# ELE Strategic Research, Innovation and Implementation Agenda – Revisited

Georg Rehm (DFKI, Germany)
georg.rehm@dfki.de

27-06-2023 META-FORUM 2023 – Digital Language Equality for a Multilingual Europe
http://european-language-equality.eu

# ELE1: Recommendations towards DLE by 2030 – Scope

- **Strategic agenda and roadmap towards Digital Language Equality in Europe by 2030**

- Informed by 66 project reports (2400+ pages with condensed findings)

- 92 languages taken into account

- The plan represents voices from research, industry and civil society

1. Research:    30+ reports on situation of individual languages; various surveys; 60+ interviews

2. Industry:    SME partners produced four technical deep dives; various surveys

   approx. 500-750 people contributed

3. Civil society:  European citizen survey with approx. 20,000 responses

# SRIA: Version History

**The Strategic Research, Innovation and Implementation Agenda for achieving full digital language equality in Europe by 2030**

- Version 0.9     Available on the ELE website (early Nov. 2022)
- Version 1.0     Available on the ELE website (late Nov. 2022)
- Version 1.5     The ELE book, especially Chapter 45 (March 2023)
- Version 2.0     Version 1.5 plus *supplementary document* available on the ELE website (June 2023)

ELE2 brought to light several new suggestions and revisions (including insights and ideas from the FSTP projects and the endorsement form).
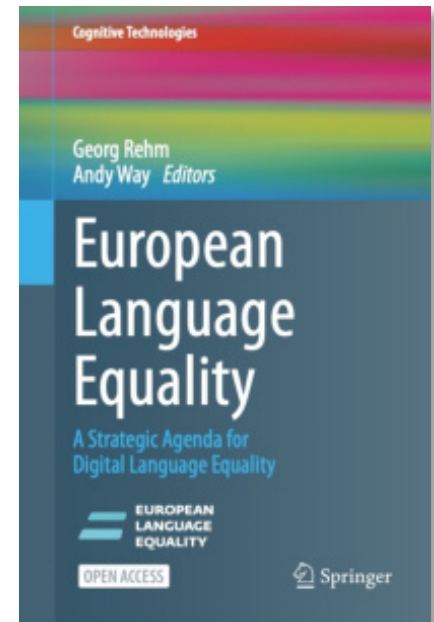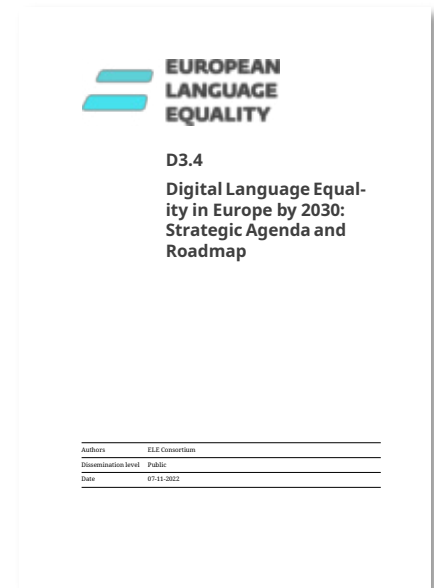
# Executive Summary

- META-NET study (2012): At least 21 European languages in danger of digital extinction (even though, politically, all our languages are supposed to be equal)

- EP Resolution (2018, "Jill Evans report"): Europe needs to take action – EU should set up a large-scale, long-term funding program to boost the area of LT, NLP, language-centric AI

- Current ELE results (2022/2023): There's still an extremely strong imbalance in terms of technology support of Europe's languages – *all* languages except English under-resourced

- Language-related AI applications (ChatGPT etc.) are a multi-billion Dollar/Euro business

- EU is currently concentrating on data spaces and data economy but mostly ignoring our languages and NLP/LT/AI research in this area (there is a little funding but not enough)

- We recommend setting up a large-scale, long-term funding programme to digitally safeguard our languages and to put Europe back in the pole position – the scientific field and also its exploitation is currently dominated by the US and Asia (Europe: second fiddle)

