
Book Section:
Roberts, I. orcid.org/0000-0002-7296-5851, Silva, A.G., Aroca, C.B. et al. (7 more authors) (2022) Language technology tools and services. In: Rehm, G., (ed.) European Language Grid: A Language Technology Platform for Multilingual Europe. Cognitive Technologies . Springer Cham , Cham , pp. 131-149. ISBN 978-3-031-17257-1

https://doi.org/10.1007/978-3-031-17258-8_7

Chapter 7
Language Technology Tools and Services

Ian Roberts, Andres Garcia Silva, Cristian Berrìo Aroca, Jose Manuel Gómez-Pérez, Miroslav Jánošík, Dimitris Galanis, Rémi Calizzano, Andis Lagzdiņš, Milan Straka, and Ulrich Germann

Abstract At the time of writing, the European Language Grid includes more than 800 LT services of varied types, including machine translation (MT), automatic speech recognition (ASR), text-to-speech synthesis (TTS), and text analysis ranging from simple tokenisers and part-of-speech taggers through to complete named entity recognition and sentiment analysis systems. This chapter gives a high-level summary of the development of the ELG service catalogue over time and digs deeper to discuss the process of service integration by looking at a few example services.

1 Introduction

The European Language Grid platform is able to support a wide variety of different types of Language Technology tools and services (see Chapter 3 for a more detailed description). Service types are classified based on the type of data they process as

Ian Roberts
University of Sheffield, UK, i.roberts@sheffield.ac.uk

Andres Garcia Silva · Cristian Berrìo Aroca · Jose Manuel Gómez-Pérez
Expert AI, Spain, agarcia@expert.ai, cberrio@expert.ai, jmgomez@expert.ai

Miroslav Jánošík
HENSOLDT Analytics GmbH, Austria, miroslav.janosik@hensoldt-analytics.com

Dimitris Galanis
Institute for Language and Speech Processing, R. C. “Athena”, Greece, galanisd@athenarc.gr

Rémi Calizzano
Deutsches Forschungszentrum für Künstliche Intelligenz GmbH, Germany, remi.calizzano@dfki.de

Andis Lagzdiņš
Tilde, Latvia, andis.lagzdins@tilde.lv

Milan Straka
Charles University, Czech Republic, straka@ufal.mff.cuni.cz

Ulrich Germann
University of Edinburgh, UK, ulrich.germann@ed.ac.uk

G. Rehm (ed.), *European Language Grid*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-17258-8_7

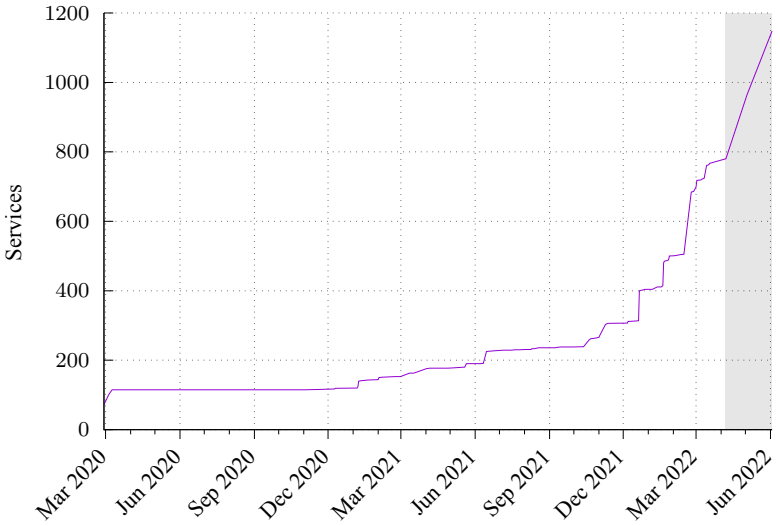

Fig. 1 Number of tools and services integrated into the European Language Grid over time; the grey shaded area denotes services whose integration is in progress at the time of writing and will be complete by the time of publication

input – text, audio, image data, etc. – and what they produce as *output* – annotations, text, audio, etc. This covers all the well-known service types such as Machine Translation (MT – text in, text out), Automatic Speech Recognition (ASR – audio in, text out), and Information Extraction/Text Analysis (IE – text in, annotations out), but also allows for services such as entity detection in *audio* data (audio in, annotations out), text-to-speech synthesis (TTS – text in, audio out), or optical character recognition (OCR – images in, text out).

Over the course of the original ELG EU project (Figure 1) the platform has grown from around 100 services available in the initial alpha release in 2020 to over 500 at the start of 2022 and almost 800 at the time of writing, with more being added all the time. The early stages of the project concentrated on services supplied by the ELG project consortium partners – such as ASR from HENSOLDT Analytics, MT from the University of Edinburgh and Tilde, TTS from Tilde, and a wide variety of Text Analysis services from Expert.AI, the University of Sheffield and DFKI (Roberts et al. 2020). More recently, an increasing number of services have been supplied by the ELG-funded pilot projects (see Part IV) and the platform has also begun to see contributions from third parties with no direct connection to the ELG consortium itself (Roberts et al. 2021, 2022). Of particular note is a set of over 500 MT services covering all pairs of EU official languages from the Neural Translation for the EU project, discussed in more detail in Section 2.[1] One third of these services have been integrated to date, with the remaining two thirds scheduled for integration during April and May 2022 (the grey shaded region in the graph), bringing the total number

