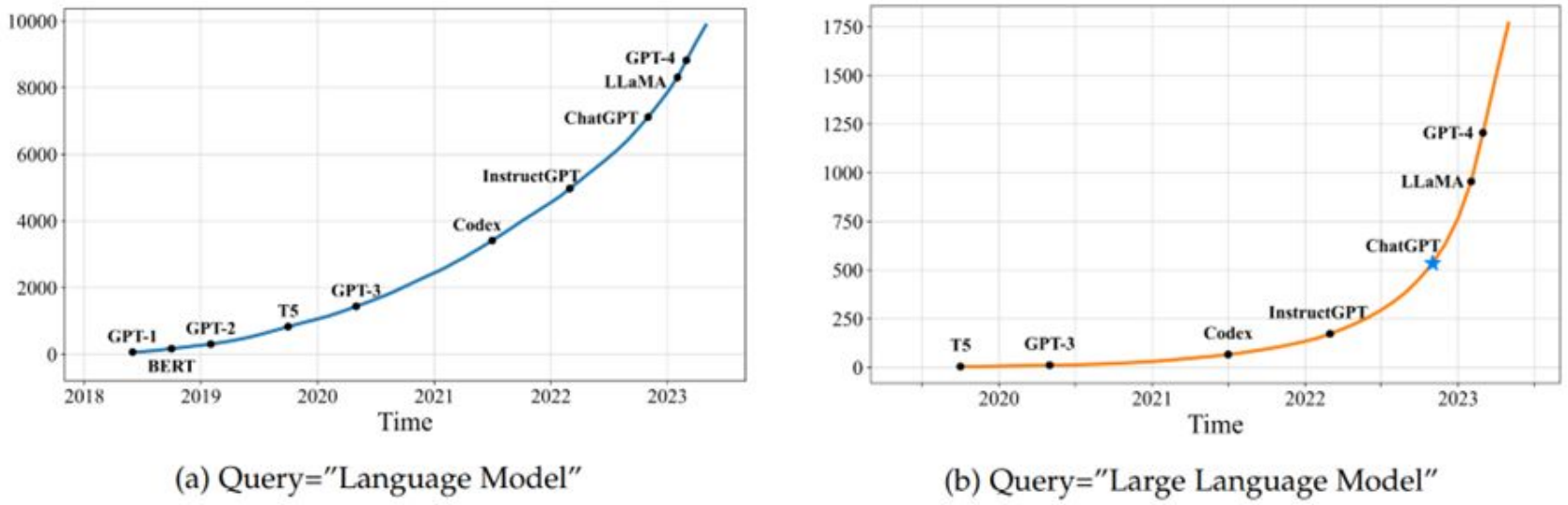


LANGUAGE TECHNOLOGY LANDSCAPE CONFERENCE

Key takeaways 4/8

AI is eating the world of (language) technology

Although not all LTs are AI-based, LLMs – as a form of foundational AI – took the world by storm, in user accumulation, business adoption, LT investment, product development, and also in research.



Number of research papers with LMs and LLMs show exponential growth

Source: A Survey of Large Language Models, Zhao et al, November 2023 https://arxiv.org/pdf/2303.18223

The incredibly rapid development in and the hype around LLMs imply that **all previous technology trends and predictions are annulled.**