# Current Situation: Support for Data and AI from the EU

Data

AI

# Largely missing: Support for Language & Language Technology
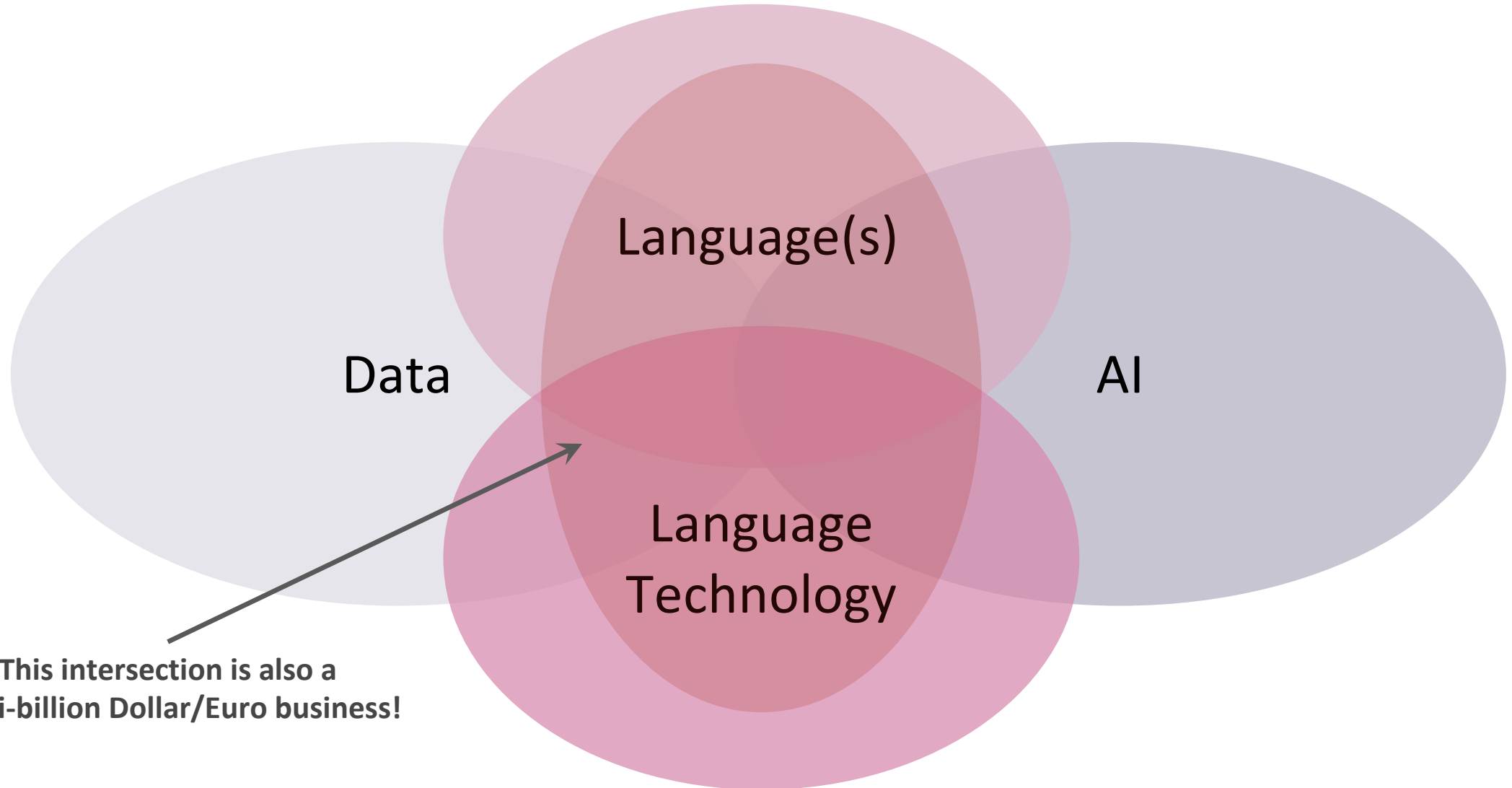
Language(s)

Data

AI
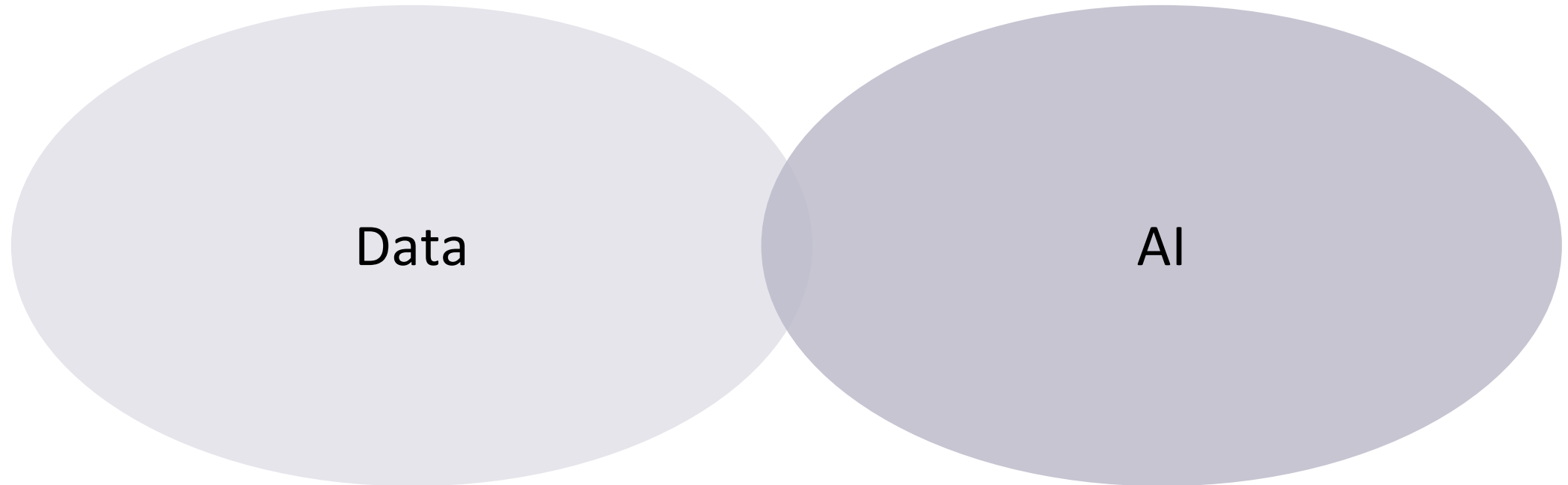
Language
Technology
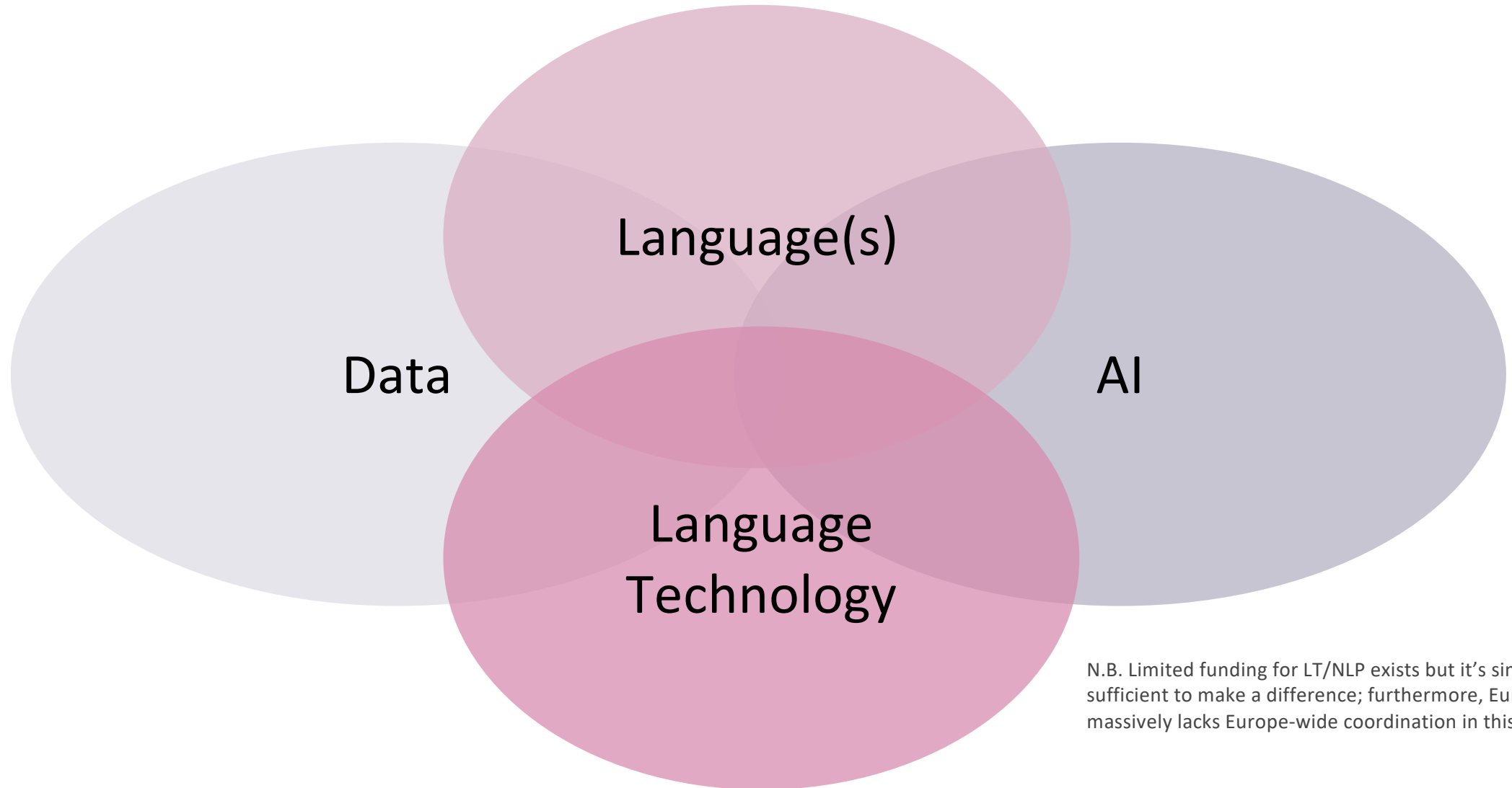
N.B. Limited funding for LT/NLP exists but it's simply not sufficient to make a difference; furthermore, Europe massively lacks Europe-wide coordination in this area.

