

# ELRC White Paper: Language data sharing & language-centric Artificial Intelligence (AI)

---

Eileen Marra, ELRC Communications Manager (DFKI)

6<sup>th</sup> ELRC Conference (31 March 2022)

## ELRC White Paper: Key facts

### Sustainable Language Data Sharing to Support Language Equality in Multilingual Europe

#### WHY LANGUAGE DATA MATTERS

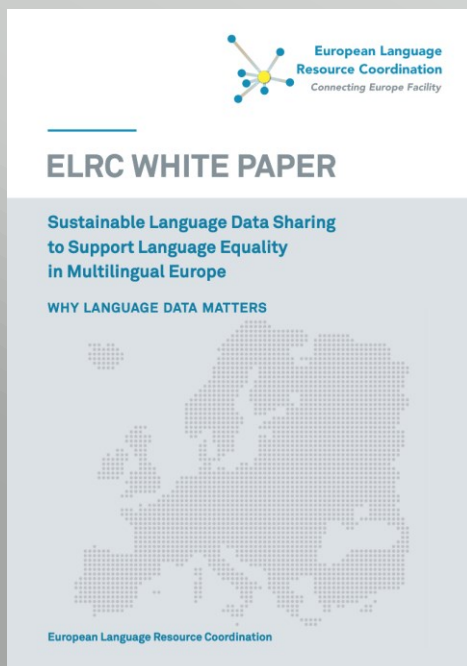
- Published in 2019
- European practices for sharing language data
- Challenges for language data sharing
- Recommendations to address the challenges
- Annex: Country Profile for each CEF-affiliated country

#### NEW

- Not only covering the public sector, but also **small and medium-sized enterprises (SMEs)**
- Going “beyond simple machine translation”: Analysing the use of **additional language technology tools** in public administrations and SMEs (e.g. Automatic Content Classification, Neural Machine Translation, etc.)
- Analysing the role of LT and language data in the digital economy

New title:










“AI for Multilingual Europe – Why language data matters”



# Why the extension?



Connecting Europe  
Language Tools

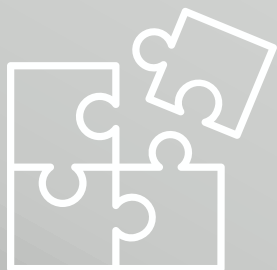
 <b>eTranslation</b>	 <b>Multilingual Tweet</b>	 <b>Speech-to-Text</b>
 <b>NLP Tools</b>	 <b>Interactive Terminology for Europe</b>	 <b>European Language Resource Coordination (ELRC)</b>
 <b>Catalogue of services</b>	 <b>CEF Building Block Information</b>	 <b>Developer's Corner</b>

Access to some of these tools requires registration. EU staff are pre-registered.  
Please visit the registration page: <https://webgate.ec.europa.eu/etranslation/public/welcome.html>.  
For any other issues, please contact [help@cefat-tools-services.eu](mailto:help@cefat-tools-services.eu).

# Approach

- Update of country profiles based on latest ELRC Workshop Round (in progress)
- Update and extension of white paper based on the analysis of
  - ✓ AI Watch Report
  - ✓ National AI Strategies
  - ELRC White Paper Survey (in progress)
- Finalisation and publication in autumn 2022

## The missing piece: ELRC White Paper Survey



- EU-wide collection of national insights on:
  - Common European practices regarding translation, data management and sharing
  - The current use and importance of language technology (LT)
  - Contents of national regulations related to LT and Artificial Intelligence (AI)
  - Ideas and priorities to facilitate data sharing and LT development for Europe's Multilingual Future
- Available online, targeting EU representatives of public sector, research/academia and SMEs
- First insights provided by National Anchor Points

