



EUROPEAN LANGUAGE EQUALITY

META[≡]NET
META[≡]FORUM 2022

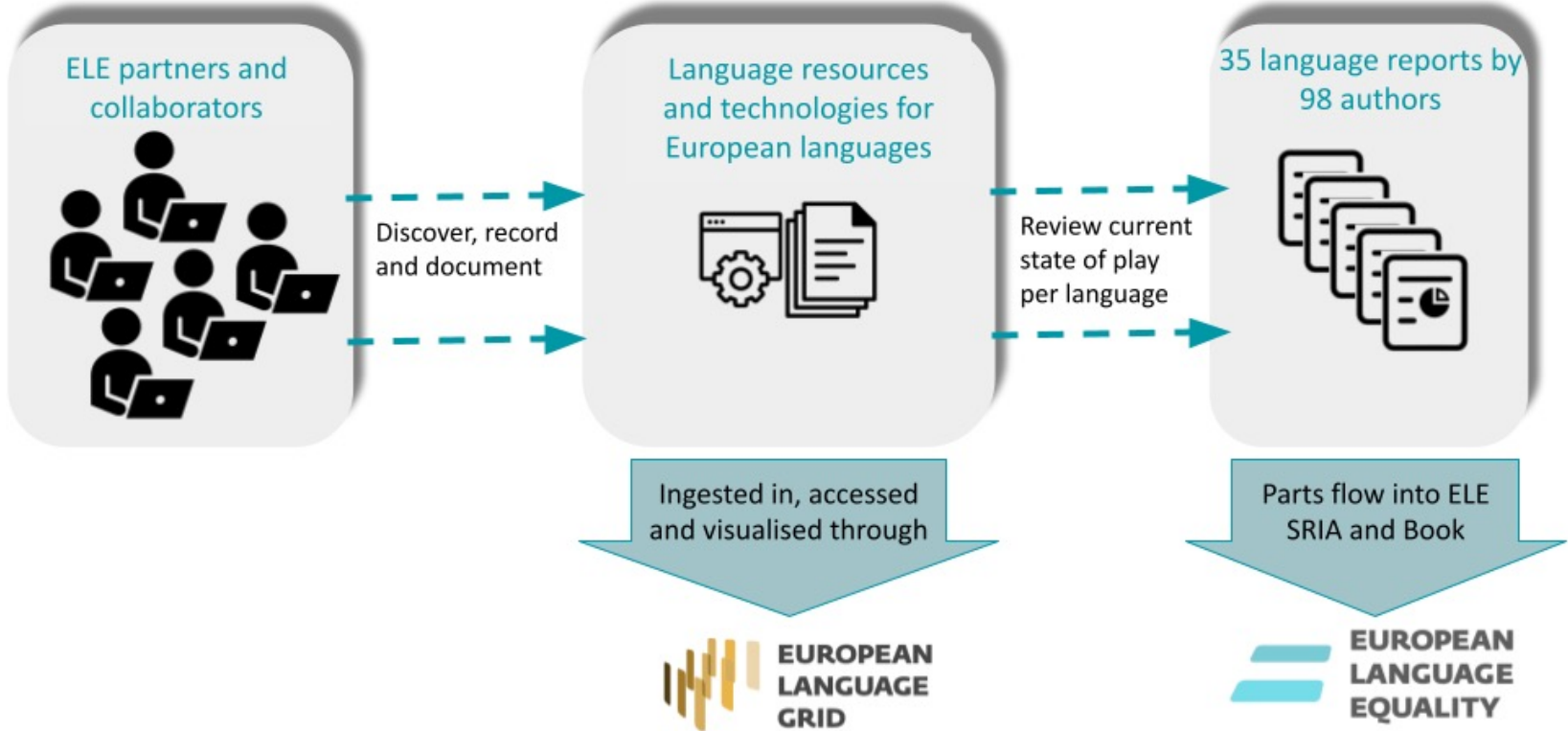


Language Technology for Europe's Languages in 2022 Cross-language Comparison

Maria Giagkou (ILSP, R.C. "Athena")
mgiagkou@athenarc.gr


08/09-06-2022 META-FORUM 2022 – Joining the European Language Grid (hybrid conference)
<http://www.european-language-grid.eu> – <http://european-language-equality.eu>