# Largely missing: Support for Language & Language Technology

Language(s)

Data

AI

Language Technology
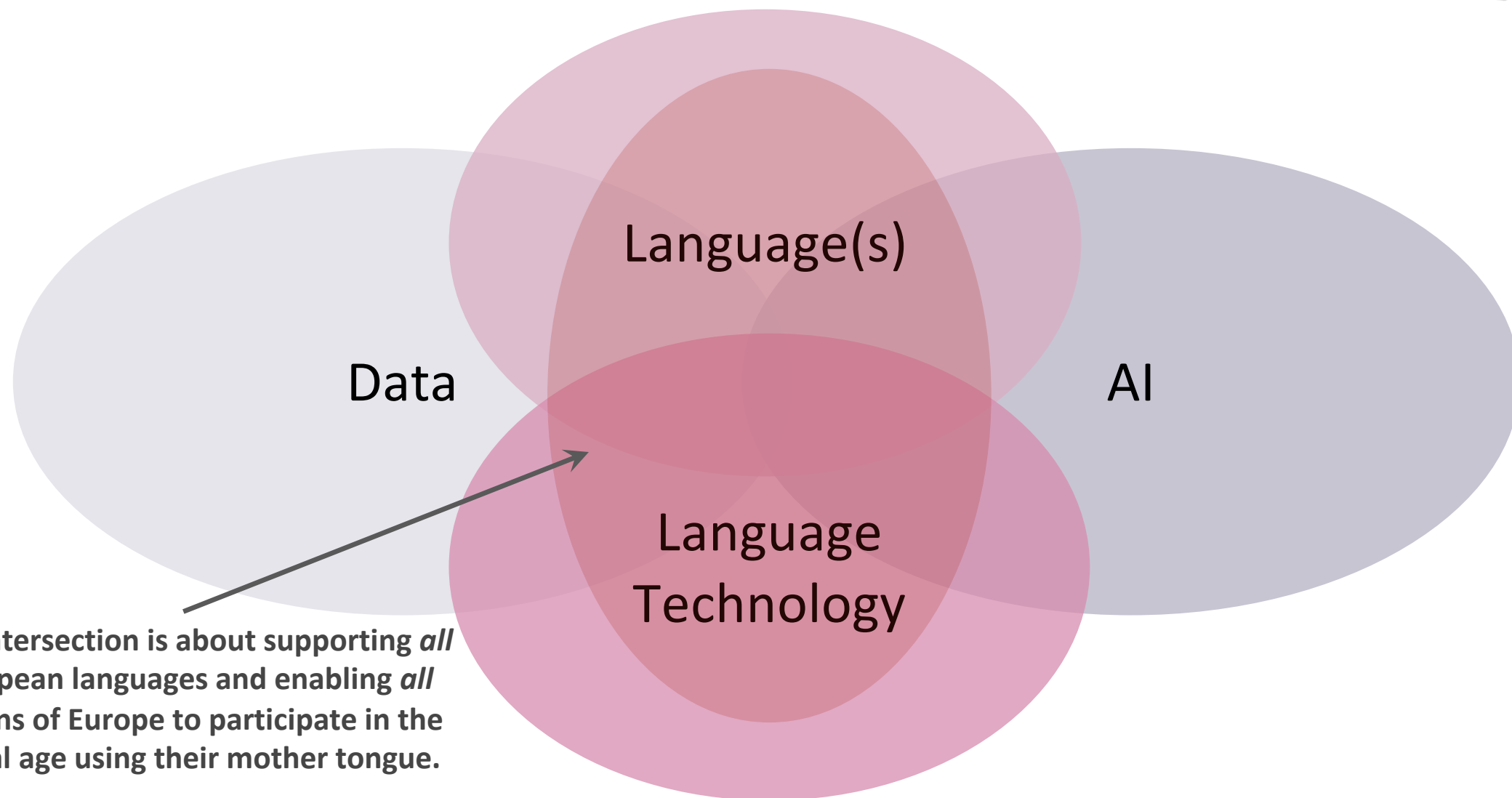
This intersection is about supporting *all* European languages and enabling *all* citizens of Europe to participate in the digital age using their mother tongue.
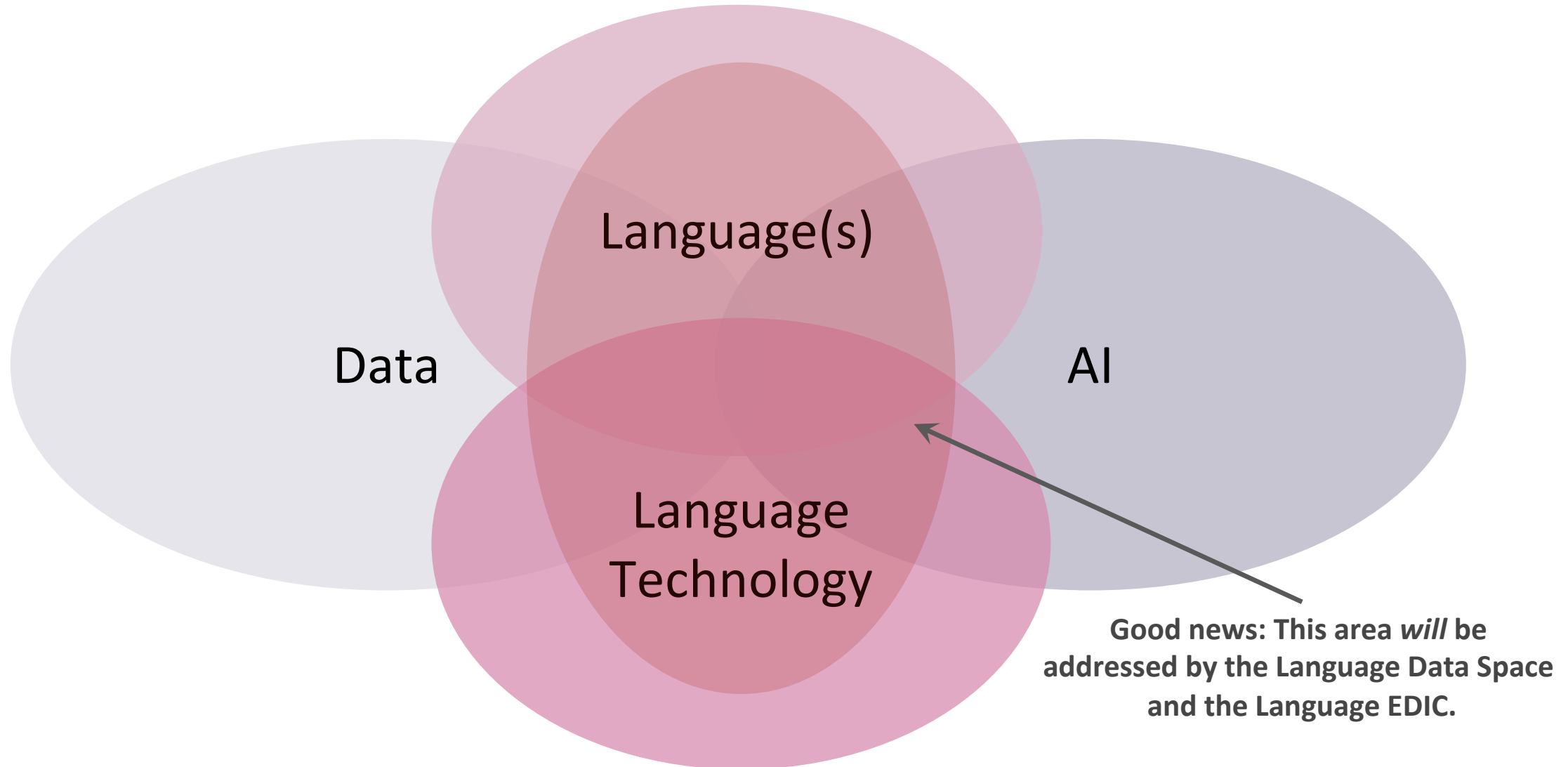
# Largely missing: Support for Language & Language Technology

