EUROPEAN LANGUAGE GRID

D4.3 Grid Content: Services, Tools & Components (Final Release)

Authors:	Ian Roberts (USFD), Andres Garcia Silva (EXPSYS) et al.
Dissemination Level:	Public
Date:	31-01-2022

About this document

Project	ELG – European Language Grid
Grant agreement no.	825627 – Horizon 2020, ICT 2018-2020 – Innovation Action
Coordinator	Prof. Dr. Georg Rehm (DFKI)
Start date, duration	01-01-2019, 42 months (GA amendment version: AMD-825627-7)
Deliverable number	D4.3
Deliverable title	Services, Tools & Components (Final Release)
Type	Report
Number of pages	61
Status and version	Final – Version 1.0
Dissemination level	Public
Date of delivery	Contractual: 31-01-2022 – Actual: 31-01-2022
WP number and title	WP4: Grid Content – Services, Tools & Components
Task number and title	All WP4 tasks
Authors	Ian Roberts (USFD), Andres Garcia Silva (EXPSYS), Miroslav Janosik (HENS), Nils Feldhus (DFKI), Dimitris Galanis (ILSP), Andis Lagzdiņš (Tilde), Rémi Calizzano (DFKI)
Reviewers	Jan Hajic (CUNI), Andrejs Vasiļjevs (Tilde)
Consortium	Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI), Germany
	Institute for Language and Speech Processing (ILSP), Greece
	University of Sheffield (USFD), United Kingdom
	Charles University (CUNI), Czech Republic
	Evaluations and Language Resources Distribution Agency (ELDA), France
	Tilde SIA (TILDE), Latvia
	Hensoldt Analytics (HENS), Austria
	Expert System Iberia SL (EXPSYS), Spain
	University of Edinburgh (UEDIN), United Kingdom
EC project officers	Philippe Gelin, Miklos Druskoczi
For copies of reports and other ELG-related information, please contact:	DFKI GmbH European Language Grid (ELG) Alt-Moabit 91c D-10559 Berlin Germany Prof. Dr. Georg Rehm, DFKI GmbH georg.rehm@dfki.de Phone: +49 (0)30 23895-1833 Fax: +49 (0)30 23895-1810 http://european-language-grid.eu © 2022 ELG Consortium

Table of Contents

List of Figures

List of Tables

Abstract

This document is an update to deliverable 4.2, describing the additional services, tools and components that have been integrated into the ELG platform since the second platform release. The focus of this part of the project has been the integration of more services provided by the WP6 pilot projects, along with additional services from consortium members to cover a wider range of non-EU languages. Significant work has also been undertaken to develop tools and libraries to simplify the process of service integration for consortium members, pilot projects and third parties alike.

1 Introduction

This document describes tools, services and components that have been developed and made available in preparation for the third public release of the ELG platform at the end of February 2022, the final release within the project runtime. It should be read in conjunction with the previous deliverables D4.1 and D4.2, which described the equivalent elements of the first two releases.

D4.1 detailed work carried out under Task 4.1 to identify existing tools, services and components from among the ELG project partners and prioritise which tools to integrate at which stages of the ELG platform development cycle. D4.2 gave updates to that plan at the time of the second release, and section 2 of this document describes how that plan has evolved since release 2.

Sections 3 to 6 describe the specific tools that have been integrated by consortium members and the ELG pilot projects for the third release, grouped as per the WP4 task breakdown – Automatic Speech Recognition (Task 4.2), Information Extraction and Text Analysis (Task 4.3), Machine Translation (Task 4.4) and other types of tools (Task 4.5). As expected the pilot projects have been the richest source of new services since release 2, particularly in the machine translation class.

Section 7 discusses the helper libraries that have been created by the ELG project to simplify the service integration process for third party LT service developers who want to add their tools to the ELG platform (Task 4.6). A summary of the final state of the catalogue as at the date of this deliverable (end of January 2022) is given in section 8.

2 Tools, Services and Components: Updates to the integration plan

Before starting the integration of the tools and services planned for release 3 we carried out a survey with the partners about the execution of the plan defined for release 2 and the current plan for release 3. We asked them to indicate whether the plan for release 1, 2 and 3 was modified and to what extent. In Table 1, Table 2, and Table 3 we present the modifications to the plan to integrate IE, MT, and ASR tools in ELG[1].

[1] Note that the plan to integrate tools in the category “Other tools” has not been modified since D4.2. Such plan does not include any tool to be integrated in R3 by consortium partners, though some services in this category have been supplied by pilot projects.

	R1	R2	R3	Total
Services in D4.2	152	295	166	613
New services	0	0	19	19
Removed services	0	1	11	12
Services in D4.3	152	294	174	620
Delta of services	0	+1	+8	+7

Table 1: Modifications to the integration plan of IE tools defined in D4.2

	R1	R2	R3	Total
Services in D4.2	15	31	9	55
New services	0	6	0	6
Removed services	0	8	3	11
Services in D4.3	15	29	6	50
Delta of services	0	-2	-3	-5

Table 2: Modifications to the integration plan of MT tools defined in D4.2.[2]

	R1	R2	R3	Total
Services in D4.2	7	8	15	30
New services	0	0	0	0
Removed services	0	0	1	1
Services in D4.3	7	8	14	29
Delta of services	0	0	-1	-1

Table 3: Modifications to the integration plan of ASR tools defined in D4.2.[3]

In short, the list of IE services increased by 7 the total number of services to be integrated in the project, while the lists of services for MT decreased by 5 and the ASR by 1. For MT, the six services added to R2 in this deliverable are ones that were described in D4.2 in the "detail" section but were inadvertently omitted from the integration plan tables. The MT and IE services removed from the plan were (a) some that had been expected to be delivered by other research projects, but those projects' work plans were disrupted by COVID, and (b) some that were based on now obsolete technologies and have been superseded by other services (e.g., MT models for the same language pairs supplied by the OPUS-MT pilot project, see section 5.2.1). In the case of the ASR tool removed from the R3 plan, this was in fact a duplicate entry in the database that should not have been counted in the first place.

In the following, we present the final timeline for the integration of IE tools and MT tools by members of the ELG consortium. In all these tables, we have grouped the supported human languages into four categories: (A) the 24 EU official languages; (B) other EU languages without official status, plus languages from candidate

[2] Tools removed from R2 were obsolote tools which have since been superseded by the work of the OPUS-MT pilot project. Tools removed in R3 are attached to ongoing projects that have not produced the expected tools yet.

[3] The only ASR service removed from R3 is due to a duplication of the entry in the database.

countries and free trade partners; (C) languages spoken by immigrants or important trade and political partners; (D) languages that do not fit (A), (B), or (C).

First in Table 4 we present the general overview of the services to be integrated in ELG broken down by service type and language category. In addition, Table 5 shows the overall number of *distinct languages* within each of the four categories that are covered by each type of service across all partners.

	A	B	C	D	Total
ASR	**12**	**3**	**11**	**1**	**27**[4]
HENS	**9**	**3**	**11**	**1**	**24**
Tilde	**2**				**2**
UEDIN	**1**				**1**
IE & Text Analysis	**381**	**58**	**153**	**26**	**618**[5]
CUNI	**122**	**35**	**64**		**221**
Dependency Parsing	24	7	13		44
Lemmatization	24	7	12		43
Morphological analyser	24	7	13		44
Named Entity Recognition	2				2
Part of Speech tagging	24	7	13		44
Tokenization	24	7	13		44
DFKI	**50**	**6**	**13**	**13**	**82**
Categorization	1				1
Date detection	2				2
Discourse Parsing	1				1
Language identification	22	6	13	13	54
Morphological analyser	6				6
Named Entity Recognition	3				3
Parsing	1				1
Sentence splitting	3				3
Summarization	5				5
Tokenization	3				3
Fake news dectection	1				1
Classification	2				2
Expert System	**73**	**4**	**42**	**13**	**132**
Keyword extraction	7		5		12
Language identification	22	4	12	13	51
Lemmatization	7		5		12
Named Entity Recognition	7		5		12
Part-of-Speech Tagging	7		5		12
Semantic annotation	7		5		12
Sentiment Analysis	7				7

[4] The difference of 2 between this number (27) and the total number of ASR services in Table 2 (29) is accounted for by 2 services that are considered language independent and therefore do not fit into any the A-D language cateogories.

[5] The difference of 2 between this number (618) and the total number of IE services given in Table 1 (620) is accounted for by 2 services that are considered language-independent and therefore do not fit into any of the A-D language categories.

	A	B	C	D	Total
Summarization	7		5		12
Text categorization	2				2
ILSP	**9**				**9**
Dependency Parsing	1				1
Information Extraction	2				2
Named Entity Recognition	3				3
Sentiment Analysis	1				1
Classification	1				1
Chunking	1				1
USFD	**80**	**3**	**4**		**87**
Categorization	6				6
Entity linking	2				2
Language identification	5				5
Measurement annotation	1				1
Measurement normalisation	1				1
Morphological analyser	1				1
Named Entity Recognition	15	1	2		18
NER Disambiguation	7				7
Noun phrase extraction	1				1
Number annotation	1				1
Number normalisation	1				1
Opinion Mining	2				2
Part of Speech tagging	22	2	2		26
Sentence splitting	3				3
Sentiment Analysis	3				3
Summarization	2				2
Tokenization	4				4
Extraction of domain-specific information	2				2
Event detection	1				1
HENS	**47**	**10**	**30**		**87**
Keyword extraction	9	3	9		21
Language identification	15	3	8		26
Named Entity Recognition	16	4	9		29
Sentiment Analysis	7		4		11
Other	**6**	**1**	**0**	**1**	**8**
DFKI	**4**	**1**	**0**	**1**	**6**
Text to Speech		1	0	1	6
Tilde	**2**				**2**
Text to Speech	2				2
MT (↓ From \ To →)	**46**		**3**	**1**	**50**
CUNI	**9**		**1**	**1**	**11**
A	8		1	1	10
C	1				1

	A	B	C	D	Total
ILSP	**2**				**2**
A	2				2
Tilde	**19**		**2**		**21**
A	18		2		20
C	1				1
UEDIN	**16**				**16**
A	16				16
Grand Total	**445**	**62**	**167**	**29**	**703**

Table 4: Services and supported languages per ELG partner.

	A	B	C	D
ASR				
ASR	12	3	11	1
IE & Text Analysis				
Morphological analyser	24	7	13	
Lemmatisation	24	7	12	
Part of Speech tagging	24	7	13	
Dependency Parsing	24	7	13	
Tokenization	24	7	13	
Language identification	22	6	14	13
Named Entity Recognition	16	5	11	
Keyword extraction	10	3	11	
Sentiment Analysis	9		4	
Summarization	7		5	
Part-of-Speech Tagging	7		5	
Semantic annotation	7		5	
Lemmatization	7		5	
Sentence splitting	4			
NER Disambiguation	4			
Entity linking	2			
Classification	2			
Text categorization	2			
Date detection	2			
Information Extraction	1			
Extraction of domain-specific information	1			
Categorization	1			
Noun phrase extraction	1			
Chunking	1			
Number annotation	1			
Event detection	1			
Number normalisation	1			
Discourse Parsing	1			

	A	B	C	D
Opinion Mining	1			
Fake news dectection	1			
Parsing	1			
Measurement annotation	1			
Measurement normalisation	1			
MT (↓ From \ To →)				
A	45		2	1
C	1			
Other				
Text to Speech	6	1		1

Table 5: Number of supported languages in each language category per service type.

2.1 IE tools updated plan

In this section we describe the the plan for the integration of IE tools after the partner modifications. An overview of the services that will be integrated in each release is presented in the following tables:

- Table 6 and Table 7 show the number of services integrated at the time of release 1 for each of the supported languages. Table 6 gives the overall total and Table 7 breaks this total down by project partner.
- Table 8 and Table 9 show the number of services that had been integrated by the time of release 2 for each supported language. This time there is no breakdown by partner, instead Table 8 lists the category A languages (EU official) and Table 9 lists the category B languages plus an overall total for categories C & D (which for this release are all language identification).
- Table 10 lists the remaining services for category C and D languages, which are being integrated as part of release 3.

In each table the cell number is the number of tools that will be integrated providing support for the service type. The full list of tools and their corresponding services to integrate in each release is in appendix A.

	Czech	Dutch	English	French	German	Greek	Italian	Latvian	Spanish	Total
Dependency Parsing	1		1	1	1	1		1	1	7
Information Extraction						2				2
Language identification	2		3	3	3	2		1	3	17
Lemmatization	1		2	2	2	1		1	2	11
Morphological analyser	1	1	3	2	2	1	1	1	2	14
Named Entity Recognition	2		12	4	8	2			2	30
NER Disambiguation			4	1	1				1	7
Number annotation			1							1
Opinion Mining			2							2
Part-of-Speech Tagging	2		5	3	3	2		2	3	20
Sentence splitting			2		2		1			5
Sentiment Analysis			2	2	2	1			2	9
Summarization			3	1	1				2	7
Text categorization			6						1	7
Tokenization	1		4	1	3	1	1	1	1	13
Total	10	1	50	20	28	13	3	7	20	152

Table 6: Overview of IE and Text Analysis services for the first release

	Czech	Dutch	English	French	German	Greek	Italian	Latvian	Spanish	Total
CUNI	**6**		**6**	**5**	**5**	**5**		**5**	**5**	**37**
Dependency Parsing	1		1	1	1	1		1	1	7
Lemmatization	1		1	1	1	1		1	1	7
Morphological analyser	1		1	1	1	1		1	1	7
Named Entity Recognition	1		1							2
Part-of-Speech Tagging	1		1	1	1	1		1	1	7
Tokenization	1		1	1	1	1		1	1	7
DFKI		**1**	**6**	**1**	**5**		**3**		**1**	**17**
Morphological analyser		1	1	1	1		1		1	6
Named Entity Recognition			1		2					3
Sentence splitting			1		1		1			3
Summarization			1							1
Text categorization			1							1
Tokenization			1		1		1			3
Expert System	**1**		**7**	**6**	**6**	**1**		**1**	**7**	**29**
Language identification	1		1	1	1	1		1	1	7
Lemmatization			1	1	1				1	4
Named Entity Recognition			1	1	1				1	4
Part-of-Speech Tagging			1	1	1				1	4
Sentiment Analysis			1	1	1				1	4
Summarization			1	1	1				1	4
Text categorization			1						1	2

	Czech	Dutch	English	French	German	Greek	Italian	Latvian	Spanish	Total
ILSP			**1**			**4**				**5**
Information Extraction						2				2
Named Entity Recognition			1			1				2
Sentiment Analysis						1				1
SAIL LABS	**2**		**3**	**3**	**3**	**2**			**3**	**16**
Language identification	1		1	1	1	1			1	6
Named Entity Recognition	1		1	1	1	1			1	6
Sentiment Analysis			1	1	1				1	4
USFD	**1**		**27**	**5**	**9**	**1**		**1**	**4**	**48**
Language identification			1	1	1				1	4
Morphological analyser			1							1
Named Entity Recognition			7	2	4					13
NER Disambiguation			4	1	1				1	7
Number annotation			1							1
Opinion Mining			2							2
Part-of-Speech Tagging	1		3	1	1	1		1	1	9
Sentence splitting			1		1					2
Summarization			1						1	2
Text categorization			4							4
Tokenization			2		1					3
Total	**10**	**1**	**50**	**20**	**28**	**13**	**3**	**7**	**20**	**152**

Table 7: IE and Text Analysis services provided by ELG partners for the first release

Service/Language	**Bulgarian**	**Croatian**	**Czech**	**Danish**	**Dutch**	**English**	**Estonian**	**Finnish**	**French**	**German**	**Greek**	**Hungarian**	**Irish**	**Italian**	**Latvian**	**Lithuanian**	**Maltese**	**Polish**	**Portuguese**	**Romanian**	**Slovak**	**Slovenian**	**Spanish**	**Swedish**	**Total**
Date detection						1				1															**2**
Dependency Parsing	1	1		1	1		1	1				1	1	1		1	1	1	1	1	1	1		1	**17**
Discourse Parsing										1															**1**
Entity linking						1				1															**2**
Keyword extraction					2	2			2	2				2				1	1	1			2		**15**
Language identification	3	2	1	2	4	1	2	2	1	1	1	3		3	1	2		3	3	3	3	2	1	3	**47**
Lemmatisation	1	1		1	1		1	1				1	1	1		1	1	1	1	1	1	1		1	**17**
Lemmatization					1									1					1						**3**
Measurement annotation						1																			**1**
Measurement normalization						1																			**1**
Morphological analyser	1	1		1	1		1	1				1	1	1		1	1	1	1	1	1	1		1	**17**
Named Entity Recognition	1	1			3	1						1		2				1	2	2	1			1	**16**
Noun phrase extraction						1																			**1**
Number normalisation						1																			**1**
Part of Speech tagging	2	2		2	3		2	2				1	1	1		1	1	2	2	2	2	2		2	**30**
Part-of-Speech Tagging					1									1					1						**3**
Semantic annotation					1	1			1	1				1					1				1		**7**
Sentence splitting					1																				**1**
Sentiment Analysis					1									2				1	2						**6**
Summarization					1					1				1					1						**4**
Tokenization	1	1		1	2		1	1				1	1	1		1	1	1	1	1	1	1		1	**18**
Grand Total	**10**	**9**	**1**	**8**	**23**	**11**	**8**	**8**	**4**	**8**	**1**	**9**	**5**	**18**	**1**	**7**	**5**	**12**	**18**	**12**	**10**	**8**	**4**	**10**	**210**

Table 8: Overview of IE and Text Analysis services integrated in the second release (category A)