An evidence-based investigation



Language Reports

<https://european-language-equality.eu/deliverables/>

 EUROPEAN LANGUAGE EQUALITY		About ▾ Strategic Agenda Deliverables Events ▾ News ▾ Consortium Contact			
D1.4	Report on Basque	UPV/EHU	February 2022	26 pages	
D1.5	Report on Bulgarian	IBL	February 2022	27 pages	
D1.6	Report on Catalan	BSC	February 2022	24 pages	
D1.7	Report on Croatian	FFZG	February 2022	30 pages	
D1.8	Report on Czech	CUNI	February 2022	23 pages	
D1.9	Report on Danish	UCPH	February 2022	26 pages	
D1.10	Report on Dutch	INT	February 2022	23 pages	
D1.11	Report on English	USFD	February 2022	21 pages	
D1.12	Report on Estonian	UTART	February 2022	20 pages	
D1.13	Report on Finnish	UHEL	February 2022	24 pages	
D1.14	Report on French	CNRS	February 2022	42 pages	
D1.15	Report on Galician	UVIGO	February 2022	20 pages	
D1.16	Report on German	DFKI	February 2022	25 pages	
D1.17	Report on Greek	ILSP	February 2022	30 pages	
D1.18	Report on Hungarian	NYTK	February 2022	26 pages	
D1.19	Report on Icelandic	SAM	February 2022	22 pages	
D1.20	Report on Irish	DCU	February 2022	30 pages	
D1.21	Report on Italian	FBK	February 2022	24 pages	
D1.22	Report on Latvian	IMCS	February 2022	29 pages	
D1.23	Report on Lithuanian	LKI	February 2022	24 pages	

Language Reports

Cross-language comparison

- Based on number of LRTs catalogued in the ELG platform
- Comparison across languages as per
 - Tools/services broadly categorised into a number of core LT application areas
 - Resources that can be used as training or evaluation data (indication of the potential for LT development) with regard to a small number of basic types
- Four bands:
 - Good support (green)
 - Moderate (light green)
 - Fragmentary (yellow)
 - Weak or no support (light yellow)

		Tools and Services						Language Resources						
		Text Processing	Speech Processing	Image/Video Processing	Information Extraction and IR	Human-Computer Interaction	Translation Technologies	Natural Language Generation	Text Corpora	Multimodal Corpora	Parallel Corpora	Models	Lexical Resources	Overall
EU official languages	Bulgarian													
	Croatian													
	Czech													
	Danish													
	Dutch													
	English													
	Estonian													
	Finnish													
	French													
	German													
	Greek													
	Hungarian													
	Irish													
	Italian													
	Latvian													
	Lithuanian													
	Maltese													
Polish														
Portuguese														
Romanian														
Slovak														
Slovenian														
Spanish														
Swedish														
(Co-)official languages	National level	Albanian												
		Bosnian												
		Icelandic												
		Luxembourgish												
		Macedonian												
		Norwegian												
	Regional level	Serbian												
		Basque												
		Catalan												
		Faroese												
		Frisian (Western)												
		Galician												
		Jerriais												
		Low German												
		Manx												
		Mirandese												
		Occitan												
		Sorbian (Upper)												
		Welsh												
All other languages														

Language Reports

Cross-language comparison

- **Translation technologies:** many languages are at least moderately supported

		Tools and Services					Language Resources							
		Text Processing	Speech Processing	Image/Video Processing	Information Extraction and IR	Human-Computer Interaction	Translation Technologies	Natural Language Generation	Text Corpora	Multimodal Corpora	Parallel Corpora	Models	Lexical Resources	Overall
(Co-)official languages	EU official languages	Bulgarian												
		Croatian												
		Czech												
		Danish												
		Dutch												
		English												
		Estonian												
		Finnish												
		French												
		German												
		Greek												
		Hungarian												
		Irish												
		Italian												
		Latvian												
		Lithuanian												
		Maltese												
		Polish												
		Portuguese												
		Romanian												
Slovak														
Slovenian														
Spanish														
Swedish														
(Co-)official languages	National level	Albanian												
		Bosnian												
		Icelandic												
		Luxembourgish												
		Macedonian												
		Norwegian												
	Regional level	Serbian												
		Basque												
		Catalan												
		Faroese												
		Frisian (Western)												
		Galician												
		Jerriais												
		Low German												
		Manx												
		Mirandese												
		Occitan												
		Sorbian (Upper)												
		Welsh												
All other languages														

