EUROPEAN LANGUAGE GRID

D4.2

Grid Content: Services, Tools and Components (Interim Release)

Authors:	Ian Roberts (USFD), Andres Garcia Silva (EXPSYS), Miroslav Janosik (HENS), Andis Lagzdiņš (Tilde), Nils Feldhus (DFKI), Georg Rehm (DFKI), Dimitris Galanis (ILSP), Dusan Varis (CUNI), Ulrich Germann (UEDIN)

Dissemination Level:	Public
Date:	28-02-2021

About this document

Project	ELG – European Language Grid
Grant agreement no.	825627 – Horizon 2020, ICT 2018-2020 – Innovation Action
Coordinator	Dr. Georg Rehm (DFKI)
Start date, duration	01-01-2019, 42 months (GA amendment version: AMD-825627-7)
Deliverable number	D4.2
Deliverable title	Grid Content: Services, Tools and Components (Interim Release)
Type	Other (report plus deployed software)
Number of pages	41
Status and version	Final
Dissemination level	Public
Date of delivery	Contractual: 28-02-2021 – Actual: 27-02-2021
WP number and title	WP4: Grid Content – Services, Tools, Components
Task number and title	Tasks 4.1: Identify and collect existing services, tools, and components (Section 2), Tasks 4.2–4.5: Integration of existing tools, services, and components (Sections 3–6), Task 4.6 Assist 3rd party providers (Section 7)
Authors	Ian Roberts (USFD), Andres Garcia Silva (EXPSYS), Miroslav Janosik (HENS), Andis Lagzdiņš (Tilde), Nils Feldhus (DFKI), Georg Rehm (DFKI), Dimitris Galanis (ILSP), Dusan Varis (CUNI), Ulrich Germann (UEDIN)
Reviewers	Jan Hajic (CUNI), Victoria Arranz (ELDA)
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), formerly Sail Labs Tech. GmbH (SAIL), Austria
	Expert System Iberia SL (EXPSYS), Spain
	University of Edinburgh (UEDIN), United Kingdom
EC project officers	Philippe Gelin, Alexandru Ceausu
For copies of reports and other ELG-related information, please contact:	DFKI GmbH European Language Grid (ELG) Alt-Moabit 91c D-10559 Berlin Germany 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 © 2021 ELG Consortium

Table of Contents

List of Tables 4
List of Abbreviations 4
Abstract 5
1 Introduction 5
2 Tools, Services and Components: Updates to the integration plan 5
2.1 IE tools updated plan 9
2.2 Machine Translation Tools Updated Plan 14
3 ASR Tools, Services and Components (Task 4.2) 15
3.1 ASR Tools Integrated by HENS (formerly SAIL) 15
3.2 ASR Tools Integrated by TILDE 17
4 IE Tools, Services and Components (Task 4.3) 17
4.1 IE Tools Integrated by EXPSYS 17
4.2 IE Tools Integrated by USFD 19
4.2.1 GATE Cloud 19
4.2.2 Integration approach 19
4.2.3 Deployed components 20
4.3 IE Tools Integrated by DFKI 20
4.4 IE Tools Integrated by HENS 21
4.5 IE Tools Integrated by CUNI 23
5 MT Tools, Services and Components (Task 4.4) 24
5.1 MT and related services integrated by DFKI 24
5.2 MT Tools and Models integrated by UEDIN 24
5.2.1 Dockerized Marian MT REST server 24
5.2.2 MT Models 24
5.3 MT Tools Integrated by TILDE 25
5.4 MT Services integrated by CUNI 25
6 Integration of other Types of Tools, Services, Components (Task 4.5) 25
6.1 TTS tools integrated by DFKI 25
7 Status of tools and services from ELG Pilot Projects and other third parties (Task 4.6) 26
7.1 OPUS-MT Pilot Project 27
7.2 Archaeology NER services from ARIADNEplus 27
8 Conclusion 28
A. Appendix 28

List of Tables

Table 1: Modifications to the integration plan of I.E tools defined in D4.1 5
Table 2: Modifications to the integration plan of MT tools defined in D4.1 6
Table 3: Services and supported language per ELG partner. 8
Table 4: Number of supported languages in each language category per service. 9
Table 5: Overview of IE and Text Analysis services for the first release 9
Table 6: IE and Text Analysis services provided by ELG partners for the first release 10
Table 7: Overview of IE and Text Analysis services to integrate in the second release (category A) 11
Table 8: Overview of IE and Text Analysis services to integrate in the second release (category B, C & D) 12
Table 9: Overview of IE and Text Analysis services to integrate in the third release (category C and D) 13
Table 10: MT tools and services in the first release 14
Table 11: MT tools and services to integrate in the second release 15
Table 12: MT tools and services to integrate in the third release 15
Table 13: HENS ASR tools integrated in R2 16
Table 14: Cogito Discover software components for deployment in ELG 19
Table 15: GATE Cloud services for general linguistic pre-processing and named entity recognition 20
Table 16: DFKI services to be integrated for Release 2 21
Table 17: IE tools provided by HENS 23
Table 18: Voices of DFKI MaryTTS (aka MARY Text-to-Speech Synthesis) integrated for Release 2. 26
Table 19: ARIADNEplus NER services deployed on ELG 27
Table 20: IE and Text Analysis tools and services to integrate in the first release (full list) 31
Table 21: IE and Text Analysis tools and services to integrate in the second release (full list) 38
Table 22: IE and Text Analysis tools and services to integrate in the third release (full list) 41

List of Abbreviations

API	Application Programming Interface
ASR	Automatic Speech Recognition
GPU	Graphics Processing Unit
IE	Information Extraction
JSON	JavaScript Object Notation
LT	Language Technology
MT	Machine Translation
NER	Named Entity Recognition
NLP	Natural Language Processing
TF-IDF	Term Frequency and Inverse Document Frequency
TTS	Text-to-speech, i.e., speech synthesis
XGAPP	XML GATE Application
XML	Extensible Markup Language

Abstract

This document is an update to deliverable 4.1, describing the additional service, tools and components that have been integrated into the ELG platform since the first platform release. The original plan laid out in D4.1 for the timetable of service integration has remained largely stable, with a few services moving from one release to another but overall the project is on track to deliver all planned services by the time of the final release. In addition, we have begun work to integrate the first batch of services from third party providers, in particular the WP6 pilot projects.

1 Introduction

This document describes tools, services and components that have been developed and made available for the second public release (M27) of the ELG platform. It should be read in conjunction with the previous deliverable D4.1, which described the equivalent elements of the first release.

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. Section 2 of this document describes how that plan has evolved since release 1.

The remainder of the document details the tools that have been integrated by consortium members for release 2, 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).

The final section describes those tools and services contributed by third parties, primarily the first round of ELG pilot projects (WP6), that have so far been integrated into the platform, and summarises those expected to be delivered by the end of those pilots (Task 4.6).

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

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

	R1	R2	R3	Total
Services in D4.1	154	277	192	623
New services	6	35	5	46
Removed services	8	17	31	56
Services in D4.2	152	295	166	613
Delta of services	-2	+18	-26	-10

Table 1: Modifications to the integration plan of I.E tools defined in D4.1

When a service was moved from one release to another then it was counted as removed from the source release and as added to the target release. Most of the changes in the plan were due to these movements. In total 42 services were moved: 7 services from R1 to R2, 4 services from R2 to R1, 5 services from R2 to R3, and 26 services from R3 to R2. In addition, 4 new services, which were not part of the plan in D4.1, were integrated to the plan in D4.2: 2 services in R1 and 2 services in R2. Finally, 3 services were removed from the original plan: 1 service in R1, 1 service in R2, and 1 service in R3. In addition, 11 services were removed since they were duplicated in the plan defined in D4.1 Overall, after removal of these 11 duplicates the total number of services integrated across the three releases decreased by 10 from the original plan.