[1] https://nteu.eu

		English	German	Italian	Spanish	French	Dutch	Swedish	Finnish	Polish	Czech	Greek	Portuguese	Danish	Bulgarian	Romanian	Estonian	Latvian	Slovenian	Croatian	Lithuanian	Slovak	Hungarian	Maltese	Irish	**Total A**	Total B (24 langs.)	Others (69 langs.)	**Total**
Text analysis	Linguistic pre-processing: Part-of-Speech Tagging	8	3	3	3	3	4	3	3	2	2	2	3	3	2	2	2	2	2	2	1	2	1	1	1	**60**	15	34	**109**
	Linguistic pre-processing: Morphology	5	2	2	2	2	2	3	3	1	1	1	1	3	1	1	5	1	1	1	1	1	1	1	1	**43**	13	27	**83**
	Linguistic pre-processing: Lemmatization	3	2	2	2	2	2	2	2	1	1	1	2	2	1	1	2	1	1	1	1	1	1	1	1	**36**	11	32	**79**
	Linguistic pre-processing: Tokenization	6	4	3	2	2	3	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	**39**	10	27	**76**
	Linguistic pre-processing: Sentence splitting	1	1				1																			**3**			**3**
	Linguistic pre-processing: Chunking	1										1														**2**			**2**
	Total pre-processing	**24**	**12**	**10**	**9**	**9**	**12**	**9**	**9**	**5**	**5**	**6**	**7**	**9**	**5**	**5**	**11**	**5**	**5**	**5**	**4**	**5**	**4**	**4**	**4**	**183**	**49**	**120**	**352**
	+ Classification	16	6	16	5	4	3	3	2	3	2	4	3	2	2	2	1	1	1	2	1	2	2			**83**	8	40	**131**
	+ Entity annotation	17	7	4	4	6	5	5	2	1	2	3	2	1	1	2				1		1	1			**65**	14	16	**95**
	+ Linking & disambiguation	7	2	3	4	4	1						1													**22**	2	5	**29**
	+ Sentiment/Opinion mining	13	3	2	2	2	1	1		1		1	2						1	1						**30**		10	**40**
	+ Text transformation	5	3	1	2	2	1						1													**15**	1	5	**21**
	+ Parsing	1	2	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	**26**	10	27	**63**
	+ Other text analysis	14	8	2	3	1	2	6	6			3	1	2												**48**	4	5	**57**
Total Text Analysis		**97**	**43**	**39**	**30**	**29**	**26**	**25**	**20**	**11**	**10**	**19**	**18**	**15**	**9**	**10**	**13**	**7**	**8**	**10**	**6**	**9**	**8**	**5**	**5**	**472**	**88**	**228**	**788**
+ Machine Translation into …		90	42	27	31	29	35	32	33	36	36	24	27	27	32	29	26	26	28	25	26	25	26	24	24	**760**	88	83	**931**
+ Speech recognition & analysis		2	2	2	3	2	2	1	2	2	1	2				2	1	2			1					**27**	9	21	**57**
+ Other services		11	7	5	4	8	4	2	1	4	2	3	2	4	4	2	3	4	2	2	4	2	2	2	1	**85**	10	77	**172**
Grand Total		**200**	**94**	**73**	**68**	**68**	**67**	**60**	**56**	**53**	**49**	**48**	**47**	**46**	**45**	**43**	**43**	**39**	**38**	**37**	**37**	**36**	**36**	**31**	**30**	**1344**	**195**	**409**	**1948**

Table 1 A snapshot of all services in the ELG platform, grouped by function and supported language. This includes all services integrated as at the end of March 2022, plus 368 additional MT services whose integration is ongoing. EU official languages (type A) are listed individually; type B represents other languages used in the EU, accession candidate countries, or EEA/EFTA members; "others" refers to languages from the rest of the world. For Machine Translation, the columns in this table represent the *target* language, see Table 2 for a breakdown by *source*.

of integrated service entries in ELG up to at least 1,148 by June 2022. We hope this trend will accelerate now that the third platform release is complete.

Furthermore, the figure of 1,148 hides the fact that a number of services combine several different functions (such as tokenisation, sentence splitting, part-of-speech tagging, entity detection, linking and disambiguation) into a single process and/or offer the same function in more than one language. Counting each language/function pair individually gives a more informative picture of the scope and coverage of ELG. For example, the platform currently provides one service that does dependency parsing for Portuguese; it also provides one service that does lemmatisation for Portuguese. The user who is looking for these two functions does not care whether they are implemented by one service or by two, only whether or not the European Language Grid can meet their needs.

By this measure, as of the end of March 2022, ELG offers 1,576 distinct service function/language combinations – already exceeding the 1,300 predicted by the project in mid-2021 (Rehm et al. 2021) – and is on track to offer at least 1,948 by June, which are summarised in Table 1. Reading from the bottom up, the 1,948 total breaks down into 931 MT (47.7% of the total), 788 text analysis (40%), 57 speech recognition and audio analysis, and 172 services of other types such as text to speech and OCR. The middle section of Table 1 breaks the 788 text analysis services down into broad sub-categories, and the top section breaks the largest sub-category (linguistic pre-processing) down into individual functions.

The largest *single* category of services is MT, with 770 catalogue entries representing 931 actual translation services (since some of the models are multilingual, with the same endpoint accepting input in several different languages and translating them all to the same target). The available text analysis services range from low-level text processing tasks such as tokenisation, part-of-speech tagging or morphological analysis, through named entity annotation and on to higher-level services such as parsing, sentiment analysis and entity linking against knowledge bases. Dependency parsing in particular is supported for 60 languages courtesy of the UDPipe parser from Charles University in Prague. For speech, the platform currently supports speech transcription for 31 languages thanks to tools from HENSOLDT Analytics and Tilde, alongside other speech processing tools such as the keyword spotting tool described in Section 3.

Breaking the numbers down on another dimension, the ELG platform now hosts at least one service providing support for each of 114 distinct languages. English is unsurprisingly the most highly represented, but there is good support for other major EU languages – German, French, Spanish, and Italian all have support for at least 20 service functions aside from machine translation – and in total 28 languages have support for at least ten functions.

Of course there is a long tail on both axes, with 16 of the 48 distinct service functions available in only one language each and 25 in fewer than five languages. On the other hand 39 out of the 114 languages are supported by only one function, and 51 by fewer than three. Full multilinguality is still in the future, but for the languages with larger numbers of speakers at least, significant progress has been and is being made.

Target → Source ↓	English	German	Czech	Polish	Dutch	Finnish	Swedish	Bulgarian	Spanish	Romanian	French	Slovenian	Italian	Danish	Portuguese	Latvian	Estonian	Lithuanian	Hungarian	Croatian	Slovak	Greek	Irish	Maltese	Total A	Total B (20)	Total Other (7)	Total
English		7	4	4	3	4	5	3	3	4	3	2	2	3	3	3	3	3	2	2	2	2	2	2	71	20	11	102
German	6		1	1	2	3	1	1	1	2	1	1	1	1	1	1	1	1	2	2	1	1	1	1	34	7	3	44
Czech	5	1		2	1	1	1	2	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	30	2	5	37
Polish	4	1	2		1	1	1	2	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	29	2	6	37
Dutch	2	2	1	1		1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	25	4	2	31
Finnish	4	3	1	1	1		3	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	31	1	1	33
Swedish	4	1	1	1	1	3		1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	29	5	1	35
Bulgarian	3	1	2	2	1	1	1		1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	27	2	3	32
Spanish	3	1	1	1	1	1	1	1		1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	25	3	1	29
Romanian	2	1	1	1	1	1	1	1	1		1	1	1	1	1	1	1	1	1	1	1	1	1	1	24		1	25
French	3	1	1	1	1	2	1	1	1	1		1	1	1	1	1	1	1	1	1	1	1	1	1	26	1	4	31
Slovenian	2	1	2	2	1	1	1	1	1	1	1		1	1	1	1	1	1	1	1	1	1	1	1	26		2	28
Italian	2	1	1	1	1	1	1	1	1	1	1	1		1	1	1	1	1	1	1	1	1	1	1	24		4	28
Danish	2	1	1	1	1	1	2	1	1	1	1	1	1		1	1	1	1	1	1	1	1	1	1	25	4	1	30
Portuguese	3	1	1	1	1	1	1	1	1	1	1	1	1	1		1	1	1	1	1	1	1	1	1	25		1	26
Latvian	3	1	1	1	1	1	1	1	1	1	1	1	1	1	1		1	1	1	1	1	1	1	1	25		2	27
Estonian	3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1		1	1	1	1	1	1	1	25		1	26
Lithuanian	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1		1	1	1	1	1	1	24		2	26
Hungarian	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1		1	1	1	1	1	24		1	25
Croatian	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1		1	1	1	1	24			24
Slovak	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1		1	1	1	24			24
Greek	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1		1	1	24			24
Irish	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1		1	24			24
Maltese	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1		24			24
Total A	65	32	29	29	26	31	30	27	25	27	26	26	24	26	25	25	25	25	25	25	24	24	24	24	669	51	52	772
Total B	16	4	2	3	4	1	1	2	3																36	20	12	68
Total Other	9	6	5	4	5	1	1	3	3	2	3	2	3	1	2	1	1	1	1		1				55	17	19	91
Grand Total	90	42	36	36	35	33	32	32	31	29	29	28	27	27	27	26	26	26	26	25	25	24	24	24	760	88	83	931