	Albanian	**Basque**	**Catalan**	**Galician**	**Norwegian**	**Serbian**	**Turkish**	**Ukrainian**	**Welsh**	**Other (C&D)**	**Total**
Dependency Parsing		1	1	1	1	1	1	1			**7**
Keyword extraction	1				1		1				**3**
Language identification	3		1		3		3	2	1	26	**39**
Lemmatization		1	1	1	1	1	1	1			**7**
Morphological analyser		1	1	1	1	1	1	1			**7**
Named Entity Recognition	1		1		1		1		1		**5**
Part-of-Speech Tagging		2	2	1	1	1	1	1			**9**
Tokenization		1	1	1	1	1	1	1			**7**
Total	**5**	**6**	**8**	**5**	**10**	**5**	**10**	**7**	**2**	**26**	**84**

Table 9: Overview of IE and Text Analysis services integrated in the second release (category B, C & D)

Service/Language	**Afrikaans**	**Arabic**	**Bengali**	**Chinese**	**English**	**French**	**German**	**Greek**	**Gujarati**	**Hebrew**	**Hindi/Urdu**	**Indonesian**	**Japanese**	**Kannada**	**Korean**	**Language Ind.**	**Latin**	**Macedonian**	**Malay**	**Marahati**	**Nepali**	**Panjabi**	**Pashto**	**Persian**	**Russian**	**Somali**	**Swahili**	**Tagalog**	**Tamil**	**Telugu**	**Thai**	**Urdu**	**Vietnamese**	**Total**
Categorization					2																													**2**
Chunking								1																										**1**
Classification							2	1																										**3**
Dependency Parsing	1	1		1				1		1	1	1	1		1		1							1	1				1				1	**14**
Event detection					1																													**1**
Extraction of domain-specific information					2																													**2**
Fake news dectection							1																											**1**
Keyword extraction		2		2						1	1	1	1		1				1				1	1	2									**14**
Language identification	1	2	1	1					1	2	2	2	1	1	1	1		1	1	1	1	1	1	2	2	1	1	1	1	1	1	1	1	**34**
Lemmatisation	1	1		1						1	1	1	1				1							1	1				1				1	**12**
Lemmatization		1		1									1		1										1									**5**
Morphological analyser	1	1		1						1	1	1	1		1		1							1	1				1				1	**13**
Named Entity Recognition		2		2	1					1	1	1	1		1				1				1	1	4									**17**
Parsing							1																											**1**
Part of Speech tagging	1	1		1						1	1	2	1		1		1							1	2				1				1	**15**
Part-of-Speech Tagging		1		1									1		1										1									**5**
Semantic annotation		1		1									1		1										1									**5**
Sentiment Analysis		1			3							1							1						1									**7**
Summarization		1		1	1	1	1						1		1	1									1									**9**
Tokenization	1	1		1						1	1	1	1		1		1							1	1				1				1	**13**
Grand Total	**6**	**16**	**1**	**14**	**10**	**1**	**5**	**3**	**1**	**9**	**9**	**11**	**12**	**1**	**11**	**2**	**5**	**1**	**4**	**1**	**1**	**1**	**3**	**9**	**19**	**1**	**1**	**1**	**6**	**1**	**1**	**1**	**6**	**174**

Table 10: Overview of IE and Text Analysis services integrated in the third release (category C and D)

2.2 Machine Translation Tools Updated Plan

The list of MT tools and services in each stage after the partner modification is presented in Table 11, Table 12 and Table 13 (for releases 1, 2 and 3 respectively).

Release 1					
Provider	**Service**	**From**	**Category**	**To**	**Category**
CUNI	Machine Translation	Czech	A	English	A
CUNI	Machine Translation	English	A	Czech	A
CUNI	Machine Translation	English	A	French	A
CUNI	Machine Translation	French	A	English	A
UEDIN	Machine Translation	Czech	A	English	A
UEDIN	Machine Translation	English	A	Czech	A
UEDIN	Machine Translation	English	A	German	A
UEDIN	Machine Translation	German	A	English	A
Tilde	Machine Translation	English	A	Bulgarian	A
Tilde	Machine Translation	English	A	Latvian	A
Tilde	Machine Translation	English	A	Polish	A
Tilde	Machine Translation	Latvian	A	English	A
Tilde	Machine Translation	Polish	A	English	A
ILSP	Machine Translation	Greek	A	English	A
ILSP	Machine Translation	English	A	Greek	A

Table 11: MT tools and services in the first release

Release 2					
Provider	**Service**	**From**	**Category**	**To**	**Category**
CUNI	Machine Translation	German	A	English	A
CUNI	Machine Translation	English	A	German	A
CUNI	Machine Translation	Polish	A	English	A
CUNI	Machine Translation	English	A	Polish	A
CUNI	Machine Translation	Russian	C	English	A
CUNI	Machine Translation	English	A	Russian	C
Tilde	Machine Translation	Bulgarian	A	English	A
Tilde	Machine Translation	Danish	A	English	A
Tilde	Machine Translation	English	A	Danish	A
Tilde	Machine Translation	English	A	Estonian	A
Tilde	Machine Translation	English	A	Finnish	A
Tilde	Machine Translation	English	A	Lithuanian	A
Tilde	Machine Translation	English	A	Swedish	A
Tilde	Machine Translation	Estonian	A	English	A
Tilde	Machine Translation	Finnish	A	English	A
Tilde	Machine Translation	Lithuanian	A	English	A
Tilde	Machine Translation	Swedish	A	English	A
Tilde	Machine Translation	English	A	German	A

Tilde	Machine Translation	German	A	English	A
ILSP	Machine Translation	Greek	A	English	A
ILSP	Machine Translation	English	A	Greek	A
UEDIN	Machine Translation	English	A	Estonian	A
UEDIN	Machine Translation	English	A	Latvian	A
UEDIN	Machine Translation	English	A	Portuguese	A
UEDIN	Machine Translation	English	A	Romanian	A
UEDIN	Machine Translation	English	A	Spanish	A
UEDIN	Machine Translation	Estonian	A	English	A
UEDIN	Machine Translation	Latvian	A	English	A
UEDIN	Machine Translation	Portuguese	A	English	A
UEDIN	Machine Translation	Romanian	A	English	A
UEDIN	Machine Translation	Spanish	A	English	A

Table 12: MT tools and services integrated in the second release

Release 3					
Provider	**Service**	**From**	**Category**	**To**	**Category**
CUNI	Machine Translation	English	A	Hindi	D
UEDIN/Bergamot	Machine Translation	English	A	Polish	A
UEDIN/Bergamot	Machine Translation	Polish	A	English	A
Tilde	Machine Translation	English	A	Arabic	C
Tilde	Machine Translation	Russian	C	English	A
Tilde	Machine Translation	English	A	Russian	C
ILSP	Machine Translation	Greek	A	English	A
ILSP	Machine Translation	English	A	Greek	A

Table 13: MT tools and services integrated in the third release

3 ASR Tools, Services and Components (Task 4.2)

3.1 ASR services from HENSOLDT Analytics

HENSOLDT Analytics brings a set of ASR components to the ELG platform which are based on HENS's Media Mining Indexer (MMI, now part of the HENSOLDT Analytics System[6]).

The MMI has been described more fully in D4.1, in summary it is a component combining a set of audio- and text-based processing sub-components which are connected internally in a pipelined manner (these connections can be configured to support different setups). All models can be updated in a transparent manner even during processing, though to date within the ELG, only a single model per container and no dynamic updating of models is supported. For the dockerized version of this component, dynamic updating and refreshing may easily be implemented by the creation of new versions of the container.

[6] https://www.hensoldt-analytics.com/hensoldt-analytics-system/

For the third release, ASR was integrated for 12 additional languages: Arabic, Egyptian Arabic, Levantine Arabic, Indonesian, Malay, Urdu, Persian (Farsi), Hebrew, Russian, Pashto, Mexican Spanish, Chinese (Mandarin). All these models have been developed by HENSOLDT Analytics.

ASR can be configured based on a set of parameters and optimized for speed or accuracy, and a variety of different domain-independent or domain-specific models are available. The current implementation uses the "base settings". In the future, it is planned to provide a series of domain-dependent models (for improved accuracy) as well as to allow settings for real-time (or faster if run from file) transcription.

License: HENS ASR is a commercial component that requires a commercial license to use it. While the ELG platform is under construction and there is no billing and payment option in place, the services are available free for testing and development purposes, request limits can be introduced at any time.

Deployed components in previous releases

Six of these components were deployed in the first release (French, English, German, Spanish, Greek, Latvian, Czech), seven more were added for second release (Norwegian, Romanian, Albanian, Italian, Turkish, Polish, Dutch).

Table 14 shows the ASR software components that were added in the third release. The code and Docker images are hosted in the ELG project repository and the corresponding container registry in gitlab[7], however all are private projects as they are commercial software.

Type	Tool Type	Image name	Service	Languages/variants supported	Licence
tool	ASR	sail-asr-ar-ar	Automatic Speech Recognition	Arabic	HENS
tool	ASR	sail-asr-ar-eg	Automatic Speech Recognition	Egyptian Arabic	HENS
tool	ASR	sail-asr-ar-lb	Automatic Speech Recognition	Levantine Arabic	HENS
tool	ASR	sail-asr-id-id	Automatic Speech Recognition	Indonesian	HENS
tool	ASR	sail-asr-ms-my	Automatic Speech Recognition	Malay	HENS
tool	ASR	sail-asr-ur-pk	Automatic Speech Recognition	Urdu	HENS
tool	ASR	sail-asr-fa-ir	Automatic Speech Recognition	Persian (Farsi)	HENS
tool	ASR	sail-asr-he-il	Automatic Speech Recognition	Hebrew	HENS
tool	ASR	sail-asr-ru-ru	Automatic Speech Recognition	Russian	HENS
tool	ASR	sail-asr-ps-af	Automatic Speech Recognition	Pashto	HENS
tool	ASR	sail-asr-es-mx	Automatic Speech Recognition	Mexican Spanish	HENS
tool	ASR	sail-asr-zh-cn	Automatic Speech Recognition	Chinese (Mandarin)	HENS[8]

Table 14: HENS ASR tools integrated in R3

3.2 ASR services from pilot projects

In addition to the services from within the ELG project consortium, ASR services have been provided by two pilot project organizations.

[7] Code is at https://gitlab.com/european-language-grid/sail/<image name> and images are registry.gitlab.com/european-language-grid/sail/<image name> for each image name in the table – repositories retain the "sail" name as it is technically difficult to rename them without adversely affecting other areas of the ELG platform

[8] Following the acquisition of SAIL LABS by HENSOLDT, the licenses were modified but are identical except for the naming.

Elhuyar Fundazioa have contributed a single ASR service for the Basque language. This service is a proxy for Elhuyar's own aditu[9] commercial speech recognition service, and is intended to allow users to access a subset of aditu functionality for evaluation and research purposes.

Lingsoft contributed four ASR services, one for Swedish, one for Norwegian (transcribing in Bokmål), and two for Finnish. As with Elhuyar, the Lingsoft services in ELG are proxies to an underlying service hosted on the provider's own hardware, and are intended for evaluation and research use only.

Particularly noteworthy is the fact that one of the Finnish services is for general-domain speech but the other is tuned specifically for medical dictation. This is one of the first examples of domain-specific (as opposed to just language-specific) ASR in the ELG catalogue.

4 IE Tools, Services and Components (Task 4.3)

4.1 IE tools from DFKI

For Release 3, the DFKI team in Berlin (Speech and Language Technology) provided five more tools.

These tools originate from three German projects coordinated by DFKI: QURATOR, PANQURA, and MACSS. The ELG team of DFKI assisted in service deployment with understanding REST API specifications and making metadata records available.

Table 15 describes the tools by DFKI deployed for Release 3. It presents the five services. Three of the four services are text classifiers: the first one gives credibility scores for English news articles, the second one focuses on topic classification for German texts, and the last one classifies the political bias of a German text into 5 classes. One of the services provided by DFKI is a summarization service for English, German, and French focusing on the generation of extended summaries similar to small news articles. The last service is a dependency tree parser for German clinical texts.

Name	Description	Languages supported	Licence	Code repository
Credibility score	Computes credibility scores for a given news article, especially for content related to COVID-19.	English	Creative Commons Zero v1.0 Universal	https://github.com/konstantinschulz/alpaca
Document classification	The service classifies German texts into various topics. It was trained on book cover texts (blurbs) from the GermEval 2019 Shared Task 1 (Hierarchical Classification of Blurbs). The topic labels originate from the Random House book collection. Each output contains confidence levels for the various topics. Higher values correspond to higher confidence, so the topic with the highest score is the predicted choice.	German	Creative Commons Zero v1.0 Universal	https://github.com/konstantinschulz/text_type_classifier

[9] https://aditu.eus

Name	Description	Languages supported	Licence	Code repository
Multi-document summarizer	Summarization service using DistilBart (sequence-to-sequence Transformers) trained on WikinewsSum. WikinewsSum is a corpus created from Wikinews where the texts to summarize are the texts of the sources of the Wikinews article, and the summary to obtain is the Wikinews article. The service creates long summaries.	English, French, German	MIT License	https://github.com/airKlizz/MultiDocMultiLingualSum
Political Bias Classifier	The model classifies the political bias of a German text into 5 classes: far-left, center-left, center, center-right, far-right. It uses a TF-IDF vectorizer to preprocess documents. Then, a Random Forest classifier is applied on the resulting vectors to determine the final class.	German	Creative Commons Zero v1.0 Universal	https://github.com/konstantinschulz/politik-news
Dependency Tree Parser for German Clinical Text	The dependency parsing model was re-trained with the neurail network dependency tree parser using the clinical data gathered during the project.	German	MIT License	https://gitlab.com/european-language-grid/dfki/dependency-tree-parser-for-german-clinical-text

Table 15: IE tools provided by DFKI for R3

4.2 IE Tools Integrated by EXPSYS

For the third release, services integrated in release 1 and 2 (NER, PoS tagging, lemmatization, summarization, sentiment analysis, keyword extraction and semantic annotation) have been enriched with support for new languages in Category C: Arabic, Chinese, Japanese, Korean and Russian. New languages in Categories C and D are also included for the language identification service (see Table 16).

Deployed components update

Table 16 shows an update of the software components described in deliverable D4.2. New supported languages from the previous release are marked in **bold.**

Type	Image name	Service	Languages	Container registry	Code repository
Tool	cogito-discover	Provides different services via adapters	see adapters	expert-system/cogito-discover	Non-available (Commercial software)
Adapter	cogito-discover-general-adapter	Named Entity recognition	English, French, German, Spanish, Italian, Portuguese, Dutch, **Arabic, Chinese, Japanese, Korean, Russian**	expertsystem/cogito-discover-general-adapter	expertsystem/tree/master/cogito-discover-general-adapter

Type	Image name	Service	Languages	Container registry	Code repository
Adapter	cogito-discover-general-adapter	Part-of-speech tagging	English, French, German, Spanish, Italian, Portuguese, Dutch, , **Arabic, Chinese, Japanese, Korean, Russian**	expertsystem/cogito-discover-general-adapter	expertsystem/tree/master/cogito-discover-general-adapter
Adapter	cogito-discover-general-adapter	Text categorization	English, Spanish	expertsystem/cogito-discover-general-adapter	expertsystem/tree/master/cogito-discover-general-adapter
Adapter	cogito-discover-general-adapter	Lemmatization	English, French, German, Spanish, Italian, Portuguese, Dutch, **Arabic, Chinese, Japanese, Korean, Russian**	expertsystem/cogito-discover-general-adapter	expertsystem/tree/master/cogito-discover-general-adapter
Adapter	cogito-discover-general-adapter	Summarization	English, French, German, Spanish, Italian, Portuguese, Dutch, **Arabic, Chinese, Japanese, Korean, Russian**	expertsystem/cogito-discover-general-adapter	expertsystem/tree/master/cogito-discover-general-adapter
Adapter	cogito-discover-general-adapter	Sentiment analysis	English, French, German, Spanish, Italian, Portuguese, Dutch, **Arabic, Chinese, Japanese, Korean, Russian**	expertsystem/cogito-discover-general-adapter	expertsystem/tree/master/cogito-discover-general-adapter

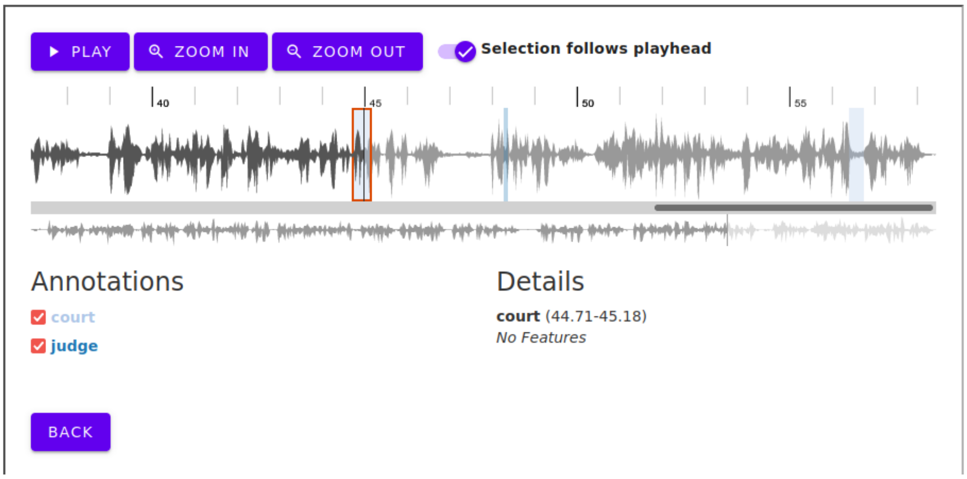
Type	**Image name**	**Service**	**Languages**	**Container registry**	**Code repository**
Adapter	cogito-discover-general-adapter	Language detection	Czech, English, French, German, Spanish, Latvian, Greek, Bulgarian, Croatian, Danish, Dutch, Estonian, Finnish, Hungarian, Italian, Lithuanian, Polish, Portuguese, Romanian, Slovak, Slovenian, Swedish, Albanian, Norwegian, Turkish, Ukrainian, **Afrikaans, Arabic, Chinese, Hebrew, Hindi/Urdu, Indonesian, Japanese, Korean, Persian, Russian, Tamil, Vietnamese, Bengali, Gujarati, Kannada, Macedonian, Marahati, Nepali, Panjabi, Somali, Swahili, Tagalog, Telugu, Thai, Urdu**	expertsystem/cogito-discover-general-adapter	expertsystem/tree/master/cogito-discover-general-adapter
Adapter	cogito-discover-general-adapter	Keyword extraction	English, French, German, Spanish, Italian, Portuguese, Dutch, **Arabic, Chinese, Japanese, Korean, Russian**	expertsystem/cogito-discover-general-adapter	expertsystem/tree/master/cogito-discover-general-adapter
Adapter	cogito-discover-general-adapter	Semantic annotation	English, French, German, Spanish, Italian, Portuguese, Dutch, **Arabic, Chinese, Japanese, Korean, Russian**	expertsystem/cogito-discover-general-adapter	expertsystem/tree/master/cogito-discover-general-adapter

Table 16: Cogito Discover software components for deployment in ELG

4.3 IE services from HENSOLDT Analytics

HENSOLDT Analytics brings to the ELG platform a set of IE components which are based on the same SAIL Media Mining Indexer (MMI) that is the basis for the ASR tools introduced in section 3.1, as well as an individual tool for the identification of language from text and for text summarization.

