EUROPEAN LANGUAGE GRID

Introducing the European Language Grid: Deep Dive 3 – ELG Content

Khalid Choukri (ELDA), Ian Roberts (USFD), Kalina Bontcheva (USFD)
(additional content from Andres Garcia Silva, Expert System)

08/09-10-2019, Brussels – META-FORUM 2019
https://www.european-language-grid.eu

Introducing the European Language Grid

Deep Dive 3 – ELG Content

08/09-10-2019, Brussels – META-FORUM 2019 – Khalid Choukri (ELDA), Ian Roberts (USFD), Kalina Bontcheva (USFD)

Part 1) Language Resources

Part 2) Tools, Services, Components

Introducing the European Language Grid
Deep Dive 3 – ELG Content
08/09-10-2019, Brussels – META-FORUM 2019 – Khalid Choukri (ELDA), Ian Roberts (USFD), Kalina Bontcheva (USFD)

Overview – Language Resources

- ELG Content: Goals and Objectives
- Market Place for Language Resources
- Identification of existing Repositories and Resources
- Contributions of the National Competence Centres (NCCs)

ELG Content: Goals and Objectives

- Establish the ELG as an important market place and broker for LRs and LTs
- Identify and negotiate necessary rights on existing Language Resources
- Provide support to address the identified gaps for some resources and languages
- Use the ELG platform to produce models based on identified resources

➢ *Language Resources:* data sets (raw data, annotated data), models for existing LTs

ELG Content: Market Place

- Establish the ELG as an important market place and broker for Language Resources
  - Liaise with and capitalize on existing activities to negotiate/ingest Language Resources repositories into the ELG.
  - Initial providers: ELRA, META-SHARE, ELRC-SHARE, consortium members
  - Develop and promote efficient mechanisms for integration of LRs into the ELG
  - Promote market place related features: upload/download, licensing, billing, payment, etc.
- Offer an additional channel for users and suppliers:
  - Research organizations that develop or use LTs or LRs
  - Companies that develop, integrate, use, deploy LTs or LRs
  - Users of technologies (private and public sectors)
- ELG will host commercial and non-commercial LTs and LRs
- Management of transactions to be specified including legal, financial, logistical issues

ELG Content: Identification of existing Language Resources

- Identification of major LR repositories (research and industry suppliers)
- Great support from the NCC network

Statistics about the identified ones (internal and feedback from the NCCs):

- About 220 Repositories
- Large data centres (CLARIN, ELRA, ELRC-SHARE, LDC, META-SHARE, SADILAR, etc.)
- Commercial and academic repositories
- Harvesters (e.g., OLAC, META-SHARE)
- All modalities (audio, texts, etc.) but also language documentations
- Local and global players

ELG Content: LRs that the ELG consortium will provide

		Language Group A			Language Group B			Language Group C			
	Open Access	Corpora	Lexicons	Models	Corpora	Lexicons	Models	Corpora	Lexicons	Models	Totals
META-SHARE	yes	617	447	16	55	54	0	84	51	1	1325
	no	582	550	1	44	65	0	198	94	0	1534
ELRC-SHARE	yes	317	114	0	3	1	0	0	0	0	435
	no	74	16	0	2	1	0	0	0	0	93
ELDA	no	563	1012	0	35	18	0	250	54	0	1932
ELG	mixed	74	108	43	0	0	12	4	1	21	263
Totals		**2227**	**2247**	**60**	**139**	**139**	**12**	**536**	**200**	**22**	**5582**

- **Group A:** Official EU languages
- **Group B:** Other EU (and EU candidate) and related under-resourced languages
- **Group C:** Languages spoken by Immigrant, Trade and Political partners

ELG Content: What can be expected for the first release of ELG

- **ELRC-SHARE** – Over 200 language resources:
  - More than 100 TMX files for MT development, mostly EU languages
  - Over 4M pairs all together (largest is 700k TMXs)
  - More than 30 terminological databases, multiple domains (law, industry, education), over 400k terms
  - Many other resources under clearing and cleaning (expected by December 2019)
- **ELRA** – Over 200 language resources owned by ELRA/ELDA
  - All modalities (speech/video, text corpora, OCR etc.)
  - Many EU (national/regional) languages and non-EU ones
  - Many evaluation packages for LT benchmarking
- **META-SHARE** – Over 250 Language Resources
  - Many modalities (annotated corpora, treebanks, transcribed broadcast news, etc.)
  - Many resources tuned for research purposes