LANGUAGE TECHNOLOGY LANDSCAPE CONFERENCE

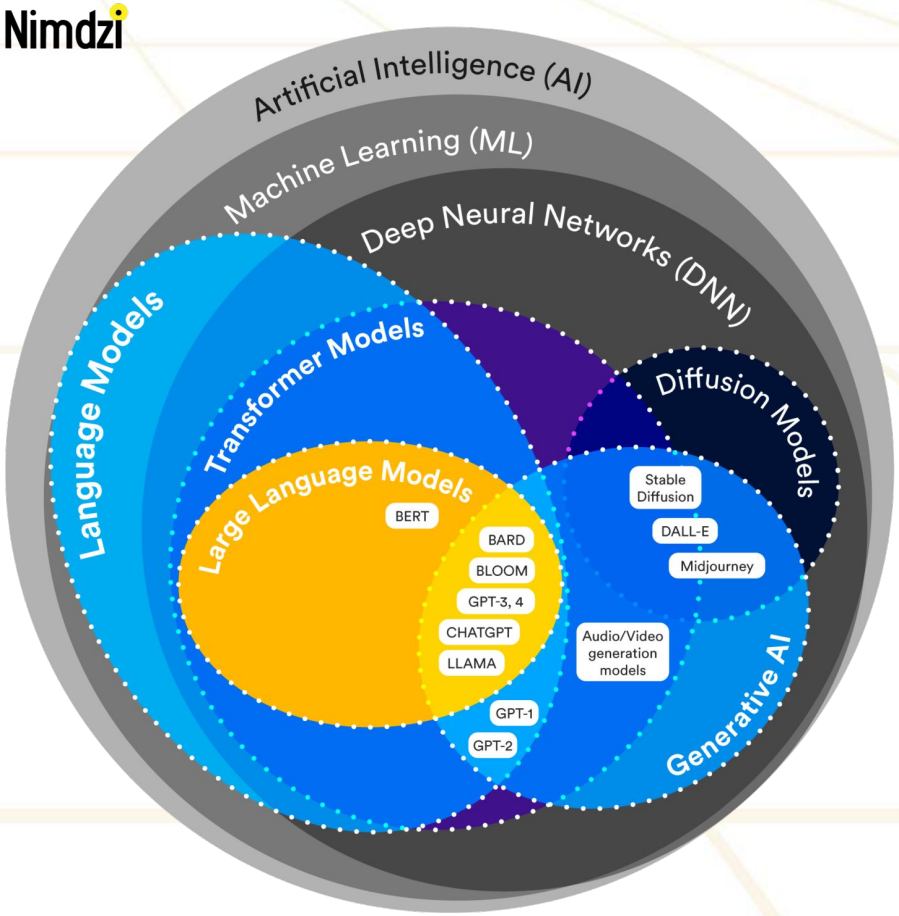
Key takeaways 4/8

AI is eating the world of (language) technology

More often than not, AI, generative AI (GenAI), and large language models (LLM) are used interchangeably.

This results in a confusion, especially in public discourse, but also in LT actor communications, investment announcements, and market estimations.

LLMs are a form of AI, but so are many other language technologies: neural machine translation, speech recognition, or speech synthesis are also AI.



Source: The Language AI Alphabet: Transformers, LLMs, Generative AI, and ChatGPT
https://www.nimdzi.com/the-language-ai-alphabet-transformers-llms-generative-ai-and-chatgpt/

LANGUAGE TECHNOLOGY LANDSCAPE CONFERENCE

Key takeaways 4/8

AI is eating the world of (language) technology

Value models of new technologies	EXAMPLE: LLMs and automated translation
Replace existing stack Better, faster, cheaper	LLM instead of traditional neural MT
Augment current technologies Enhance efficiencies	Automated post-editing
Create new value Automate what couldn't be done before	Source optimization Large-scale translation memory cleanups

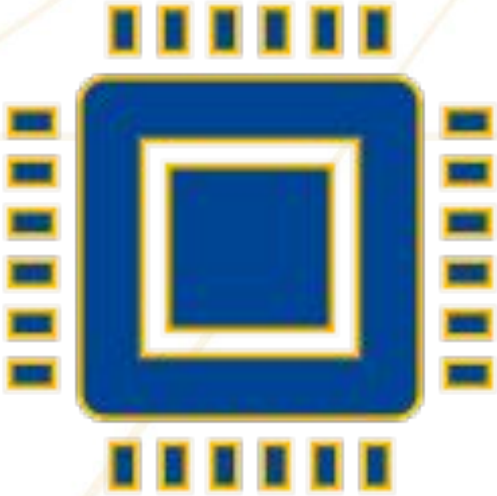
Harvesting wide-scale benefits of LLMs is more challenging than initially anticipated.
The limitations of LLMs are far from being understood, and so is the ROI.

LANGUAGE TECHNOLOGY LANDSCAPE CONFERENCE

Key takeaways 5/8

Data, data, data, talent, and compute power

Computational resources



Cloud big-tech companies (Microsoft, Google, Amazon, IBM, Nvidia, Baidu, etc) and large other SaaS and PaaS players such as Meta, Uber, Adobe Intel, Dell, HP, SAP, Databricks, and Snowflake, have the capacity to train, host, and deploy AI models including LLMs at scale for their own internal use and for their corporate clients. Even ‘independent’ actors (Hugging Face, Mistral, Anthropic, or Cohere) make significant business deals with cloud.