Within the set of IE tools of ELG, HENS provides the MMI for the recognition of named-entities (NER) and for Sentiment Analysis (SA), Keyword Spotting/Extraction (KWS) for detecting keywords in speech, Gender-detection (GEN) and Age-detection (AGE) for detecting Gender or Age of speaker from speech. Note that while we describe the KWS, GEN and AGE services in this "information extraction" section, they are not *text* analytics services; they extract information from *audio* rather than text and return their annotations with start and end points measured in seconds of audio rather than numbers of characters. Figure 1 shows how this appears in the "try out" tab of the ELG catalogue.



☑ court
☑ judge

court (44.71-45.18)
No Features

Figure 1: "Try out" GUI for KWS audio annotation service

The component for Summarization of text (SUM) is based on a component part of a suite of tools for the processing of (textual) data from the Internet and Social Media.

The component for Language identification (LID) in this release is based on the fasttext library and trained model located at https://fasttext.cc/docs/en/language-identification.html - this is a change from the version included in the previous release, which used internally developed HENS components. It is based on papers by Joulin et al (2016 & 2017). LID returns *up to* three most-probable languages for a given text, possibly fewer than three or even none at all, if no language meets the probability threshold.

In the first ELG release HENS (then SAIL) provided:

- NER for 6 languages: Czech, English, French, German, Greek, and Spanish.
- SA for 4 languages: English, French, German and Spanish.
- LID for 6 languages: Czech, English, French, German, Greek, and Spanish.

In second release there was added:

- NER for an additional 14 languages: Bulgarian, Croatian, Dutch, Hungarian, Italian, Polish, Portuguese, Romanian, Slovak, Swedish, Albanian, Catalan, Norwegian, Turkish.
- SA for 3 languages: Italian, Polish, Portuguese.
- LID for +25 languages: Bulgarian, Dutch, Hungarian, Italian, Polish, Portuguese, Romanian, Slovak, Swedish, Albanian, Norwegian, Turkish
- KWS for 12 languages: Dutch, English, French, German, Greek, Italian, Polish, Romanian, Spanish, Albanian, Norwegian, Turkish

In current third release there was added:

- NER for an additional 9 languages: Arabic, Chinese, Hebrew, Hindi/Urdu, Indonesian, Malay, Pashto, Persian, Russian
- SA for an additional 4 languages: Arabic, Indonesian, Malay, Russian.
- LID for detecting languages: Arabic, Hebrew, Hindi/Urdu, Indonesian, Malay, Pashto, Persian, Russian
- KWS for an additional 9 languages: Arabic, Chinese, Hebrew, Hindi/Urdu, Indonesian, Malay, Pashto, Persian, Russian
- SUM – Summarization: language-independent
- GEN – Gender-detection: language-independent
- AGE – Age-detection: language-independent

NER can be configured to be based on a set of patterns as well as to work on a set of features based on the sequence of words and morphological features. Currently, only the pattern-based functionality has been included. It contains rudimentary morphological processing. NER within the Media Mining environment is typically used in combination with tokenization and text-cleaning components which have not been integrated into NER implementation for ELG (but this is planned for future updates of the respective components).

SA is based on a set of patterns of “sentiment carrying words and expressions” and performs a 4-way categorization into the classes “positive”, “negative”, “neutral” and “mixed”.

KWS is based on analysing the speech-to-text lattice generated from an audio segment and searching through it for specific words. These keywords are specified as a parameter for the service with input audio file and lattice cutoff thresholds.

SUM is based on algorithm detecting patterns in text to select key points into smaller summary text containing only few sentences.

AGE and GEN are based on detection of speaker voice properties which is employed within the HENSOLDT Media Mining System to attempt to determine gender and age of speaker and confidence value of the assumption.

Deployed components

Table 17 shows the software components that are deployed in ELG platform in third release. The code and Docker images are hosted in the ELG project repository and the corresponding container registry in gitlab[10], however all are private projects as they are commercial software.

Type	Tool Type	Image name	Service	Languages/ variants supported	Licence
tool	NER	sail-ned-ar-ar	Detection of named-entities	Arabic	HENS
tool	NER	sail-ned-zh-cn	Detection of named-entities	Chinese	HENS
tool	NER	sail-ned-he-il	Detection of named-entities	Hebrew	HENS
tool	NER	sail-ned-ur-pk	Detection of named-entities	Hindi/Urdu	HENS
tool	NER	sail-ned-id-id	Detection of named-entities	Indonesian	HENS
tool	NER	sail-ned-ms-my	Detection of named-entities	Malay	HENS

[10] Code is at https://gitlab.com/european-language-grid/sail/<image name> and images are registry.gitlab.com/european-language-grid/sail/<image name> for each image name in the table – repositories retain the “sail” name as it is technically difficult to rename them without adversely affecting other areas of the ELG platform

Type	Tool Type	Image name	Service	Languages/ variants supported	Licence
tool	NER	sail-ned-ps-af	Detection of named-entities	Pashto	HENS
tool	NER	sail-ned-fa-ir	Detection of named-entities	Persian	HENS
tool	NER	sail-ned-ru-ru	Detection of named-entities	Russian	HENS
tool	SA	sail-sed-ar-ar	Sentiment Detection	Arabic	HENS
tool	SA	sail-sed-id-id	Sentiment Detection	Indonesian	HENS
tool	SA	sail-sed-ms-my	Sentiment Detection	Malay	HENS
tool	SA	sail-sed-ru-ru	Sentiment Detection	Russian	HENS
tool	LID	sail-lid	Language ID from text	English, German, French, Spanish, Greek, Czech, Bulgarian, Dutch, Hungarian, Italian, Polish, Portuguese, Romanian, Slovak, Swedish, Albanian, Norwegian, Turkish, Arabic, Hebrew, Hindi/Urdu, Indonesian, Malay, Pashto, Persian, Russian	HENS
tool	KWS	sail-kws-ar-ar	Keyword extraction	Arabic	HENS
tool	KWS	sail-kws-zh-cn	Keyword extraction	Chinese	HENS
tool	KWS	sail-kws-he-il	Keyword extraction	Hebrew	HENS
tool	KWS	sail-kws-ur-pk	Keyword extraction	Hindi/Urdu	HENS
tool	KWS	sail-kws-id-id	Keyword extraction	Indonesian	HENS
tool	KWS	sail-kws-ms-my	Keyword extraction	Malay	HENS
tool	KWS	sail-kws-ps-af	Keyword extraction	Pashto	HENS
tool	KWS	sail-kws-fa-ir	Keyword extraction	Persian	HENS
tool	KWS	sail-kws-ru-ru	Keyword extraction	Russian	HENS
tool	SUM	sail-sum	Text summarization	independent	HENS
tool	AGE	sail-age	Detection of speaker age	independent	HENS
tool	GEN	sail-gen	Detection of speaker gender	independent	HENS[11]

Table 17: HENS IE tools integrated in R3

4.4 IE Tools Integrated by ILSP

In the first of release of the ELG platform, the Insitute of Language and Speech Processing (ILSP) contributed 5 IE tools. No tools were added in the second release. For the third release, ILSP has registered to the platform the following 4 IE tools:

Service/Application	Language	Proxy image location	Docker image location	License
ILSP neural dependency parser https://live.european-language-grid.eu/catalogue/tool-service/16516	Greek	registry.gitlab.com/ilsp-nlpli-elg/elg-ilsp-lt-services/proxy	registry.gitlab.com/ilsp-nlpli-elg/elg-ilsp-lt-services/neural	ILSP ToS[12]

[11] Following the acquisition of SAIL LABS by HENSOLDT, the licenses will be modified but are expected to remain identical except for the naming.

[12] https://gitlab.com/european-language-grid/ilsp/elg-ilsp-lt-services-info/-/blob/master/LICENSE

ILSP neural named entity recognizer https://live.european-language-grid.eu/catalogue/tool-service/9480	Greek	registry.gitlab.com/ilsp-nlpli-elg/elg-ilsp-lt-services/proxy	registry.gitlab.com/ilsp-nlpli-elg/elg-ilsp-lt-services/neural	ILSP ToS
ILSP neural text classifier https://live.european-language-grid.eu/catalogue/tool-service/9481	Greek	registry.gitlab.com/ilsp-nlpli-elg/elg-ilsp-lt-services/proxy	registry.gitlab.com/ilsp-nlpli-elg/elg-ilsp-lt-services/neural	ILSP ToS
ILSP neural chunker https://live.european-language-grid.eu/catalogue/tool-service/9479	Greek	registry.gitlab.com/ilsp-nlpli-elg/elg-ilsp-lt-services/proxy	registry.gitlab.com/ilsp-nlpli-elg/elg-ilsp-lt-services/neural	ILSP ToS

Table 18. IE tools integrated by ILSP in R3

The tools are based on neural nets and have been trained on appropriate data. All of them have been packaged to the same Docker image (registry.gitlab.com/ilsp-nlpli-elg/elg-ilsp-lt-services/neural) that follows the respective ELG specifications; i.e., the tools are offered as REST services that produce and consume messages in the ELG format. The image is deployed at a dedicated server in ILSP and the ELG platform communicates with it via a proxy container (registry.gitlab.com/ilsp-nlpli-elg/elg-ilsp-lt-services/proxy) that has been deployed to the ELG cluster, specifically in “elg-srv-ilsp”, a private namespace which was created for deploying only ILSP services/components.

4.5 IE Tools Integrated by USFD

The University of Sheffield (USFD) provides a variety of text analysis tools based on the open-source GATE text processing framework[13], alongside a public platform (GATE Cloud) through which some of these tools can be called as services on the web[14]. Much of the work in the first year of the ELG project involved creating tools to allow for any GATE Cloud service to be exposed via the ELG platform, then using these tools to expose a subset of existing GATE cloud pipelines on the ELG. For the second and third ELG releases these tools were used to extend the set of services made available on the ELG platform.

4.5.1 Integration approach

As part of the first release, two solutions were developed in order to integrate GATE tools into the ELG framework. These were

A proxy component: effectively a bridge between a GATE Cloud pipeline and the ELG. The component accepts requests in the ELG API format, converts the request to a GATE-compatible request and dispatches it to the appropriate GATE Cloud endpoint. The result of this is then translated to an ELG compatible response. This component is packaged as a docker image that runs as a container within the ELG cluster. The code for this is publicly available in the ELG GitLab namespace[15], but requires GATE Cloud access credentials to operate. For the final release this proxy has been re-implemented on top of a new underlying framework[16] to be a much smaller

[13] https://gate.ac.uk
[14] https://cloud.gate.ac.uk
[15] https://gitlab.com/european-language-grid/usfd/elg-gate-cloud-bridge
[16] Micronaut – see section 7.2 for details

and more efficient Docker container, as well as being more secure by compiling to native code to avoid a whole class of Java "remote code execution" vulnerabilities.

Direct integration: rather than having a proxy container delegate calls to GATE Cloud, a GATE pipeline can also be directly run within a Docker container in the ELG infrastructure and accessed directly via the ELG specified API. This solution relies on the fact that the GATE framework is designed around reusable NLP components which can be built into pipelines necessary to extract relevant textual metadata ("annotations"). The configuration of a pipeline (the set of components needed alongside their parameters) can be stored in an XML format called "XGAPP", which can then be used to recreate the same pipeline automatically in any software based on a compatible version of GATE.

The "gate-ie-worker" tool can load any XGAPP and expose it as an ELG compatible endpoint. This component is available on GitLab[17], and is released as a public Docker image containing the Java runtime and the GATE worker software – but no XGAPP. This means that integrating a new GATE application is simply a matter of creating a child image which embeds the relevant XGAPP at a known location – new code does not need to be written.

The majority of GATE tools in the ELG are currently integrated via the GATE Cloud bridge but if a particular tool attracts significant interest then it is simple to migrate it to run directly in the ELG cluster via the gate-ie-worker and such a change would be transparent to the user.

In addition, USFD has recently started developing a number of new tools in other research projects that are not based on the GATE framework at all, but which still needed to be made available via the GATE Cloud platform and integrated as part of other GATE-based processing pipelines. To address this need USFD has re-used the "internal" LT service API specifications created as part of the ELG project and created a bridge component that operates in the opposite direction, to call an ELG-compatible LT service API endpoint as one step in a GATE pipeline[18]. New Python-based tools created at USFD are wrapped as ELG-compatible services and then published to GATE Cloud using the GATE-to-ELG bridge. Some of these ELG-compatible Docker images are open source and have been released to ELG in their own right.

4.5.2 Deployed components

As part of the first release, 36 GATE Cloud services were integrated into the ELG platform: 34 via the bridge component and 2 through direct integration. The second release added seventeen additional services via the ELG-to-GATE-Cloud bridge plus one via direct integration. Further services added for the third release are listed in Table 19 and include a number of new tools that have been created since the initial integration plan was formulated in 2019.

Name	Description	Language(s)
RussIE	Two versions of a Russian named entity recognition service modelled after the ANNIE rule-based English NER pipeline. The advanced version augments the basic NER system with an inflexional gazetteer designed to pick up more inflected forms and morphological variants of target entity names,	Russian

[17] https://gitlab.com/european-language-grid/usfd/gate-ie-worker
[18] https://gitlab.com/european-language-grid/usfd/gate-elg-client

Name	Description	Language(s)
	and an "orthomatcher" to provide limited coreference resolution based on orthographic similarity.	
Universal Dependencies POS Tagger	A POS tagger for various languages using the Universal Dependencies POS tagset. This tagger is based on a simple maximum entropy model trained on the corpus from the universal dependencies collection using the GATE Learning Framework plugin.	Individual pipelines for Russian and Indonesian, in addition to the languages already integrated
"GATE Hate"	A service that tags abusive utterances in any text. It includes a feature, "type", indicating the type of abuse if any, such as sexist, racist etc, and a "target" feature that indicates if the abuse was aimed at the addressee or some other party. This can be run on any English language text. GATE Hate is a generalized version of a pipeline specific to UK politics which USFD has been using for several years in other projects, so it also annotates a range of topics relevant to UK politics, and mentions of any UK members of parliament elected in the 2015, 2017 and 2019 UK general elections.	English
COVID-19 claim categoriser	A machine learning classifier trained to categorise claims about COVID-19 into 10 categories proposed by the Reuters Institute for the Study of Journalism.	English
COVID-19 vaccine text categoriser	A machine learning classifier trained to categorise COVID-19 vaccine text into 6 categories.	English
Toxic language classifier	A fine-tuned Roberta-base model using the simpletransformers toolkit for classifying toxic language. We use the Kaggle Toxic Comments Challenge dataset as training data. This dataset contains Wikipedia comments classified as toxic or non-toxic.	English
Offensive language classifier	The same Roberta-base model fine tuned using the simpletransformers toolkit using the OLIDv1 dataset from OffensEval 2019 as training data. This dataset contains tweets classified as offensive or non-offensive. Tells you if text is offensive, and the probability of confidence in it.	English
ChemDataExtractor	Wrapper for the Python-based chemical information extraction tool ChemDataExtractor[19] which resolves chemical names and abbreviations	English
OSCAR4 chemical NER	Chemical named entity recogniser based on the OSCAR4 toolkit (Open Source Chemistry Analysis Routines)[20]. It can be used to identify chemical names, reaction names, ontology terms, enzymes and chemical prefixes and adjectives, and chemical data such as state, yield, IR, NMR and mass spectra and elemental analyses	English
Journalist Safety Analyser	An application to recognise information related to descriptions of violations against journalists such as killings, threats etc. It indicates the kind of event(s) that took place, the names and other characteristics of journalists involved (such as their job role, organisation they work for) and other event details such as organisations involved, weapons used etc.	English

Table 19: GATE Cloud services integrated in the third ELG release

[19] http://chemdataextractor.org
[20] https://github.com/BlueObelisk/oscar4

The toxic & offensive classifiers and ChemDataExtractor are ELG-native Python services that run within the ELG cluster, the remaining services are implemented via the bridge to GATE Cloud.

4.6 IE tools supplied by pilot projects

Several of the ELG pilot projects have supplied IE & text analysis services for this release.