Machine Translation plan was less modified. As shown in Table 2, 8 services moved from R1 to R2 and 4 new MT services were added to the original plan: 2 services in R1, and 2 services in R2. The plan for the integration of ASR tools and the so called Other tools were not modified. The reasons behind the modification of the plans for MT and IE tools were varied. For example, some tools were removed as they were discontinued when producing this deliverable. Moving tools from one release to another was mainly due to the maturity level of the tools. For example some tools that were in preliminary beta versions were postponed while tools that were already fully developed and easy to integrate were moved to earlier releases.

	R1	R2	R3	Total
Services in D4.1 plan	21	21	9	51
New services	2	10	0	12
Removed services	8	0	0	8
Services in D4.2	15	31	9	55
Delta of services	-6	+10	0	+4

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

In the following we present the updated plan for the integration of IE tools and MT tools. In all these tables we have grouped the supported human languages into four categories: (A) official EU languages; (B) other EU languages without official status, plus languages from candidate countries and free trade partners; (C) languages spoken by immigrants or important trade and political partners; (D) languages that do not fit (A), (B), (C).

First in Table 3 we present the general overview of the services to be integrated in ELG broken down by service type and language category. In addition, Table 4 shows the overall number of distinct languages covered within each category of services across all partners.

	A	B	C	D	Total
ASR	**12**	**3**	**12**	**1**	**28**
HENS (formerly SAIL)	**9**	**3**	**12**	**1**	**25**
ASR	9	3	12	1	25
Tilde	**2**				**2**
ASR	2				2
UEDIN	**1**				**1**
ASR	1				1
IE & Text Analysis	**368**	**58**	**153**	**26**	**605**[1]
CUNI	**122**	**35**	**64**		**221**

[1] The difference of 8 between this number (605) and the total number of IE services given in Table 1 (613) is accounted for by 8 services that are language-independent and therefore do not fit into any of the A-D language categories

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	**48**	**6**	**13**	**13**	**80**
Date detection	2				2
Discourse Parsing	1				1
Language identification	22	6	13	13	54
Morphological analyser	6				6
Named Entity Recognition	3				3
Negation Detection	1				1
Parsing	1				1
Sentence splitting	3				3
Summarization	2				2
Text categorization	1				1
Textual Entailment	3				3
Tokenization	3				3
Expert System	**73**	**4**	**43**	**13**	**133**
Keyword extraction	7		5		12
Language identification	22	4	13	13	52
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
Summarization	7		5		12
Text categorization	2				2
ILSP	**5**				**5**
Information Extraction	2				2
Named Entity Recognition	2				2
Sentiment Analysis	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
USFD	**73**	**3**	**3**		**79**
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	1		17
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
Summarization	2				2
Text categorization	4				4
Text extraction	1				1

Tokenization	4				4
Other	**7**	**1**	**1**	**2**	**11**
DFKI	**5**	**1**	**1**	**2**	**9**
Text to Speech	5	1	1	2	9
Tilde	**2**				**2**
Text to Speech	2				2
MT	**42**		**3**	**1**	**46**
CUNI	**4**			**1**	**5**
A	4			1	5
DFKI	**1**				**1**
A	1				1
ILSP	**2**				**2**
A	2				2
Tilde	**19**		**2**		**21**
A	18		2		20
C	1				1
UEDIN	**16**		**1**		**17**
A	15		1		16
C	1				1
Total general	**429**	**62**	**169**	**30**	**690**

Table 3: Services and supported language per ELG partner.

	A	**B**	**C**	**D**
ASR				
ASR	12	3	12	1
IE & Text Analysis				
Date detection	2			
Dependency Parsing	24	7	13	
Discourse Parsing	1			
Entity linking	2			
Information Extraction	1			
Keyword extraction	10	3	11	
Language identification	22	6	14	13
Lemmatization	24	7	13	
Measurement annotation	1			
Measurement normalisation	1			
Morphological analyser	24	7	13	
Named Entity Recognition	16	5	11	
Negation Detection	1			
NER Disambiguation	4			
Noun phrase extraction	1			
Number annotation	1			
Number normalisation	1			
Opinion Mining	1			
Parsing	1			
Part-of-Speech Tagging	24	7	13	
Semantic annotation	7		5	
Sentence splitting	4			
Sentiment Analysis	9		4	
Summarization	7		5	
Text categorization	2			
Text extraction	1			
Textual Entailment	3			
Tokenization	24	7	13	

	A	B	C	D
MT (↓ From – To →)				
A	30		2	1
C	1			
Other				
Text to Speech	7	1	1	2

Table 4: Number of supported languages in each language category per service.

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 5 and Table 6 show the number of services integrated at the time of release 1. Table 5 gives the overall total and Table 6 breaks this total down by project partner.
- Table 7 and Table 8 show the number of services that will be integrated at the time of release 2 for each supported language. This time there is no breakdown by partner, instead Table 7 lists the category A languages (EU official) and Table 8 lists the category B languages plus an overall total for categories C & D (which for this release are all language identification).
- Table 9 lists the remaining services for category C and D languages, which are scheduled for integration by release 3.

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

	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 5: 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
NER	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
NER			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
NER			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
ILSP			**1**			**4**				**5**
Information Extraction						2				2
NER			1			1				2
Sentiment Analysis						1				1
HENS	**2**		**3**	**3**	**3**	**2**			**3**	**16**
Language identification	1		1	1	1	1			1	6
NER	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
NER			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 6: IE and Text Analysis services provided by ELG partners for the first release

	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	1			2				1	1	1			2		16
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
Lemmatization	1	1		1	2		1	1				1	1	2		1	1	1	2	1	1	1		1	20
Measurement annotation						1																			1
Measurement normalisation						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		2				1	2	2	1			1	15
Noun phrase extraction						1																			1
Number normalisation						1																			1
Part-of-Speech Tagging	2	2		2	4		2	2				1	1	2		1	1	2	3	2	2	2		2	33
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
Text extraction						1																			1
Tokenization	1	1		1	2		1	1				1	1	1		1	1	1	1	1	1	1		1	18
Total general	**10**	**9**	**1**	**8**	**23**	**11**	**8**	**8**	**4**	**8**	**2**	**9**	**5**	**18**	**1**	**7**	**5**	**12**	**18**	**12**	**10**	**8**	**4**	**10**	**211**

Table 7: Overview of IE and Text Analysis services to integrate 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
NER	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 8: Overview of IE and Text Analysis services to integrate in the second release (category B, C & D)