Language Reports

Cross-language comparison

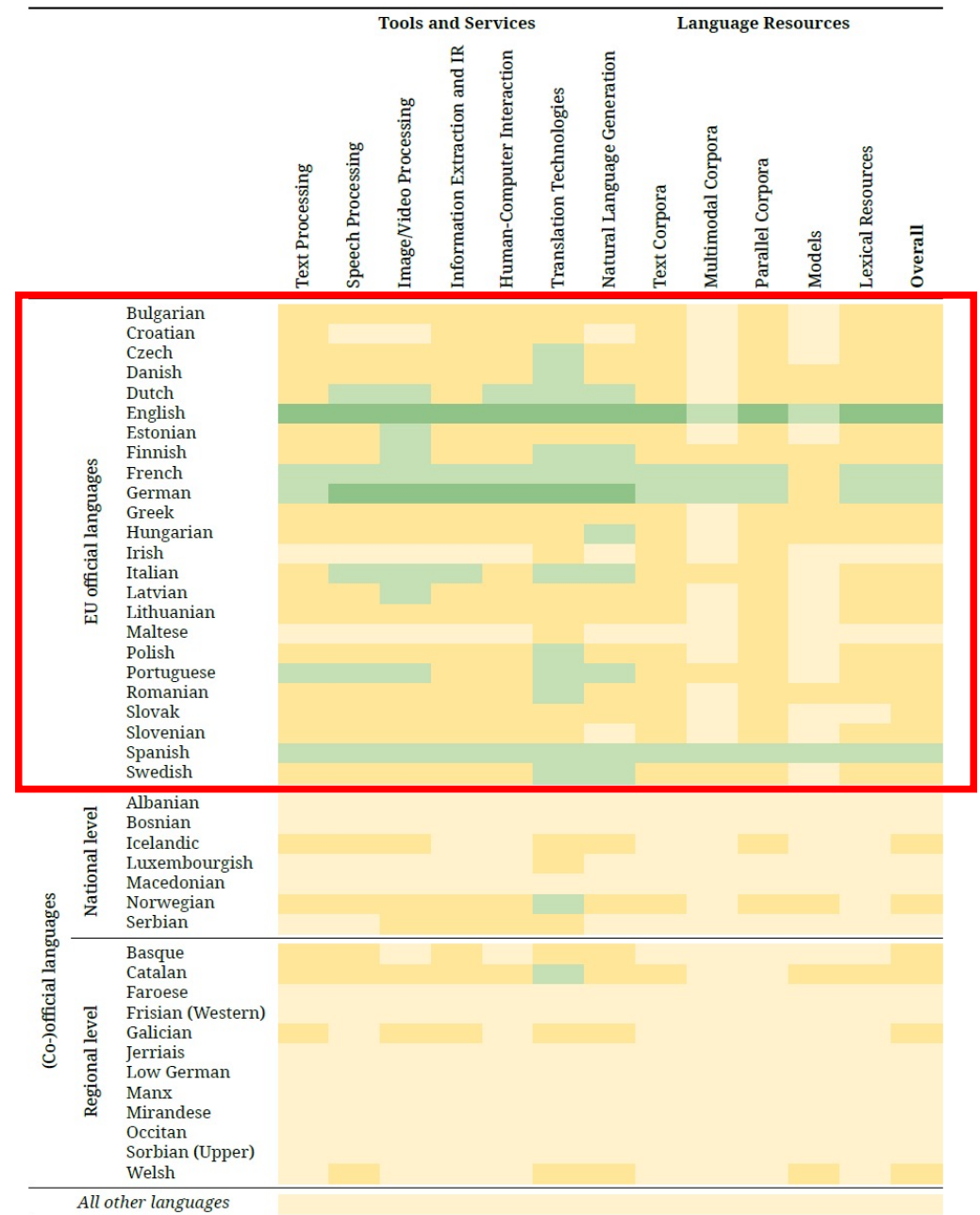
- **Multimodal corpora** and **Models**: many languages are weakly supported