Introducing the European Language Grid
Deep Dive 3 – ELG Content
08/09-10-2019, Brussels – META-FORUM 2019 – Khalid Choukri (ELDA), Ian Roberts (USFD), Kalina Bontcheva (USFD)

Overview – Tools, Services and Components

- Addressing Heterogeneity
- Summary of existing Tools to be integrated
- API Design Principles

ELG Functional Content: Tools, Services, and Components

- Collecting information on all major existing tools, services and components (TSCs)
  - Ensure maximum coverage of EU languages through prioritisation
- Consult target user groups on what TSCs they need most
- Ensure commercial-grade service integration and robustness by having industrial and open-source NLP/LT leaders responsible for integration
- (Starting soon) Integrate results from ELG pilots, other ICT-29 projects, relevant e-infrastructures, and European and national projects

ELG Functional Content: Addressing Heterogeneity

- Variety of inputs
  - Text – at least plain text, some services can parse JSON/HTML/XML/PDF as well
  - Audio – 16 bit WAV is the de facto standard, MP3 supported by some APIs (lower bandwidth)
- Variety of outputs
  - “annotations” – standoff markup over regions of text/audio
  - Text, e.g., translations/transcriptions
  - Classifications – e.g., language ID
  - Audio (for Text-to-Speech)
- ELG approach – define common API for each “class” of services
  - Text to Text (MT/summarisation); Text to Annotations (IE/NER); Speech to Text (ASR), etc.

Existing tools: Automatic Speech Recognition

A – Official EU Languages (24)

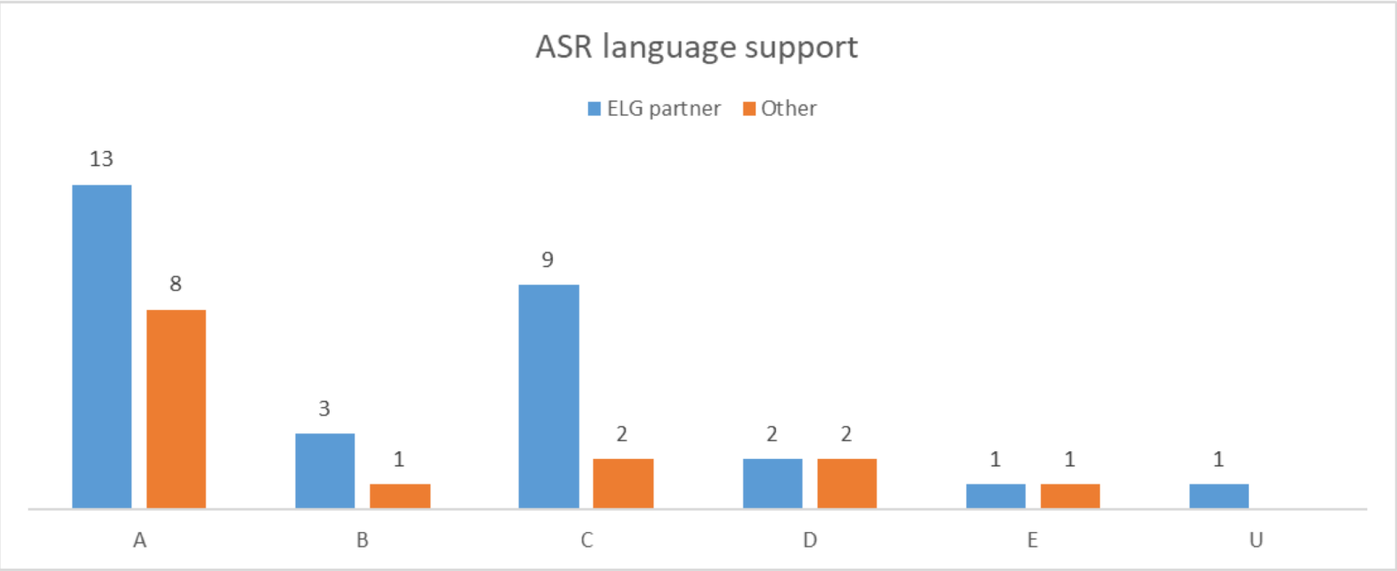
B – Other EU languages; languages from EU candidate countries and trade partners (11)