Table 2 A snapshot of supported MT language pairs as at the end of March 2022, with the addition of the remaining NTEU services for all pairs of EU official languages

2 Machine Translation

The ELG platform includes MT tools for 781 individual source/target language pairs, totalling 931 distinct services. Table 2 shows the breakdown; while English still dominates, it is much less ubiquitous than in the past, with only 21% of services involving English (102 from English, 90 into English, for a total of 192 out of the 931 available services). All pairs of EU official languages ("type A" in Table 2) are supported. In addition there is support for unofficial or regional European languages such as Basque, Galician and Luxembourgish and languages of accession candidates or free trade partners such as Icelandic, Norwegian[2] and Serbian[3] as well as languages important for trade and political reasons such as Modern Standard Arabic, Hindi, Ukrainian and Russian.

[2] Both Nynorsk and Bokmål varieties.

[3] Both Latin and Cyrillic script.

In addition to the MT services contributed by the ELG consortium partners Tilde (Pinnis and Bergmanis 2020) and University of Edinburgh (Junczys-Dowmunt et al. 2018; Germann et al. 2020; Germann 2020), two contributors in particular deserve a special mention here: the OPUS-MT ELG pilot project and the EU project Neural Translation for the European Union (NTEU).

The OPUS-MT ELG pilot project (Chapter 24, p. 325 ff., also see Tiedemann and Thottingal 2020) is responsible for 312 of the total 931 translation service options. To reduce the overall load on the ELG computing infrastructure, many of these language pairs are supported by multilingual models, where a single Docker container can accept input and/or produce output in many related languages. For example, there is a single OPUS model for "West Germanic", which can translate either way between any pair of English, German, Dutch, Luxembourgish, Afrikaans, Low Saxon, Gronings and Hunsrik. Some language pairs are supported by multiple models with different performance characteristics, for example, English to German is supported by a monolingual English-German model, a one-to-many "English to West Germanic", and the aforementioned many-to-many West Germanic model. Which model is most appropriate for a given task will vary, for example, if the input is known to be good-quality English then the monolingual model may be best, but if the input is a mix of languages, or English written by native speakers of other Germanic languages, then the multilingual model may be more accurate. Enabling users to test out different services on their own real data and switch between them with no technical changes to their code is one of the greatest benefits of the ELG approach.

NTEU is a project with a different focus, it was funded to produce high-quality translation tools for *all* possible pairs of EU official languages, to reduce the need for relay translation through a better-resourced language such as English (Bié et al. 2020; García-Martínez et al. 2021). This gives a total of 552 translation models (24 source languages each translating into the other 23 targets), so to spread the load of developing the models, NTEU involved three partner organisations, each responsible for models translating into eight target languages (one third of the total EU24). At the time of writing, one of the three sets of models has been published as ELG-integrated services and the other two sets are expected to be available by the time this book is published. The inclusion of these services marks an important milestone for ELG for two key reasons. First it shows the strong commitment of ELG to full multilinguality in the European Digital Single Market, and second it is the single largest contribution to the ELG platform originating outside the original ELG project consortium and pilot project ecosystem, demonstrating that ELG truly is a platform for the whole EU language technology community.

3 Automatic Speech Recognition

For automatic speech recognition, ELG currently hosts 48 services covering 30 languages and dialects. The majority of these have been provided by HENSOLDT Analytics, the speech recognition specialist in the ELG project consortium. In addition,

there have also been important contributions from Tilde for the Baltic languages, and from two of the pilot project organisations: Elhuyar for Basque (see Chapter 15, p. 271 ff.) and Lingsoft for Scandinavian languages (see Chapter 20, p. 301 ff.). Lingsoft have also begun to deliver *domain-specific* ASR services, for example a service tuned to recognise clinical speech in Finnish. As general purpose ASR systems increasingly become commodities, the creation and provision of domain-specific models provides an important niche for smaller ASR providers.

These organisations are all commercial service providers; though the tools themselves are based on open source frameworks such as Kaldi[4], the models are the proprietary intellectual property of the respective provider.

3.1 Case Study: Speech Tools from HENSOLDT

In addition to the actual ASR, the components provided by HENSOLDT also perform several preprocessing steps: audio is downsampled and converted to the native format of the respective models (typically 16kHz, 16 bit, mono, signed). Segmentation and classification of the input audio is carried out next. Any segment classified as containing an insufficient amount of speech is discarded and not processed by the ASR. Disfluencies and non-speech within segments identified as audio-segments are processed by the ASR system via specific non-speech models. Segmentation as well as classification are parameterised and can be adapted to specific audio conditions (the components provided within ELG use standard settings). Processing within the HENSOLDT ASR is staged in a pipelined manner for optimal throughput. Processing parameters can be employed to balance processing speed and accuracy. Like Lingsoft, HENSOLDT also provides *domain-specific* models which can be included in the respective Docker components. The ASR engine itself is *aware* of processing throughput as well as of the various models used. It can be adjusted to provide realtime processing as well as to reload different sub-models as soon as they become available. While the current services use one standard model, this allows for future updates of vocabularies and language models in a transparent manner. Output of the HENSOLDT ASR component can be provided in 1-best, n-best or lattice formats. The former is currently used in the deployed components, however, lattice-based output is used indirectly for use of the ASR component for keyword-spotting (KWS) applications only. A sample result of the detection of keywords via ASR can be seen in Figure 2.