	Afrikaans	Arabic	Bengali	Chinese	English	German	Gujarati	Hebrew	Hindi/Urdu	Indonesian	Italian	Japanese	Kannada	Korean	Language ind.	Latin	Macedonian	Malay	Marahati	Nepali	Panjabi	Pashto	Persian	Russian	Somali	Swahili	Tagalog	Tamil	Telugu	Thai	Urdu	Vietnamese	Total
Dependency Parsing	1	1		1				1	1	1		1		1	1	1							1	1				1				1	14
Entity linking															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	2	1	1	1	1	2	2	1	1	1	1	1	1	1	1	35
Lemmatization	1	2		2				1	1	1		2		1		1							1	2				1				1	17
Morphological analyser	1	1		1				1	1	1		1		1		1							1	1				1				1	13
NER		2		2				1	1	1		1		1				1				1	1	3									15
Negation Detection						1																											1
Parsing						1																											1
Part-of-Speech Tagging	1	2		2				1	1	2		2		2	1	1							1	3				1				1	21
Semantic annotation		1		1								1		1										1									5
Sentence splitting															1																		1
Sentiment Analysis		1								1								1						1									4
Summarization		1		1								1		1	1									1									6
Text extraction															1																		1
Textual Entailment					1	1					1																						3
Tokenization	1	1		1				1	1	1		1		1	1	1							1	1				1				1	14
Total	**6**	**16**	**1**	**14**	**1**	**3**	**1**	**9**	**9**	**11**	**1**	**12**	**1**	**11**	**8**	**5**	**1**	**5**	**1**	**1**	**1**	**3**	**9**	**18**	**1**	**1**	**1**	**6**	**1**	**1**	**1**	**6**	**16**

Table 9: Overview of IE and Text Analysis services to integrate in the third release (category C and D)

2.2 Machine Translation Tools Updated Plan

The lists of MT tools and services to integrate in each stage after the partner modification are presented in Table 10, Table 11 and Table 12 (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
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
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
ILSP	Machine Translation	Greek	A	English	A
ILSP	Machine Translation	English	A	Greek	A

Table 10: MT tools and services in the first release

Release 2

Provider	Service	From	Category	To	Category
DFKI	TQ-AutoTest	German	A	English	A
DFKI	Quality Estimation	German	A	English	A
DFKI	Quality Estimation	English	A	German	A
DFKI	Quality Estimation	Spanish	A	English	A
DFKI	Quality Estimation	English	A	Spanish	A
DFKI	Quality Estimation	French	A	English	A
DFKI	Quality Estimation	English	A	French	A
DFKI	Machine Translation	German	A	English	A
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
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

Release 2					
Provider	**Service**	**From**	**Category**	**To**	**Category**
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
Tilde	Machine Translation	English	A	German	A
Tilde	Machine Translation	German	A	English	A

Table 11: MT tools and services to integrate in the second release

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

Table 12: MT tools and services to integrate in the third release

3 ASR Tools, Services and Components (Task 4.2)

Release 2 of the ELG platform includes additional ASR tools by HENSOLDT Analytics.

3.1 ASR Tools Integrated by HENS (formerly SAIL)

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[2]).

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.

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

For the first release ASR was integrated for 6 languages: Czech, English, French, German, Greek, and Spanish. This second release adds an additional 7 languages: Norwegian, Romanian, Albanian, Italian, Turkish, Polish, Dutch.

All models have been developed by SAIL LABS with the exception of Czech, which was developed by the University of Edinburgh. As the same underlying technology (KALDI[3]) as well as the same *recipe* for model-building are used, it was possible to integrate the ASR models for Czech in a seamless manner into the MMI.

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

Six of these components were deployed in the first release (French, English, German, Spanish, Greek, Latvian, Czech), seven more were added for release 2 (Norwegian, Romanian, Albanian, Italian, Turkish, Polish, Dutch).

Table 13 shows the ASR software components that are deployed in ELG platform. The code and Docker images are hosted in the ELG project repository and the corresponding container registry in gitlab[4], however all are private projects as they are commercial software.

Type	Tool Type	Image name	Service	Languages/variants supported	Licence
tool	ASR	sail-asr-it	Automatic Speech Recognition	Italian	SAIL LABS
tool	ASR	sail-asr-nl	Automatic Speech Recognition	Dutch	SAIL LABS
tool	ASR	sail-asr-no	Automatic Speech Recognition	Norwegian	SAIL LABS
tool	ASR	sail-asr-pl	Automatic Speech Recognition	Polish	SAIL LABS
tool	ASR	sail-asr-ro	Automatic Speech Recognition	Romanian	SAIL LABS
tool	ASR	sail-asr-sq	Automatic Speech Recognition	Albanian	SAIL LABS
tool	ASR	sail-asr-tr	Automatic Speech Recognition	Turkish	SAIL LABS[5]

Table 13: HENS ASR tools integrated in R2

[3] http://kaldi-asr.org
[4] 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
[5] Following the acquisition of SAIL LABS by HENSOLDT, the licenses will be modified but are expected to remain identical except for the naming.

3.2 ASR Tools Integrated by TILDE

Under the original integration plan based on prioritizing the core consortium languages for release 1, Tilde expected to deliver Latvian ASR for release 1 and Lithuanian & Estonian for release 2. However all three languages were in fact delivered in release 1, and have already been documented in deliverable 4.1.

4 IE Tools, Services and Components (Task 4.3)

Several project partners contributed additional IE & Text Analysis services to release 2, some additional language support for the pre-existing tools from release 1 and some brand new services for this release. Most of these services process text as their input, but a few – the HENS keyword spotting services – extract information from audio.

4.1 IE Tools Integrated by EXPSYS

For the second release, two new Cogito Discover services are added following essentially the same integration approach described in D4.1 with a slight change in the adapters. In addition, services integrated in release 1 and the new services integrated in release 2 are enriched with support to new languages in category A.

Cogito Discover

New Cogito Discover services that are included in the second release are:

- Keyword extraction: Annotation of the most relevant information: main syncons, main lemmas, main multiword expressions. It is related to the summarization service, since both services are offered in ELG through the service “Cogito Discover Summarizer”.
- Semantic annotation: This service returns the concepts spotted in a text which are modelled in the Cogito Discover knowledge graph. It is related with the lemmatization service, since both services are offered in ELG through the service “Cogito Discover Semantic Annotator”.

Integration approach update

The only change regarding the integration approach from D4.1, is that now there is a general adapter that supports all the services. The reason behind this change is to offer all the different services in a REST API, using only one port to access all the services.

Deployed components update

Table 14 shows an update of the software components described in deliverable D4.1. New services and 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)

Type	Image name	Service	Languages	Container registry	Code repository
Adapter	cogito-discover-general-adapter	Named Entity recognition	English, French, German, Spanish, **Italian, Portuguese, Dutch**	expertsystem/cogito-discover-general-adapter	expertsystem/tree/master/cogito-discover-general-adapter
Adapter	cogito-discover-general-adapter	Part-of-speech tagging	English, French, German, Spanish, **Italian, Portuguese, Dutch**	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**	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**	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**	expertsystem/cogito-discover-general-adapter	expertsystem/tree/master/cogito-discover-general-adapter
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**	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**	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	**Semantic annotation**	**English**, **French**, **German**, **Spanish**, **Italian, Portuguese, Dutch**	expertsystem/cogito-discover-general-adapter	expertsystem/tree/master/cogito-discover-general-adapter

Table 14: Cogito Discover software components for deployment in ELG

4.2 IE Tools Integrated by USFD

4.2.1 GATE Cloud

The University of Sheffield (USFD) provides both a variety of text analysis tools based on the open-source GATE text processing framework[6], alongside a public platform (GATE Cloud) through which some of these tools can be called as services on the web[7]. 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. In the second year these tools were used to extend the set of services made available on the ELG platform.