Language(s)

Data

AI

Language Technology

**This intersection is also a multi-billion Dollar/Euro business!**

# Largely missing: Support for Language & Language Technology

Language(s)

Data

AI

Language Technology

**Good news: This area *will* be addressed by the Language Data Space and the Language EDIC.**

# State of Play in 2023

# "Language Equality" EP Resolution (2018)

European Parliament Resolution "Language equality in the digital age" P8_TA(2018)0332 – Voting (11 Sept. 2018): **592 yes** – 45 no

Recommendation: Establish a large-scale, long-term coordinated funding programme for research, development and innovation in the field of language technologies, at European, national and regional levels, tailored specifically to Europe's needs and demands

**Council of the European Union – Education, Youth, Culture and Sport Council, 4-5 April 2022:**

The conclusions call for the development of an ambitious digital policy for the development of language technologies [...].
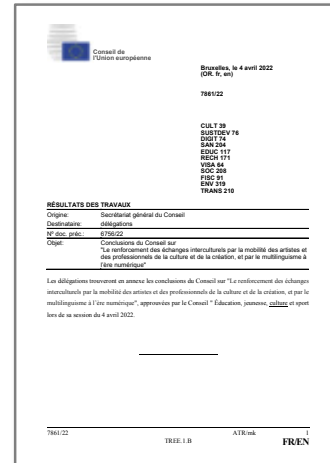
European Parliament
2014-2019

**TEXTS ADOPTED**
*Provisional edition*

**P8_TA-PROV(2018)0332**

**Language equality in the digital age**

**European Parliament resolution of 11 September 2018 on language equality in the digital age (2018/2028(INI))**

*The European Parliament,*

– having regard to Articles 2 and 3(3) of the Treaty on the Functioning of the European Union (TFEU),

– having regard to Articles 21(1) and 22 of the Charter of Fundamental Rights of the European Union,

– having regard to the 2003 UNESCO Convention for the Safeguarding of the Intangible Cultural Heritage,

– having regard to Directive 2003/98/EC of the European Parliament and of the Council of 17 November 2003 on the re-use of public sector information[1],

– having regard to Directive 2013/37/EU of the European Parliament and of the Council of 26 June 2013 amending Directive 2003/98/EC on the re-use of public sector information[2],