[4] http://kaldi-asr.org

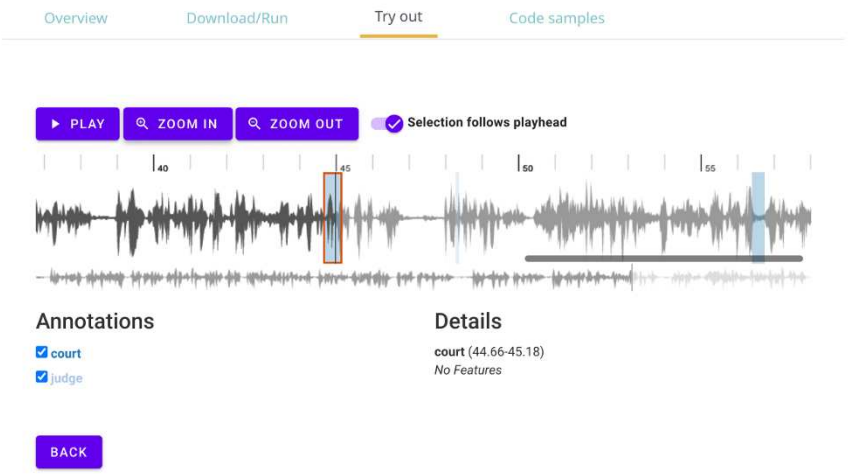

Fig. 2 Example of the word "court" having been detected as a keyword using HENSOLDT ASR

4 Text Analytics

After the set of MT services, the second largest group of services in the ELG platform are concerned in one way or another with the analysis and annotation of text, as discussed in Section 1. These cover a wide range from low-level text pre-processing tasks such as tokenisation and sentence splitting, through named entity annotation and linking tools (in many languages and domains), to dependency parsing, summarisation, sentiment analysis, and special purpose services such as the detection of misinformation or hate speech, and spelling and grammar checking.

Text analysis services have been provided by most members of the ELG project consortium, Expert.AI contributing their Cogito Discover toolkit, the University of Sheffield providing many services based on their GATE framework, Charles University providing their UDPipe dependency parser and other tools (e. g., Straka and Straková 2020; Straka et al. 2019b; Straka 2018; Straková et al. 2019; Straka et al. 2019a) and HENSOLDT (Dikici et al. 2019), ILSP (e. g., Prokopis and Piperidis 2020; Pontiki et al. 2018; Papanikolaou et al. 2016; Pontiki and Papageorgiou 2015) and DFKI (e. g., Schulz et al. 2022; Aksenov et al. 2021; Leitner et al. 2019) providing a variety of tools from their respective inventories. In addition, several of the pilot projects have contributed services in this class, notably

- *European Clinical Case Corpus* (Chapter 17, p. 283 ff.) – Fondazione Bruno Kessler. Clinical named entity recognisers in six languages.
- *Italian EVALITA Benchmark Linguistic Resources, NLP Services and Tools* (Chapter 19, p. 295 ff.) – University of Turin. A variety of services based on systems that participated in the various EVALITA shared tasks throughout the

years such as misogyny and hate speech detection and gender prediction, all in the Italian language.

- *Lingsoft Solutions as Distributable Containers* (Chapter 20, p. 301 ff.) – Lingsoft. General text analysis, proofing tools (spelling and grammar checking) and morphology analysis, in English and Scandinavian languages. This includes regional variations, such as distinct services for Swedish as used in Sweden and Swedish as used in Finland, and domain variations with specific services for medical domain text.
- *Universal Semantic Annotator* (Chapter 28, p. 349 ff.) – Sapienza University of Rome. This service performs word sense disambiguation, semantic role labelling and parsing for a wide variety of different languages.

4.1 Case Study: Cogito Discover from Expert.AI

Cogito Discover is Expert.AI's scalable software platform for automatic semantic metadata generation and auto-classification that can be easily integrated in the production environment of document-processing applications or workflows. It can be deployed on premise and in cloud environments and is available for both Linux and Windows systems. Cogito Discover services that are included in ELG are:

- Language detection: Identify the main language used in a text.
- Part-of-speech annotation: Annotations at different levels (token, word/compound word, group, clause, sentence) with grammatical types.
- Named Entity Recognition: Annotation of entities, i. e., people, organisations, places, known concepts, unknown concepts and also tags, i. e., URLs, email addresses, phone numbers, addresses, dates, time, measures, money, percentage, file folder.
- Semantic annotation: This service returns the concepts spotted in a text which are modelled in the Cogito Discover knowledge graph.
- Lemmatisation: This service returns the lemma of each concept spotted in the text that is modelled in the Cogito Discover knowledge graph.
- Keyword extraction: Annotation of the most relevant information, i. e., main syncons, main lemmas, main multiword expressions.
- Sentiment analysis: Provides a sentiment score (positive or negative) for the entities recognised in the text, and an overall score for the whole set of entities in the document.
- Summarisation: Annotation of the most relevant information, i. e., main syncons, main lemmas, main multiword expressions, main sentences and main domains.
- Categorisation: Classify documents using the IPTC taxonomy.

Most services are available in 12 languages: English, Italian, Spanish, German, French, Dutch, Portuguese, Chinese, Arabic, Russian, Japanese and Korean.

For its deployment in ELG, Expert.AI generated a Docker image containing a Cogito Discover installation, the linguistic packages, and a general adapter that manages the communication between the ELG platform and Cogito Discover. The general adapter was developed using the ELG Spring Boot Starter described in Chapter 4 (Part I, p. 67 ff.)[5], which makes it as easy as possible to create ELG-compliant tools in Java using Spring Boot.

4.2 Case Study: GATE from University of Sheffield

The University of Sheffield has been developing and maintaining the GATE framework for Natural Language Processing[6] for over 20 years. The basic framework is open source software written in Java and comes with a wide variety of plugins, some implementing specific NLP algorithms and some providing the generic base on which other specific rule-based and machine learning-based tools can be built.

The GATE ecosystem includes its own software-as-a-service platform called GATE Cloud (Tablan et al. 2013). An early focus of Sheffield's work in the ELG project was to develop a bridge to GATE Cloud, i. e., a proxy that accepts ELG API requests and dispatches them to a service endpoint on GATE Cloud, translating the resulting annotations into the ELG API response format. The development of this bridge has enabled the rapid deployment of many GATE Cloud hosted services into the ELG catalogue with little demand on the computing capacity of the ELG platform itself. At the time of writing, there are 66 GATE-based services integrated in ELG via the bridging proxy.

However, GATE Cloud itself has rate limits, so alongside the bridge component, Sheffield has developed a generic tool that can take any NLP application built against the GATE framework and bundle the application and all the plugins on which it depends as a Docker image that can run the application in-process within the ELG infrastructure. This mechanism has been used to wrap up certain particularly significant GATE-based applications so they can run directly in the ELG Kubernetes cluster and take advantage of the ELG platform's auto-scaling capabilities (see Chapter 5).

As the ELG EU project draws to a close, things have started to come full circle, as a number of recent additions *to* GATE Cloud have in fact been implemented as ELG-compatible Docker images, with a bridge in the other direction to enable a GATE application to call out to an endpoint that exposes the ELG internal LT service API. Some of these ELG-compatible images have been contributed back to ELG.

In addition, Sheffield has promoted the use of ELG-compatible services and Docker images in a number of other projects, notably the Horizon 2020 projects WeVerify[7] and RISIS2[8]. Many of Sheffield's contributions to these projects have

[5] https://gitlab.com/european-language-grid/platform/elg-spring-boot-starter

[6] General Architecture for Text Engineering, https://gate.ac.uk, see Cunningham et al. (2013).

[7] Wider and Enhanced Verification For You, https://weverify.eu, see Marinova et al. (2020).