4.2.2 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[8], but requires GATE Cloud access credentials to operate.

Direct integration: rather than getting a proxy container to 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[9], 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.

[6] https://gate.ac.uk
[7] https://cloud.gate.ac.uk
[8] https://gitlab.com/european-language-grid/usfd/elg-gate-cloud-bridge
[9] https://gitlab.com/european-language-grid/usfd/gate-ie-worker

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.

4.2.3 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. Those exposed as part of the second ELG release are listed in Table 15. Apart from the CymrIE service which has been directly integrated and is running in the cluster, all services in this release were initially exposed via the bridge container.

Name	Description	
CymrIE	CYMRIE is a Welsh language version of GATE's prototypical information extraction pipeline, ANNIE. It is part of the Welsh Natural Language Toolkit, a Welsh Government funded project. CYMRIE is distributed with the GATE framework.	Welsh
NP Chunker	An implementation of the Ramshaw and Marcus BaseNP chunker, which marks noun phrases with a NounChunk annotation. This application also includes a tokeniser, sentence splitter and POS tagger (required by the chunking algorithm).	English
OpenNLP (Dutch)	The Dutch tokeniser, sentence splitter, POS tagger, phrase chunker and named-entity recogniser from Apache OpenNLP. The components are based on the maxent machine learning algorithm, and produce Token and Sentence annotations in a form compatible with other standard GATE tools.	Dutch
Romanian NER	A named entity recognition service for documents in the Romanian language. Based on ANNIE, it identifies names of persons, locations, organizations, as well as money amounts, time and date expressions.	Romanian
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: Bulgarian, Croatian, Danish, Dutch, Estonian, Finnish, Polish, Portuguese, Romanian, Slovak, Slovenian, Swedish, Basque and Catalan

Table 15: GATE Cloud services for general linguistic pre-processing and named entity recognition

In addition to these services developed by the GATE team at USFD, we have also integrated a set of six third-party GATE pipelines provided by the ARIADNEplus project[10] which are described in section 7.

4.3 IE Tools Integrated by DFKI

For Release 2, the DFKI team in Berlin (Speech and Language Technology) provides 75 tools in nine separate service categories Discourse Parsing, Summarization, Date Detection, Language Identification, Dependency Parsing, Morphological Annotation, Named Entity Recognition, Part-of-Speech Tagging and Tokenization.[11]

[10] https://ariadne-infrastructure.eu
[11] For changes to the original schedule of D4.1, please refer to Section 2 of this document.

The integration approach remains the same as documented by D4.1: These tools originate from past projects of a large group of developers and the ELG team of DFKI helps them understanding the REST API specifications and the provision of metadata records.

Table 16 describes the tools by DFKI scheduled for Release 2; this table does not include new language support added for R1 services. QURATOR-LangIdent and Lynx/QURATOR Summarization already appeared in the corresponding table of D4.1. Their metadata did not change. Only the latter now has German as an additional language to English.

Name	Description	Languages	Licence	Code repository
GSDP: an end-to-end Shallow Discourse Parser for German	Identifying shallow discourse relations (following the Penn Discourse TreeBank paradigm) in German text. This Shallow Discourse Parser for German is the (practical) outcome of a PhD Thesis on that very topic.	German	GNU Lesser General Public License (LPGL v3.0)	https://gitlab.com/qurator-platform/dfki/srv-germanshallowdiscourseparser
Lynx-TIMEX	Text analyser to annotate temporal expressions in English, German, Spanish and Dutch.	English, German, Spanish, Dutch	GPL-3.0	https://gitlab.com/qurator-platform/dfki/srv-timex-elg
MunderLine	MunderLine provides a pipeline that applies the following processing steps on an input text: tokenization, part-of-speech tagging, morphology tagging, named entity recognition, dependency parsing. Tokenization (and sentence border recognition) is done by a simple tokenizer. Part-of-speech tagging as well as named entity recognition is done by two instances of the GNTagger using statistical-based models. Dependency parsing is done by the MDParser, again using a statistical-based model.	English, German, Greek, Spanish	GNU Lesser General Public License (LPGL v3.0)	https://iread.dfki.de/#munderline---the-multilingual-universal-dependency-and-relation-pipeline

Table 16: DFKI services to be integrated for Release 2

4.4 IE Tools Integrated by HENS

HENSOLDT Analytics (formerly SAIL LABS) 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.

The component for the identification of language from text (*langClassifier*) is based on a component which forms part of a suite of tools for the processing of (textual) data from the Internet and Social Media.

Within the set of IE tools of ELG, HENS provides the MMI for the recognition of named-entities (NER) and for Sentiment Analysis (SA) and the langClassifier (LID) for the detection of language from text and Keyword Spotting/Extraction (KWS) for detecting keywords in speech.

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 (or *unknown*).

In the current 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 (or *unknown*).
- KWS is supported for 12 languages: Dutch, English, French, German, Greek, Italian, Polish, Romanian, Spanish, Albanian, Norwegian, Turkish

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". LID is based on a set of components which are employed within the HENSOLDT Media Mining System for the identification as well as the verification of languages of text-content. The underlying models combine TF-IDF-derived word-based features with character-based n-gram features.

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.

Deployed components

Table 17 shows the software components that are deployed in ELG platform in second release. The code and Docker images are hosted in the ELG project repository and the corresponding container registry in gitlab[12], however all are private projects as they are commercial software.

Type	Tool Type	Image name	Service	Languages	Licence
tool	NER	sail-ned-bg	Detection of a set of named-entities	Bulgarian	SAIL LABS
tool	NER	sail-ned-hr	Detection of a set of named-entities	Croatian	SAIL LABS
tool	NER	sail-ned-nl	Detection of a set of named-entities	Dutch	SAIL LABS
tool	NER	sail-ned-hu	Detection of a set of named-entities	Hungarian	SAIL LABS
tool	NER	sail-ned-it	Detection of a set of named-entities	Italian	SAIL LABS
tool	NER	sail-ned-pl	Detection of a set of named-entities	Polish	SAIL LABS

[12] 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	Licence
tool	NER	sail-ned-pt	Detection of a set of named-entities	Portuguese	SAIL LABS
tool	NER	sail-ned-ro	Detection of a set of named-entities	Romanian	SAIL LABS
tool	NER	sail-ned-sk	Detection of a set of named-entities	Slovak	SAIL LABS
tool	NER	sail-ned-sv	Detection of a set of named-entities	Swedish	SAIL LABS
tool	NER	sail-ned-sq	Detection of a set of named-entities	Albanian	SAIL LABS
tool	NER	sail-ned-ca	Detection of a set of named-entities	Catalan	SAIL LABS
tool	NER	sail-ned-no	Detection of a set of named-entities	Norwegian	SAIL LABS
tool	NER	sail-ned-tr	Detection of a set of named-entities	Turkish	SAIL LABS
tool	SA	sail-sed-it	Sentiment Detection	Italian	SAIL LABS
tool	SA	sail-sed-pl	Sentiment Detection	Polish	SAIL LABS
tool	SA	sail-sed-pt	Sentiment Detection	Portuguese	SAIL LABS
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	SAIL LABS
tool	KWS	sail-kws-nl	Keyword extraction	Dutch	SAIL LABS
tool	KWS	sail-kws-en	Keyword extraction	English	SAIL LABS
tool	KWS	sail-kws-fr	Keyword extraction	French	SAIL LABS
tool	KWS	sail-kws-de	Keyword extraction	German	SAIL LABS
tool	KWS	sail-kws-el	Keyword extraction	Greek	SAIL LABS
tool	MTKWS	sail-kws-it	Keyword extraction	Italian	SAIL LABS
tool	KWS	sail-kws-pl	Keyword extraction	Polish	SAIL LABS
tool	KWS	sail-kws-ro	Keyword extraction	Romanian	SAIL LABS
tool	KWS	sail-kws-es	Keyword extraction	Spanish	SAIL LABS
tool	KWS	sail-kws-sq	Keyword extraction	Albanian	SAIL LABS
tool	KWS	sail-kws-no	Keyword extraction	Norwegian	SAIL LABS
tool	KWS	sail-kws-tr	Keyword extraction	Turkish	SAIL LABS[13]