– having regard to Decision (EU) 2015/2240 of the European Parliament and of the Council of 25 November 2015 establishing a programme on interoperability solutions and common frameworks for European public administrations, businesses and citizens (ISA2 programme) as a means for modernising the public sector[3],

– having regard to the Council resolution of 21 November 2008 on a European strategy for multilingualism (2008/C 320/01)[4],

– having regard to the Council decision of 3 December 2013 establishing the specific programme implementing Horizon 2020 – the Framework Programme for Research and

[1] OJ L 345, 31.12.2003, p. 90.
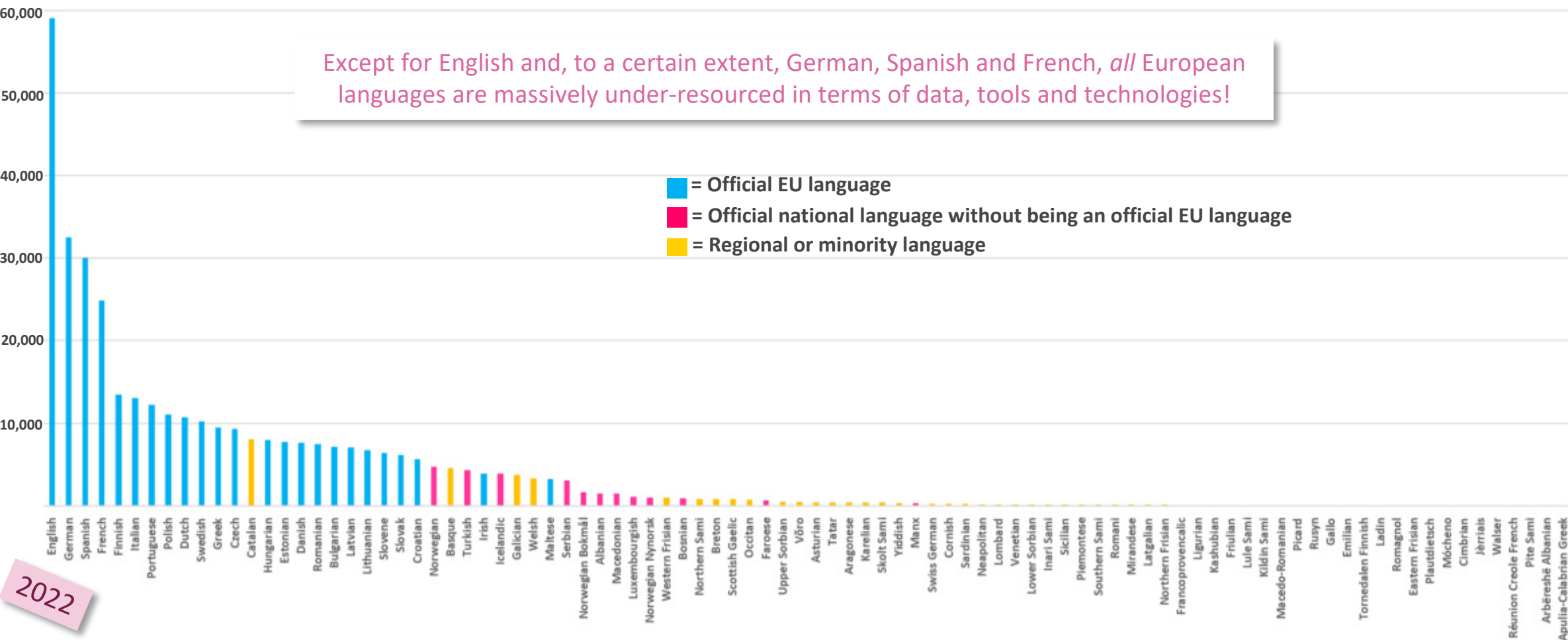[2] OJ L 175, 27.6.2013, p. 1.
[3] OJ L 318, 4.12.2015, p. 1.
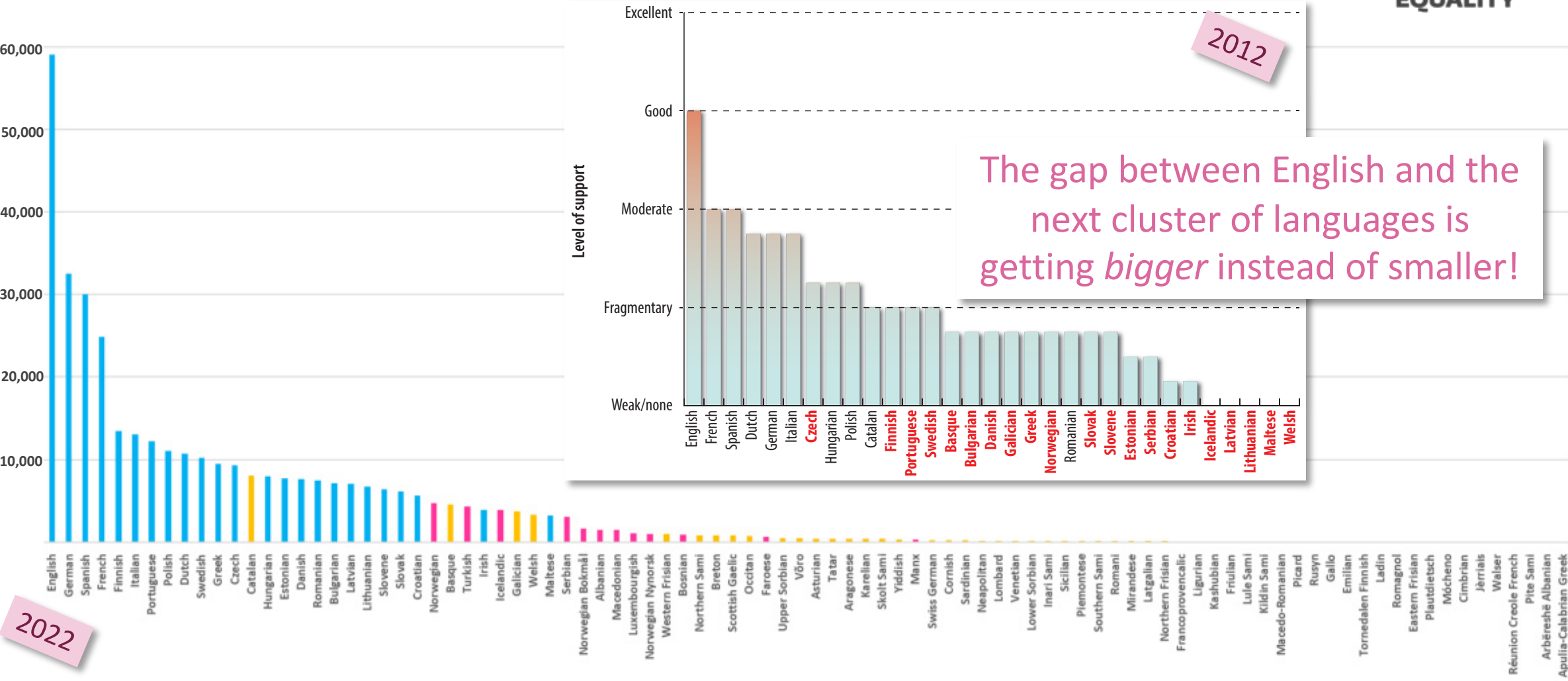[4] OJ C 320, 16.12.2008, p. 1.