		Tools and Services								Language Resources				
		Text Processing	Speech Processing	Image/Video Processing	Information Extraction and IR	Human-Computer Interaction	Translation Technologies	Natural Language Generation	Text Corpora	Multimodal Corpora	Parallel Corpora	Models	Lexical Resources	Overall
EU official languages	Bulgarian													
	Croatian													
	Czech													
	Danish													
	Dutch													
	English													
	Estonian													
	Finnish													
	French													
	German													
	Greek													
	Hungarian													
	Irish													
	Italian													
	Latvian													
	Lithuanian													
	Maltese													
	Polish													
	Portuguese													
	Romanian													
Slovak														
Slovenian														
Spanish														
Swedish														
(Co-)official languages	National level	Albanian												
		Bosnian												
		Icelandic												
		Luxembourgish												
		Macedonian												
		Norwegian												
		Serbian												
	Regional level	Basque												
		Catalan												
		Faroese												
		Frisian (Western)												
		Galician												
		Jerriais												
		Low German												
		Manx												
		Mirandese												
		Occitan												
		Sorbian (Upper)												
		Welsh												
All other languages														

Language Reports

Cross-language comparison

- **EU official languages** better supported than other European languages



Language Reports

Cross-language comparison

- **Best supported:** English, followed by Spanish, German and French

		Tools and Services							Language Resources					
		Text Processing	Speech Processing	Image/Video Processing	Information Extraction and IR	Human-Computer Interaction	Translation Technologies	Natural Language Generation	Text Corpora	Multimodal Corpora	Parallel Corpora	Models	Lexical Resources	Overall
EU official languages	Bulgarian													
	Croatian													
	Czech													
	Danish													
	Dutch													
	English													
	Estonian													
	Finnish													
	French													
	German													
	Greek													
	Hungarian													
	Irish													
	Italian													
	Latvian													
	Lithuanian													
	Maltese													
	Polish													
	Portuguese													
Romanian														
Slovak														
Slovenian														
Spanish														
Swedish														
(Co-)official languages	National level	Albanian												
		Bosnian												
		Icelandic												
		Luxembourgish												
		Macedonian												
		Norwegian												
		Serbian												
	Regional level	Basque												
		Catalan												
		Faroese												
		Frisian (Western)												
		Galician												
		Jerriais												
		Low German												
		Manx												
		Mirandese												
		Occitan												
		Sorbian (Upper)												
		Welsh												
All other languages														

Language Reports

Cross-language comparison

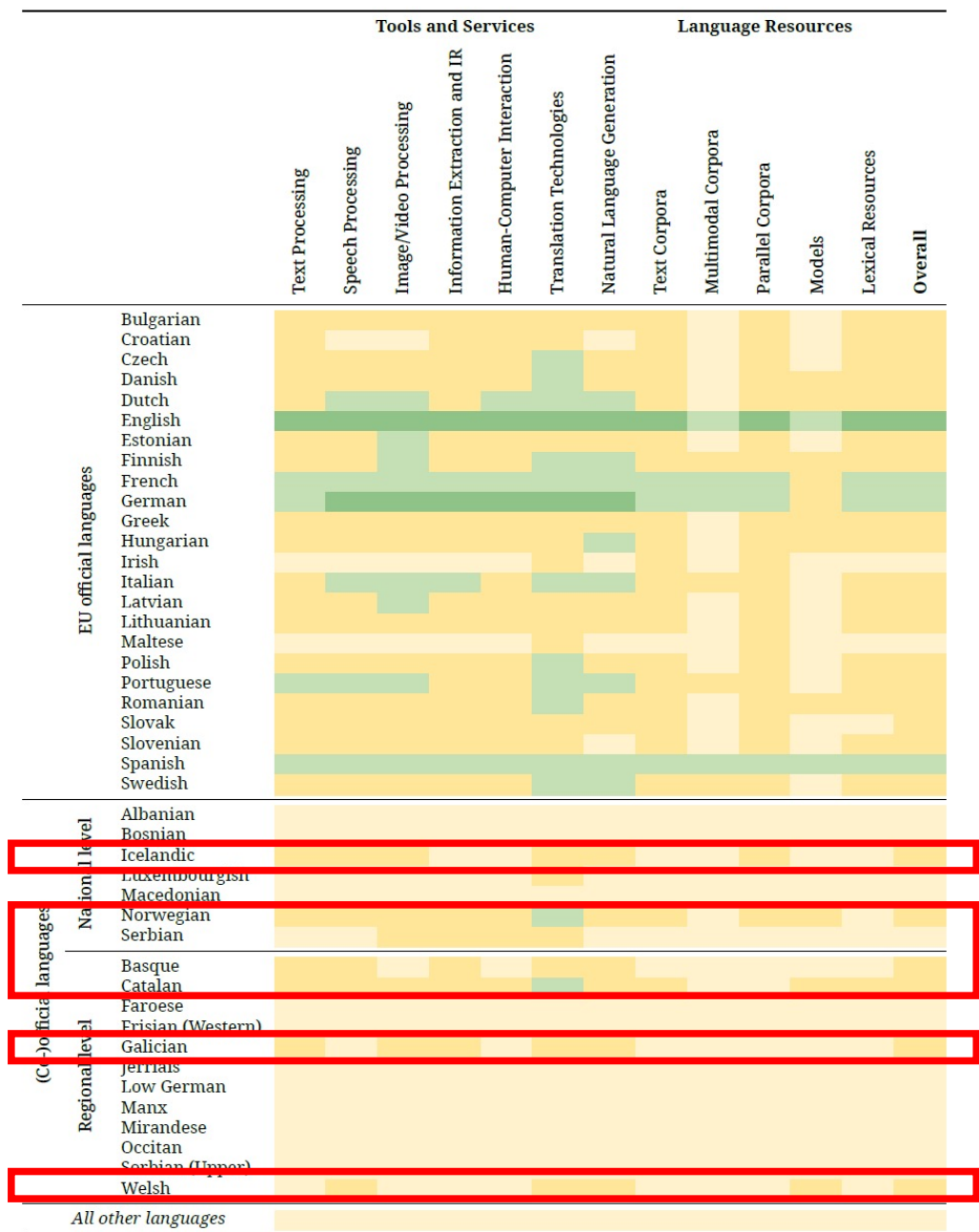
- **Least supported** among the EU official: Irish and Maltese