4.6.1 EVALITA tools for Italian

The EVALITA initiative is a periodic evaluation campaign of Natural Language Processing (NLP) and speech tools for the Italian language. It is structured around a set of shared tasks, and a number of system submissions from recent EVALITA evaluations have been published as ELG-compatible LT services, four from the 2018 EVALITA campaign and four from 2020.

Five tools were submissions to the HaSpeeDe tasks on hate speech detection. Hate Speech can be defined as any expression “that is abusive, insulting, intimidating, harassing, and/or incites to violence, hatred, or discrimination. It is directed against people on the basis of their race, ethnic origin, religion, gender, age, physical condition, disability, sexual orientation, political conviction, and so forth” (Erjavec and Kovačič, 2012). The basic HaSpeeDe task in both 2018 and 2020 was a binary classification of Italian Tweets as to whether they do or do not contain hate speech. Four tools for this task were published as ELG services, derived from the submissions by INRIA & Fondazione Bruno Kessler (Corazza et al, 2018), Rikjuniversiteit Groningen (Bai et al, 2018), the “HanSEL” system from the University of Bari (Polignano & Basile, 2018) and the 2020 submission “Montanti” (Bisconti & Montagnani, 2020) from the University of Pisa. A separate system, again from the “Montanti” team, was a submission to a sub-task to detect the use of stereotyping, or an “oversimplified opinion, prejudiced attitude, or uncritical judgment, toward a given target” (Bisconti & Montagnani, 2020).

Three other systems from three other shared tasks have also been published as ELG services:

- A misogyny identification tool “AlBERTo” from Università di Bologna (Muti & Barrón-Cedeño, 2020)
- A gender prediction system using simple n-gram models (Basile et al, 2018)
- A part of speech tagger “KLUMSy”, based on the “SoMeWeTa” tagger but trained for use on transcriptions of spoken Italian (Proisl & Lapesa, 2020)

4.6.2 E3C clinical entity recognisers

The primary output of the E3C “European Clinical Case Corpus” was, as the name suggests, a corpus of clinical case notes in five European languages English, Italian, Spanish, French and Basque. The corpus includes manual annotations of various entity types of clinical interest, so in addition to the corpus itself the project contributed six clinical entity recognition services, trained by fine tuning an XLM-RoBERTa multilingual base model using the annotated data from the E3C corpus. Five of the services are mono-lingual models trained on one of the five language sub-corpora, the sixth is a multi-lingual model trained on all five data sets in one and able to accept input in any of the five languages.

4.6.3 Lingsoft

Lingsoft contributed a set of tools covering English and the Scandinavian languages Danish, Swedish, Finnish and Norwegian. The tools cover:

- Proofing tools (spelling and grammar checking) for Danish, English, Finnish, Norwegian (both Bokmål and Nynorsk) and Swedish, plus a separate service tuned for Swedish as spoken in Finland, and another for medical-domain Finnish text.
- Morphology analysis and Named Entity Recognition in Danish, English, Finnish, Norwegian (both Bokmål and Nynorsk) and Swedish
- Named entity linking services for Swedish (one service) and Finnish (two services using different ontologies)

The services are implemented as proxies to an underlying service hosted on the provider's own hardware, and are intended for evaluation and research use only.

4.6.4 EDIA

The CEFR Labelling and Assessment Services pilot project by EDIA has developed corpora manually annotated with the readability level of each text based on the A1/A2/B1/B2/C1/C2 scale specified by the Common European Framework of Reference for Languages[21]. These corpora have been used to train a variety of classifiers that have been supplied as ELG-compatible services. The following services have been published so far:

- CEFR Readability Classification for English and German, with Dutch to follow after this deliverable. These services classify an input text on a six, nine or twelve-class CEFR scale
- Word difficulty taggers for English and German, that tag specific words in a text that are at a higher difficulty level than that of the text as a whole.

These services are implemented as proxies to an underlying service running on external infrastructure. In addition, metadata records have been published for other tools such as an annotation user interface developed using pilot project funding but which are not LT services based on the ELG API specifications.

5 MT Tools, Services and Components (Task 4.4)

5.1 MT tools Integrated by Tilde

The integration of Tilde's commercial MT platform into the ELG is described in Section 6.5 of Deliverable D4.1 – Tilde offers a cloud-based translation service hosted remotely from the ELG, which is integrated into the ELG ecosystem by means of a proxy component running in the ELG cluster to translate ELG API requests into calls out to the Tilde cloud service. Since the first and second release of the ELG platform the following translation directions have been added to the repertoire of MT services proxied through the ELG platform:

- English → Russian
- Russian → English

5.2 MT tools provided by pilot projects

The remaining Machine Translation tools included in this release have been supplied by two of the ELG pilot projects. A large number of language pairs have been contributed by OPUS-MT, and a few have also been supplied by Lingsoft.

[21] https://www.coe.int/en/web/common-european-framework-reference-languages

5.2.1 OPUS-MT

The OPUS-MT project aims to train MT models on *openly available* parallel data, producing models that can be used in both commercial and non-commercial settings without onerous licensing requirements.

The project has developed a Docker image which runs one or more instances of the standard Marian-NMT REST server each with one translation model, and then a lightweight Python Tornado application that acts as a service adapter, listening for ELG API requests and transcoding and dispatching each one to the relevant Marian endpoint. Variants of this base image can then be created to wrap any translation model for deployment into the ELG infrastructure. The translation runs on regular CPUs with minimal resource requirements thanks to the efficient decoder implementation in Marian-NMT.

Using this framework we have deployed 41 Docker images to the ELG infrastructure, but many of these models are multi-lingual, with one image able to translate to and/or from multiple languages. For example, there is one image for “West Germanic” languages which can translate either way among a set of 8 related languages. All told, these 41 Docker images result in 98 registered metadata records in the catalogue, and 187 distinct source-target translation services. For multilingual models each *target* language is represented as a separate metadata record so the user does not have to pass the required target language as a parameter, but many *source* languages can be supported on the same endpoint since the model does not need to be told in advance which of its trained languages is the source for a particular call. In fact, the endpoint can accept a single document containing a mix of sentences in different languages, and as long as all those languages are understood by the model it will still produce a reasonable translation.

Table 20 shows the full matrix of translation services provided so far to the ELG by OPUS-MT. There are a few language pairs that are provided more than once; for example there are three models that cover English to German, one is a mono-lingual EN-DE model, one is a one-to-many model from English into West Germanic languages, and the third is the many-to-many West Germanic model described above, which includes English and German in both its source and target language lists. The OPUS-MT team are keen to integrate further models between this deliverable and the end of the project, to fill more of the cells in the language matrix.

5.2.2 Lingsoft

Lingsoft contributed a set of neural MT services centred on the Finnish language. Their services translate Finnish to and from English, Swedish and German, and an additional pair of services have been supplied for English to and from Swedish.

Target → Source ↓	English	German	Luxembourgish	Dutch	Scots	Western Frisian	Low German	Afrikaans	Finnish	Northern Sami	Spanish	Russian	Swedish	Norwegian Nynorsk	Norwegian Bokmål	Ukrainian	Serbian	Breton	Belarusian	Icelandic	Romanian	Basque	Galician	French	Croatian	Faroese	Hungarian	Catalan	Danish	Italian	Portuguese	Polish	Bulgarian	Czech	Arabic	Turkish	Gaelic	Welsh	Slovak	Slovenian	Maltese	Bosnian	Macedonian	Irish	**Total**
English		3	2	2	1	2	2	2	1		1	1	1	1	1	1	1	2	1	1	1	1	1	1	1		1	1	1	1	1	1	1	1			1	1	1	1	1	1	1	1	**47**
German	2		1	1	1	1	1	1	1								2				1				1		1								1	1									**16**
Hunsrik	1	1	1	1	1	1	1	1																																					**8**
Swedish	1								1	1				1	1					1						1			1																**8**
Gronings		1	1	1	1	1	1	1																																					**7**
Luxembourgish	1	1		1	1	1	1	1																																					**7**
Dutch	1	1	1		1	1	1	1																																					**7**
Low German	1	1	1	1	1	1		1																																					**7**
Western Frisian	1	1	1	1	1		1	1																																					**7**
Afrikaans	1	1	1	1	1	1	1																																						**7**
Finnish	1	1								1		1	1											1																					**6**
Danish	1												1	1	1					1						1																			**6**
Russian	1								1							1			1																										**4**
Northern Sami									1				1	1	1																														**4**
Spanish	1																					1	1					1																	**4**
Ukrainian	1											1							1																										**3**
French	1								1									1																											**3**
Belarusian	1											1				1																													**3**
Catalan	1										1																																		**2**
Breton	2																																												**2**
Basque	1										1																																		**2**
Galician	1										1																																		**2**
Croatian	1																																												**1**
Serbo-Croatian	1																																												**1**
Norwegian										1																																			**1**
Welsh	1																																												**1**

Target → **Source ↓**	English	German	Luxembourgish	Dutch	Scots	Western Frisian	Low German	Afrikaans	Finnish	Northern Sami	Spanish	Russian	Swedish	Norwegian Nynorsk	Norwegian Bokmål	Ukrainian	Serbian	Breton	Belarusian	Icelandic	Romanian	Basque	Galician	French	Croatian	Faroese	Hungarian	Catalan	Danish	Italian	Portuguese	Polish	Bulgarian	Czech	Arabic	Turkish	Gaelic	Welsh	Slovak	Slovenian	Maltese	Bosnian	Macedonian	Irish	**Total**
Occitan (post 1500)	1																																											**1**
Irish	1																																											**1**
Macedonian	1																																											**1**
Hungarian	1																																											**1**
Swiss German	1																																											**1**
Romanian	1																																											**1**
Polish	1																																											**1**
Rusyn	1																																											**1**
Czech	1																																											**1**
Slovak	1																																											**1**
Chavacano	1																																											**1**
Norwegian Nynorsk											1																																		**1**
Bulgarian	1																																											**1**
Gaelic	1																																											**1**
Norwegian Bokmål											1																																		**1**
Maltese	1																																											**1**
Aragonese	1																																											**1**
Portuguese	1																																											**1**
Slovenian	1																																											**1**
Ladino	1																																											**1**
Italian	1																																											**1**
Total	**43**	**11**	**9**	**9**	**9**	**9**	**9**	**9**	**6**	**5**	**4**	**4**	**4**	**4**	**4**	**3**	**3**	**3**	**3**	**3**	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**187**

Table 20: Matrix of source and target languages supported by OPUS-MT translation services

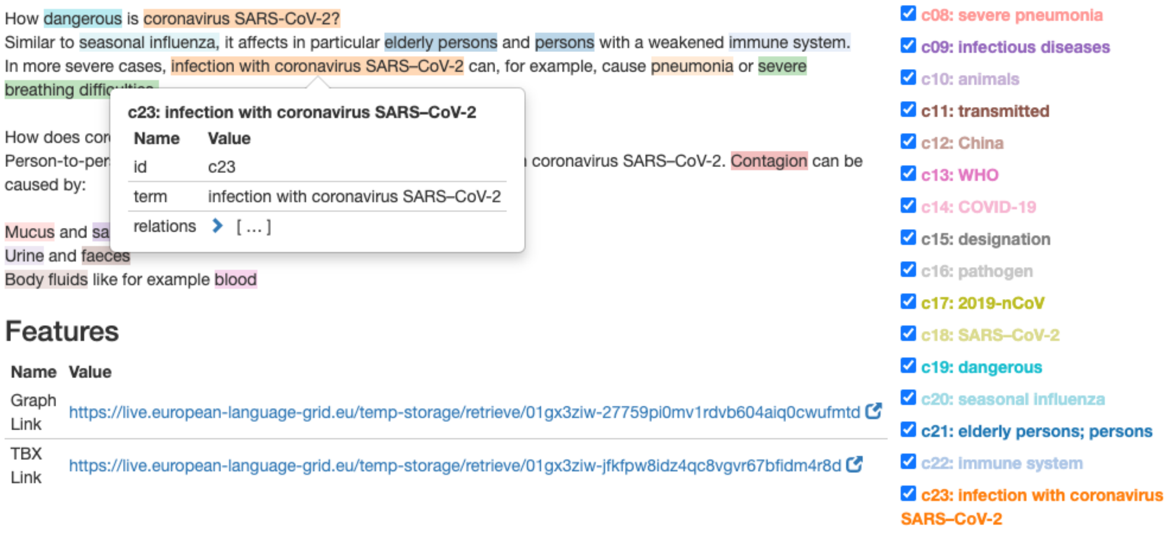
6 Integration of other Types of Tools, Services, Components (Task 4.5)

Beyond the initial three principal service classes of ASR, MT and IE/text analysis, a number of other service types have emerged over the course of the project. Elhuyar have delivered Basque language text-to-speech alongside their ASR service described in section 3.2 as part of their pilot project, and two of the other pilots have introduced entirely new service types: Text2TCS have developed a service that extracts terms and relations between them and generates a TermBase in the standard TBX file format, and Coreon have deployed a SPARQL endpoint along with a simple web UI to query it.

6.1 Text2TCS

The Text2TCS pilot project has developed a service that automatically extracts a Terminological Concept System (TCS) from domain-specific texts in multiple languages. A TCS is a terminological resource that conceptually structures domain-specific terms and provides hierarchical and non-hierarchical relations between them. The tool implements a pipeline consisting of preprocessing, term extraction, relation extraction, and post-processing steps. The pipeline takes domain-specific natural language sentences or text as input and outputs a TCS in the ISO standard TermBase eXchange (TBX) format[22], and a concept map as a PNG image.

These output formats do not fit naturally into the JSON-based structure used by all previous ELG LT services, but adding specific support for each file format that a service might wish to return would not be practical. Instead, the solution adopted by the ELG platform was to introduce a “helper” service for temporary data storage. This is an endpoint that is accessible at a known fixed URL of http://storage.elg/store by any LT service running within the ELG Kubernetes cluster, to which the LT service can POST arbitrary data with any MIME type in the Content-Type request header. The service stores that data in a temporary location and returns a URL that is resolvable from outside the cluster, which the LT service can then include in its response to the user.



How dangerous is coronavirus SARS-CoV-2?
Similar to seasonal influenza, it affects in particular elderly persons and persons with a weakened immune system.
In more severe cases, infection with coronavirus SARS–CoV-2 can, for example, cause pneumonia or severe breathing diffic

How does cor
Person-to-per coronavirus SARS–CoV-2. Contagion can be caused by:

Mucus and sa
Urine and faeces
Body fluids like for example blood

c23: infection with coronavirus SARS–CoV-2

Name	Value
id	c23
term	infection with coronavirus SARS–CoV-2
relations	› […]

Features

Name	Value
Graph Link	https://live.european-language-grid.eu/temp-storage/retrieve/01gx3ziw-27759pi0mv1rdvb604aiq0cwufmtd
TBX Link	https://live.european-language-grid.eu/temp-storage/retrieve/01gx3ziw-jfkfpw8idz4qc8vgvr67bfidm4r8d

☑ c08: severe pneumonia
☑ c09: infectious diseases
☑ c10: animals
☑ c11: transmitted
☑ c12: China
☑ c13: WHO
☑ c14: COVID-19
☑ c15: designation
☑ c16: pathogen
☑ c17: 2019-nCoV
☑ c18: SARS–CoV-2
☑ c19: dangerous
☑ c20: seasonal influenza
☑ c21: elderly persons; persons
☑ c22: immune system
☑ c23: infection with coronavirus SARS–CoV-2

Figure 2: Results from Text2TCS service showing temporary URLs to the graph and TBX

[22] https://www.iso.org/standard/62510.html

Any errors from the storage service are returned in the same failure message structure as the LT service API itself uses, and these errors can be validly echoed directly back to the caller in the LT service response.

The user receiving one of these temporary storage URLs can perform a GET request on the URL to retrieve the stored data with the original Content-Type. Stored data expires and is deleted after a short time – by default 15 minutes from the time of upload but the LT service can specify a longer TTL up to 24 hours when it initially stores the data. The URLs are un-guessable, generated by a cryptographically-secure random number generator, so authentication is not required in order to download – knowledge of the correct URL is sufficient authentication in itself.

The Text2TCS service makes use of the temporary storage solution to store the TBX and PNG files that it generates, and is thus able to leverage the standard ELG "standoff annotations" response type. Figure 2 shows the annotations it returns, which are the mentions of the extracted terms within the original text, and the response also includes top-level "features" giving the temporary storage URLs from which the TBX and PNG (Figure 3) can be retrieved.

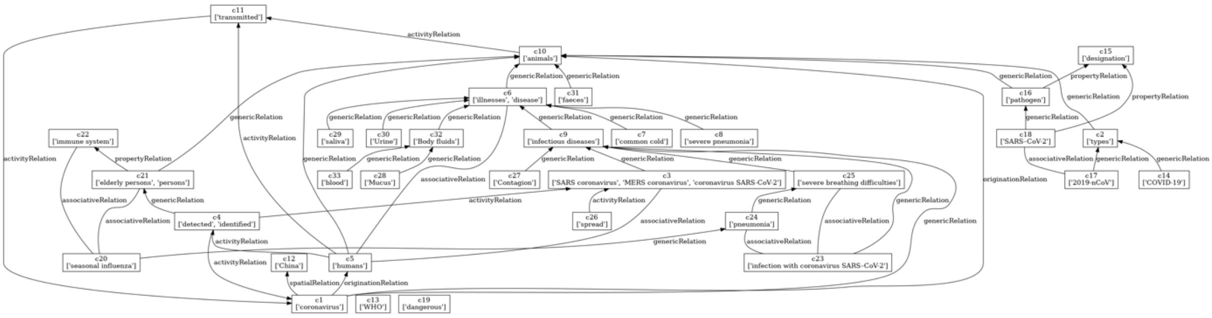


Figure 3: Terminology graph generated by Text2TCS

6.2 Coreon SPARQL endpoints

The "MKS as Linguistic Linked Open Data" project by Coreon aimed to make Coreon's Multilingual Knowledge System (MKS) knowledge bases available as Linked Open Data. It is unclear conceptually whether this project's outputs fit here in WP4 or whether they belong in WP5 – the object of the project was to deploy *resources* (the knowledge bases) but the mechanism for doing so was to make them available as *services*, so it exists at the junction between the two work packages. Given the effort required on the part of the ELG technical team was more akin to deploying a service than simply uploading a resource we have chosen to document it here rather than in D5.3.