# DLE Metric: Technological Scores



Except for English and, to a certain extent, German, Spanish and French, *all* European languages are massively under-resourced in terms of data, tools and technologies!

- = Official EU language
- = Official national language without being an official EU language
- = Regional or minority language

2022

# DLE Metric: Technological Scores



The gap between English and the next cluster of languages is getting *bigger* instead of smaller!

2012

2022

# DLE Metric: Contextual Scores



Nearly all European languages exist in socio-political ecosystems that do _not_ incentivise, encourage or foster the development of technologies for these languages.

■ = Official EU language
■ = Official national languages without being an official EU language
■ = Regional and minotrity languages
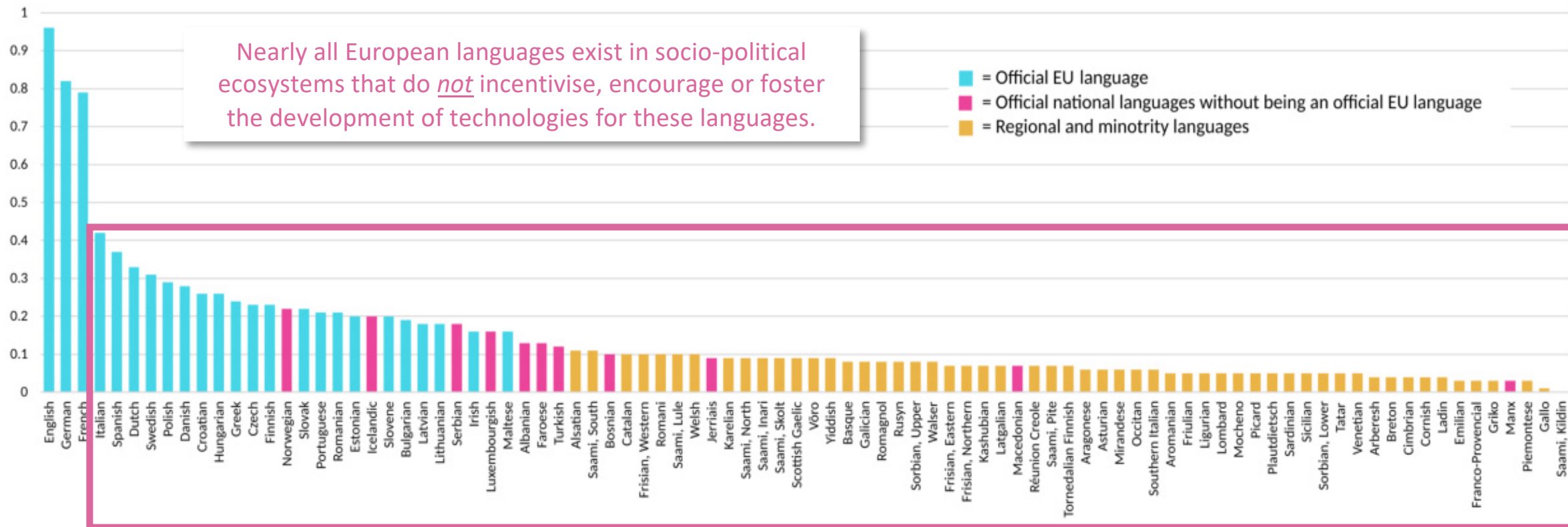
# Recommendations

# Strategic Agenda (SRIA): Outline

**Strategic Research, Innovation and Implementation Agenda and Roadmap for Achieving full Digital Language Equality in Europe by 2030 (Version 1.0)**

1. Multilingual Europe and Digital Technologies

2. Trends and Mega-Trends in Digital Technologies

3. Language Technology and Language-Centric Artificial Intelligence

4. Language Technology and Digital Language Equality in 2022

5. Language Technology and Digital Language Equality in 2030: The ELE Technology Vision and Priority Research Themes

6. A Shared European Programme for Language Technology and Digital Language Equality in Europe by 2030: Recommendations

7. Roadmap towards Digital Language Equality in Europe by 2030
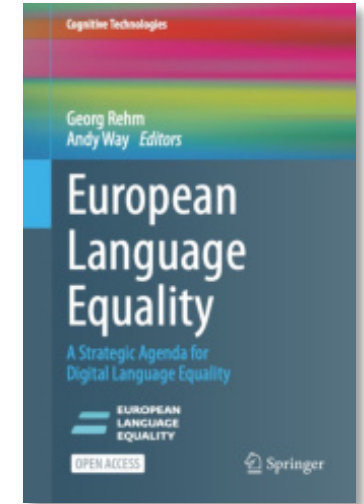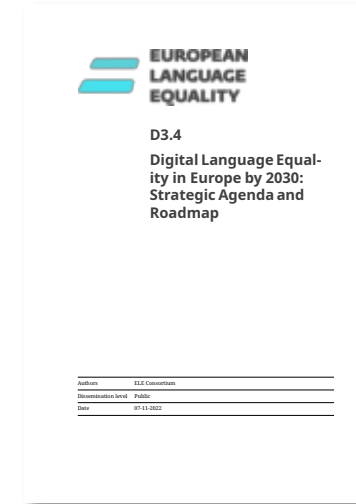
1. Overview and Main Concept