C – Languages spoken by EU immigrants; languages of important trade and political partners (18)

D – Other

E – Language independent

U – Not specified



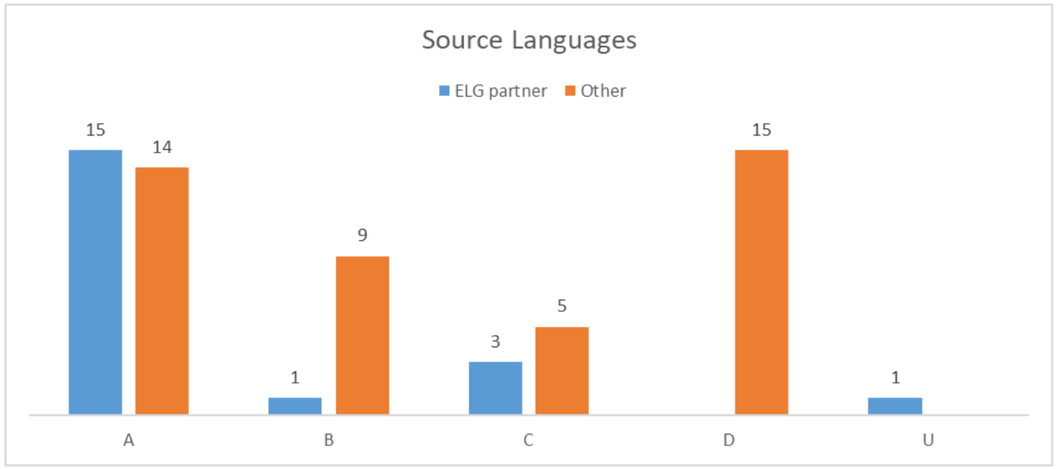
Language Category	ELG partner	Other	Disct. Languages	Coverage
A	13	8	**13**	**72%**
B	3	1	**4**	**22%**
C	9	2	**9**	**50%**
D	2	2	**4**	-
E	1	1	**1**	-
U	1	0	**1**	-
Total general	**29**	**14**	**32**	-