[8] Research Infrastructure for Science and Innovation Policy Studies, https://www.risis2.eu, see Reale et al. (2019).

been implemented as ELG-compatible Docker images, with bridging components written for those projects to act as clients of the ELG API. The same mechanism has been used as part of a long-term collaboration between the University of Sheffield and King's College London, to integrate medical domain LT services developed in Python at King's into an existing GATE-based processing workflow. The use of the ELG standardised API makes it easy to integrate a variety of services implemented in different programming languages in a minimally-invasive way.

4.3 Case Study: Microservices At Your Service

With the third release in 2022, the ELG platform has begun to see contributions from third parties beyond the initial ELG consortium and pilot projects. One notable source is the project Microservices At Your Service[9], funded by the European Commission's Connecting Europe Facility (CEF) programme and led by Lingsoft (one of the organisations funded for a pilot project in the first ELG open call, see Chapter 20, p. 301 ff.). The project describes its mission as "bridging the gap between NLP research and industry" and it aims to identify open source text analysis tools that could benefit the community, package them as Docker images, and publish them for wider use. The project has selected the ELG platform as its primary vehicle for publication of the tools, and uses the ELG API as its standard specification for interoperability.

The project concentrates primarily on Finnish, Estonian, Icelandic, Spanish and Portuguese, plus some tools for minority languages from the same regions such as Faroese, Galician and Catalan. So far more than 14 services have been published, including:

- A proxy to the Finto-AI subject indexing service[10], in Finnish, Swedish and English (Suominen et al. 2022)
- Named entity recognition tools for Swedish and Norwegian, originally from the respective national libraries of the two countries (Kummervold et al. 2021)
- A tokeniser and morphological analysis tool for Estonian (Kaalep and Vaino 2001)
- A variety of tools for Icelandic from the University of Reykjavík, including a tokeniser, part-of-speech tagger, shallow parser and named entity recogniser, as well as machine translation models between Icelandic and English

One of the Icelandic services, a part-of-speech tagger and lemmatizer, is shown in Figure 3.

[9] https://www.lingsoft.fi/en/microservices-at-your-service-bridging-gap-between-nlp-research-and-industry

[10] https://ai.finto.fi

Fig. 3 Icelandic lemmatizer and part-of-speech tagger from Microservices At Your Service

5 Other Service Types

Right from the start of the ELG project, it was clear that the three principal service classes (ASR, MT, Text Analytics), while significant, would never be exhaustive. An important goal of ELG was to remain flexible enough to be able to easily integrate new classes of services and tools that had not been foreseen in the original proposal. The API specifications were designed with this flexibility in mind, being based solely on the kinds of data each service expects and returns, rather than placing any requirements on what the service *does* with that data.

Three classes of "other" services have emerged since the beginning of the project:

- *Text-to-speech* services that take text and synthesise audio.
- *Audio analysis* services that take audio input and return standoff annotations over time segments of the audio stream.
- *Image analysis* services, in particular optical character recognition (OCR).

Text-to-speech services have been provided by Tilde within the ELG project consortium (for Latvian and Lithuanian), and by the Elhuyar pilot project (for Basque). The audio analysis services are the keyword spotting tools from HENSOLDT Analytics described along with their speech recognition systems in Section 3.

The University of Sheffield has contributed a multilingual image OCR service developed as part of the Horizon 2020 EU project WeVerify. The service is based on a multi-step pipeline of neural models, first running a segmentation model to identify regions within the image that contain text, then a classifier to identify the writing system and language of each text block, and finally an appropriate text recognition model on each block depending on the identified script (Arabic, Bengali-Assamese, Chinese, Latin, Devanagari, Kanna, Hangul or Cyrillic). An example can be seen in Figure 4. The models have been deliberately designed *not* to use the "attention" mechanism typical of other deep neural models, as this was found to give only marginal improvements in performance at the cost of significantly increased memory and compute requirements.

Part of the reason for ELG funding the open call for pilot projects was precisely to elicit suggestions of new classes of services that were not previously known to the project consortium. Two pilots in particular delivered on this: Text2TCS (Section 5.1) and Coreon's MKS as LLOD (Section 5.2).

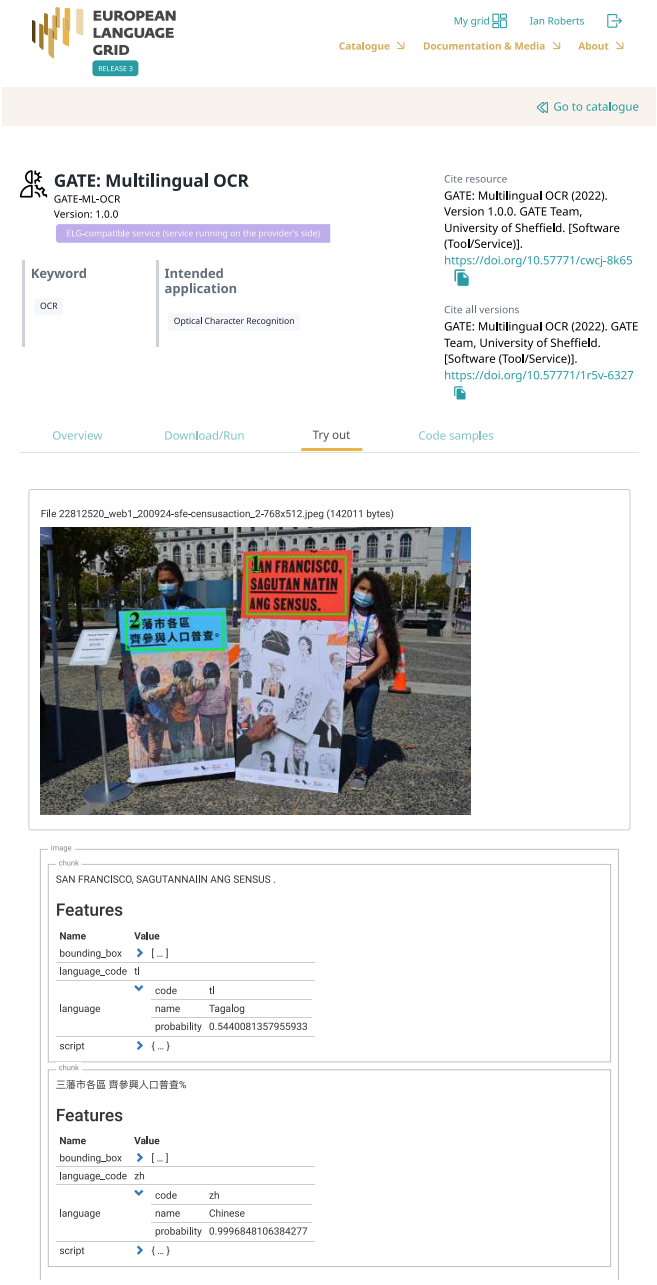

Fig. 4 The Multilingual OCR service showing detection of two blocks of text in different scripts (the bounding boxes *are* part of the "try out" UI, they have not been added to this figure)