2. Research Recommendations

3. Technology and Data Recommendations

4. Infrastructure Recommendations

5. Policy Recommendations

6. Governance Model

7. Implementation Recommendations

EUROPEAN LANGUAGE EQUALITY

D3.4
Digital Language Equality in Europe by 2030: Strategic Agenda and Roadmap

| Authors | ELE Consortium | |
| Dissemination level | Public | |
| Date | 07-11-2022 | |

Cognitive Technologies

Georg Rehm
Andy Way *Editors*

European Language Equality

A Strategic Agenda for Digital Language Equality

EUROPEAN LANGUAGE EQUALITY

OPEN ACCESS

Springer

# SRIA: Policy, Governance, Technology, Data, Infrastructure

- **ELE Programme –** shared, long-term, coordinated, collaborative LT funding programme:

  - *Tailored to Europe's needs, demands and values – EU's multilingualism and language equality*

  - **EU:** Coordination, shared infrastructures, scientific goals, ELE principles etc.; flex funds

  - **Participating countries:** funding for the development of technologies for their languages

- Reduce the technology gap between English and all other European languages

- Address the lack of available data (relevant for *all* European languages except English)

- Allocate multilingualism and Language Technology to the portfolio of a Commissioner (again)

- Focus upon open source, open access, open standards, interoperability; foster standardisation

- Strengthen existing and emerging infrastructures, platforms and data spaces

- Ensure access to suitable HPC infrastructure

# SRIA: Research Recommendations

- **Strengthen and refocus European LT/NLP research through the ELE Programme.**

  - Scientific goal: *Deep Natural Language Understanding by 2030*.

- **Create large open access language models for all European languages**: *datasets*, *multilingual models*, models that include **symbolic knowledge**, **discourse features**, etc.

- **Machine Translation:** Better *context modeling:* how do *visual features* improve MT? *MT quality:* more focus upon research that looks into the root causes of MT performance issues.

- **Speech:** Development of *natural synthetic voices;* ASR, MT and text-to-speech should focus on a wide variety of languages, language variants and dialects, to keep those alive.

- **Text Understanding:** Research on *DNLU which integrates speech, NLP and contextual information* and *additional modalities*; support basic research in *neurosymbolic approaches*, incl. *grounding*.

# SRIA: Implementation Recommendations

- **ELE Programme:** 9 years (3 + 3 + 3 years) – approx. 2024-2032
  - EU/EC and participating countries to discuss needs & goals and to define financial setup
- **Shared programme:**

| Participating Countries | European Union |
|---|---|
| Invest into the development of resources and technologies for their own languages | Invest into pan-European coordination of all language-specific projects and initiatives, support mechanisms, infrastructures, data procedures, cross-cutting projects etc. |
| EU should establish binding legislation to encourage or ensure participation | |
| National projects: development of LLMs, data sets, technologies, tools | Provide flexible funds for bootstrapping poorly supported languages |

# SRIA: Implementation Recommendations – Setup & Themes

- **Coordination of ELE Programme:** Maintenance of ELE Principles, quality assurance approaches, shared tasks etc.

- **Language Modelling theme – coordination:** Training of multilingual and multimodal LLMs using text, speech, image, video etc., experiments, new approaches, shared tasks etc.

- **Data and Knowledge theme – coordination:** Collection, production, annotation, curation, quality assessment, standardisation etc. with regard to text data, spoken data, multimodal data, other modalities; inclusion of knowledge

- **Language Modelling theme – research:** Combine national projects and data sets with international consortia

- **Language Modelling theme – innovation and deployment:** Apply LLMs in industrial sectors and use cases
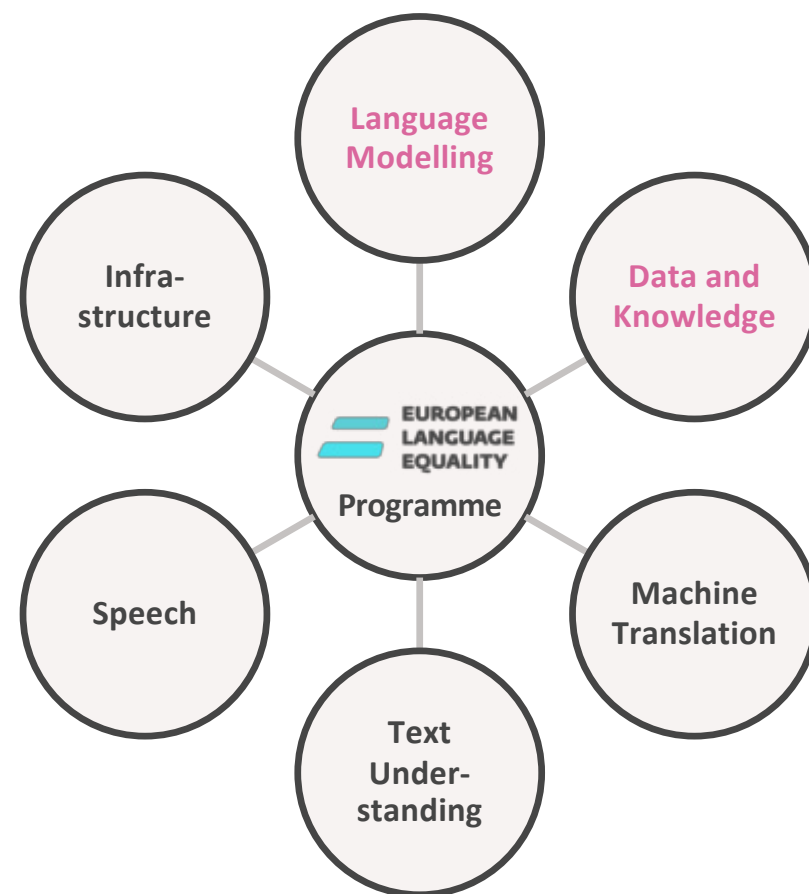
# ELE Programme – Budget Breakdown (EU)