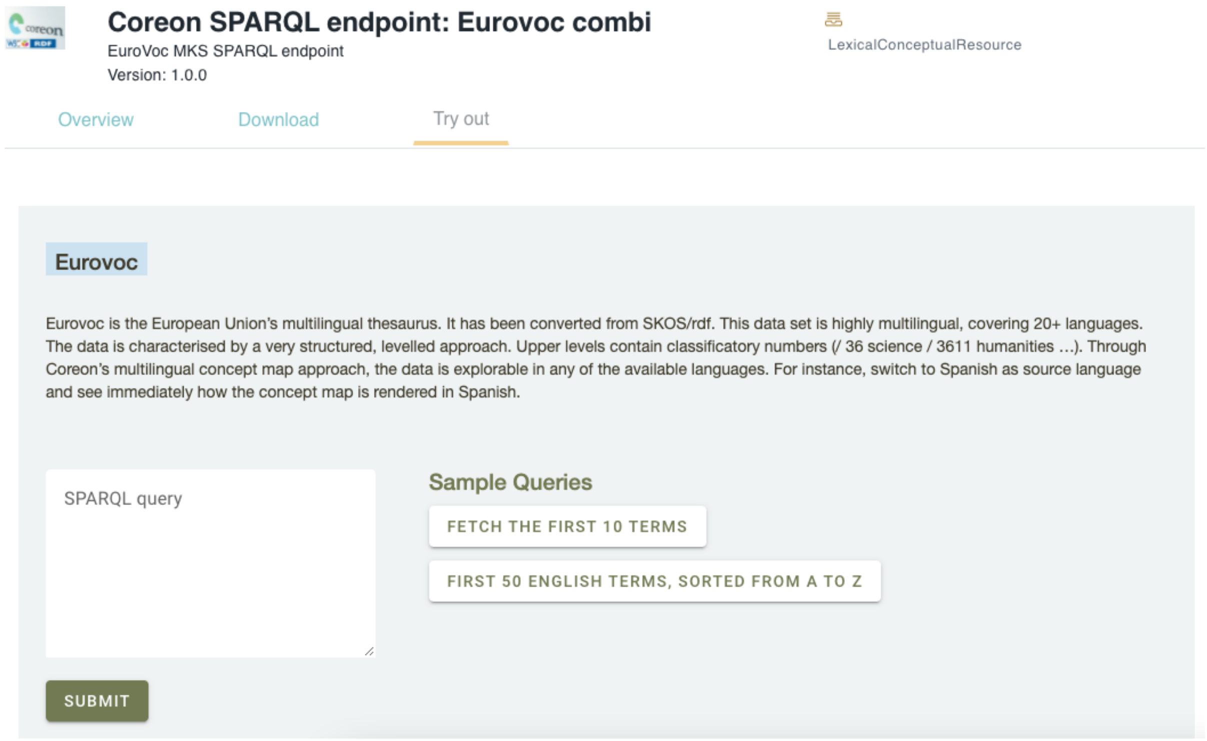
Unlike the case for standard LT services of the kinds discussed up until now in ELG, the Linked Open Data community *already* had a widely accepted de facto standard API for access to LOD resources which pre-dates ELG, namely the RDF query language SPARQL. Thus rather than define a new API for LOD under the aegis of the ELG project, the ELG chose simply to adopt SPARQL as the approved API for this service class.

To make the Coreon SPARQL endpoints available within ELG, a proxy mechanism was developed to permit access to the endpoints with the use of ELG Keycloak access tokens for authentication, the same mechanism as is used by the restserver for standard LT services. In order to showcase the MKS resources in the ELG portal, a "try out" GUI mechanism was created along exactly the same lines as for standard LT services. The ELG catalogue team introduced two new concepts to the catalogue model, first a *resource access* service type to denote a service whose role is to access other resources, and second an "ELG-compatible Lexical/Conceptual Resource" (LCR) which links in its metadata to the service used to access it. When a user browses to an ELG-compatible LCR in the catalogue, the UI adds a "try out" tab which includes the trial GUI registered against the relevant resource access service, and configures this UI via JavaScript message passing in exactly the same way as for LT services. The configuration message includes an access token and the API URL of the LCR's own metadata record, from which the UI can then retrieve the address of the actual SPARQL endpoint. The same access token is then used by the UI to make its SPARQL requests.

Thus Coreon delivered to the ELG catalogue:

- One metadata record for their SPARQL query tool (a "resource access" service)
- One metadata record for each SPARQL endpoint (an ELG-compatible LCR linked to the access service)

For the Coreon case, each SPARQL endpoint uses the "sample data" slot in its metadata record to link to an HTML page showing sample queries. This HTML is human-readable, but it is also marked up with special CSS classes that allow the query UI JavaScript to parse the "sample" file, extract the relevant queries for this endpoint, and configure the interface appropriately. Figure 4 shows the resulting view in the catalogue UI – the title, description and sample queries have been parsed out of the HTML.


Coreon SPARQL endpoint: Eurovoc combi
EuroVoc MKS SPARQL endpoint
Version: 1.0.0

Eurovoc

Eurovoc is the European Union's multilingual thesaurus. It has been converted from SKOS/rdf. This data set is highly multilingual, covering 20+ languages. The data is characterised by a very structured, levelled approach. Upper levels contain classificatory numbers (/ 36 science / 3611 humanities ...). Through Coreon's multilingual concept map approach, the data is explorable in any of the available languages. For instance, switch to Spanish as source language and see immediately how the concept map is rendered in Spanish.

Figure 4: Coreon SPARQL query UI

7 Helper tools for LT service developers

As part of task 4.6 "assist 3rd party service, tool, component providers with integration and deployment", the ELG team has developed a set of "helper" libraries to enable potential LT service developers to package their tools in an ELG-compatible way with the minimum of effort, and to enforce (or at least encourage) best practice in the use of the LT service APIs. The most popular programming language by far for LT service developers both within and outside of the ELG consortium is Python, and there is also significant interest in Java and other JVM-based languages due to the extensive legacy of language technology frameworks such as GATE and UIMA. This section describes the helper tools that have been developed for those two languages.

7.1 ELG Python SDK for LT service developers

The ELG team at DFKI, assisted by developers at other partners, have developed a set of modules for Python covering many tasks related to the ELG platform and APIs. They include the following tasks for *consumers* of ELG resources and services:

- A set of Pydantic[23] model classes to represent the JSON message structures of the ELG APIs
- Classes to represent the ELG metadata model, including languge resources, LT services, etc.
- A client to authenticate against the ELG Keycloak identity provider, search the catalogue, retrieve metadata records, and invoke ELG-compatible LT services
- A command line interface to the above functions

But also two helper classes to support the creation of ELG-compatible LT service Docker images. These classes are called FlaskService and QuartService, the former is more appropriate when developing CPU-bound services, the latter for I/O-bound services (such as proxies to remote web services) that can take advantage of Python's asyncio framework.

Users only need to extend one of these classes and implement the `process_text` or `process_audio` method. This method contains the code of the LT tool, it takes as input an ELG request object (one of the Pydantic model classes referred to above) and needs to return a valid ELG response object or raise an error. This allows the users to focus on their LT tools and to not have to worry about the creation of the HTTP server.

```
from elg import FlaskService
from elg.model import TextRequest, AnnotationsResponse
import langdetect

class ELGService(FlaskService):
    def process_text(self, request: TextRequest):
        langs = langdetect.detect_langs(request.content)
        ld = {}
        for l in langs:
            ld[l.lang] = l.prob
        return AnnotationsResponse(features=ld)

service = ELGService("LangDetection")
app = service.app
```

[23] https://github.com/samuelcolvin/pydantic

More advanced functionality is available for services that need to return ongoing progress updates to their caller. Once the service class has been created, the developer can use the elg command line tool to generate an optimized Dockerfile for the services created with the FlaskService or the QuartService class, plus the other required "boilerplate" files ("entrypoint" script, requirements.txt file detailing the dependencies of the service, etc.) required to build their service class into a Docker image that will run in ELG.

The "elg" library is available to install via `pip` from the standard PyPi Python package repository, and is documented in the ELG platform documentation at https://european-language-grid.readthedocs.io/en/stable/all/A1_PythonSDK/TutoServiceIntegration.html.

7.2 ELG helper libraries for Java

In a similar vein to the Python library described in the previous section, ELG developers at USFD have built a set of helper libraries for Java developers:

- Java model classes to represent the JSON API messages, annotated for use with the Jackson[24] JSON binding library
- Spring Boot Starter for LT service developers who use the Spring Boot[25] framework
- Helper library for LT service developers who use the Micronaut[26] microservices framework

These are all published to the Central Maven repository (the de facto standard place where all Java-based build tools look for their dependencies), under the `eu.european-language-grid` "group ID".

The Java model library (artifact ID `elg-java-bindings`) includes JSON mappings for all the request, response and error messages. It also includes the definitions of the "standard" ELG status message codes (service not found, internal error during processing, etc.) along with their translations into other languages. These model classes are used by the Spring Boot and Micronaut helpers, as well as two core ELG platform components, namely the LT service execution restserver and the i18n message resolver service.

The Spring Boot Starter[27] (artifact ID `elg-spring-boot-starter`) and Micronaut helper[28] (artifact ID `lt-service-micronaut`) are intended for use by LT service developers to build services in Java, Groovy, Kotlin or any other JVM-based language. They operate in slightly different ways but the basic principle is the same – a developer must

1. Create a blank application using the framework's standard launcher tool https://start.spring.io or https://micronaut.io/launch/.
2. Add a dependency on the ELG helper (which in turn depends on `elg-java-bindings`).
3. Create the appropriate class for the framework – in the case of Spring Boot this means a class with the `@Component` and `@ElgHandler` annotations, for Micronaut it means a `@Controller` which is a subclass of the `LTService` base class provided by the helper.
4. Implement a request handling method.

[24] https://github.com/FasterXML/jackson
[25] https://spring.io/projects/spring-boot
[26] https://micronaut.io
[27] https://gitlab.com/european-language-grid/platform/elg-spring-boot-starter
[28] https://gitlab.com/european-language-grid/platform/lt-service-micronaut

In both frameworks the request handler is a method that receives as input the elg-java-bindings object representing the request, and which must either return a corresponding response object or throw an exception to signal failure. The helper handles all the surrounding boilerplate to parse the request, and serialize the response or convert the exception into a valid failure message. Both framework helpers include support for asynchronous non-blocking request handlers using Reactive Streams[29], a mechanism to issue partial progress messages (which are delivered to the caller in the ELG specified server-sent events format), and a simple client to enable the LT service to store data via the ELG temporary storage mechanism described in section 6.1.

The choice of framework is up to the developer, but if they do not already have reason to choose one over the other then we recommend the use of Micronaut, since it is specifically optimized for micro-services and can produce much smaller Docker images with faster startup time. Micronaut is designed to use virtually no dynamic class loading, so it is able to leverage GraalVM[30] to compile Java down to native code ahead of time. These native images are only a few MB in size (so quicker to pull onto the cluster nodes), start up from cold in a matter of milliseconds (so fit very will to the scale-on-demand model of knative), and are immune to a whole class of remote code execution security vulnerabilities (since dynamic class loading is restricted to specific classes that were allow-listed at the time the image was built).

8 Conclusions and future work

The ELG project has successfully delivered almost all the services that were envisaged in our original D4.1 integration plan, plus a number of additional services and languages that were not foreseen at the start of the project. The period since the second platform release has seen many more services published by the pilot projects, greatly improving the ELG platform's service offering in both depth and breadth. Special mention goes to OPUS-MT, who have greatly increased the machine translation language coverage of the ELG platform; the addition of the OPUS-MT services means that across the whole ELG suite of translation services (from project partners, pilot projects, and elsewhere) all 24 EU official languages are now represented as the source and as the target of at least one translation service each. We hope to integrate more OPUS models in the remainder of the project runtime to futher broaden our coverage, particularly in services that do not need to pivot via English.

The pilot projects have also been instrumental in broadening the scope of the *types* of services that can be supported by the ELG platform. The temporary storage mechanism introduced to support Text2TCS opens the door for other services that generate outputs other than text and audio, and the "ELG compatible LCR" concept introduced for Coreon will permit the inclusion of other resource types that are "queried" via standard protocols rather than "downloaded" as a unit.

Table 21 shows a snapshot of the state of the ELG services catalogue at 28th January 2022 when this deliverable was completed, with one month still to go until the target date of release 3. This includes all ELG-compatible services integrated at this date whether from the consortium, the pilot projects, or outside contributors. As in

[29] https://www.reactive-streams.org, and in particular https://projectreactor.io
[30] https://www.graalvm.org

section 2 the languages are grouped with category A being the EU official languages and B being other languages used in EU member states, accession candidate countries, or the wider EFTA/EEA free trade area. Again following the pattern of section 2, where one metadata record covers several service functions and/or several languages, it is counted in each relevant cell – this gives a more accurate picture of the *true* scope of the ELG service offering than would be found by simply counting metadata records.

At the end of January 2022 the ELG platform offered 1,116 distinct service function/language combinations, two thirds are text analysis services, 21.6% machine translation, and the remainder split between speech recognition and other service types.

358 (32%) of these have been contributed by the pilot projects:

- 5 ASR services (4 Lingsoft, 1 Elhuyar)
- 195 MT services (187 OPUS-MT, 8 Lingsoft)
- 90 IE/Text analysis functions (15 EVALITA, 20 E3C, 50 Lingsoft, 5 EDIA)
- 68 others (Elhuyar Basque TTS, 22 languages of “term extraction” and 22 of “relation extraction” for Text2TCS, and Coreon SPARQL access to knowledge bases covering 23 languages)

In addition to these raw numbers, the act of working with and supporting the pilot projects to develop their services has been hugely valuable to the ELG technical team, in finding and fixing bugs within the platform, addressing limitations of the metadata model and API specifications, and generally giving an honest “view from outside” to prompt the team to consider situations that had not previously been foreseen. While most of the pilot project service development pre-dated the introduction of the ELG Python SDK, our experiences with and feedback from the pilot project developers have been instrumental in driving the development of that tool, and at least one pilot project has adopted the ELG SDK for use in updated versions of their services.

Looking ahead, there are further services to be integrated from both ELG partners and pilot projects, in particular USFD intends to introduce a service for image OCR that was developed as part of the WeVerify project – an API has already been specified for ELG services that process images. Aside from this, now that the ELG has a broad portfolio of services deployed, our focus will start to shift towards ways to promote the uptake and *use* of those services by potential consumers, to build a sustainable customer base for the ELG platform beyond the end of the Horizon 2020 project itself.