Table 17: IE tools provided by HENS

4.5 IE Tools Integrated by CUNI

For release 1 of the ELG platform, CUNI integrated their UDPipe system which includes segmentation, tokenization, POS tagging, morphological analysis, lemmatization and dependency parsing. The original plan envisaged the integration of just the seven priority languages of consortium member organisations in release 1, with the other category A and B languages to follow in release 2 and the category C and D ones to be added in release 3, but in practice it proved simpler to integrate *all* the UDPipe supported language models in one bundle for release 1.

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

5 MT Tools, Services and Components (Task 4.4)

The MT services integrated in release 2 are mostly additional language models for tools that were originally integrated for a limited number of language pairs in release 1. There is also a suite of new MT services provided by the OPUS-MT pilot project, described in section 7.1.

5.1 MT and related services integrated by DFKI

DFKI have integrated one MT service (for German to English) and a suite of translation quality estimation tools into release 2 – these were originally scheduled for release 1 and have been documented in deliverable D4.1.

5.2 MT Tools and Models integrated by UEDIN

For release 2 UEDIN has released an updated version of the Marian MT server tool, and deployed new models for a number of additional language pairs.

5.2.1 Dockerized Marian MT REST server

UEDIN is a major contributor to the open-source Marian Toolkit for Neural Machine Translation[14]. While most toolkits for neural MT are based on python, Marian is written in C++. It is one of the fastest neural MT toolkits on the market. Within the context of ELG, major development effort at UEDIN went into the design and implementation of a Marian-based REST translation service that can be deployed in a Docker container, full details of which are given in section 6.1.1 of the previous deliverable D4.1.

The dockerized Marian-based server has so far performed well and only few bug fixes were necessary. We recently overhauled the sentence splitting module after discovering an inefficiency in sentence splitting of very long text chunks. One area of concern is the cost of running MT servers in the platform infrastructure, because conventional MT models require GPUs for satisfactory translation speed. In collaboration with the Bergamot Project[15], which develops an MT component for the Firefox Web Browser, we are therefore working towards deploying models developed for translation with low computational requirements in the ELG infrastructure. These models are smaller models obtained by knowledge distillation from larger teacher models, which are then quantized to allow faster, integer-based matrix multiplication.[16] This will allow for translation with lower computational overhead, so that acceptable translation speeds can be realized without requiring a GPU.

5.2.2 MT Models

In addition to the models deployed in Release 1 of the platform (English ↔ German, Czech → English), UEDIN has contributed translation services for the following language pairs for Release 2:

- English ↔ Czech (models from the Bergamot Project)
- English ↔ Estonian (models from the Bergamot Project)
- English ↔ Latvian
- English ↔ Portuguese
- English ↔ Romanian
- English ↔ Spanish (models from the Bergamot project)

[14] https://marian-nmt.github.io
[15] https://browser.mt
[16] U. Germann et al., 2020. “Speed-optimized, Compact Student Models that Distill Knowledge from a Larger Teacher Model: the UEDIN-CUNI Submission to the WMT 2020 News Translation Task.” Fifth Conference on Machine Translation (WMT), pp 191-196.

5.3 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 release of the ELG platform the following translation directions have been added to the repertoire of MT services proxied through the ELG platform:

- Bulgarian → English (English → Bulgarian was already included in Release 1)
- English ↔ Danish
- English ↔ Estonian
- English ↔ Finnish
- English ↔ German
- English ↔ Lithuanian
- English ↔ Swedish

5.4 MT Services integrated by CUNI

CUNI provides a number of MT models at the LINDAT web site[17], and has packaged a number of them as Docker images deployed in the ELG platform. The integration is based on a Docker image containing TensorFlow model server[18] as a model backend combined with a lightweight Python front-end to adapt the TensorFlow server to the ELG API – full details are in Section 6.3 of Deliverable 4.1. In addition to the translation directions English ↔ Czech, English ↔ French and English → Hindi that were deployed in release 1, the following services have been added for release 2:

- English ↔ German
- English ↔ Polish
- English ↔ Russian

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

As with release 1 the only "other" service type that has been integrated so far is text-to-speech. It is likely that more service types will be added with or shortly after release 2 (at the end of M27, March 2021) as the first round of pilot projects draw to a close but these were not ready at the deadline for this deliverable (M26, February 2021).

6.1 TTS tools integrated by DFKI

DFKI has integrated **MaryTTS**[19] (or MARY Text-to-Speech System) for Release 2. This service covering 17 distinct voices, male and female, across six languages is the result of a collaborative project by DFKI's Language Technology Lab (Saarbrücken/Berlin) and the Institute of Phonetics at Saarland University.

[17] https://lindat.cz
[18] https://www.tensorflow.org/tfx/guide/serving
[19] http://mary.dfki.de

The DFKI team in Berlin together with the University of Sheffield developed a Spring Boot version of MaryTTS for Release 2, and as with most ELG tools the code and Docker images are hosted on GitLab[20].

Table 18 lists the voices provided to the ELG in Release 2. Note that MARY is a special case in terms of licences in that there are multiple and they can differ. This is due to the speech synthesis tool itself being licensed as LPGL v3.0 but the voices stemming from various sources being licensed separately.

Voice	Language	License
English (UK) female (dfki-poppy-hsmm)	English	CC BY-ND 3.0, LGPL v3.0
English (UK) male (dfki-prudence-hsmm)	English	CC BY-ND 3.0, LGPL v3.0
English (UK) male (dfki-obadiah-hsmm)	English	CC BY-ND 3.0, LGPL v3.0
English (UK) male (dfki-spike-hsmm)	English	CC BY-ND 3.0, LGPL v3.0
English (US) female (cmu-slt-hsmm)	English	ARCTIC, LGPL v3.0
English (US) male (cmu-bdl-hsmm)	English	ARCTIC, LGPL v3.0
English (US) male (cmu-rms-hsmm)	English	ARCTIC, LGPL v3.0
French female (enst-camille-hsmm)	French	CC BY-ND 3.0, LGPL v3.0
French female (upmc-jessica-hsmm)	French	CC BY-ND 3.0, LGPL v3.0
French male (enst-dennys-hsmm)	French	CC BY-ND 3.0, LGPL v3.0
French male (upmc-pierre-hsmm)	French	CC BY-ND 3.0, LGPL v3.0
German female (bits1-hsmm)	German	CC BY-ND 3.0, LGPL v3.0
German male (bits3-hsmm)	German	CC BY-ND 3.0, LGPL v3.0
German male (dfki-pavoque-neutral-hsmm)	German	CC BY-ND 3.0, LGPL v3.0
Italian female (istc-lucia-hsmm)	Italian	CC BY-ND 3.0, LGPL v3.0
Telugu female (cmu-nk-hsmm)	Telugu	CC BY-ND 3.0, LGPL v3.0
Turkish male (dfki-ot-hsmm)	Turkish	CC BY-ND 3.0, LGPL v3.0

Abbreviations: CC BY-ND 3.0 = Creative Commons Attribution No Derivatives 3.0 Unported ; LPGL v3.0 = GNU Lesser General Public License v3.0

Table 18: Voices of DFKI MaryTTS (aka MARY Text-to-Speech Synthesis) integrated for Release 2.