		Tools and Services							Language Resources					
		Text Processing	Speech Processing	Image/Video Processing	Information Extraction and IR	Human-Computer Interaction	Translation Technologies	Natural Language Generation	Text Corpora	Multimodal Corpora	Parallel Corpora	Models	Lexical Resources	Overall
EU official languages	Bulgarian													
	Croatian													
	Czech													
	Danish													
	Dutch													
	English													
	Estonian													
	Finnish													
	French													
	German													
	Greek													
	Hungarian													
	Irish													
	Italian													
	Latvian													
	Lithuanian													
	Maltese													
	Polish													
	Portuguese													
	Romanian													
Slovak														
Slovenian														
Spanish														
Swedish														
(Co-)official languages	National level	Albanian												
		Bosnian												
		Icelandic												
		Luxembourgish												
		Macedonian												
		Norwegian												
		Serbian												
	Regional level	Basque												
		Catalan												
		Faroese												
		Frisian (Western)												
		Galician												
		Jerriais												
		Low German												
		Manx												
		Mirandese												
		Occitan												
		Sorbian (Upper)												
		Welsh												
All other languages														

Language Reports

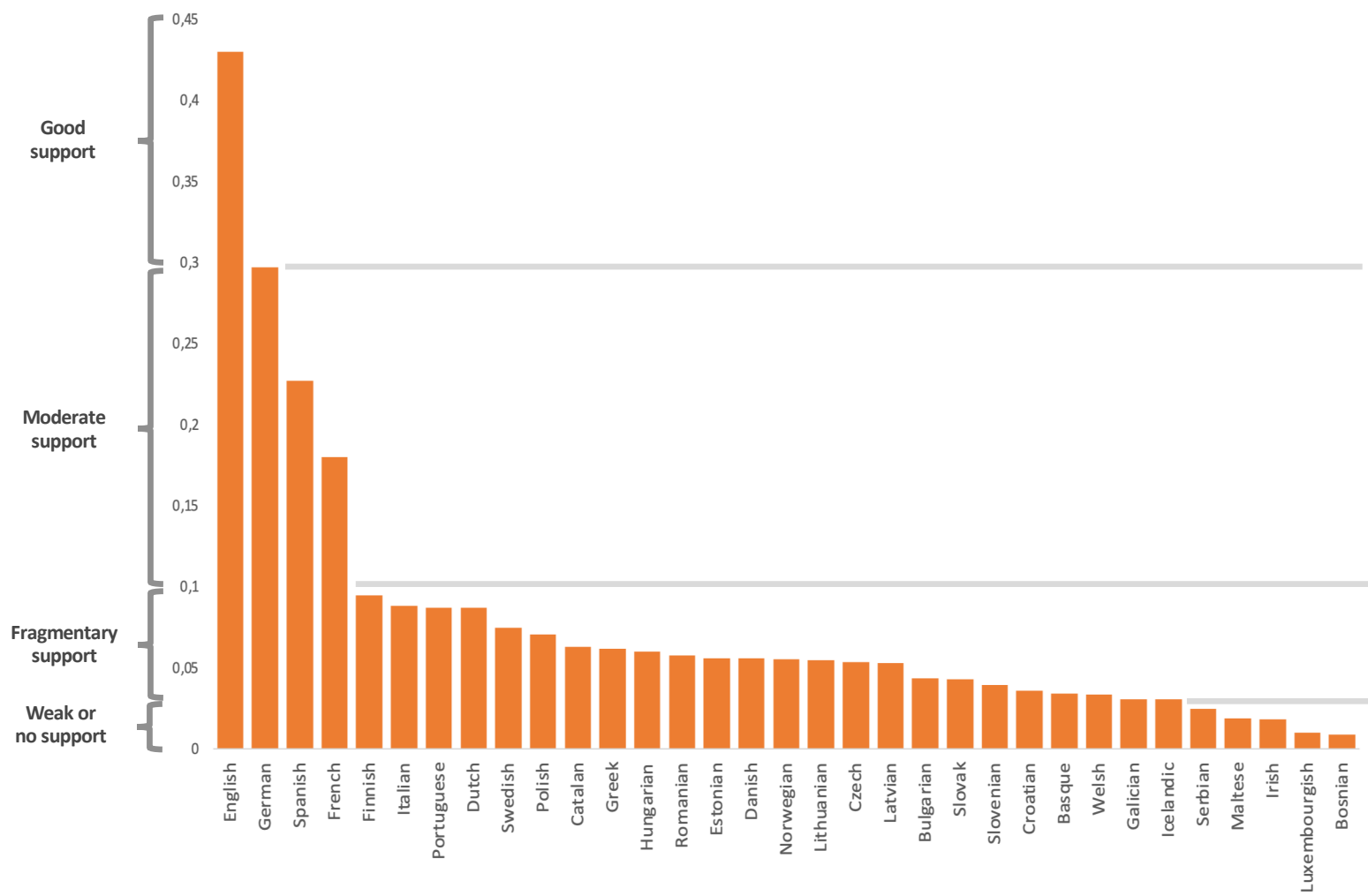
Cross-language comparison

- Best supported among languages with official status at the national or regional level (but not EU): Norwegian, Icelandic, Serbian, Basque, Catalan, Galician, Welsh