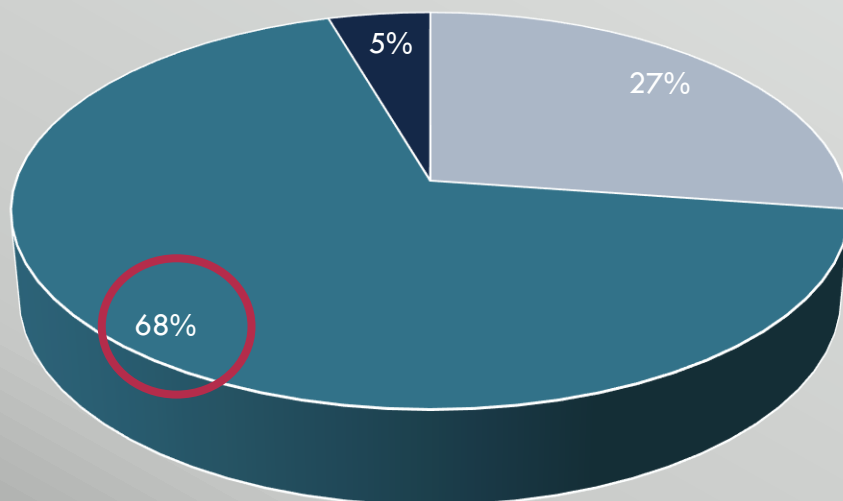


# Preliminary Findings: Translation Practices

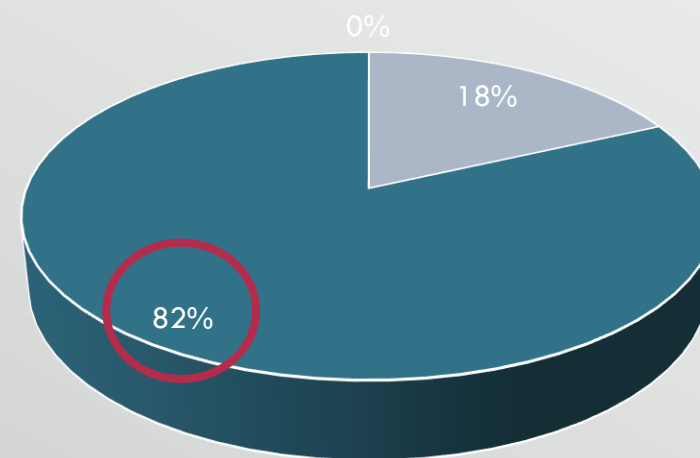
---

## Translation practices: In-house vs. outsourced translation

2022



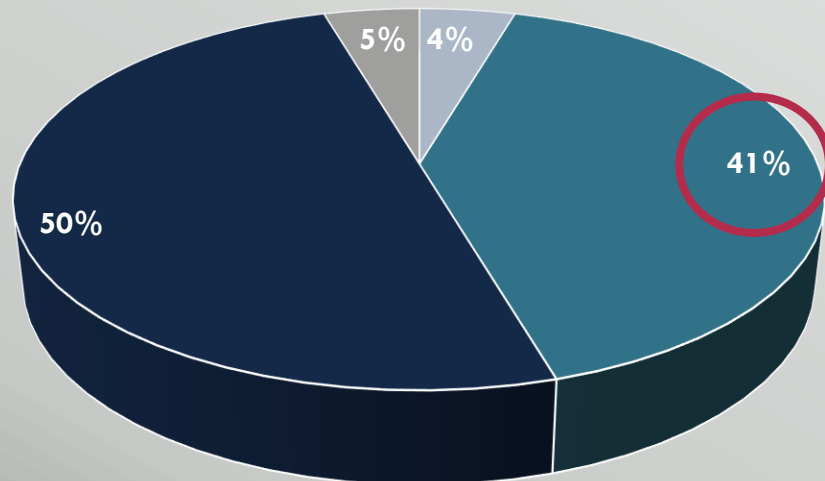
2019



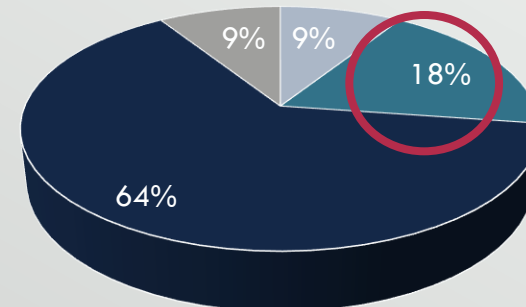
- More than 50% in-house
- Mostly outsourcing
- All outsourced