7 Status of tools and services from ELG Pilot Projects and other third parties (Task 4.6)

Task 4.6 in the ELG work plan concerns the support that the ELG project partners provide to third parties to integrate their tools and services into the ELG platform. The primary target for this task is the ELG-funded pilot projects managed by WP6 – at the time of writing OPUS-MT has completed the integration of their first ser-

[20] https://gitlab.com/PolinaGusenkova/elg-spring-boot-marytts/

vices, EVALITA4ELG has begun the process and expects to have at least some services active by the time of release 2 in M27, and discussions are ongoing with the remaining projects. In addition there have been a few other third party providers who have reached out to the ELG to ask for their services to be included.

7.1 OPUS-MT Pilot Project

The primary goal of the OPUS-MT ELG pilot project aims to develop MT models and services for a number of European minority languages. The OPUS MT system is based on the same open-source Marian MT system as is already used by UEDIN, but with some additional pre- and post-processing steps beyond the pure Marian workflow. These additional steps are encapsulated in a separate Python-based HTTP server component, which acts as an "adapter" delegating to the standard Marian MT server for the actual translation step. This system has already been published on GitHub[21].

For the ELG deployment, a modified version of the adapter has been created that exposes ELG-compliant endpoints, and this is built into a single Docker image along with the Marian server and the relevant set of models. On startup the adapter process launches the Marian server on a local loopback port and then calls it in response to ELG API requests received from the platform. ELG team members from USFD were closely involved in debugging the technical integration aspects, and this has led to the introduction of additional optional features in the Helm charts used to deploy all services to the ELG cluster, which will also benefit other service providers in future. At the time of writing ten services have so far been contributed to the ELG platform, centred on the Finnish language, translating Finnish to and from English, French, German, Swedish and Russian. Notably these are the first MT services integrated in ELG that do not use English as either the source or that target language.

7.2 Archaeology NER services from ARIADNEplus

As mentioned in section 4.2, USFD were approached in 2019 by researchers from the ARIADNEplus research project[22], who had developed some GATE-based tools for named entity recognition in the domains of archaeology and dendrochronology and were keen to make these available to a wider audience. These tools have been deployed on the GATE Cloud platform, and the developers also expressed an interest in having their services made available via the ELG.

Table 19 shows the six services that have been deployed to the ELG, with the assistance of the USFD team, covering two domains with three languages per domain.

Name	Description	Languages/variants support
Archeology Named Entity Recogniser	The entity recognizer identifies terms and phrases for analysing archaeological text. The pipeline delivers named entities of archaeological context, physical object, material, time appellation and structure.	Individual pipelines for: English, Dutch, Swedish
Dendrochronology Named Entity Recogniser	A named entity recognition service for archaeology documents. Identifies named entities of various types, terms and sentences relevant to dendrochronology	Individual pipielines for English, Dutch, Swedish

Table 19: ARIADNEplus NER services deployed on ELG

[21] https://github.com/Helsinki-NLP/Opus-MT
[22] https://ariadne-infrastructure.eu

8 Conclusion

The ELG project remains on track to deliver the services and tools that were envisaged in the original prioritised plan that was outlined in D4.1 with minimal modifications as described in section 2. The vast majority of services originally scheduled for release 2 have either been deployed recently and described in this document, or were delivered earlier than planned and included in D4.1. A small number of MT services have been added that were not in the original plan at all, and a few services have had to be shifted later in the release schedule, but all services that were included in the original plan are still on track to be delivered by the time of the final release. The project has also begun to include services and tools from outside the immediate project consortium, and we expect this process to accelerate as the first round of pilot projects reach their conclusions and the second round of projects begin to release their outputs in the run up to release 3.

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).

Table 20 lists the services from release 1 (D4.1 in M14), Table 21 for the current release 2 (now M26 following the project extension) and Table 22 for release 3 (D4.3 in M37).

		M14		
Provider	**Tool**	**Service**	**Language**	**Cat.**
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	Lemmatisation	Czech	A
CUNI	UDPipe tagger	Lemmatisation	English	A
CUNI	UDPipe tagger	Lemmatisation	French	A
CUNI	UDPipe tagger	Lemmatisation	German	A
CUNI	UDPipe tagger	Lemmatisation	Greek	A
CUNI	UDPipe tagger	Lemmatisation	Latvian	A
CUNI	UDPipe tagger	Lemmatisation	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

M14

Provider	Tool	Service	Language	Cat.
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
DFKI	geolocator	Categorization	English	A
DFKI	JTok	Sentence splitting	English	A
DFKI	JTok	Sentence splitting	German	A
DFKI	JTok	Sentence splitting	Italian	A
DFKI	JTok	Tokenization	English	A
DFKI	JTok	Tokenization	German	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 Sum	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
Expert System	Cogito Discover	Categorization	English	A
Expert System	Cogito Discover	Categorization	Spanish	A
Expert System	Cogito Discover	Language identification	Czech	A
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	Lemmatisation	English	A
Expert System	Cogito Discover	Lemmatisation	French	A
Expert System	Cogito Discover	Lemmatisation	German	A
Expert System	Cogito Discover	Lemmatisation	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	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
ILSP	ILSP-ABSA	Sentiment Analysis	Greek	A
ILSP	ILSP-Events-physical-attack	Information Extraction	Greek	A

M14

Provider	Tool	Service	Language	Cat.
ILSP	ILSP-Events-protest	Information Extraction	Greek	A
ILSP	ILSP-NER	Named Entity Recognition	English	A
ILSP	ILSP-NER	Named Entity Recognition	Greek	A
SAIL LABS	SAIL language ID	Language identification	Czech	A
SAIL LABS	SAIL language ID	Language identification	English	A
SAIL LABS	SAIL language ID	Language identification	French	A
SAIL LABS	SAIL language ID	Language identification	German	A
SAIL LABS	SAIL language ID	Language identification	Greek	A
SAIL LABS	SAIL language ID	Language identification	Spanish	A
SAIL LABS	SAIL NER	Named Entity Recognition	Czech	A
SAIL LABS	SAIL NER	Named Entity Recognition	English	A
SAIL LABS	SAIL NER	Named Entity Recognition	French	A
SAIL LABS	SAIL NER	Named Entity Recognition	German	A
SAIL LABS	SAIL NER	Named Entity Recognition	Greek	A
SAIL LABS	SAIL NER	Named Entity Recognition	Spanish	A
SAIL LABS	SAIL polarity analysis	Sentiment Analysis	English	A
SAIL LABS	SAIL polarity analysis	Sentiment Analysis	French	A
SAIL LABS	SAIL polarity analysis	Sentiment Analysis	German	A
SAIL LABS	SAIL 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: Language ID for Tweets	Language identification	English	A
USFD	GATE Cloud: Language ID for Tweets	Language identification	French	A
USFD	GATE Cloud: Language ID for Tweets	Language identification	German	A
USFD	GATE Cloud: Language 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