In Europe, OVH Cloud hosts limited models and develops none. EuroHPC leads the way.

In the meantime, AI hardware (GPU/TPU) stocks and investments soar, with new startups aiming to disrupt.

Compute is scarce, almost monopolized, and expensive.

LANGUAGE TECHNOLOGY LANDSCAPE CONFERENCE

Key takeaways 5/8

Data, data, data, talent, and compute power

Talent

LT actors compete for the same talent (machine learning, data science and engineering) as other segments of ML and data technologies, such as analytics and business intelligence, cybersecurity, drug discovery, and computer vision.

Demand for “AI talent” is probably the highest ever, and growing as adoption widens. Investment trust goes to big-tech alumni.

Europe is subject to brain drain by the US, even locally. OpenAI, Meta, MS, Google all have AI labs in Europe.

China is trying with acquisitions to grab a global footprint, and MEA is on the hiring map too.

Talent (and talent cost) is major differentiator.

LANGUAGE TECHNOLOGY LANDSCAPE CONFERENCE

Key takeaways 5/8

Data, data, data, talent, and compute power

Data, data, data

The quality and amount of data critical to successful LT development, but there is no uniformity of data availability across languages and modalities. “High” and “low” resource languages is a serious divide, and so is high and low business opportunity.

US labs follow “fair use” and obscurity, as data and training recipes are core differentiators. While multiple copyright lawsuits and off-court agreements are ongoing, it’s hard to sell tech in the US without AI in the mix.

The EU aims to lead with quality over quantity and the establishment of a single market for data. Data sovereignty is seen as critical by the more conservative and (rightly) cautious European enterprises.

Data quality and availability make a big difference.

LANGUAGE TECHNOLOGY LANDSCAPE CONFERENCE

Key takeaways 6/8

Big tech’s big role in the LT market

Free consumer apps drive low-cost expectations.

Big-tech / APIs make LTs a commodity.

Enable ecosystems of startups.

Platformisation, lock-in, and the paradox of choice.

LANGUAGE TECHNOLOGY LANDSCAPE CONFERENCE

Key takeaways 7/8

Why big tech is in the language technology field.

Big tech companies have the capabilities.

Funding, data, compute, and talent.
R&D teams are result oriented.
The need for innovation signaling.

Localisation efficiencies.

Millions of users, billions of words, hundreds of millions of **dollars'** of language services in ecommerce, office suites, social media.
Lots to streamline.

Technology platform lock-in strategy.

LT's "complete" the platform offering of cloud, machine learning, application deployment.
No 3rd party required.

Protective attitude towards the core products and services.

Enhance experience while eliminating dependencies and building moats.
YouTube is transcribed. Amazon products are translated. Facebook has chatbots. MS Office is stacked with LTs.

LANGUAGE TECHNOLOGY LANDSCAPE CONFERENCE

Key takeaways 8/8

There is a market for LTs beyond big-tech's offerings

In the language technology space, we observe multiple value creation streams:

- **Direct sales of LTs via licences**, such as per seat or per team licensing (typical of SaaS solutions).
- **Technology output revenues**, volume-based per-word or per-character billing.
- **Tech-adjacent professional services**, such as data collection / annotation, custom model training, deployment.
- **Connected human services**, such as MT post-editing or transcription correction.

Ultimately, LTs don't directly add value unless they actually solve a problem. B2B or B2G implementations are almost always accompanied by custom services.
Mid- and large enterprises require custom, tailor-made solutions rather than pieces of technology. This is where the real money, the meat of the LT industry, is.

Democratisation of technology means that tech-savvy enterprises can implement their own language technologies, competing with LT actors.
Language data expertise is often the core missing ingredient.

LANGUAGE
TECHNOLOGY
LANDSCAPE
CONFERENCE

Q&A

Thank you and see you after the lunch break.

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

Feedback survey

SCAN ME

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

Thank you for attending!

And see you next year.

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

Feedback survey

SCAN ME