/instituut voor de
Nederlandse taal/

EUROPEAN
LANGUAGE
GRID

European Language Grid Workshop Nederland

Vincent Vandeghinste
NCC lead Nederland

14.00 uur	Welkom	Vincent Vandeghinste (INT)
14.05 uur	Introducing the European Language Grid	Georg Rehm (DFKI)
14.35 uur	Demonstratie van de ELG, met focus op het Nederlands	Vincent Vandeghinste (INT)
15.00 uur	Vraag- en antwoordsessie en discussie	
15.10 uur	Pauze	
15.25 uur	Taalmaterialen voor het Nederlands	Bob Boelhouwer (INT)
15.35 uur	Presentaties van verschillende bedrijven (3) ● EDIA (Mark Breuker) ● Telecats (Arjan van Hessen) ● Y.digital (Ian Fitzpatrick)	Oele Koornwinder (NOTaS)
16.25 uur	Vraag- en antwoordsessie en discussie	

EUROPEAN
LANGUAGE
GRID

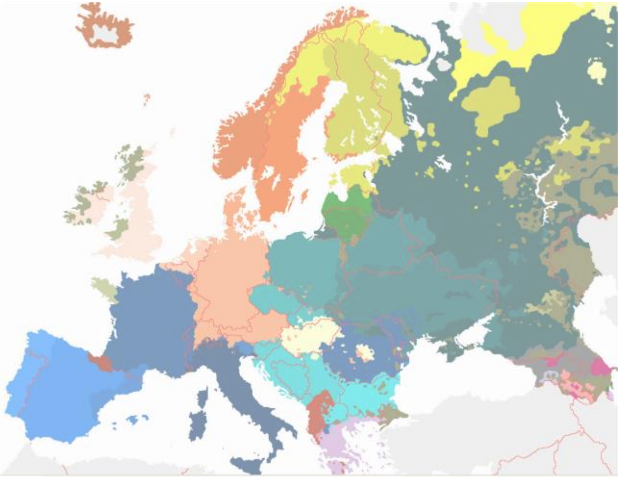
European Language Grid: Broad Overview

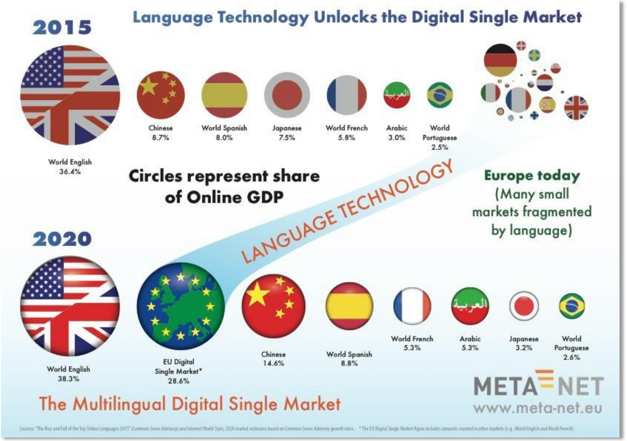
Prof. Dr. Georg Rehm (DFKI GmbH, Germany) Coordinator
ELG, Co-coordinator ELE

03-12-2021 ELG NCC Workshop Netherlands
https://www.european-language-grid.eu
https://www.european-language-equality.eu

Multilingualism in Europe

- Multilingualism is at the heart of the European idea
- 24 EU languages – they all have the same status
- Dozens of regional and minority languages as well as languages of immigrants and trade partners
- Many economic, social and technical challenges:
  - The Digital Single Market needs to be multilingual
  - Cross-border, cross-lingual, cross-cultural communication
  - Fragmentation of the LT market and landscape (including availability of LTs and LRs)





Sources: "The Rise and Fall of the Top Online Languages 2015" (Common Sense Advisory) and Internet World Stats; 2020 market estimates based on Common Sense Advisory growth rates. * The EU Digital Single Market figure includes amounts counted in other markets (e.g., World English and World French).