## Translation practices: Use of CAT Tools

2022



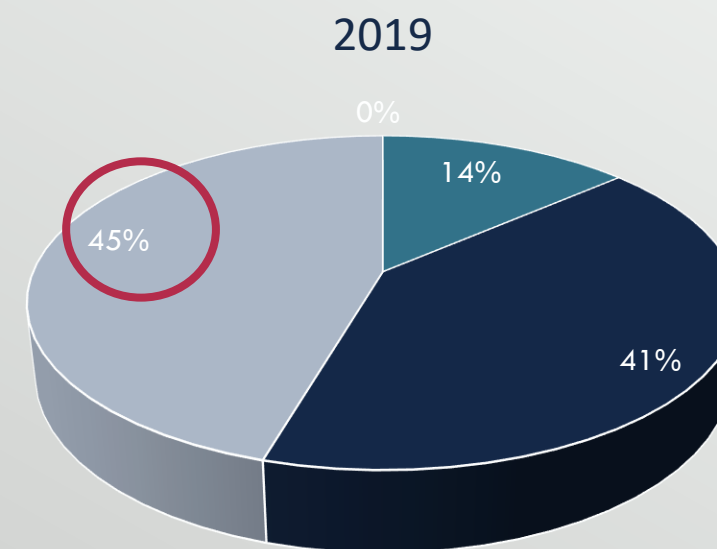
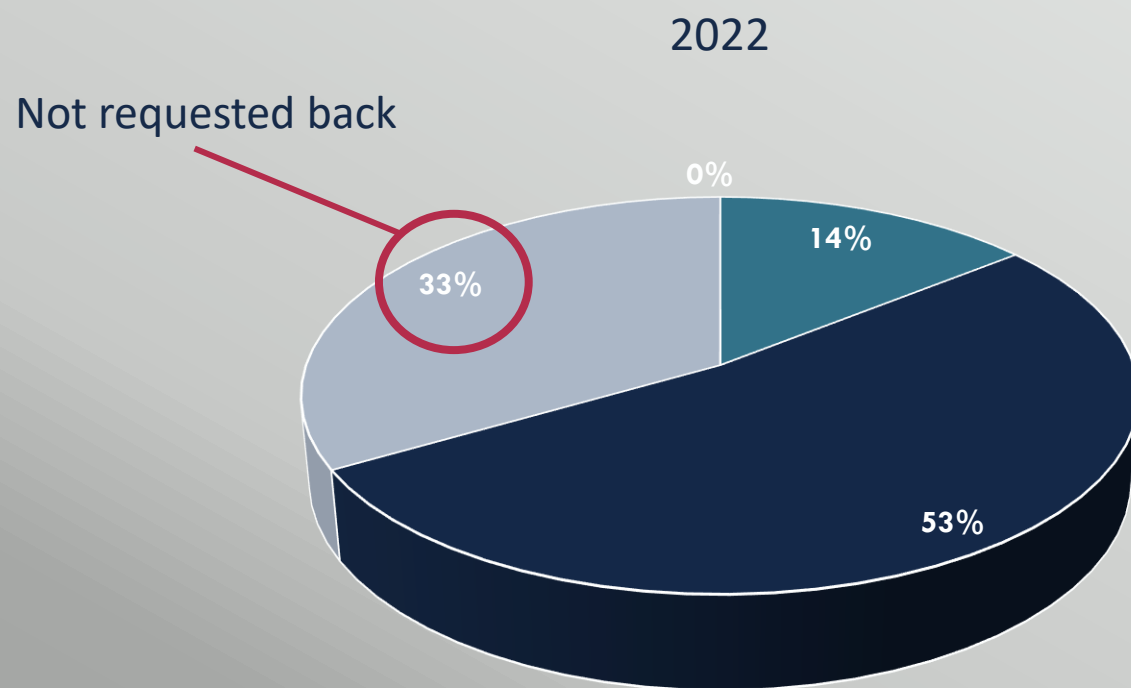
2019



- All translations carried out with CAT tools
- Common practice that translation services / translation professionals use CAT tools
- Only single translation services or translators use CAT
- No use of CAT