| | Project Type and Number | | Phase 1 | | | Phase 2 | | | | Phase 3 | | Budget | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Type | Number | 2024 | 2025 | 2026 | 2027 | 2028 | 2029 | 2030 | 2031 | 2032 | Each | Sum |
| ELE Programme – overall coordination | CSA | 3 | | | | | | | | | | 20M€ | 60M€ |
| Theme Data and Knowledge – coordination | CSA | 3 | | | | | | | | | | 15M€ | 45M€ |
| Theme Language Modelling – coordination | CSA | 3 | | | | | | | | | | 15M€ | 45M€ |
| Theme Language Modelling – research | RIA | 15 | | | | | | | | | | 5M€ | 75M€ |
| Theme Language Modelling – innovation and deployment | IA | 15 | | | | | | | | | | 5M€ | 75M€ |
| Theme Machine Translation – research | RIA | 12 | | | | | | | | | | 5M€ | 60M€ |
| Theme Machine Translation – innovation and deployment | IA | 12 | | | | | | | | | | 5M€ | 60M€ |
| Theme Text Understanding – research | RIA | 12 | | | | | | | | | | 5M€ | 60M€ |
| Theme Text Understanding – innovation and deployment | IA | 12 | | | | | | | | | | 5M€ | 60M€ |
| Theme Speech – research | RIA | 12 | | | | | | | | | | 5M€ | 60M€ |
| Theme Speech – innovation and deployment | IA | 12 | | | | | | | | | | 5M€ | 60M€ |
| Theme Infrastructure – support | CSA | 3 | | | | | | | | | | 10M€ | 30M€ |
| | | | | | | | | | | | | | 690M€ |
| *Flexible funds for languages with fragmentary, weak or no technological support.* | | | | | | | | | | | | | 150M€ |

# ELE Programme – Budget Breakdown (Countries)

- Investments needed on the individual language level are *extremely difficult to predict*.

- These national/regional investments are – mostly – complementary to the EU/EC funding.

- We currently group the surveyed languages into three clusters:

  1. Languages with *weak or no* support:      40-50M€ each

  2. Languages with *fragmentary* support:      30-40M€ each

  3. Languages with *moderate* support:      20-30M€ each

- This funding is supposed to be provided by the respective country.

- However, the EU should help bootstrap the development of technologies for languages that are not doing well digitally (using flexible funds, see previous slide).
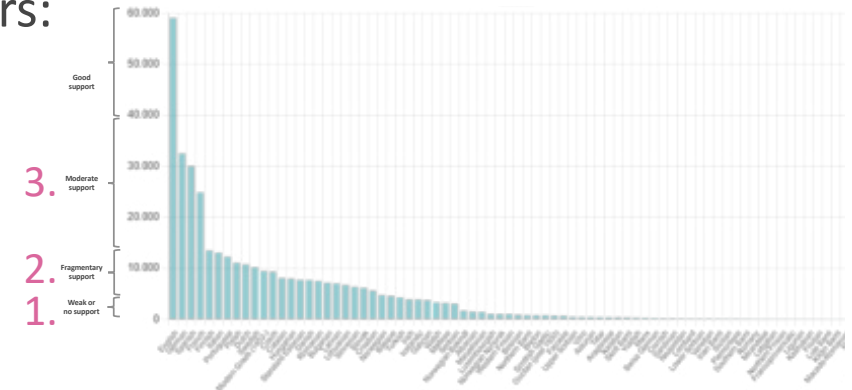
| Data Spaces and Data Infrastructures | Research and Research Data Infrastructures | Various AI Initiatives | AI on Demand Platform | High Performance Computing | Standardisation |
|---|---|---|---|---|---|
| **EU/EC Data Spaces**<br><br>EUROPEAN LANGUAGE DATA SPACE<br><br>**Language EDIC**<br><br>BDV BIG DATA VALUE ASSOCIATION<br><br>gaia-x<br><br>INTERNATIONAL DATA SPACES ASSOCIATION | eosc<br><br>nfdi Nationale Forschungsdaten Infrastruktur<br><br>CLARIN ERIC<br>Common Language Resources and Technology Infrastructure<br><br>RDA RESEARCH DATA ALLIANCE<br><br>... | Adra<br><br>CLAIRE<br><br>LEAM:AI<br><br>HUMANE AI NET<br><br>openGPT-X<br><br>... | AI-on-Demand Platform<br><br>AI4EU<br><br>EUROPEAN LANGUAGE GRID<br><br>... | EuroHPC Joint Undertaking<br><br>... | W3C®<br><br>DIN<br><br>... |

EUROPEAN LANGUAGE EQUALITY **Programme**

**Positioning and Foreseen Collaborations:**

- The ELE Programme complements these related initiatives and organisations.

- The ELE Programme will make use of the services and resources provided by these initiatives.

# ELE 2: SRIA Supplementary Document

- ELE 2 had the objective to extend and to revise the SRIA produced by ELE 1.

- Our additional consultations have not brought to light any need for substantial revisions.

- Strong endorsement by all stakeholders!

- We complied a document with the various pieces of additional evidence (among others, from the FSTP projects) and additional ideas

- ELE 2 D4.2 (see the cover page on the right)

**EUROPEAN LANGUAGE EQUALITY** 2

**D4.2**

**Digital Language Equality in Europe by 2030: Strategic Agenda and Roadmap – Supplementary Document**

| Authors | ELE 2 Consortium |
|---|---|
| Dissemination level | Public |
| Date | 30-06-2023 |

# Summary and Final Remarks

- META-NET study (2012): At least 21 European languages in danger of digital extinction (even though, politically, all our languages are supposed to be equal)

- EP Resolution (2018, "Jill Evans report"): Europe needs to take action!

- Current ELE results (2022/2023): There's still an extremely strong imbalance in terms of technology support of Europe's languages – *all* languages except English under-resourced!

- EU concentrates on data spaces and data economy but mostly ignoring our languages

- ELE recommends setting up a large-scale, long-term funding programme to digitally safeguard our languages and to put Europe back in the pole position – the scientific field and also its exploitation is currently dominated by the US and Asia

- Europe-wide coordination, shared scientific goals, focus on data and language models (LDS, EDIC)

- Impact – Our vision for 2030: Full Digital Language Equality in Europe by 2030 – transparent, barrier-free communication using Europe's own sovereign LT and NLP infrastructure.

**European Language Equality**

# Thank you!

Georg Rehm (DFKI, Germany)
georg.rehm@dfki.de

27-06-2023 META-FORUM 2023 – Digital Language Equality for a Multilingual Europe
http://european-language-equality.eu