Language Reports

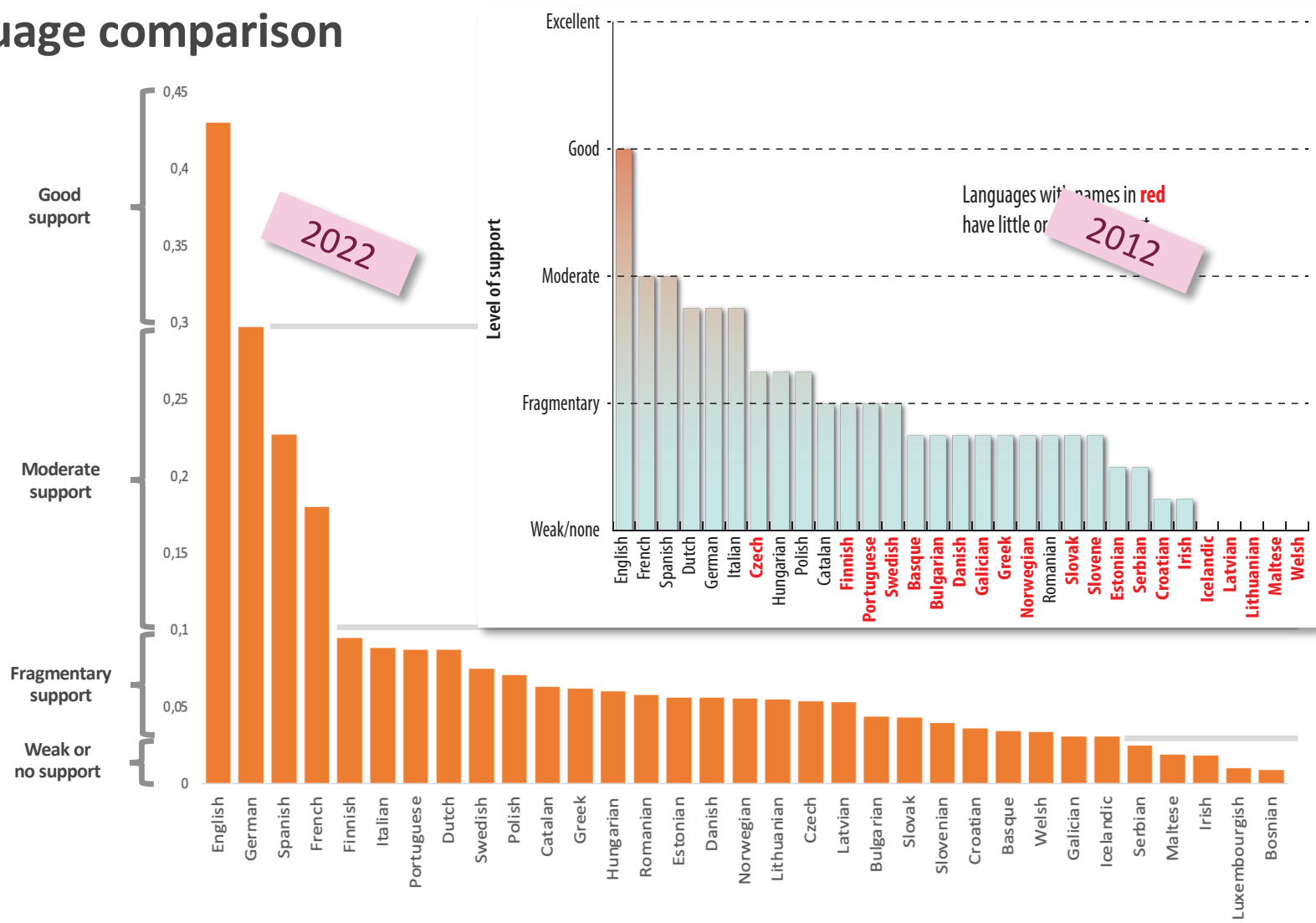
Cross-language comparison



Results based on raw counts of the 11,000+ language resources and language technologies currently described with metadata records in the ELG platform.

Language Reports

Cross-language comparison



Results based on raw counts of the 11,000+ language resources and language technologies currently described with metadata records in the ELG platform.

Indicative recurring issues – LRTs needed

Most frequent mentions across the reports:

- **Data:** multimodal, models, domain-specific, social media language, speech, semantic resources (annotations, knowledge bases ...)
- **Technologies:** conversational systems, QA, NLG, summarisation, NLU, discourse processing



Language Reports

Indicative recurring issues

Strengths:

- general positive tone for the achievements in the last decade
- national support for LT and/or language resources has already yielded significant results; maintain and build on previous achievements (e.g. Spain, Iceland, Estonia, Denmark...)
- positive impact of participation to LRIs

Opportunity:

- **transfer learning** methodologies promise results for small languages without huge investment (BUT consider their **limitations**; language- and culture-specific approaches are still necessary)

Weaknesses:

- unclear **legal** status/ licencing
- **fragmented** and discontinued funding streams with one-off grants