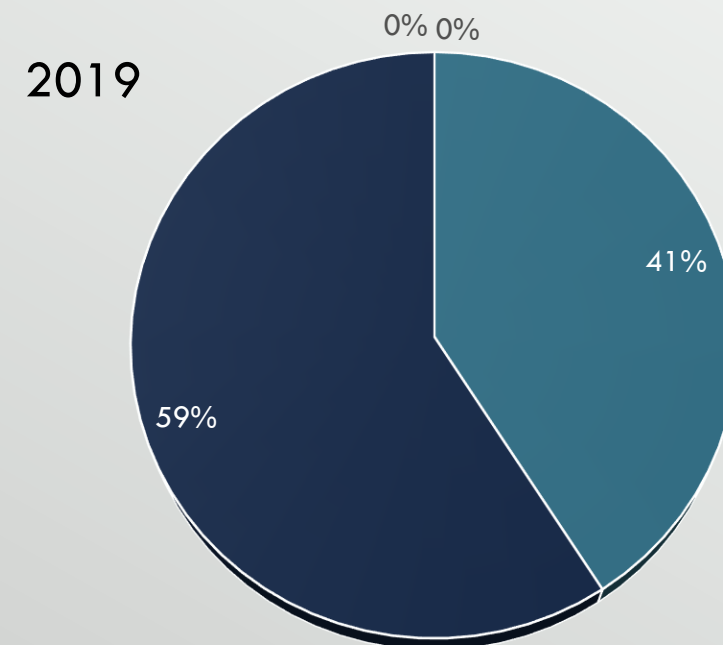
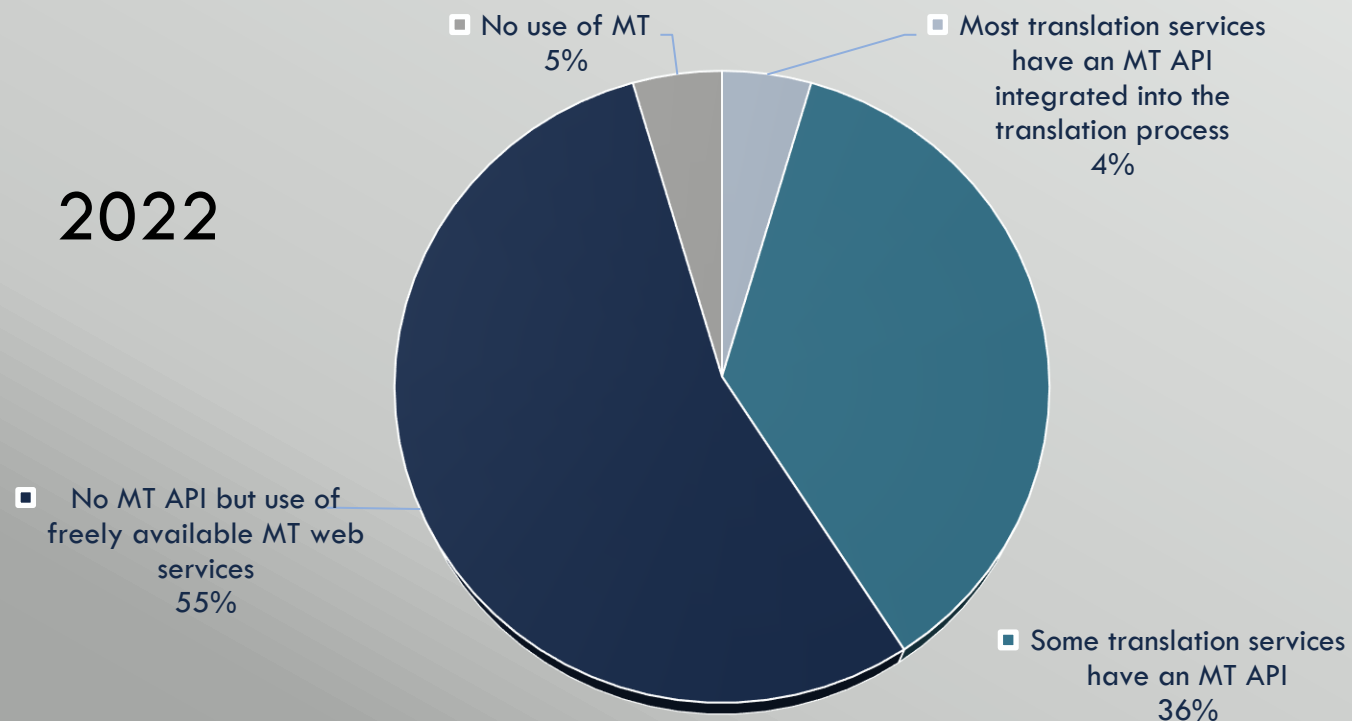