M14				
Provider	**Tool**	**Service**	**Language**	**Cat.**
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
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

Table 20: IE and Text Analysis tools and services to integrate in the first release (full list)

M26				
Provider	**Tool**	**Service**	**Language**	**Cat.**
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

M26

Provider	Tool	Service	Language	Cat.
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
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	Lemmatisation	Bulgarian	A
CUNI	UDPipe tagger	Lemmatisation	Croatian	A
CUNI	UDPipe tagger	Lemmatisation	Danish	A
CUNI	UDPipe tagger	Lemmatisation	Dutch	A
CUNI	UDPipe tagger	Lemmatisation	Estonian	A
CUNI	UDPipe tagger	Lemmatisation	Finnish	A
CUNI	UDPipe tagger	Lemmatisation	Hungarian	A
CUNI	UDPipe tagger	Lemmatisation	Irish	A
CUNI	UDPipe tagger	Lemmatisation	Italian	A
CUNI	UDPipe tagger	Lemmatisation	Lithuanian	A
CUNI	UDPipe tagger	Lemmatisation	Maltese	A
CUNI	UDPipe tagger	Lemmatisation	Polish	A
CUNI	UDPipe tagger	Lemmatisation	Portuguese	A
CUNI	UDPipe tagger	Lemmatisation	Romanian	A
CUNI	UDPipe tagger	Lemmatisation	Slovak	A
CUNI	UDPipe tagger	Lemmatisation	Slovenian	A
CUNI	UDPipe tagger	Lemmatisation	Swedish	A
CUNI	UDPipe tagger	Lemmatisation	Basque	B
CUNI	UDPipe tagger	Lemmatisation	Catalan	B
CUNI	UDPipe tagger	Lemmatisation	Galician	B
CUNI	UDPipe tagger	Lemmatisation	Norwegian	B
CUNI	UDPipe tagger	Lemmatisation	Serbian	B
CUNI	UDPipe tagger	Lemmatisation	Turkish	B
CUNI	UDPipe tagger	Lemmatisation	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

M26

Provider	Tool	Service	Language	Cat.
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
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
CUNI	UDPipe tokenizer	Tokenization	Hungarian	A

M26

Provider	Tool	Service	Language	Cat.
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
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	Lynx-TIMEX	Time annotation	English	A
DFKI	Lynx-TIMEX	Time annotation	German	A
DFKI	MunderLine	Dependency Parsing	English	A
DFKI	MunderLine	Part of Speech tagging	English	A
DFKI	MunderLine	Morphological annotation	English	A
DFKI	MunderLine	Tokenization	English	A
DFKI	MunderLine	Named entity recognition	English	A
DFKI	MunderLine	Dependency Parsing	German	A
DFKI	MunderLine	Part of Speech tagging	German	A
DFKI	MunderLine	Morphological annotation	German	A
DFKI	MunderLine	Tokenization	German	A
DFKI	MunderLine	Named entity recognition	German	A
DFKI	MunderLine	Dependency Parsing	Greek	A
DFKI	MunderLine	Part of Speech tagging	Greek	A
DFKI	MunderLine	Morphological annotation	Greek	A
DFKI	MunderLine	Tokenization	Greek	A
DFKI	MunderLine	Named entity recognition	Greek	A
DFKI	MunderLine	Dependency Parsing	Spanish	A
DFKI	MunderLine	Part of Speech tagging	Spanish	A
DFKI	MunderLine	Morphological annotation	Spanish	A
DFKI	MunderLine	Tokenization	Spanish	A
DFKI	MunderLine	Named entity recognition	Spanish	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

M26

Provider	Tool	Service	Language	Cat.
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
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
Expert System	Cogito Discover	Key phrase Extraction	Dutch	A

M26

Provider	Tool	Service	Language	Cat.
Expert System	Cogito Discover	Key phrase Extraction	English	A
Expert System	Cogito Discover	Key phrase Extraction	French	A
Expert System	Cogito Discover	Key phrase Extraction	German	A
Expert System	Cogito Discover	Key phrase Extraction	Italian	A
Expert System	Cogito Discover	Key phrase Extraction	Portuguese	A
Expert System	Cogito Discover	Key phrase 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
Expert System	Cogito Discover	Lemmatisation	Dutch	A
Expert System	Cogito Discover	Lemmatisation	Italian	A
Expert System	Cogito Discover	Lemmatisation	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	Summarization	Dutch	A
Expert System	Cogito Discover	Summarization	Italian	A
Expert System	Cogito Discover	Summarization	Portuguese	A
Expert System	Cogito Discover	Word sense disambiguation	Dutch	A
Expert System	Cogito Discover	Word sense disambiguation	English	A
Expert System	Cogito Discover	Word sense disambiguation	French	A
Expert System	Cogito Discover	Word sense disambiguation	German	A
Expert System	Cogito Discover	Word sense disambiguation	Italian	A
Expert System	Cogito Discover	Word sense disambiguation	Portuguese	A
Expert System	Cogito Discover	Word sense disambiguation	Spanish	A
SAIL LABS	SAIL KWS	Keyword extraction	Dutch	A
SAIL LABS	SAIL KWS	Keyword extraction	English	A
SAIL LABS	SAIL KWS	Keyword extraction	French	A
SAIL LABS	SAIL KWS	Keyword extraction	German	A

M26

Provider	Tool	Service	Language	Cat.
SAIL LABS	SAIL KWS	Keyword extraction	Greek	A
SAIL LABS	SAIL KWS	Keyword extraction	Italian	A
SAIL LABS	SAIL KWS	Keyword extraction	Polish	A
SAIL LABS	SAIL KWS	Keyword extraction	Romanian	A
SAIL LABS	SAIL KWS	Keyword extraction	Spanish	A
SAIL LABS	SAIL KWS	Keyword extraction	Albanian	B
SAIL LABS	SAIL KWS	Keyword extraction	Norwegian	B
SAIL LABS	SAIL KWS	Keyword extraction	Turkish	B
SAIL LABS	SAIL language ID	Language identification	Bulgarian	A
SAIL LABS	SAIL language ID	Language identification	Dutch	A
SAIL LABS	SAIL language ID	Language identification	Hungarian	A
SAIL LABS	SAIL language ID	Language identification	Italian	A
SAIL LABS	SAIL language ID	Language identification	Polish	A
SAIL LABS	SAIL language ID	Language identification	Portuguese	A
SAIL LABS	SAIL language ID	Language identification	Romanian	A
SAIL LABS	SAIL language ID	Language identification	Slovak	A
SAIL LABS	SAIL language ID	Language identification	Swedish	A
SAIL LABS	SAIL language ID	Language identification	Albanian	B
SAIL LABS	SAIL language ID	Language identification	Norwegian	B
SAIL LABS	SAIL language ID	Language identification	Turkish	B
SAIL LABS	SAIL NER	Named Entity Recognition	Bulgarian	A
SAIL LABS	SAIL NER	Named Entity Recognition	Croatian	A
SAIL LABS	SAIL NER	Named Entity Recognition	Dutch	A
SAIL LABS	SAIL NER	Named Entity Recognition	Hungarian	A
SAIL LABS	SAIL NER	Named Entity Recognition	Italian	A
SAIL LABS	SAIL NER	Named Entity Recognition	Polish	A
SAIL LABS	SAIL NER	Named Entity Recognition	Portuguese	A
SAIL LABS	SAIL NER	Named Entity Recognition	Romanian	A
SAIL LABS	SAIL NER	Named Entity Recognition	Slovak	A
SAIL LABS	SAIL NER	Named Entity Recognition	Swedish	A
SAIL LABS	SAIL NER	Named Entity Recognition	Albanian	B
SAIL LABS	SAIL NER	Named Entity Recognition	Catalan	B
SAIL LABS	SAIL NER	Named Entity Recognition	Norwegian	B
SAIL LABS	SAIL NER	Named Entity Recognition	Turkish	B
SAIL LABS	SAIL polarity analysis	Polarity detection	English	A
SAIL LABS	SAIL polarity analysis	Polarity detection	French	A
SAIL LABS	SAIL polarity analysis	Polarity detection	German	A
SAIL LABS	SAIL polarity analysis	Polarity detection	Italian	A
SAIL LABS	SAIL polarity analysis	Polarity detection	Polish	A
SAIL LABS	SAIL polarity analysis	Polarity detection	Portuguese	A
SAIL LABS	SAIL polarity analysis	Polarity detection	Spanish	A
SAIL LABS	SAIL polarity analysis	Sentiment Analysis	Italian	A
SAIL LABS	SAIL polarity analysis	Sentiment Analysis	Polish	A
SAIL LABS	SAIL 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: Language 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