Languages: • **A = EU official** • **B = other languages used by EU members, accession candidates, or EEA/EFTA members**		Category A – EU official languages																											
		English	German	Italian	French	Spanish	Dutch	Swedish	Finnish	Greek	Danish	Portuguese	Polish	Romanian	Czech	Bulgarian	Latvian	Croatian	Slovak	Estonian	Lithuanian	Hungarian	Slovenian	Maltese	Irish	**Total Category A**	Total Category B	Others (C & D)	**Grand Total**
Source ↓	**Machine Translation Target →**																												
Category A – EU official languages	English		6	1	2	2	2	4	3	1	2	2	3	2	3	2	2	1	1	2	1	1	1	1	1	**46**	20	8	**74**
	German	5					1		2					1				1				1				**11**	7	2	**20**
	Swedish	3							2		1															**6**	5		**11**
	Finnish	3	2		1			2																		**8**	1	1	**10**
	Dutch	1	1																							**2**	4	1	**7**
	Danish	1						1																		**2**	4		**6**
	Spanish	2																								**2**	3		**5**
	French	2							1																	**3**	1		**4**
	Czech	4																								**4**			**4**
	Polish	3																								**3**			**3**
	Portuguese	2																								**2**			**2**
	Latvian	2																								**2**			**2**
	Estonian	2																								**2**			**2**
	Bulgarian	2																								**2**			**2**
	Slovenian	1																								**1**			**1**

Languages: • **A = EU official** • **B = other languages used by EU members, accession candidates, or EEA/EFTA members**	Category A – EU official languages																										
	English	German	Italian	French	Spanish	Dutch	Swedish	Finnish	Greek	Danish	Portuguese	Polish	Romanian	Czech	Bulgarian	Latvian	Croatian	Slovak	Estonian	Lithuanian	Hungarian	Slovenian	Maltese	Irish	**Total Category A**	Total Category B	Others (C & D)	**Grand Total**
Romanian	2																								**2**			**2**
Greek	1																								**1**			**1**
Lithuanian	1																								**1**			**1**
Slovak	1																								**1**			**1**
Maltese	1																								**1**			**1**
Italian	1																								**1**			**1**
Croatian	1																								**1**			**1**
Irish	1																								**1**			**1**
Hungarian	1																								**1**			**1**
Total Category A	**43**	**9**	**1**	**3**	**2**	**3**	**7**	**8**	**1**	**3**	**2**	**3**	**3**	**3**	**2**	**2**	**2**	**1**	**2**	**1**	**2**	**1**	**1**	**1**	**106**	**45**	**12**	**163**
Total Category B	17	4			3	4	1	1																	**30**	18	4	**52**
Total Others (C & D)	7	2				2		1																	**12**	8	7	**27**
Total Machine Translation	**67**	**15**	**1**	**3**	**5**	**9**	**8**	**10**	**1**	**3**	**2**	**3**	**3**	**3**	**2**	**2**	**2**	**1**	**2**	**1**	**2**	**1**	**1**	**1**	**148**	**71**	**23**	**242**
Speech recognition and audio analysis																												
Speech Recognition	1	1	1	1	2	1	1	2	1			1	1	1		2			1	1					**18**	6	11	**35**
Term extraction from audio	1	1	1	1	1	1			1			1	1												**9**	3	9	**21**
Total Speech & audio	**2**	**2**	**2**	**2**	**3**	**2**	**1**	**2**	**2**			**2**	**2**	**1**		**2**			**1**	**1**					**27**	**9**	**20**	**56**

Languages: • **A = EU official** • **B = other languages used by EU members, accession candidates, or EEA/EFTA members**	Category A – EU official languages																										
	English	German	Italian	French	Spanish	Dutch	Swedish	Finnish	Greek	Danish	Portuguese	Polish	Romanian	Czech	Bulgarian	Latvian	Croatian	Slovak	Estonian	Lithuanian	Hungarian	Slovenian	Maltese	Irish	**Total Category A**	Total Category B	Others (C & D)	**Grand Total**
Information Extraction & Text Analysis																												
Part-of-Speech Tagging	8	3	3	3	3	4	3	3	2	3	3	2	2	2	2	2	2	2	2	1	1	2	1	1	**60**	13	29	**102**
Language identification	3	3	2	3	3	3	2	1	2	2	2	2	2	2	2	1	2	2	1	1	2	1			**44**	8	36	**88**
Named Entity Recognition	17	7	4	6	4	5	4	2	3	1	2	1	2	2	1		1	1			1				**64**	9	11	**84**
Tokenization	6	4	3	2	2	3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	**38**	9	27	**74**
Lemmatization	3	2	2	2	2	2	2	2	1	2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	**35**	11	27	**73**
Morphological annotation	3	2	2	2	2	2	2	2	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	**34**	11	27	**72**
Dependency parsing	1	2	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	**26**	9	27	**62**
Sentiment analysis	13	3	2	2	2	1			1		2	1					1					1			**29**		5	**34**
Summarization	4	3	1	2	2	1					1														**14**			**14**
Text categorization	4		7		1				1																**13**			**13**
Polarity labelling	1	1	1	1	1						1	1													**7**		4	**11**
Entity linking	2		2	2	2																				**8**	2		**10**
Keyword extraction	1	1	1	1	1	1	1	1			1														**9**			**9**
Spell checking	1						2	2		1															**6**	2		**8**
Morphological analysis	2						1	1		1															**5**	2		**7**
Named Entity Disambiguation	4	1		1	1																				**7**			**7**
Grammar checking	1						2	2		1															**6**	1		**7**

Languages: • **A = EU official** • **B = other languages used by EU members, accession candidates, or EEA/EFTA members**	Category A – EU official languages																								**Total Category A**	Total Category B	Others (C & D)	**Grand Total**
	English	German	Italian	French	Spanish	Dutch	Swedish	Finnish	Greek	Danish	Portuguese	Polish	Romanian	Czech	Bulgarian	Latvian	Croatian	Slovak	Estonian	Lithuanian	Hungarian	Slovenian	Maltese	Irish				
Semantic annotation	1	1	1	1	1	1					1														**7**			**7**
Hate speech recognition			6																						**6**			**6**
Readability annotation	3	2																							**5**			**5**
Structural annotation	1	1			1				1																**4**			**4**
Temporal Expression Analysis	1	1			1	1																			**4**			**4**
Topic Detection	2	1																							**3**			**3**
Information Extraction	3																								**3**			**3**
Sentence splitting	1	1				1																			**3**			**3**
Event detection									2																**2**			**2**
Political Bias Classification	1	1																							**2**			**2**
ontology							1	1																	**2**			**2**
Document Classification		1																							**1**			**1**
Truth labelling	1																								**1**			**1**
Fake news detection	1																								**1**			**1**
Speech annotation			1																						**1**			**1**
Text analysis	1																								**1**			**1**
Anonymization	1																								**1**			**1**
Extraction of domain-specific information	1																								**1**			**1**

Languages: • **A = EU official** • **B = other languages used by EU members, accession candidates, or EEA/EFTA members**	Category A – EU official languages																											
	English	German	Italian	French	Spanish	Dutch	Swedish	Finnish	Greek	Danish	Portuguese	Polish	Romanian	Czech	Bulgarian	Latvian	Croatian	Slovak	Estonian	Lithuanian	Hungarian	Slovenian	Maltese	Irish	**Total Category A**	Total Category B	Others (C & D)	**Grand Total**
Chunking									1																**1**			**1**
Topic Modelling		1																							**1**			**1**
Polarity detection									1																**1**			**1**
Discourse annotation		1																							**1**			**1**
Annotation of measurements	1																								**1**			**1**
Noun phrase chunking	1																								**1**			**1**
Total Text Analysis	**94**	**43**	**39**	**29**	**30**	**26**	**22**	**19**	**19**	**15**	**18**	**11**	**10**	**10**	**9**	**7**	**10**	**9**	**7**	**6**	**8**	**8**	**5**	**5**	**459**	**77**	**193**	**729**
Other service types																												
SPARQL Endpoint	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1		**23**			**23**
Term extraction	1	1	1	1	1	1			1	1		1			1										**10**		12	**22**
Relation Extraction	1	1	1	1	1	1			1	1		1			1										**10**		12	**22**
Speech Synthesis	7	3	1	4												2				2					**19**	2	1	**22**
Total Other	**10**	**6**	**4**	**7**	**3**	**3**	**1**	**1**	**3**	**3**	**1**	**3**	**1**	**1**	**3**	**3**	**1**	**1**	**1**	**3**	**1**	**1**	**1**		**62**	**2**	**25**	**89**
Grand Total	**173**	**66**	**46**	**41**	**41**	**40**	**32**	**32**	**25**	**21**	**21**	**19**	**16**	**15**	**14**	**14**	**13**	**11**	**11**	**11**	**11**	**10**	**7**	**6**	**696**	**159**	**261**	**1116**

Table 21: Current state of ELG services at 2022-01-28

9 References

Xiaoyu Bai, Flavio Merenda, Claudia Zaghi, Tommaso Caselli, and Malvina Nissim. 2018. RuG @ EVALITA 2018: Hate Speech Detection In Italian Social Media. In *Proceedings of Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018).* http://ceur-ws.org/Vol-2263/paper042.pdf

Angelo Basile, Gareth Dwyer, and Chiara Rubagotti. 2018. CapetownMilanoTirana for GxG at Evalita2018. Simple n-gram based models perform well for gender prediction. Sometimes. In *Proceedings of Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018),* Turin, Italy. http://ceur-ws.org/Vol-2263/paper028.pdf

Elia Bisconti and Matteo Montagnani. 2020. Montanti @ HaSpeeDe2 EVALITA 2020: Hate Speech Detection in online contents. In *Proceedings of the 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2020)*. http://ceur-ws.org/Vol-2765/paper177.pdf

Michele Corazza, Stefano Menini, Pinar Arslan, Rachele Sprugnoli, Elena Cabrio, Sara Tonelli, and Serena Villata. 2018. Comparing Different Supervised Approaches to Hate Speech Detection. In *Proceedings of Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018).* http://ceur-ws.org/Vol-2263/paper039.pdf

Karmen Erjavec and Melita Poler Kovačič. 2012. ”You Don’t Understand, This is a New War!” Analysis of Hate Speech in News Web Sites’ Comments. Mass Communication and Society, **15**(6).

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hervé Jégou, Tomás Mikolov. 2016. FastText.zip: Compressing text classification models. arXiv pre-print, https://arxiv.org/abs/1612.03651

Armand Joulin, Edouard Grave, Piotr Bojanowski and Tomas Mikolov. 2017. Bag of Tricks for Efficient Text Classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. https://aclanthology.org/E17-2068

Arianna Muti and Alberto Barròn-Cedeño. 2020. UniBO@AMI: A Multi-Class Approach to Misogyny and Aggressiveness Identification on Twitter Posts Using AlBERTo. In *Proceedings of Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020),* Bologna, Italy. http://ceur-ws.org/Vol-2765/paper117.pdf

Marco Polignano and Pierpaolo Basile. 2018. HanSEL: Italian Hate Speech Detection through Ensemble Learning and Deep Neural Networks. In *Proceedings of Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018).* http://ceur-ws.org/Vol-2263/paper038.pdf

Thomas Proisl and Gabriella Lapesa. 2020. KLUMSy@KIPoS: Experiments on Part-of-Speech Tagging of Spoken Italian. In *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*, Online. http://ceur-ws.org/Vol-2765/paper140.pdf

A. Appendix

This appendix provides the full list of IE & Text Analysis services targeted for inclusion in the three ELG platform releases, listed by project partner, and target language (and A-D language category).

R1

Provider	Tool	Service	Lang./Category	
CUNI	NameTag	Named Entity Recognition	Czech	A
CUNI	NameTag	Named Entity Recognition	English	A
CUNI	UDPipe parser	Dependency Parsing	Czech	A
CUNI	UDPipe parser	Dependency Parsing	English	A
CUNI	UDPipe parser	Dependency Parsing	French	A
CUNI	UDPipe parser	Dependency Parsing	German	A
CUNI	UDPipe parser	Dependency Parsing	Greek	A
CUNI	UDPipe parser	Dependency Parsing	Latvian	A
CUNI	UDPipe parser	Dependency Parsing	Spanish	A
CUNI	UDPipe tagger	Lemmatization	Czech	A
CUNI	UDPipe tagger	Lemmatization	English	A
CUNI	UDPipe tagger	Lemmatization	French	A
CUNI	UDPipe tagger	Lemmatization	German	A
CUNI	UDPipe tagger	Lemmatization	Greek	A
CUNI	UDPipe tagger	Lemmatization	Latvian	A
CUNI	UDPipe tagger	Lemmatization	Spanish	A
CUNI	UDPipe tagger	Morphological analyser	Czech	A
CUNI	UDPipe tagger	Morphological analyser	English	A
CUNI	UDPipe tagger	Morphological analyser	French	A
CUNI	UDPipe tagger	Morphological analyser	German	A
CUNI	UDPipe tagger	Morphological analyser	Greek	A
CUNI	UDPipe tagger	Morphological analyser	Latvian	A
CUNI	UDPipe tagger	Morphological analyser	Spanish	A
CUNI	UDPipe tagger	Part of Speech tagging	Czech	A
CUNI	UDPipe tagger	Part of Speech tagging	English	A
CUNI	UDPipe tagger	Part of Speech tagging	French	A
CUNI	UDPipe tagger	Part of Speech tagging	German	A
CUNI	UDPipe tagger	Part of Speech tagging	Greek	A
CUNI	UDPipe tagger	Part of Speech tagging	Latvian	A
CUNI	UDPipe tagger	Part of Speech tagging	Spanish	A
CUNI	UDPipe tokenizer	Tokenization	Czech	A
CUNI	UDPipe tokenizer	Tokenization	English	A
CUNI	UDPipe tokenizer	Tokenization	French	A
CUNI	UDPipe tokenizer	Tokenization	German	A
CUNI	UDPipe tokenizer	Tokenization	Greek	A
CUNI	UDPipe tokenizer	Tokenization	Latvian	A
CUNI	UDPipe tokenizer	Tokenization	Spanish	A
Expert System	Cogito Discover	Text categorization	English	A
Expert System	Cogito Discover	Text categorization	Spanish	A
Expert System	Cogito Discover	Language identification	Czech	A

R1

Provider	Tool	Service	Lang./Category	
Expert System	Cogito Discover	Language identification	English	A
Expert System	Cogito Discover	Language identification	French	A
Expert System	Cogito Discover	Language identification	German	A
Expert System	Cogito Discover	Language identification	Greek	A
Expert System	Cogito Discover	Language identification	Latvian	A
Expert System	Cogito Discover	Language identification	Spanish	A
Expert System	Cogito Discover	Lemmatization	English	A
Expert System	Cogito Discover	Lemmatization	French	A
Expert System	Cogito Discover	Lemmatization	German	A
Expert System	Cogito Discover	Lemmatization	Spanish	A
Expert System	Cogito Discover	Named Entity Recognition	English	A
Expert System	Cogito Discover	Named Entity Recognition	French	A
Expert System	Cogito Discover	Named Entity Recognition	German	A
Expert System	Cogito Discover	Named Entity Recognition	Spanish	A
Expert System	Cogito Discover	Part-of-Speech Tagging	English	A
Expert System	Cogito Discover	Part-of-Speech Tagging	French	A
Expert System	Cogito Discover	Part-of-Speech Tagging	German	A
Expert System	Cogito Discover	Part-of-Speech Tagging	Spanish	A
Expert System	Cogito Discover	Sentiment Analysis	English	A
Expert System	Cogito Discover	Sentiment Analysis	Spanish	A
Expert System	Cogito Discover	Sentiment Analysis	German	A
Expert System	Cogito Discover	Sentiment Analysis	French	A
Expert System	Cogito Discover	Summarization	English	A
Expert System	Cogito Discover	Summarization	French	A
Expert System	Cogito Discover	Summarization	German	A
Expert System	Cogito Discover	Summarization	Spanish	A
DFKI	geolocator	Categorization	English	A
DFKI	JTok	Sentence splitting	English	A
DFKI	JTok	Tokenization	English	A
DFKI	JTok	Sentence splitting	German	A
DFKI	JTok	Tokenization	German	A
DFKI	JTok	Sentence splitting	Italian	A
DFKI	JTok	Tokenization	Italian	A
DFKI	Lynx-Legal NER	Named Entity Recognition	German	A
DFKI	Lynx/QURATOR BERTNER	Named Entity Recognition	English	A
DFKI	Lynx/QURATOR BERTNER	Named Entity Recognition	German	A
DFKI	Lynx/QURATOR Summarization	Summarization	English	A
DFKI	MMorph3	Morphological analyser	English	A
DFKI	MMorph3	Morphological analyser	French	A
DFKI	MMorph3	Morphological analyser	German	A
DFKI	MMorph3	Morphological analyser	Spanish	A
DFKI	MMorph3	Morphological analyser	Dutch	A
DFKI	MMorph3	Morphological analyser	Italian	A
HENS	HENS language ID	Language identification	Czech	A

R1

Provider	Tool	Service	Lang./Category	
HENS	HENS language ID	Language identification	English	A
HENS	HENS language ID	Language identification	French	A
HENS	HENS language ID	Language identification	German	A
HENS	HENS language ID	Language identification	Greek	A
HENS	HENS language ID	Language identification	Spanish	A
HENS	HENS NER	Named Entity Recognition	Czech	A
HENS	HENS NER	Named Entity Recognition	English	A
HENS	HENS NER	Named Entity Recognition	French	A
HENS	HENS NER	Named Entity Recognition	German	A
HENS	HENS NER	Named Entity Recognition	Greek	A
HENS	HENS NER	Named Entity Recognition	Spanish	A
HENS	HENS polarity analysis	Sentiment Analysis	English	A
HENS	HENS polarity analysis	Sentiment Analysis	French	A
HENS	HENS polarity analysis	Sentiment Analysis	German	A
HENS	HENS polarity analysis	Sentiment Analysis	Spanish	A
USFD	BioYODIE (Full)	NER Disambiguation	English	A
USFD	BioYODIE (MeSH Only)	NER Disambiguation	English	A
USFD	BioYODIE (Snomed)	NER Disambiguation	English	A
USFD	Brexit Analyzer	Categorization	English	A
USFD	Brexit Analyzer	Named Entity Recognition	English	A
USFD	DecarboNET Environmental Annotator	Named Entity Recognition	English	A
USFD	DecarboNET Environmental Annotator	Named Entity Recognition	German	A
USFD	GATE Cloud: ANNIE	Named Entity Recognition	English	A
USFD	GATE Cloud: French NER	Named Entity Recognition	French	A
USFD	GATE Cloud: French NER for Tweets	Named Entity Recognition	French	A
USFD	GATE Cloud: Generic Opinion Mining	Opinion Mining	English	A
USFD	GATE Cloud: Generic Opinion Mining for Tweets	Opinion Mining	English	A
USFD	GATE Cloud: German NER	Named Entity Recognition	German	A
USFD	GATE Cloud: German NER for Tweets	Named Entity Recognition	German	A
USFD	GATE Cloud: Langugage ID for Tweets	Language identification	English	A
USFD	GATE Cloud: Langugage ID for Tweets	Language identification	French	A
USFD	GATE Cloud: Langugage ID for Tweets	Language identification	German	A
USFD	GATE Cloud: Langugage ID for Tweets	Language identification	Spanish	A
USFD	GATE Cloud: Measurement Annotator	Number annotation	English	A
USFD	GATE Cloud: OpenNLP Pipelines	Named Entity Recognition	English	A
USFD	GATE Cloud: OpenNLP Pipelines	Named Entity Recognition	German	A
USFD	GATE Cloud: OpenNLP Pipelines	Part of Speech tagging	English	A
USFD	GATE Cloud: OpenNLP Pipelines	Part of Speech tagging	German	A
USFD	GATE Cloud: OpenNLP Pipelines	Sentence splitting	English	A
USFD	GATE Cloud: OpenNLP Pipelines	Sentence splitting	German	A
USFD	GATE Cloud: OpenNLP Pipelines	Tokenization	English	A
USFD	GATE Cloud: OpenNLP Pipelines	Tokenization	German	A
USFD	GATE Cloud: POS and Morph	Morphological analyser	English	A