## Translation practices: TM files and by-products from outsourced translations

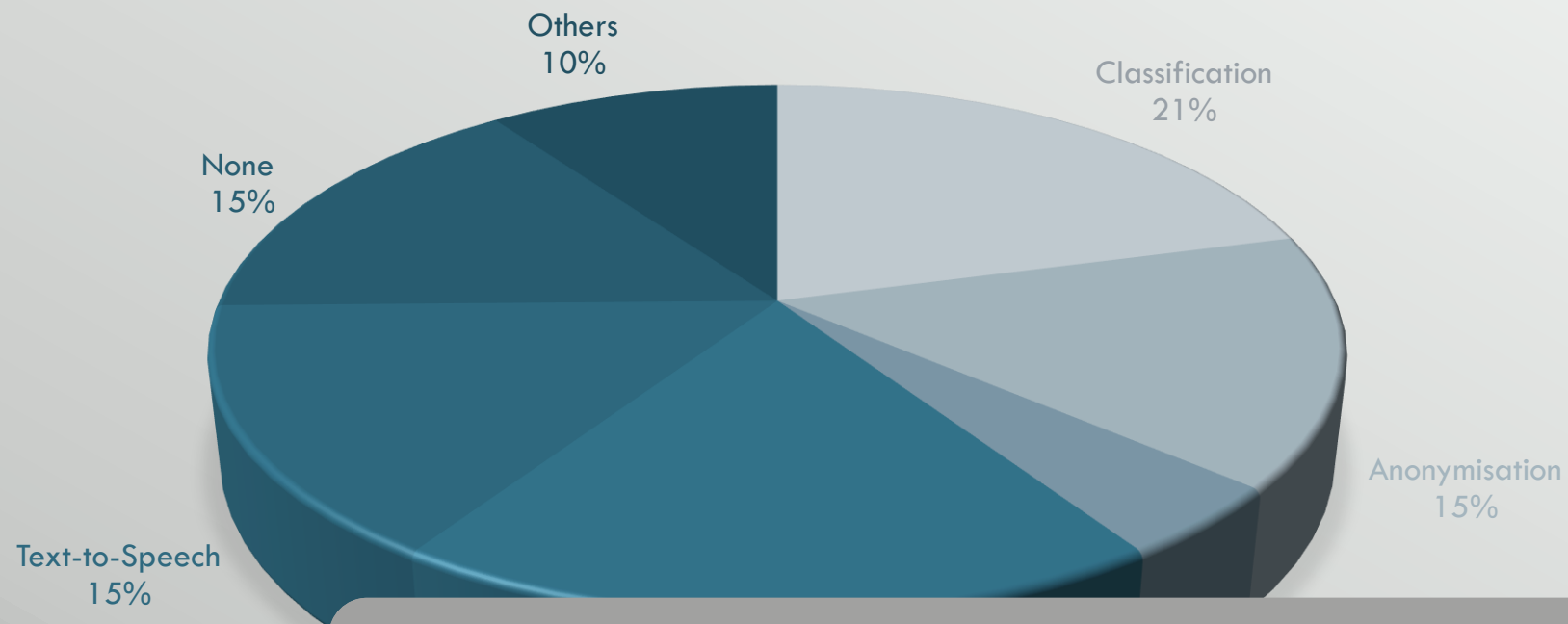


- Requested back by default
- Requested back for most outsourced translations
- Requested by some PAs
- Not requested back

## Use of machine translation in Public Administrations



## Use of LT Tools in Public Administrations/ SMEs



All of them require data!

## Challenges for language data sharing (2019 and 2022)

- Raising awareness on the value of language data
- Increasing interest in MT/LT as part of the digital policy
- Establishing good data management practices
- Tackle legal concerns
- Identify and gain access to outsourced translations
- Others