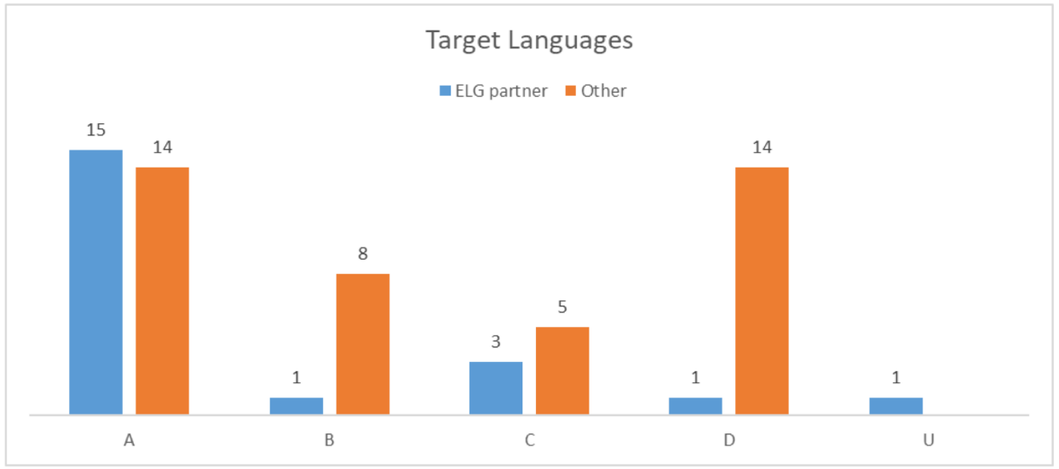
Language supported (total): 32

ELG partners: 29

Other: 14

Existing tools: Machine Translation

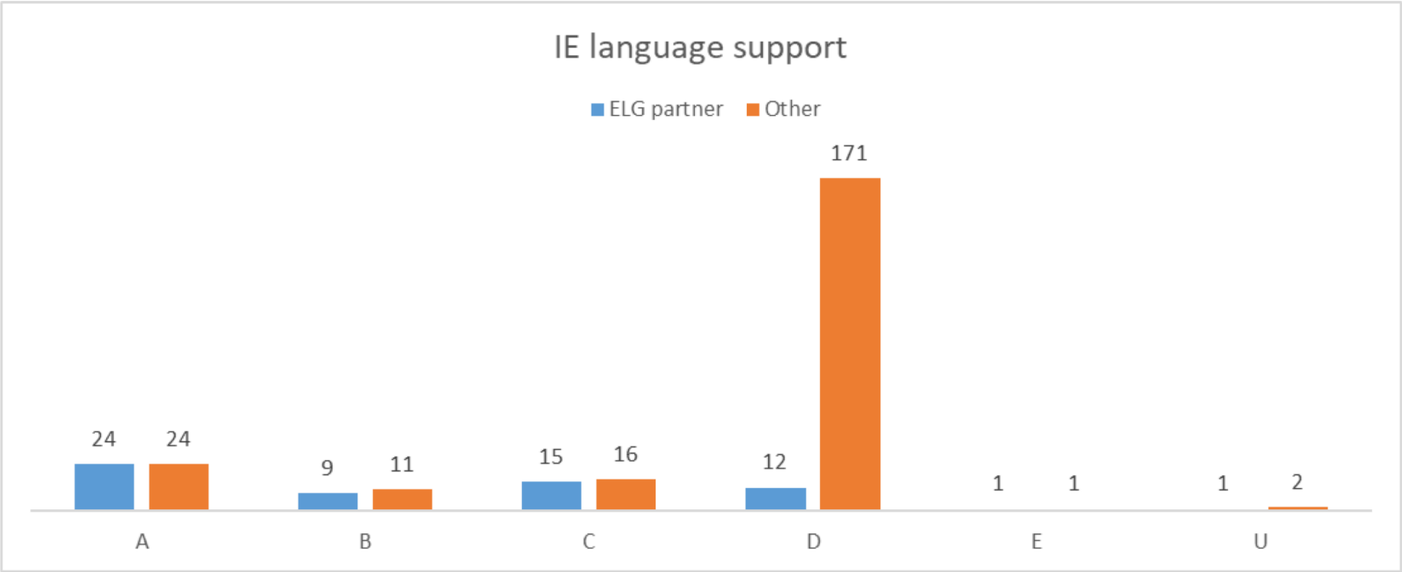




Categories	ELG partner	Other	Distinct Lang	Coverage
A	15	14	19	79%
B	1	9	9	82%
C	3	5	6	33%
D		15	15	-
U	1		1	-
Total	**20**	**43**	**50**	-

Categories	ELG partner	Other	Distinct Lang	Coverage
A	15	14	**19**	79%
B	1	8	**9**	82%
C	3	5	**6**	33%
D	1	14	**14**	-
U	1		**1**	-
Total	**21**	**41**	**49**	-