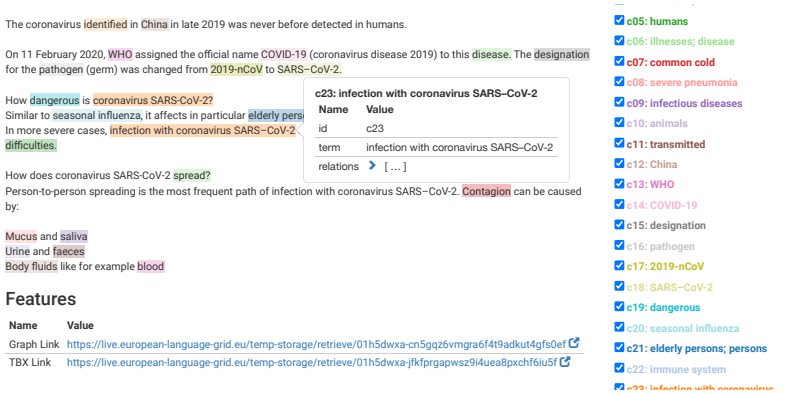

Fig. 5 Text2TCS service results in the "try out" GUI, showing links to the termbase and graph

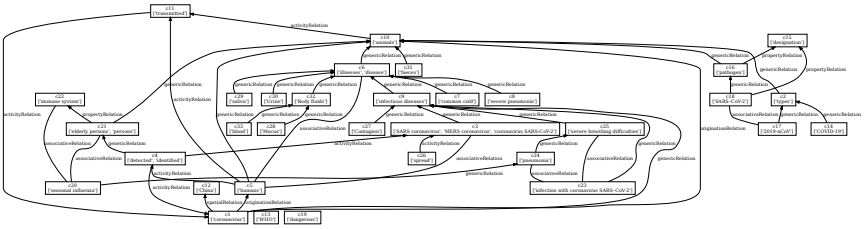

Fig. 6 The termbase graph generated from the sample input text (Figure 5)

5.1 Pilot Project: Terminological Concept Systems from Natural Language Text from University of Vienna

The Text2TCS project (see Chapter 18, Part IV, p. 289 ff.) aimed to develop a tool for deriving terminological concept systems from natural language text. This required the generation not only of typical standoff annotations representing the mentions of the detected terms in the source text, but also two additional output files for the termbase in TBX format[11] and a visualisation of the terminology as a PNG image.

These additional outputs did not naturally fit the JSON-based data interchange formats of the ELG API. It would have been possible to force them into this format by, for example, encoding the PNG data in base 64 encoding, but instead the ELG team took this as the impetus to introduce the "temporary storage" helper service for use by LT service containers. The operation of the temporary storage service is very simple. LT services can send arbitrary binary data to a well-known URL `http://storage.elg/store` (a private host name that resolves only within the ELG Kubernetes cluster), and will receive in return a publicly-resolvable URL which can be returned to the caller of the LT service for them to use to retrieve the same

[11] https://www.tbxinfo.net

data. Storage URLs include a cryptographically-secure random token to make them un-guessable, and they expire by default 15 minutes from their generation, at which time the stored data is permanently deleted.

Figures 5 and 6 show how this appears in the ELG portal when a user tests the Text2TCS service using the "try out" mechanism.

The temporary storage service provides an elegant solution to the problem of allowing LT services to return binary data without introducing additional complexity for the majority of services that do not have this requirement.

5.2 Pilot Project: MKS as Linguistic Linked Open Data from Coreon

The pilot project MKS as LLOD by knowledge management company Coreon (see Chapter 23, Part IV, 319 ff.) is an interesting case that in some ways sits at the boundary between services and resources. The aim of the project was to take Coreon's existing knowledge representation systems, known as MKS for Multilingual Knowledge System, and expose them as Linguistic Linked Open Data (LLOD). There is already a (de jure *and* de facto) standard API for querying linked (open) data resources, i. e., the SPARQL query language[12], so rather than defining a new format under the ELG umbrella, we decided to adopt the existing standard.

For ELG, the question was how best to represent this kind of resource in the ELG metadata scheme. On the one hand, the object that was being provided by Coreon was conceptually a data resource, albeit one accessed via a query API rather than via direct download, but on the other hand the technical method of integration would be through providing a SPARQL *service* for users to query. The eventual solution was in fact a mixture of both.

The Coreon SPARQL endpoint was integrated into the ELG infrastructure and set up so that SPARQL queries could be authenticated using access tokens issued by the ELG Keycloak identity provider, exactly as for other ELG LT services. In parallel, Coreon developed a "try out" UI to allow users to make test queries through the ELG catalogue interface. The two were then tied together as follows:

1. The "try out" UI was registered in its own right as a "service" in the ELG catalogue, whose function is "resource access".
2. Each SPARQL endpoint was then registered as an individual "ELG-compatible Lexical or Conceptual Resource" (LCR), with a link to the "try out" UI as "this resource is queried by that service".

Logic was introduced in the ELG catalogue to recognise when a user visits an ELG-compatible LCR that has an associated query service, and to inject the query UI as a "try out" tab which is configured with the necessary information and access token to be able to query the SPARQL endpoint (see Figure 7 for the final result).