# Preliminary Findings: The Value of Language Data

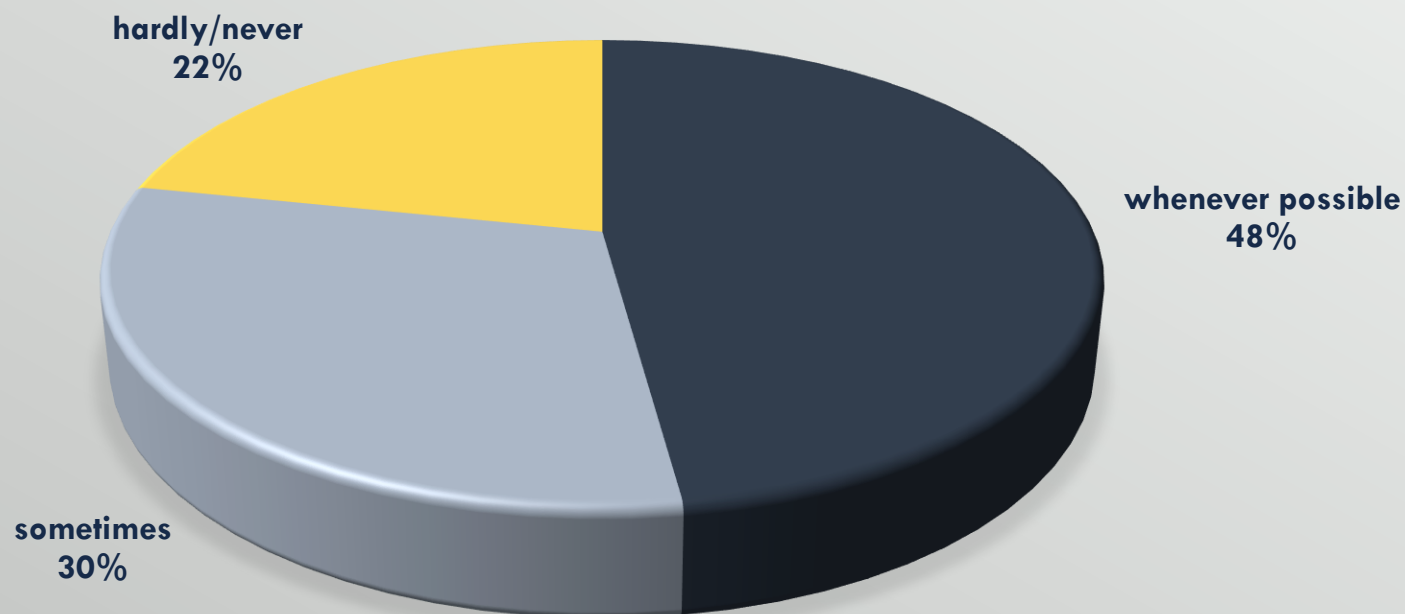
---

“There is reason to believe that the public sector possesses far more data that could be used in developing language technology than it realises.”

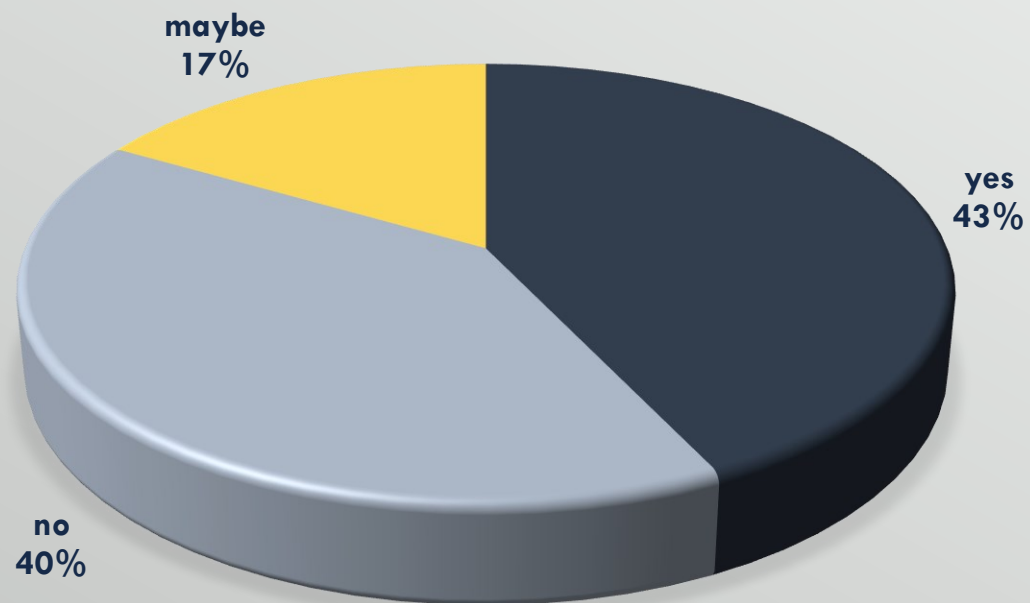
(quote from Norwegian AI Strategy, p.20)

Data has a better idea

Is your organisation storing language data like tmx files, translations, audio files, video recordings, etc.?