- inadequate focus on **non-standard or regional varieties** and dialects
- insufficiently recognised **value of language data**, huge amounts of public data remains unexploited

Threats:

- for bilingual communities **language shift** to dominant language, e.g. Irish, Welsh and Maltese vs English; Catalan, Basque and Galician vs Spanish; West Frisian vs Dutch; Nordic minority languages vs. the official state languages
- **digital extinction**



Thank you!



The European Language Grid has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement № 825627 (ELG).

The European Language Equality project has received funding from the European Union under grant agreement № LC-01641480 – 101018166 (ELE),

Maria Giagkou (ILSP, R.C. “Athena”)
mgiagkou@athenarc.gr

08/09-06-2022 META-FORUM 2022 – Joining the European Language Grid (hybrid conference)
<http://www.european-language-grid.eu> – <http://european-language-equality.eu>

Session 8: Technology Support of Europe's Languages in 2022

- 11:20 Language Report Irish – Teresa Lynn (Dublin City University, Ireland)
- 11:25 Language Report Dutch – Frieda Steurs (Instituut voor de Nederlandse Taal, Netherlands)
- 11:30 Language Report Bulgarian – Svetla Koeva (Bulgarian Academy of Sciences, Bulgaria)
- 11:35 Language Report Welsh – Gareth Watkins (Bangor University, UK)
- 11:40 Language Report Icelandic – Eiríkur Rögnvaldsson (Institute for Icelandic Studies, Iceland)
- 11:45 Discussion