Existing tools: Information Extraction and Text Analysis



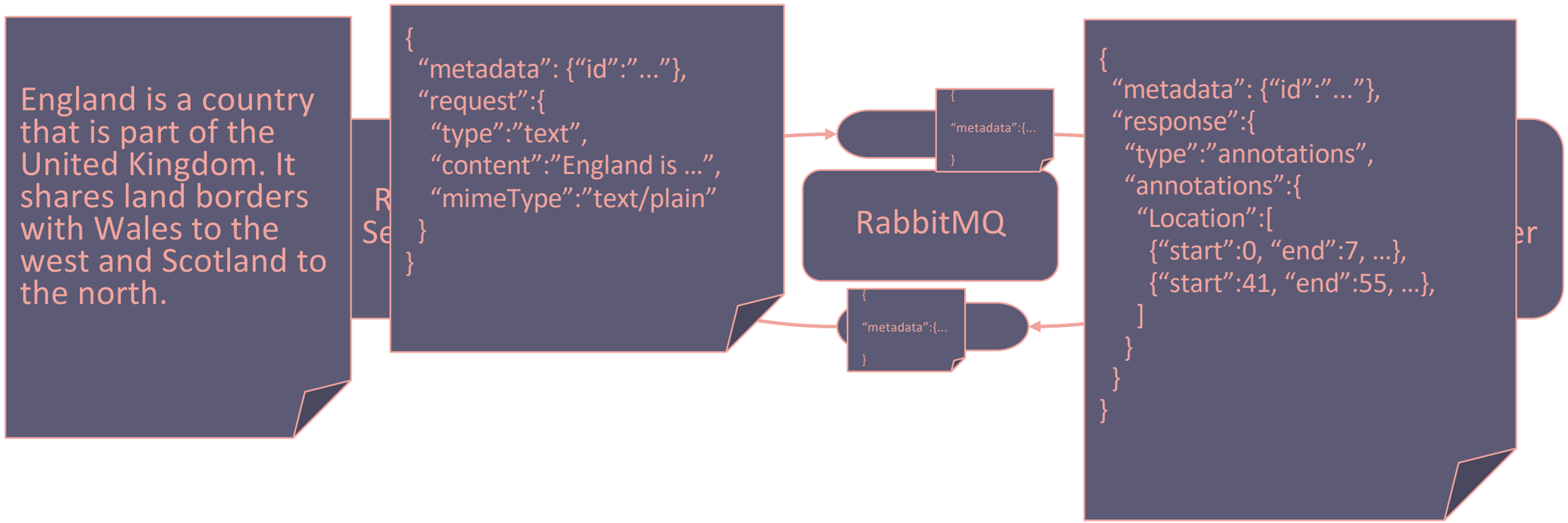
Language Category	ELG Partner	Other	Distinct Language	Coverage
A	24	24	24	**100%**
B	9	11	11	**100%**
C	15	16	17	**94%**
D	12	169	170	-
E	1	1	1	-
U	1	4	4	-
Total	**62**	**225**	227	-

"Other" tools with a large coverage of "D" languages

- Polyglot: 156 languages
  - Lang Identification (121), Morph analysis (90), Sentiment analysis (92)
- OpenNLP: 52 languages
  - Language Identification (52)
- TextBlob: 22 languages
  - Word / Noun Phrase frequencies (22)

API Design 1/3

- Message queueing approach with defined JSON schema for message content
- Message formats for requests, progress reports, successful and unsuccessful responses
  - Different message format defined for each input/output type (text, audio, annotations, etc.)
- Front end will handle all issues of user authentication, permissions, etc. – tools just need to know how to process messages



API Design 2/3

- Horizontally scalable – if too many waiting messages for service X, spin up another pod
- Long-running tools can provide progress update messages (20% done, 50%, …)
- i18n for errors – specified by code, lookup REST service to provide translations


```
{
  "code":"elg.service.internalError",
  "text":"Internal error during processing: {0}",
  "params":["IndexOutOfBounds"]
}
```


Error interno durante el procesamiento: IndexOutOfBounds

API Design 3/3

- Platform provides common public-facing APIs for each “category” of tools
  - IE – text in, annotations out
  - MT/summarisation – text in, text(s) out
  - ASR – audio in, text(s) out
- Currently exposes synchronous and polling-style APIs, future plans for batch-mode
- … but tools themselves don’t need to care – they just receive and respond to messages

Putting your own tools on ELG

- Current tools have taken anything from a few hours to a few days to integrate
  - Some are easier than others
- Hope to get this down across the board to minutes in the future
- We have helper libraries that deal with much of the RabbitMQ interaction, e.g.
  - Spring Boot Starter for Java – you provide one implementation class, the rest is boilerplate

```
@Component
@ElgHandler
public class HelloWorldHandler {

  @ElgMessageHandler
  public AnnotationsResponse process(TextRequest request) throws Exception {
    return new AnnotationsResponse().withAnnotations("Hello",
            Arrays.asList(new AnnotationObject()
                    .withOffsets(0,1)
                    .withFeatures("hello", "world")));
  }
}
```

Introducing the European Language Grid
Deep Dive 3 – ELG Content

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).

Khalid Choukri (ELDA), Ian Roberts (USFD), Kalina Bontcheva (USFD)
(additional content from Andres Garcia Silva, Expert System)

08/09-10-2019, Brussels – META-FORUM 2019
https://www.european-language-grid.eu