M26				
Provider	**Tool**	**Service**	**Language**	**Cat.**
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
USFD	GATE Cloud: Romanian NER	Named Entity Recognition	Romanian	A
USFD	GATE Cloud: Welsh NER	Named Entity Recognition	Welsh	B
USFD	TermRaider	Text extraction	English	A
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 21: IE and Text Analysis tools and services to integrate in the second release (full list)

M37				
Provider	**Tool**	**Service**	**Language**	**Cat.**
CUNI	Entity Linker	Entity linking	Lang. independent	E
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	Lemmatisation	Afrikaans	C
CUNI	UDPipe tagger	Lemmatisation	Arabic	C
CUNI	UDPipe tagger	Lemmatisation	Chinese	C
CUNI	UDPipe tagger	Lemmatisation	Hebrew	C
CUNI	UDPipe tagger	Lemmatisation	Hindi/Urdu	C
CUNI	UDPipe tagger	Lemmatisation	Indonesian	C
CUNI	UDPipe tagger	Lemmatisation	Japanese	C
CUNI	UDPipe tagger	Lemmatisation	Latin	C
CUNI	UDPipe tagger	Lemmatisation	Persian	C

M37

Provider	Tool	Service	Language	Cat.
CUNI	UDPipe tagger	Lemmatisation	Russian	C
CUNI	UDPipe tagger	Lemmatisation	Tamil	C
CUNI	UDPipe tagger	Lemmatisation	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
DFKI	Dependency Tree Parser for German Clinical Text	Parsing	German	A
DFKI	Excitement Open Platform	Textual Entailment	English	A
DFKI	Excitement Open Platform	Textual Entailment	German	A
DFKI	Excitement Open Platform	Textual Entailment	Italian	A
DFKI	Negation Detection	Negation Detection	German	A
Expert System	Cogito Discover	Key phrase Extraction	Arabic	C

		M37		
Provider	**Tool**	**Service**	**Language**	**Cat.**
Expert System	Cogito Discover	Key phrase Extraction	Chinese	C
Expert System	Cogito Discover	Key phrase Extraction	Japanese	C
Expert System	Cogito Discover	Key phrase Extraction	Korean	C
Expert System	Cogito Discover	Key phrase 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	Malay	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	Lemmatisation	Arabic	C
Expert System	Cogito Discover	Lemmatisation	Chinese	C
Expert System	Cogito Discover	Lemmatisation	Japanese	C
Expert System	Cogito Discover	Lemmatisation	Korean	C
Expert System	Cogito Discover	Lemmatisation	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
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	Text extraction	Lang. independent	E
Expert System	Cogito Discover	Word sense disambiguation	Arabic	C
Expert System	Cogito Discover	Word sense disambiguation	Chinese	C

		M37		
Provider	**Tool**	**Service**	**Language**	**Cat.**
Expert System	Cogito Discover	Word sense disambiguation	Japanese	C
Expert System	Cogito Discover	Word sense disambiguation	Korean	C
Expert System	Cogito Discover	Word sense disambiguation	Russian	C
SAIL LABS	SAIL KWS	Keyword extraction	Arabic	C
SAIL LABS	SAIL KWS	Keyword extraction	Chinese	C
SAIL LABS	SAIL KWS	Keyword extraction	Hebrew	C
SAIL LABS	SAIL KWS	Keyword extraction	Hindi/Urdu	C
SAIL LABS	SAIL KWS	Keyword extraction	Indonesian	C
SAIL LABS	SAIL KWS	Keyword extraction	Malay	C
SAIL LABS	SAIL KWS	Keyword extraction	Pashto	C
SAIL LABS	SAIL KWS	Keyword extraction	Persian	C
SAIL LABS	SAIL KWS	Keyword extraction	Russian	C
SAIL LABS	SAIL language ID	Language identification	Arabic	C
SAIL LABS	SAIL language ID	Language identification	Hebrew	C
SAIL LABS	SAIL language ID	Language identification	Hindi/Urdu	C
SAIL LABS	SAIL language ID	Language identification	Indonesian	C
SAIL LABS	SAIL language ID	Language identification	Malay	C
SAIL LABS	SAIL language ID	Language identification	Pashto	C
SAIL LABS	SAIL language ID	Language identification	Persian	C
SAIL LABS	SAIL language ID	Language identification	Russian	C
SAIL LABS	SAIL language ID	Language identification	Lang. independent	E
SAIL LABS	SAIL NER	Named Entity Recognition	Arabic	C
SAIL LABS	SAIL NER	Named Entity Recognition	Chinese	C
SAIL LABS	SAIL NER	Named Entity Recognition	Hebrew	C
SAIL LABS	SAIL NER	Named Entity Recognition	Hindi/Urdu	C
SAIL LABS	SAIL NER	Named Entity Recognition	Indonesian	C
SAIL LABS	SAIL NER	Named Entity Recognition	Malay	C
SAIL LABS	SAIL NER	Named Entity Recognition	Pashto	C
SAIL LABS	SAIL NER	Named Entity Recognition	Persian	C
SAIL LABS	SAIL NER	Named Entity Recognition	Russian	C
SAIL LABS	SAIL polarity analysis	Polarity detection	Arabic	C
SAIL LABS	SAIL polarity analysis	Polarity detection	Indonesian	C
SAIL LABS	SAIL polarity analysis	Polarity detection	Malay	C
SAIL LABS	SAIL polarity analysis	Polarity detection	Russian	C
SAIL LABS	SAIL polarity analysis	Sentiment Analysis	Arabic	C
SAIL LABS	SAIL polarity analysis	Sentiment Analysis	Indonesian	C
SAIL LABS	SAIL polarity analysis	Sentiment Analysis	Malay	C
SAIL LABS	SAIL polarity analysis	Sentiment Analysis	Russian	C
SAIL LABS	SAIL summarization	Summarization	Lang. independent	E
USFD	GATE Cloud: Russian NER	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

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