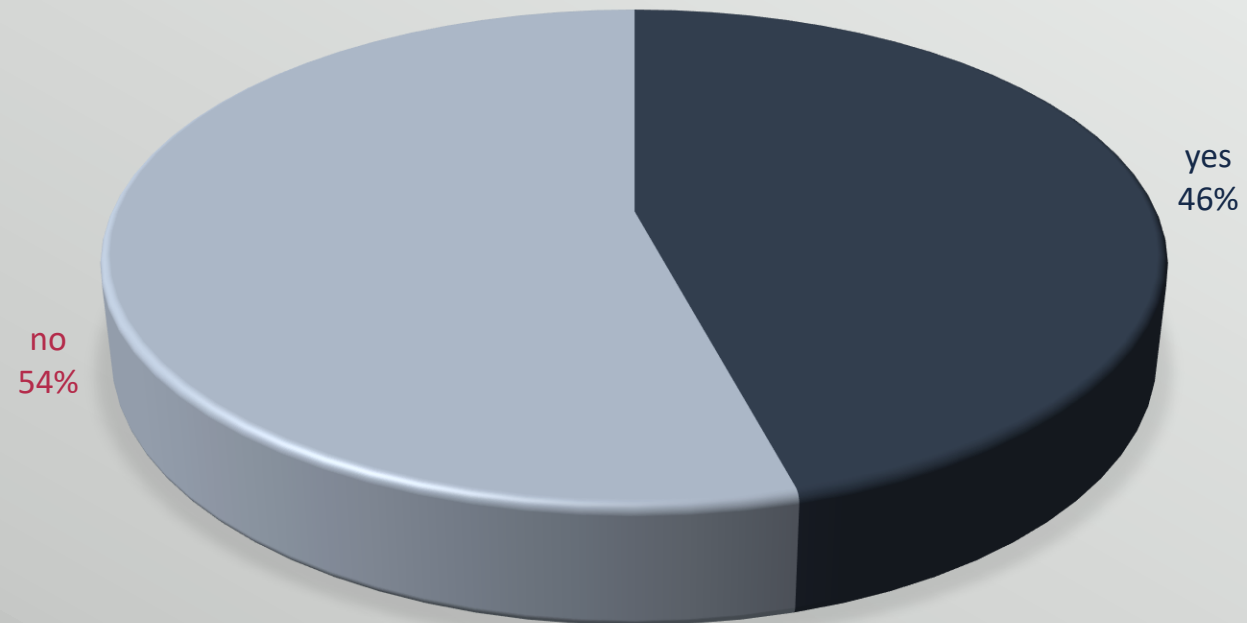


Do you think that the value of language data has been recognised in your country?



What would you say?

Is the National AI Strategy in your country referring to the importance of language data?





## The Value of Language Data: Good Practice Examples

### Ireland

- “Many of the language datasets currently used for training AI systems originate from US-based sources and may not contain common everyday terms used by people in Ireland. To render AI systems accessible to a wider range of our population, as well as to develop services in Irish based on AI for Irish language-speakers, good language technology resources need to be developed.” ([AI Strategy](#) “AI – Here for Good”, p.42)

### Norway

- Special chapter about LR & LT:  
“There is reason to believe that the public sector possesses far more data that could be used in developing language technology than it realises. The Government will therefore promote awareness of language data and language resources in the public sector” ([National Strategy for Artificial Intelligence](#), p. 20)

### Spain

- One of the action items:  
“Boosting the National Language Technology Plan and the creation of resources in the Spanish Language in AI initiative” (ES: Impulso al Plan Nacional de Tecnologías del Lenguaje y la creación de recursos en la iniciativa de Lengua Española en la IA”)

## Challenges for language data sharing (2019 and 2022)

- Raising awareness on the value of language data
- Increasing interest in MT/LT as part of the digital policy
- Establishing good data management practices
- Tackle legal concerns
- Identify and gain access to outsourced translations
- Others