R1

Provider	Tool	Service	Lang./Category	
USFD	GATE Cloud: POS and Morph	Part of Speech tagging	English	A
USFD	GATE Cloud: Tweet POS Tagger	Part of Speech tagging	English	A
USFD	GATE Cloud: Tweet tokenizer	Tokenization	English	A
USFD	Political Futures Tracker	Categorization	English	A
USFD	Political Futures Tracker	Named Entity Recognition	English	A
USFD	Rumour Veracity Classifier	Categorization	English	A
USFD	SUMMA	Summarization	English	A
USFD	SUMMA	Summarization	Spanish	A
USFD	Tweet User Classification	Categorization	English	A
USFD	Tweet User Classification	Named Entity Recognition	English	A
USFD	TwitIE	Named Entity Recognition	English	A
USFD	Universal Dependencies POS Tagger	Part of Speech tagging	Czech	A
USFD	Universal Dependencies POS Tagger	Part of Speech tagging	French	A
USFD	Universal Dependencies POS Tagger	Part of Speech tagging	Greek	A
USFD	Universal Dependencies POS Tagger	Part of Speech tagging	Latvian	A
USFD	Universal Dependencies POS Tagger	Part of Speech tagging	Spanish	A
USFD	YODIE	NER Disambiguation	English	A
USFD	YODIE	NER Disambiguation	French	A
USFD	YODIE	NER Disambiguation	German	A
USFD	YODIE	NER Disambiguation	Spanish	A
ILSP	ILSP-ABSA	Sentiment Analysis	Greek	A
ILSP	ILSP-Events-physical-attack	Information Extraction	Greek	A
ILSP	ILSP-Events-protest	Information Extraction	Greek	A
ILSP	ILSP-NER	Named Entity Recognition	English	A
ILSP	ILSP-NER	Named Entity Recognition	Greek	A

Table 22: IE and Text Analysis tools and services to integrate in the first release (full list)

R2

Provider	Tool	Service	Lang./Category	
CUNI	UDPipe parser	Dependency Parsing	Bulgarian	A
CUNI	UDPipe parser	Dependency Parsing	Croatian	A
CUNI	UDPipe parser	Dependency Parsing	Danish	A
CUNI	UDPipe parser	Dependency Parsing	Dutch	A
CUNI	UDPipe parser	Dependency Parsing	Estonian	A
CUNI	UDPipe parser	Dependency Parsing	Finnish	A
CUNI	UDPipe parser	Dependency Parsing	Hungarian	A
CUNI	UDPipe parser	Dependency Parsing	Irish	A
CUNI	UDPipe parser	Dependency Parsing	Italian	A
CUNI	UDPipe parser	Dependency Parsing	Lithuanian	A
CUNI	UDPipe parser	Dependency Parsing	Maltese	A
CUNI	UDPipe parser	Dependency Parsing	Polish	A
CUNI	UDPipe parser	Dependency Parsing	Portuguese	A
CUNI	UDPipe parser	Dependency Parsing	Romanian	A

R2

Provider	Tool	Service	Lang./Category	
CUNI	UDPipe parser	Dependency Parsing	Slovak	A
CUNI	UDPipe parser	Dependency Parsing	Slovenian	A
CUNI	UDPipe parser	Dependency Parsing	Swedish	A
CUNI	UDPipe parser	Dependency Parsing	Basque	B
CUNI	UDPipe parser	Dependency Parsing	Catalan	B
CUNI	UDPipe parser	Dependency Parsing	Galician	B
CUNI	UDPipe parser	Dependency Parsing	Norwegian	B
CUNI	UDPipe parser	Dependency Parsing	Serbian	B
CUNI	UDPipe parser	Dependency Parsing	Turkish	B
CUNI	UDPipe parser	Dependency Parsing	Ukrainian	B
CUNI	UDPipe tagger	Lemmatization	Bulgarian	A
CUNI	UDPipe tagger	Lemmatization	Croatian	A
CUNI	UDPipe tagger	Lemmatization	Danish	A
CUNI	UDPipe tagger	Lemmatization	Dutch	A
CUNI	UDPipe tagger	Lemmatization	Estonian	A
CUNI	UDPipe tagger	Lemmatization	Finnish	A
CUNI	UDPipe tagger	Lemmatization	Hungarian	A
CUNI	UDPipe tagger	Lemmatization	Irish	A
CUNI	UDPipe tagger	Lemmatization	Italian	A
CUNI	UDPipe tagger	Lemmatization	Lithuanian	A
CUNI	UDPipe tagger	Lemmatization	Maltese	A
CUNI	UDPipe tagger	Lemmatization	Polish	A
CUNI	UDPipe tagger	Lemmatization	Portuguese	A
CUNI	UDPipe tagger	Lemmatization	Romanian	A
CUNI	UDPipe tagger	Lemmatization	Slovak	A
CUNI	UDPipe tagger	Lemmatization	Slovenian	A
CUNI	UDPipe tagger	Lemmatization	Swedish	A
CUNI	UDPipe tagger	Lemmatization	Basque	B
CUNI	UDPipe tagger	Lemmatization	Catalan	B
CUNI	UDPipe tagger	Lemmatization	Galician	B
CUNI	UDPipe tagger	Lemmatization	Norwegian	B
CUNI	UDPipe tagger	Lemmatization	Serbian	B
CUNI	UDPipe tagger	Lemmatization	Turkish	B
CUNI	UDPipe tagger	Lemmatization	Ukrainian	B
CUNI	UDPipe tagger	Morphological analyser	Bulgarian	A
CUNI	UDPipe tagger	Morphological analyser	Croatian	A
CUNI	UDPipe tagger	Morphological analyser	Danish	A
CUNI	UDPipe tagger	Morphological analyser	Dutch	A
CUNI	UDPipe tagger	Morphological analyser	Estonian	A
CUNI	UDPipe tagger	Morphological analyser	Finnish	A
CUNI	UDPipe tagger	Morphological analyser	Hungarian	A
CUNI	UDPipe tagger	Morphological analyser	Irish	A
CUNI	UDPipe tagger	Morphological analyser	Italian	A
CUNI	UDPipe tagger	Morphological analyser	Lithuanian	A

R2

Provider	Tool	Service	Lang./Category	
CUNI	UDPipe tagger	Morphological analyser	Maltese	A
CUNI	UDPipe tagger	Morphological analyser	Polish	A
CUNI	UDPipe tagger	Morphological analyser	Portuguese	A
CUNI	UDPipe tagger	Morphological analyser	Romanian	A
CUNI	UDPipe tagger	Morphological analyser	Slovak	A
CUNI	UDPipe tagger	Morphological analyser	Slovenian	A
CUNI	UDPipe tagger	Morphological analyser	Swedish	A
CUNI	UDPipe tagger	Morphological analyser	Basque	B
CUNI	UDPipe tagger	Morphological analyser	Catalan	B
CUNI	UDPipe tagger	Morphological analyser	Galician	B
CUNI	UDPipe tagger	Morphological analyser	Norwegian	B
CUNI	UDPipe tagger	Morphological analyser	Serbian	B
CUNI	UDPipe tagger	Morphological analyser	Turkish	B
CUNI	UDPipe tagger	Morphological analyser	Ukrainian	B
CUNI	UDPipe tagger	Part of Speech tagging	Bulgarian	A
CUNI	UDPipe tagger	Part of Speech tagging	Croatian	A
CUNI	UDPipe tagger	Part of Speech tagging	Danish	A
CUNI	UDPipe tagger	Part of Speech tagging	Dutch	A
CUNI	UDPipe tagger	Part of Speech tagging	Estonian	A
CUNI	UDPipe tagger	Part of Speech tagging	Finnish	A
CUNI	UDPipe tagger	Part of Speech tagging	Hungarian	A
CUNI	UDPipe tagger	Part of Speech tagging	Irish	A
CUNI	UDPipe tagger	Part of Speech tagging	Italian	A
CUNI	UDPipe tagger	Part of Speech tagging	Lithuanian	A
CUNI	UDPipe tagger	Part of Speech tagging	Maltese	A
CUNI	UDPipe tagger	Part of Speech tagging	Polish	A
CUNI	UDPipe tagger	Part of Speech tagging	Portuguese	A
CUNI	UDPipe tagger	Part of Speech tagging	Romanian	A
CUNI	UDPipe tagger	Part of Speech tagging	Slovak	A
CUNI	UDPipe tagger	Part of Speech tagging	Slovenian	A
CUNI	UDPipe tagger	Part of Speech tagging	Swedish	A
CUNI	UDPipe tagger	Part of Speech tagging	Basque	B
CUNI	UDPipe tagger	Part of Speech tagging	Catalan	B
CUNI	UDPipe tagger	Part of Speech tagging	Galician	B
CUNI	UDPipe tagger	Part of Speech tagging	Norwegian	B
CUNI	UDPipe tagger	Part of Speech tagging	Serbian	B
CUNI	UDPipe tagger	Part of Speech tagging	Turkish	B
CUNI	UDPipe tagger	Part of Speech tagging	Ukrainian	B
CUNI	UDPipe tokenizer	Tokenization	Bulgarian	A
CUNI	UDPipe tokenizer	Tokenization	Croatian	A
CUNI	UDPipe tokenizer	Tokenization	Danish	A
CUNI	UDPipe tokenizer	Tokenization	Dutch	A
CUNI	UDPipe tokenizer	Tokenization	Estonian	A
CUNI	UDPipe tokenizer	Tokenization	Finnish	A

R2

Provider	Tool	Service	Lang./Category	
CUNI	UDPipe tokenizer	Tokenization	Hungarian	A
CUNI	UDPipe tokenizer	Tokenization	Irish	A
CUNI	UDPipe tokenizer	Tokenization	Italian	A
CUNI	UDPipe tokenizer	Tokenization	Lithuanian	A
CUNI	UDPipe tokenizer	Tokenization	Maltese	A
CUNI	UDPipe tokenizer	Tokenization	Polish	A
CUNI	UDPipe tokenizer	Tokenization	Portuguese	A
CUNI	UDPipe tokenizer	Tokenization	Romanian	A
CUNI	UDPipe tokenizer	Tokenization	Slovak	A
CUNI	UDPipe tokenizer	Tokenization	Slovenian	A
CUNI	UDPipe tokenizer	Tokenization	Swedish	A
CUNI	UDPipe tokenizer	Tokenization	Basque	B
CUNI	UDPipe tokenizer	Tokenization	Catalan	B
CUNI	UDPipe tokenizer	Tokenization	Galician	B
CUNI	UDPipe tokenizer	Tokenization	Norwegian	B
CUNI	UDPipe tokenizer	Tokenization	Serbian	B
CUNI	UDPipe tokenizer	Tokenization	Turkish	B
CUNI	UDPipe tokenizer	Tokenization	Ukrainian	B
Expert System	Cogito Discover	Keyword extraction	Dutch	A
Expert System	Cogito Discover	Keyword extraction	English	A
Expert System	Cogito Discover	Keyword extraction	French	A
Expert System	Cogito Discover	Keyword extraction	German	A
Expert System	Cogito Discover	Keyword extraction	Italian	A
Expert System	Cogito Discover	Keyword extraction	Portuguese	A
Expert System	Cogito Discover	Keyword extraction	Spanish	A
Expert System	Cogito Discover	Language identification	Bulgarian	A
Expert System	Cogito Discover	Language identification	Croatian	A
Expert System	Cogito Discover	Language identification	Danish	A
Expert System	Cogito Discover	Language identification	Dutch	A
Expert System	Cogito Discover	Language identification	Estonian	A
Expert System	Cogito Discover	Language identification	Finnish	A
Expert System	Cogito Discover	Language identification	Hungarian	A
Expert System	Cogito Discover	Language identification	Italian	A
Expert System	Cogito Discover	Language identification	Lithuanian	A
Expert System	Cogito Discover	Language identification	Polish	A
Expert System	Cogito Discover	Language identification	Portuguese	A
Expert System	Cogito Discover	Language identification	Romanian	A
Expert System	Cogito Discover	Language identification	Slovak	A
Expert System	Cogito Discover	Language identification	Slovenian	A
Expert System	Cogito Discover	Language identification	Swedish	A
Expert System	Cogito Discover	Language identification	Albanian	B
Expert System	Cogito Discover	Language identification	Norwegian	B
Expert System	Cogito Discover	Language identification	Turkish	B
Expert System	Cogito Discover	Language identification	Ukrainian	B

R2

Provider	Tool	Service	Lang./Category	
Expert System	Cogito Discover	Lemmatization	Dutch	A
Expert System	Cogito Discover	Lemmatization	Italian	A
Expert System	Cogito Discover	Lemmatization	Portuguese	A
Expert System	Cogito Discover	Named Entity Recognition	Dutch	A
Expert System	Cogito Discover	Named Entity Recognition	Italian	A
Expert System	Cogito Discover	Named Entity Recognition	Portuguese	A
Expert System	Cogito Discover	Part-of-Speech Tagging	Dutch	A
Expert System	Cogito Discover	Part-of-Speech Tagging	Italian	A
Expert System	Cogito Discover	Part-of-Speech Tagging	Portuguese	A
Expert System	Cogito Discover	Sentiment Analysis	Italian	A
Expert System	Cogito Discover	Sentiment Analysis	Dutch	A
Expert System	Cogito Discover	Sentiment Analysis	Portuguese	A
Expert System	Cogito Discover	Summarization	Dutch	A
Expert System	Cogito Discover	Summarization	Italian	A
Expert System	Cogito Discover	Summarization	Portuguese	A
Expert System	Cogito Discover	Semantic annotation	Dutch	A
Expert System	Cogito Discover	Semantic annotation	English	A
Expert System	Cogito Discover	Semantic annotation	French	A
Expert System	Cogito Discover	Semantic annotation	German	A
Expert System	Cogito Discover	Semantic annotation	Italian	A
Expert System	Cogito Discover	Semantic annotation	Portuguese	A
Expert System	Cogito Discover	Semantic annotation	Spanish	A
DFKI	German Shallow Discourse Parser	Discourse Parsing	German	A
DFKI	Lynx/QURATOR Summarization	Summarization	German	A
DFKI	Lynx-TIMEX	Date detection	English	A
DFKI	Lynx-TIMEX	Date detection	German	A
DFKI	Qurator-LangIdent	Language identification	Czech	A
DFKI	Qurator-LangIdent	Language identification	English	A
DFKI	Qurator-LangIdent	Language identification	French	A
DFKI	Qurator-LangIdent	Language identification	German	A
DFKI	Qurator-LangIdent	Language identification	Greek	A
DFKI	Qurator-LangIdent	Language identification	Latvian	A
DFKI	Qurator-LangIdent	Language identification	Spanish	A
DFKI	Qurator-LangIdent	Language identification	Bulgarian	A
DFKI	Qurator-LangIdent	Language identification	Croatian	A
DFKI	Qurator-LangIdent	Language identification	Danish	A
DFKI	Qurator-LangIdent	Language identification	Dutch	A
DFKI	Qurator-LangIdent	Language identification	Estonian	A
DFKI	Qurator-LangIdent	Language identification	Finnish	A
DFKI	Qurator-LangIdent	Language identification	Hungarian	A
DFKI	Qurator-LangIdent	Language identification	Italian	A
DFKI	Qurator-LangIdent	Language identification	Lithuanian	A
DFKI	Qurator-LangIdent	Language identification	Polish	A
DFKI	Qurator-LangIdent	Language identification	Portuguese	A

R2

Provider	Tool	Service	Lang./Category	
DFKI	Qurator-LangIdent	Language identification	Romanian	A
DFKI	Qurator-LangIdent	Language identification	Slovak	A
DFKI	Qurator-LangIdent	Language identification	Slovenian	A
DFKI	Qurator-LangIdent	Language identification	Swedish	A
DFKI	Qurator-LangIdent	Language identification	Albanian	B
DFKI	Qurator-LangIdent	Language identification	Catalan	B
DFKI	Qurator-LangIdent	Language identification	Norwegian	B
DFKI	Qurator-LangIdent	Language identification	Turkish	B
DFKI	Qurator-LangIdent	Language identification	Ukrainian	B
DFKI	Qurator-LangIdent	Language identification	Welsh	B
DFKI	Qurator-LangIdent	Language identification	Afrikaans	C
DFKI	Qurator-LangIdent	Language identification	Arabic	C
DFKI	Qurator-LangIdent	Language identification	Chinese	C
DFKI	Qurator-LangIdent	Language identification	Hebrew	C
DFKI	Qurator-LangIdent	Language identification	Hindi/Urdu	C
DFKI	Qurator-LangIdent	Language identification	Indonesian	C
DFKI	Qurator-LangIdent	Language identification	Japanese	C
DFKI	Qurator-LangIdent	Language identification	Korean	C
DFKI	Qurator-LangIdent	Language identification	Malay	C
DFKI	Qurator-LangIdent	Language identification	Persian	C
DFKI	Qurator-LangIdent	Language identification	Russian	C
DFKI	Qurator-LangIdent	Language identification	Tamil	C
DFKI	Qurator-LangIdent	Language identification	Vietnamese	C
DFKI	Qurator-LangIdent	Language identification	Bengali	D
DFKI	Qurator-LangIdent	Language identification	Gujarati	D
DFKI	Qurator-LangIdent	Language identification	Kannada	D
DFKI	Qurator-LangIdent	Language identification	Macedonian	D
DFKI	Qurator-LangIdent	Language identification	Marahati	D
DFKI	Qurator-LangIdent	Language identification	Nepali	D
DFKI	Qurator-LangIdent	Language identification	Panjabi	D
DFKI	Qurator-LangIdent	Language identification	Somali	D
DFKI	Qurator-LangIdent	Language identification	Swahili	D
DFKI	Qurator-LangIdent	Language identification	Tagalog	D
DFKI	Qurator-LangIdent	Language identification	Telugu	D
DFKI	Qurator-LangIdent	Language identification	Thai	D
DFKI	Qurator-LangIdent	Language identification	Urdu	D
HENS	HENS KWS	Keyword extraction	Dutch	A
HENS	HENS KWS	Keyword extraction	English	A
HENS	HENS KWS	Keyword extraction	French	A
HENS	HENS KWS	Keyword extraction	German	A
HENS	HENS KWS	Keyword extraction	Greek	A
HENS	HENS KWS	Keyword extraction	Italian	A
HENS	HENS KWS	Keyword extraction	Polish	A
HENS	HENS KWS	Keyword extraction	Romanian	A