[12] https://www.w3.org/TR/sparql11-overview/

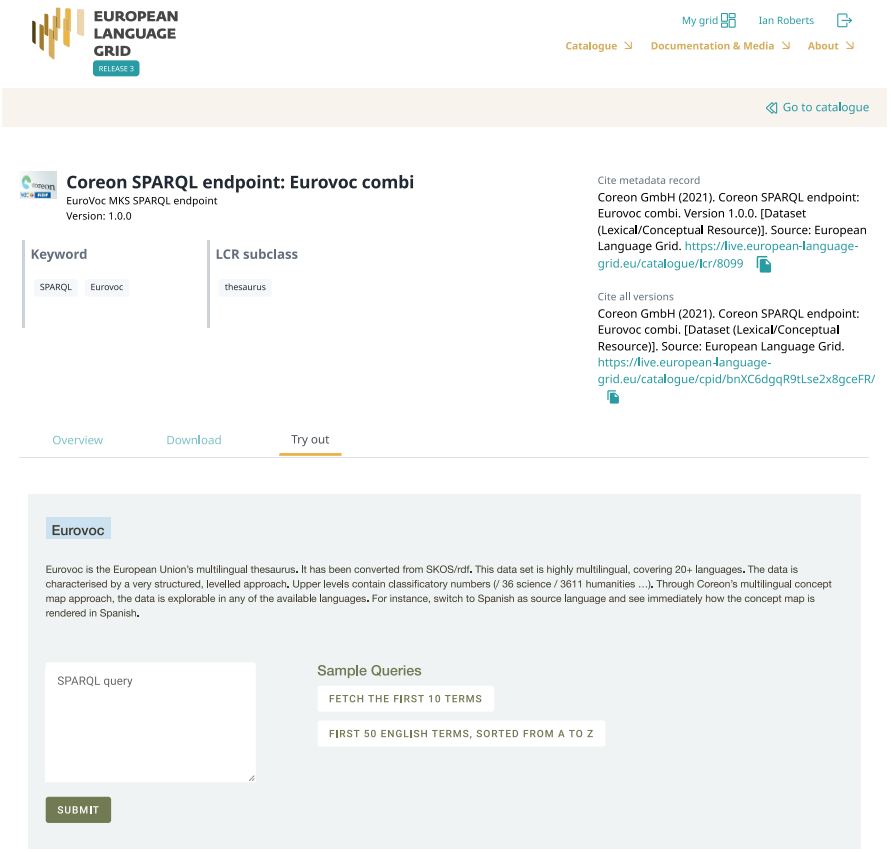

Fig. 7 Coreon SPARQL endpoint as an ELG-compatible Lexical/Conceptual Resource

6 Conclusions

Overall, the ELG project has succeeded in its aim to offer a broad variety of different service types covering many languages, and supplied by a range of different providers both academic and industrial. All the major classes of LT services are well represented in the ELG catalogue including ASR, MT and text analysis, with further classes of interest emerging during the course of the project. The generic design of the LT service execution APIs means that even services that do not exactly fit an existing class can be easily accommodated in the ELG platform, for example the HENSOLDT services for keyword spotting in audio required no API changes at all, only an adaptation of the "try out" GUI mechanism.

Inevitably, the majority of early contributions to the ELG platform were from the original ELG project consortium members. This was expected and planned for in

the original project proposal, and the pilot project funding system was designed to help broaden the contributor pool more quickly by incentivising providers to adopt the ELG formats and specifications. It has succeeded in this aim, and many more details can be found in the various pilot project chapters in Part IV. As the funded project draws to a close and the ELG platform begins to transition to its long term sustainable mode of operation, we are seeing an increasing number of third-party contributions from beyond the original consortium and pilot projects, which stands the ELG in good stead for its sustainability as a platform over the coming years.

References

Aksenov, Dmitrii, Peter Bourgonje, Karolina Zaczynska, Malte Ostendorff, Julián Moreno-Schneider, and Georg Rehm (2021). "Fine-grained Classification of Political Bias in German News: A Data Set and Initial Experiments". In: *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*. Ed. by Aida Mostafazadeh Davani, Douwe Kiela, Mathias Lambert, Bertie Vidgen, Vinodkumar Prabhakaran, and Zeerak Waseem. Bangkok, Thailand: ACL, pp. 121–131. URL: https://aclanthology.org/2021.woah-1.13.pdf.

Bié, Laurent, Aleix Cerdà-i-Cucó, Hans Degroote, Amando Estela, Mercedes García-Martínez, Manuel Herranz, Alejandro Kohan, Maite Melero, Tony O'Dowd, Sinéad O'Gorman, Mārcis Pinnis, Roberts Rozis, Riccardo Superbo, and Artūrs Vasiļevskis (2020). "Neural Translation for the European Union (NTEU) Project". In: *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*. Lisboa, Portugal: European Association for Machine Translation, pp. 477–478. URL: https://aclanthology.org/2020.eamt-1.60.

Cunningham, Hamish, Valentin Tablan, Angus Roberts, and Kalina Bontcheva (2013). "Getting More Out of Biomedical Documents with GATE's Full Lifecycle Open Source Text Analytics". In: *PLOS Computational Biology* 9.2, pp. 1–16. DOI: 10.1371/journal.pcbi.1002854.

Dikici, Erinç, Gerhard Backfried, and Jürgen Riedler (2019). "The SAIL LABS Media Mining Indexer and the CAVA Framework". In: *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association*. Ed. by Gernot Kubin and Zdravko Kacic. Graz, Austria: ISCA, pp. 4630–4631. URL: https://researchr.org/publication/DikiciBR19.

García-Martínez, Mercedes, Laurent Bié, Aleix Cerdà, Amando Estela, Manuel Herranz, Rihards Krišlauks, Maite Melero, Tony O'Dowd, Sinead O'Gorman, Marcis Pinnis, Artūrs Stafanovič, Riccardo Superbo, and Artūrs Vasiļevskis (2021). "Neural Translation for European Union (NTEU)". In: *Proceedings of Machine Translation Summit XVIII: Users and Providers Track*. Association for Machine Translation in the Americas, pp. 316–334. URL: https://aclanthology.org/2021.mtsummit-up.23.

Germann, Ulrich (2020). "The University of Edinburgh's submission to the German-to-English and English-to-German Tracks in the WMT 2020 News Translation and Zero-shot Translation Robustness Tasks". In: *Proceedings of the Fifth Conference on Machine Translation*. ACL, pp. 197–201. URL: https://aclanthology.org/2020.wmt-1.18.

Germann, Ulrich, Roman Grundkiewicz, Martin Popel, Radina Dobreva, Nikolay Bogoychev, and Kenneth Heafield (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". In: *Proceedings of the Fifth Conference on Machine Translation*. ACL, pp. 191–196. URL: https://aclanthology.org/2020.wmt-1.17.

Junczys-Dowmunt, Marcin, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch (2018). "Marian: Fast Neural Machine Translation

in C++". In: *Proceedings of ACL 2018, System Demonstrations*. Melbourne, Australia: ACL, pp. 116–121. URL: http://www.aclweb.org/anthology/P18-4020.

Kaalep, Heiki-Jaan and Tarmo Vaino (2001). "Complete Morphological Analysis in the Linguist's Toolbox". In: *Congressus Nonus Internationalis Fenno-Ugristarum Pars V*, pp. 9–16.

Kummervold, Per E, Javier De la Rosa, Freddy Wetjen, and Svein Arne Brygfjeld (2021). "Operationalizing a National Digital Library: The Case for a Norwegian Transformer Model". In: *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*. Reykjavik, Iceland: Linköping University Electronic Press, Sweden, pp. 20–29. URL: https://aclanthology.org/2021.nodalida-main.3.

Leitner, Elena, Georg Rehm, and Julián Moreno-Schneider (2019). "Fine-grained Named Entity Recognition in Legal Documents". In: *Semantic Systems. The Power of AI and Knowledge Graphs. Proceedings of the 15th International Conference (SEMANTiCS 2019)*. Ed. by Maribel Acosta, Philippe Cudré-Mauroux, Maria Maleshkova, Tassilo Pellegrini, Harald Sack, and York Sure-Vetter. Lecture Notes in Computer Science 11702. Karlsruhe, Germany: Springer, pp. 272–287. URL: https://link.springer.com/content/pdf/10.1007%2F978-3-030-33220-4_20.pdf.

Marinova, Zlatina, Jochen Spangenberg, Denis Teyssou, Symeon Papadopoulos, Nikos Sarris, Alexandre Alaphilippe, and Kalina Bontcheva (2020). "Weverify: Wider and Enhanced Verification for You Project Overview and Tools". In: *2020 IEEE International Conference on Multimedia Expo Workshops (ICMEW)*, pp. 1–4. DOI: 10.1109/ICMEW46912.2020.9106056.