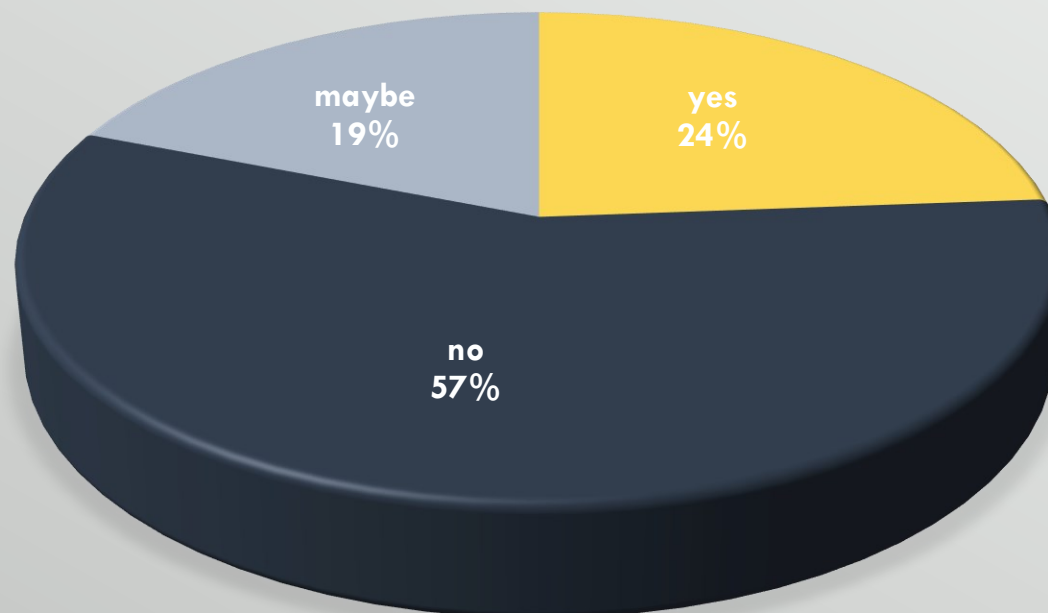
## Preliminary Findings: The role of LT in national policies

---

“In the future, citizens can receive services seamlessly in the language they need (...).”

(quote from [Finland's age of AI](#), p.54)

Do you think LT is appropriately represented in the  
AI strategy of your country?



What would you say?

## The role of LT in National AI Strategies

- 21 of the published 24 national AI Strategies mention Language Technology
  - Varying emphasis: Full chapters on LT (e.g. Malta) vs. side notes on language-centric AI (e.g. Luxembourg)
- Countries where LT is not explicitly mentioned: Sweden, Estonia (but: included LT in draft strategy of Estonian language), Netherlands (but: mentions chatbots as a useful NLP application)
- Examples:
  - Bulgaria: Use of LT to support foreign language learning  
“In practice, any formalized set of grammar rules can be considered as a resource for automatic testing of knowledge of the relevant aspects of the language, which is built into specially designed tests for verification. It would be useful for Bulgarians abroad to provide a public online interface for learning Bulgarian grammar.” ([Concept for the development of AI in Bulgaria until 2030](#), p. 44)
  - Hungary: “The application and further development of existing technologies to the Hungarian language is of significant national interest.” ([Hungary’s Artificial Intelligence Strategy](#), p. 26)

# Preliminary Findings: Recent Advances



## Recent advances

*“The **pan-European data collection campaigns** (e.g. **ELRC workshops**) represent a **major breakthrough** that will change the general attitude towards the preservation of digital textual data, mono- and multilingual.”*  
(Croatia)

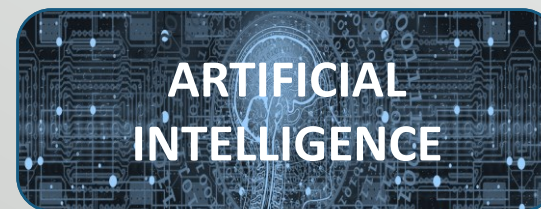
*“**Machine translation** has **progressed enormously**. Several Finnish public administration services have introduced customised neural machine translators.”* (Finland)

*“**Language corpora** are more widely used (incl. by translators), since they **have become more accessible**.”* (Latvia)

*“There is an **increase in digitalisation** due to Covid.”* (Romania)

# “Maximum success through cooperation” – and that’s where you come in!

Please help us get further insights into the developments  
in your country regarding



by participating in the White Paper Survey 2022





## European Language Resource Coordination

*Connecting Europe Facility*

### White Paper Survey 2022

Think **big**.

For Europe's multilingual **future!**

SCAN ME



# Thank you for your attention