R2

Provider	Tool	Service	Lang./Category	
HENS	HENS KWS	Keyword extraction	Spanish	A
HENS	HENS KWS	Keyword extraction	Albanian	B
HENS	HENS KWS	Keyword extraction	Norwegian	B
HENS	HENS KWS	Keyword extraction	Turkish	B
HENS	HENS language ID	Language identification	Bulgarian	A
HENS	HENS language ID	Language identification	Dutch	A
HENS	HENS language ID	Language identification	Hungarian	A
HENS	HENS language ID	Language identification	Italian	A
HENS	HENS language ID	Language identification	Polish	A
HENS	HENS language ID	Language identification	Portuguese	A
HENS	HENS language ID	Language identification	Romanian	A
HENS	HENS language ID	Language identification	Slovak	A
HENS	HENS language ID	Language identification	Swedish	A
HENS	HENS language ID	Language identification	Albanian	B
HENS	HENS language ID	Language identification	Norwegian	B
HENS	HENS language ID	Language identification	Turkish	B
HENS	HENS NER	Named Entity Recognition	Bulgarian	A
HENS	HENS NER	Named Entity Recognition	Croatian	A
HENS	HENS NER	Named Entity Recognition	Dutch	A
HENS	HENS NER	Named Entity Recognition	Hungarian	A
HENS	HENS NER	Named Entity Recognition	Italian	A
HENS	HENS NER	Named Entity Recognition	Polish	A
HENS	HENS NER	Named Entity Recognition	Portuguese	A
HENS	HENS NER	Named Entity Recognition	Romanian	A
HENS	HENS NER	Named Entity Recognition	Slovak	A
HENS	HENS NER	Named Entity Recognition	Swedish	A
HENS	HENS NER	Named Entity Recognition	Albanian	B
HENS	HENS NER	Named Entity Recognition	Catalan	B
HENS	HENS NER	Named Entity Recognition	Norwegian	B
HENS	HENS NER	Named Entity Recognition	Turkish	B
HENS	HENS polarity analysis	Sentiment Analysis	Italian	A
HENS	HENS polarity analysis	Sentiment Analysis	Polish	A
HENS	HENS polarity analysis	Sentiment Analysis	Portuguese	A
USFD	DecarboNET Environmental Annotator	Entity linking	English	A
USFD	DecarboNET Environmental Annotator	Entity linking	German	A
USFD	GATE Cloud: Langugage ID for Tweets	Language identification	Dutch	A
USFD	GATE Cloud: Measurement Annotator	Measurement annotation	English	A
USFD	GATE Cloud: Measurement Annotator	Measurement normalisation	English	A
USFD	GATE Cloud: Measurement Annotator	Number normalisation	English	A
USFD	GATE Cloud: NP Chunker	Noun phrase extraction	English	A
USFD	GATE Cloud: OpenNLP Pipelines	Named Entity Recognition	Dutch	A
USFD	GATE Cloud: OpenNLP Pipelines	Part of Speech tagging	Dutch	A
USFD	GATE Cloud: OpenNLP Pipelines	Sentence splitting	Dutch	A
USFD	GATE Cloud: OpenNLP Pipelines	Tokenization	Dutch	A

R2

Provider	Tool	Service	Lang./Category	
USFD	GATE Cloud: Romanian NER	Named Entity Recognition	Romanian	A
USFD	GATE Cloud: Welsh NER	Named Entity Recognition	Welsh	B
USFD	Universal Dependencies POS Tagger	Part of Speech tagging	Bulgarian	A
USFD	Universal Dependencies POS Tagger	Part of Speech tagging	Croatian	A
USFD	Universal Dependencies POS Tagger	Part of Speech tagging	Danish	A
USFD	Universal Dependencies POS Tagger	Part of Speech tagging	Dutch	A
USFD	Universal Dependencies POS Tagger	Part of Speech tagging	Estonian	A
USFD	Universal Dependencies POS Tagger	Part of Speech tagging	Finnish	A
USFD	Universal Dependencies POS Tagger	Part of Speech tagging	Polish	A
USFD	Universal Dependencies POS Tagger	Part of Speech tagging	Portuguese	A
USFD	Universal Dependencies POS Tagger	Part of Speech tagging	Romanian	A
USFD	Universal Dependencies POS Tagger	Part of Speech tagging	Slovak	A
USFD	Universal Dependencies POS Tagger	Part of Speech tagging	Slovenian	A
USFD	Universal Dependencies POS Tagger	Part of Speech tagging	Swedish	A
USFD	Universal Dependencies POS Tagger	Part of Speech tagging	Basque	B
USFD	Universal Dependencies POS Tagger	Part of Speech tagging	Catalan	B

Table 23: IE and Text Analysis tools and services to integrate in the second release (full list)

R3

Provider	Tool	Service	Lang./Category	
CUNI	UDPipe parser	Dependency Parsing	Afrikaans	C
CUNI	UDPipe parser	Dependency Parsing	Arabic	C
CUNI	UDPipe parser	Dependency Parsing	Chinese	C
CUNI	UDPipe parser	Dependency Parsing	Hebrew	C
CUNI	UDPipe parser	Dependency Parsing	Hindi/Urdu	C
CUNI	UDPipe parser	Dependency Parsing	Indonesian	C
CUNI	UDPipe parser	Dependency Parsing	Japanese	C
CUNI	UDPipe parser	Dependency Parsing	Korean	C
CUNI	UDPipe parser	Dependency Parsing	Latin	C
CUNI	UDPipe parser	Dependency Parsing	Persian	C
CUNI	UDPipe parser	Dependency Parsing	Russian	C
CUNI	UDPipe parser	Dependency Parsing	Tamil	C
CUNI	UDPipe parser	Dependency Parsing	Vietnamese	C
CUNI	UDPipe tagger	Lemmatization	Afrikaans	C
CUNI	UDPipe tagger	Lemmatization	Arabic	C
CUNI	UDPipe tagger	Lemmatization	Chinese	C
CUNI	UDPipe tagger	Lemmatization	Hebrew	C
CUNI	UDPipe tagger	Lemmatization	Hindi/Urdu	C
CUNI	UDPipe tagger	Lemmatization	Indonesian	C
CUNI	UDPipe tagger	Lemmatization	Japanese	C
CUNI	UDPipe tagger	Lemmatization	Latin	C
CUNI	UDPipe tagger	Lemmatization	Persian	C
CUNI	UDPipe tagger	Lemmatization	Russian	C

		R3		
Provider	**Tool**	**Service**	**Lang./Category**	
CUNI	UDPipe tagger	Lemmatization	Tamil	C
CUNI	UDPipe tagger	Lemmatization	Vietnamese	C
CUNI	UDPipe tagger	Morphological analyser	Afrikaans	C
CUNI	UDPipe tagger	Morphological analyser	Arabic	C
CUNI	UDPipe tagger	Morphological analyser	Chinese	C
CUNI	UDPipe tagger	Morphological analyser	Hebrew	C
CUNI	UDPipe tagger	Morphological analyser	Hindi/Urdu	C
CUNI	UDPipe tagger	Morphological analyser	Indonesian	C
CUNI	UDPipe tagger	Morphological analyser	Japanese	C
CUNI	UDPipe tagger	Morphological analyser	Korean	C
CUNI	UDPipe tagger	Morphological analyser	Latin	C
CUNI	UDPipe tagger	Morphological analyser	Persian	C
CUNI	UDPipe tagger	Morphological analyser	Russian	C
CUNI	UDPipe tagger	Morphological analyser	Tamil	C
CUNI	UDPipe tagger	Morphological analyser	Vietnamese	C
CUNI	UDPipe tagger	Part of Speech tagging	Afrikaans	C
CUNI	UDPipe tagger	Part of Speech tagging	Arabic	C
CUNI	UDPipe tagger	Part of Speech tagging	Chinese	C
CUNI	UDPipe tagger	Part of Speech tagging	Hebrew	C
CUNI	UDPipe tagger	Part of Speech tagging	Hindi/Urdu	C
CUNI	UDPipe tagger	Part of Speech tagging	Indonesian	C
CUNI	UDPipe tagger	Part of Speech tagging	Japanese	C
CUNI	UDPipe tagger	Part of Speech tagging	Korean	C
CUNI	UDPipe tagger	Part of Speech tagging	Latin	C
CUNI	UDPipe tagger	Part of Speech tagging	Persian	C
CUNI	UDPipe tagger	Part of Speech tagging	Russian	C
CUNI	UDPipe tagger	Part of Speech tagging	Tamil	C
CUNI	UDPipe tagger	Part of Speech tagging	Vietnamese	C
CUNI	UDPipe tokenizer	Tokenization	Afrikaans	C
CUNI	UDPipe tokenizer	Tokenization	Arabic	C
CUNI	UDPipe tokenizer	Tokenization	Chinese	C
CUNI	UDPipe tokenizer	Tokenization	Hebrew	C
CUNI	UDPipe tokenizer	Tokenization	Hindi/Urdu	C
CUNI	UDPipe tokenizer	Tokenization	Indonesian	C
CUNI	UDPipe tokenizer	Tokenization	Japanese	C
CUNI	UDPipe tokenizer	Tokenization	Korean	C
CUNI	UDPipe tokenizer	Tokenization	Latin	C
CUNI	UDPipe tokenizer	Tokenization	Persian	C
CUNI	UDPipe tokenizer	Tokenization	Russian	C
CUNI	UDPipe tokenizer	Tokenization	Tamil	C
CUNI	UDPipe tokenizer	Tokenization	Vietnamese	C
Expert System	Cogito Discover	Keyword extraction	Arabic	C
Expert System	Cogito Discover	Keyword extraction	Chinese	C
Expert System	Cogito Discover	Keyword extraction	Japanese	C

		R3		
Provider	**Tool**	**Service**	**Lang./Category**	
Expert System	Cogito Discover	Keyword extraction	Korean	C
Expert System	Cogito Discover	Keyword extraction	Russian	C
Expert System	Cogito Discover	Language identification	Afrikaans	C
Expert System	Cogito Discover	Language identification	Arabic	C
Expert System	Cogito Discover	Language identification	Chinese	C
Expert System	Cogito Discover	Language identification	Hebrew	C
Expert System	Cogito Discover	Language identification	Hindi/Urdu	C
Expert System	Cogito Discover	Language identification	Indonesian	C
Expert System	Cogito Discover	Language identification	Japanese	C
Expert System	Cogito Discover	Language identification	Korean	C
Expert System	Cogito Discover	Language identification	Persian	C
Expert System	Cogito Discover	Language identification	Russian	C
Expert System	Cogito Discover	Language identification	Tamil	C
Expert System	Cogito Discover	Language identification	Vietnamese	C
Expert System	Cogito Discover	Language identification	Bengali	D
Expert System	Cogito Discover	Language identification	Gujarati	D
Expert System	Cogito Discover	Language identification	Kannada	D
Expert System	Cogito Discover	Language identification	Macedonian	D
Expert System	Cogito Discover	Language identification	Marahati	D
Expert System	Cogito Discover	Language identification	Nepali	D
Expert System	Cogito Discover	Language identification	Panjabi	D
Expert System	Cogito Discover	Language identification	Somali	D
Expert System	Cogito Discover	Language identification	Swahili	D
Expert System	Cogito Discover	Language identification	Tagalog	D
Expert System	Cogito Discover	Language identification	Telugu	D
Expert System	Cogito Discover	Language identification	Thai	D
Expert System	Cogito Discover	Language identification	Urdu	D
Expert System	Cogito Discover	Lemmatization	Arabic	C
Expert System	Cogito Discover	Lemmatization	Chinese	C
Expert System	Cogito Discover	Lemmatization	Japanese	C
Expert System	Cogito Discover	Lemmatization	Korean	C
Expert System	Cogito Discover	Lemmatization	Russian	C
Expert System	Cogito Discover	Named Entity Recognition	Arabic	C
Expert System	Cogito Discover	Named Entity Recognition	Chinese	C
Expert System	Cogito Discover	Named Entity Recognition	Japanese	C
Expert System	Cogito Discover	Named Entity Recognition	Korean	C
Expert System	Cogito Discover	Named Entity Recognition	Russian	C
Expert System	Cogito Discover	Part-of-Speech Tagging	Arabic	C
Expert System	Cogito Discover	Part-of-Speech Tagging	Chinese	C
Expert System	Cogito Discover	Part-of-Speech Tagging	Japanese	C
Expert System	Cogito Discover	Part-of-Speech Tagging	Korean	C
Expert System	Cogito Discover	Part-of-Speech Tagging	Russian	C
Expert System	Cogito Discover	Summarization	Arabic	C
Expert System	Cogito Discover	Summarization	Chinese	C

		R3		
Provider	**Tool**	**Service**	**Lang./Category**	
Expert System	Cogito Discover	Summarization	Japanese	C
Expert System	Cogito Discover	Summarization	Korean	C
Expert System	Cogito Discover	Summarization	Russian	C
Expert System	Cogito Discover	Semantic annotation	Arabic	C
Expert System	Cogito Discover	Semantic annotation	Chinese	C
Expert System	Cogito Discover	Semantic annotation	Japanese	C
Expert System	Cogito Discover	Semantic annotation	Korean	C
Expert System	Cogito Discover	Semantic annotation	Russian	C
DFKI	Dependency Tree Parser for German Clinical Text	Parsing	German	A
DFKI	Credibility score	Fake news dectection	German	A
DFKI	Document classification	Classification	German	A
DFKI	Multi-document summarizer	Summarization	English	A
DFKI	Multi-document summarizer	Summarization	German	A
DFKI	Multi-document summarizer	Summarization	French	A
DFKI	Political Bias Classifier	Classification	German	A
HENS	HENS KWS	Keyword extraction	Arabic	C
HENS	HENS KWS	Keyword extraction	Chinese	C
HENS	HENS KWS	Keyword extraction	Hebrew	C
HENS	HENS KWS	Keyword extraction	Hindi/Urdu	C
HENS	HENS KWS	Keyword extraction	Indonesian	C
HENS	HENS KWS	Keyword extraction	Malay	C
HENS	HENS KWS	Keyword extraction	Pashto	C
HENS	HENS KWS	Keyword extraction	Persian	C
HENS	HENS KWS	Keyword extraction	Russian	C
HENS	HENS language ID	Language identification	Arabic	C
HENS	HENS language ID	Language identification	Hebrew	C
HENS	HENS language ID	Language identification	Hindi/Urdu	C
HENS	HENS language ID	Language identification	Indonesian	C
HENS	HENS language ID	Language identification	Malay	C
HENS	HENS language ID	Language identification	Pashto	C
HENS	HENS language ID	Language identification	Persian	C
HENS	HENS language ID	Language identification	Russian	C
HENS	HENS language ID	Language identification	Language independent	E
HENS	HENS NER	Named Entity Recognition	Arabic	C
HENS	HENS NER	Named Entity Recognition	Chinese	C
HENS	HENS NER	Named Entity Recognition	Hebrew	C
HENS	HENS NER	Named Entity Recognition	Hindi/Urdu	C
HENS	HENS NER	Named Entity Recognition	Indonesian	C
HENS	HENS NER	Named Entity Recognition	Malay	C
HENS	HENS NER	Named Entity Recognition	Pashto	C
HENS	HENS NER	Named Entity Recognition	Persian	C
HENS	HENS NER	Named Entity Recognition	Russian	C
HENS	HENS polarity analysis	Sentiment Analysis	Arabic	C

		R3		
Provider	**Tool**	**Service**	**Lang./Category**	
HENS	HENS polarity analysis	Sentiment Analysis	Indonesian	C
HENS	HENS polarity analysis	Sentiment Analysis	Malay	C
HENS	HENS polarity analysis	Sentiment Analysis	Russian	C
HENS	HENS summarization	Summarization	Language independent	E
USFD	GATE Cloud: Russian NER	Named Entity Recognition	Russian	C
USFD	GATE Cloud: Russian NER (basic version)	Named Entity Recognition	Russian	C
USFD	Universal Dependencies POS Tagger	Part of Speech tagging	Indonesian	C
USFD	Universal Dependencies POS Tagger	Part of Speech tagging	Russian	C
USFD	GATE: COVID-19 claim categoriser	Categorization	English	A
USFD	GATE: COVID-19 vaccine text categoriser	Categorization	English	A
USFD	GATE Hate	Sentiment analysis	English	A
USFD	GATE: Offensive language classifier	Sentiment analysis	English	A
USFD	GATE: Toxic language classifier	Sentiment analysis	English	A
USFD	GATE: ChemDataExtractor	Extraction of domain-specific information	English	A
USFD	GATE: OSCAR4 Chemical Named Entity Recognizer	Extraction of domain-specific information	English	A
USFD	GATE: Journalist Safety Analyser	Event detection	English	A
ILSP	ILSP neural named entity recognizer	Named Entity Recognition	Greek	A
ILSP	ILSP neural dependency parser	Dependency parsing	Greek	A
ILSP	ILSP neural chunker	Chunking	Greek	A
ILSP	ILSP neural text classifier	Classification	Greek	A

Table 24: IE and Text Analysis tools and services to integrate in the third release (full list)