Papanikolaou, Konstantina, Harris Papageorgiou, Nikos Papasarantopoulos, Theoni Stathopoulou, and George Papastefanatos (2016). ""Just the Facts" with PALOMAR: Detecting Protest Events in Media Outlets and Twitter". In: *Tenth International AAAI Conference on Web and Social Media*. Vol. 10. 2, pp. 135–142.

Pinnis, Mārcis and Toms Bergmanis (2020). "Tilde's Neural Machine Translation Technology". In: *Latvian Academy of Sciences Yearbook 2020*. Latvian Academy of Sciences, pp. 85–89.

Pontiki, Maria and Harris Papageorgiou (2015). "Opinion Mining and Target Extraction in Greek Review Texts". In: *Proceedings of the 12th International Conference on Greek Linguistics (ICGL 12)*. Vol. 2. Freie Universität. Berlin, Germany, pp. 871–883.

Pontiki, Maria, Konstantina Papanikolaou, and Haris Papageorgiou (2018). "Exploring the Predominant Targets of Xenophobia-motivated Behavior: A Longitudinal Study for Greece". In: *Proceedings of the Natural Language Processing meets Journalism Workshop (NLPJ 2018)*. Ed. by Octavian Popescu and Carlo Strapparava. ELRA.

Prokopis, Prokopidis and Stelios Piperidis (2020). "A Neural NLP toolkit for Greek". In: *11th Hellenic Conference on Artificial Intelligence*, pp. 125–128. URL: http://nlp.ilsp.gr/setn-2020/3411408.3411430.pdf.

Reale, Emanuela, Grazia Battiato, and Serena Fabrizio (2019). "RISIS2: an innovative research infrastructure as a support for STI research community". In: *ISSI*, pp. 2658–2659. DOI: 10.5281/zenodo.3478408.

Rehm, Georg, Stelios Piperidis, Kalina Bontcheva, Jan Hajic, Victoria Arranz, Andrejs Vasiļjevs, Gerhard Backfried, José Manuel Gómez Pérez, Ulrich Germann, Rémi Calizzano, Nils Feldhus, Stefanie Hegele, Florian Kintzel, Katrin Marheinecke, Julian Moreno-Schneider, Dimitris Galanis, Penny Labropoulou, Miltos Deligiannis, Katerina Gkirtzou, Athanasia Kolovou, Dimitris Gkoumas, Leon Voukoutis, Ian Roberts, Jana Hamrlová, Dusan Varis, Lukáš Kačena, Khalid Choukri, Valérie Mapelli, Mickaël Rigault, Jūlija Meļņika, Miro Janosik, Katja Prinz, Andres Garcia-Silva, Cristian Berrio, Ondrej Klejch, and Steve Renals (2021). "European Language Grid: A Joint Platform for the European Language Technology Community". In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations (EACL 2021)*. Kyiv, Ukraine: ACL, pp. 221–230. URL: https://www.aclweb.org/anthology/2021.eacl-demos.26.pdf.

Roberts, Ian, Andres Garcia Silva, Miroslav Janosik, Nils Feldhus, Dimitris Galanis, Andis Lagzdiņš, and Rémi Calizzano (2022). *Deliverable D4.3 Services, Tools and Components (Final Release)*. Project deliverable; EU project European Language Grid (ELG); Grant Agreement no. 825627 ELG. URL: https://www.european-language-grid.eu/wp-content/uploads/2022/04/ELG-Deliverable-D4.3-final.pdf.

Roberts, Ian, Andres Garcia Silva, Miroslav Janosik, Andis Lagzdiņš, Nils Feldhus, Georg Rehm, Dimitris Galanis, Dusan Varis, and Ulrich Germann (2020). *Deliverable D4.1 Services, Tools and Components (First Release)*. Project deliverable; EU project European Language Grid (ELG); Grant Agreement no. 825627 ELG. URL: https://www.european-language-grid.eu/wp-content/uploads/2021/02/ELG-Deliverable-D4.1-final.pdf.

Roberts, Ian, Andres Garcia Silva, Miroslav Janosik, Andis Lagzdiņš, Nils Feldhus, Georg Rehm, Dimitris Galanis, Dusan Varis, and Ulrich Germann (2021). *Deliverable D4.2 Grid Content: Services, Tools and Components (Interim Release)*. Project deliverable; EU project European Language Grid (ELG); Grant Agreement no. 825627 ELG. URL: https://www.european-language-grid.eu/wp-content/uploads/2022/04/ELG-Deliverable-D4.2-final.pdf.

Schulz, Konstantin, Jens Rauenbusch, Jan Fillies, Lisa Rutenburg, Dimitrios Karvelas, and Georg Rehm (2022). "User Experience Design for Automatic Credibility Assessment of News Content About COVID-19". In: *Proceedings of HCI International 2022 – Late Breaking Papers*. Accepted for publication. 26 June-01 July 2022.

Straka, Milan (2018). "UDPipe 2.0 Prototype at CoNLL 2018 UD Shared Task". In: *Proceedings of CoNLL 2018: The SIGNLL Conference on Computational Natural Language Learning*. Stroudsburg, PA, USA: ACL, pp. 197–207.

Straka, Milan and Jana Straková (2020). "UDPipe at EvaLatin 2020: Contextualized Embeddings and Treebank Embeddings". In: *Proceedings of LT4HALA 2020 – 1st Workshop on Language Technologies for Historical and Ancient Languages*. Marseille, France: ELRA, pp. 124–129.

Straka, Milan, Jana Straková, and Jan Hajič (2019a). "Czech Text Processing with Contextual Embeddings: POS Tagging, Lemmatization, Parsing and NER". In: *Proceedings of the 22nd International Conference on Text, Speech and Dialogue (TSD 2019)*. Cham, Heidelberg, New York etc.: Springer, pp. 137–150.

Straka, Milan, Jana Straková, and Jan Hajič (2019b). "UDPipe at SIGMORPHON 2019: Contextualized Embeddings, Regularization with Morphological Categories, Corpora Merging". In: *Proceedings of the 16th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*. Stroudsburg, PA, USA: ACL, pp. 95–103.

Straková, Jana, Milan Straka, and Jan Hajič (2019). "Neural Architectures for Nested NER through Linearization". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Stroudsburg, PA, USA: ACL, pp. 5326–5331.

Suominen, Osma, Mona Lehtinen, and Juho Inkinen (2022). *Annif and Finto AI: Developing and Implementing Automated Subject Indexing*. Macerata. DOI: 10.4403/jlis.it-12740.

Tablan, Valentin, Ian Roberts, Hamish Cunningham, and Kalina Bontcheva (2013). "GATECloud.net: A Platform for large-scale, Open-Source Text Processing on the Cloud". In: *Philosophical Transactions of the Royal Society A: Math., Phys. and Eng. Sciences* 371.20120071.

Tiedemann, Jörg and Santhosh Thottingal (2020). "OPUS-MT – Building open translation services for the World". In: *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation (EAMT)*. Lisboa, Portugal: European Association for Machine Translation, pp. 479–480. URL: https://helda.helsinki.fi/bitstream/handle/10138/327852/2020.eamt_1_499.pdf.