META FORUM 2010

T4ME

META NET

T4ME
EU-funded project (Seventh Framework Programme) working on technologies for the Mul`lingual European Informa`on Society

META-NET
Established in 2010, META-NET is a network of Excellence consis`ng of 60 research centres from 34 countries, building the technological founda` ons of a mul`lingual European informa`on society

META-FORUM 2010
"Challenges for Mul`lingual Europe"
(November 17/18, 2010)

2010

2011

2012

META NORD

METANET4U

CESAR

META-NORD
EU-funded (ICT PSP)
Project to establish an open linguistic infrastructure in the Baltic and Nordic countries (2011 – 2013)

METANET4U
EU-funded project (ICT PSP) to enhance the European Linguistic Infrastructure (2011 – 2013)

CESAR
EU-funded project (ICT PSP) functioning as a part of META-NET to standardise language resources and tools (2011 – 2013)

META-FORUM 2011
Conference "Solutions for Multilingual Europe"
(Budapest – June 27/28, 2011)

META FORUM 2011

META SHARE

META FORUM 2012

META-FORUM 2012
Conference "A Strategy for Multilingual Europe"
(Brussels – June 20/21, 2012)

META-NET White Papers
Release of 32 volumes on 31 languages, revealing that there is a severe threat of digital extinction for at least 21 European languages
(December 2012)

2013

2014

"State of the Art of Machine Translation – Current Challenges and Future Opportunities"
Workshop on "State of the Art of Machine Translation", commissioned by the EU Parliament (December 2013)

META-SHARE
Initiated in 2013, META-SHARE has functioned as an open and secure network of repositories for sharing and exchanging language data, tools and services

Strategic Research Agenda for Multilingual Europe 2020
Launch of the Strategic Research Agenda for Multilingual Europe 2020 (January 2013)

META-FORUM 2013
Conference "Connecting Europe for New Horizons"
(Berlin –September 19/20, 2013)

META FORUM 2013

META FORUM 2015

MULTILINGUAL DIGITAL SINGLE MARKET
Riga Summit
2015

CRACKER
EU-funded project CRACKER (Horizon2020) pushing towards an improvement of MT research in terms of efficiency and effectiveness (2015 – 2017)

Cracking the Language Barrier Federation
Founded in 2015, the federation has been assembling European research and innovation projects as well as all related community organisations working on multilingual technologies

META-FORUM 2015
Conference "Technologies for the Multilingual Digital Single Market" (Riga – April 27, 2015)

Strategic Agenda for the Multilingual Digital Single Market (Version 0.5)
Launch of the Strategic Agenda for the Multilingual Digital Single Market titled "Technologies for Overcoming Language Barriers towards a truly integrated European Online Market" (April 2015)

Riga Summit on the Multilingual Digital Single Market
Summit "Shape the future of the multilingual digital single market" (April 27–29,2015)

2015

2016

Strategic Research and Innovation Agenda (Version 0.9)
Launch of the Strategic Research and Innovation Agenda titled "Language as a Data Type and Key Challenge for Big Data" (July 2016)

META-FORUM 2016
Conference "Beyond Multilingual Europe"
(Lisbon – July 4/5, 2016)

CRACKER
Cracking the Language Barrier

Cracking the Language Barrier

META FORUM 2016

META FORUM 2017

META FORUM 2019

AI4EU

"Language Equality in the Digital Age"
Workshop on "Language Equality in the Digital Age", commissioned by the EU Parliament's Science and Technology Options Assessment Committee (STOA)
(January 2017)

"Language equality in the digital age: Towards a Human Language Project"
Launch of the study on "Language equality in the digital age: Towards a Human Language Project", commissioned by the EU Parliament (March 2017)

META-FORUM 2017
Conference "Towards a Human Language Project"
(Brussels – November 13/14, 2017)

Strategic Research and Innovation Agenda (V1.0)
Launch of the Strategic Research and Innovation Agenda titled "Language Technologies for Multilingual Europe – Towards a Human Language Project" (December 2017)

2017

2018

ELG Final Proposal Submission
Final submission on Feb. 20, 2018

EUROPEAN LANGUAGE EQUALITY
2021-2022

EUROPEAN LANGUAGE GRID
2019-2022

ELG

DFKI IEA ILSP The University Of Sheffield. eLDa TILDE HENSOLDT EXPERT SYSTEM THE UNIVERSITY of EDINBURGH

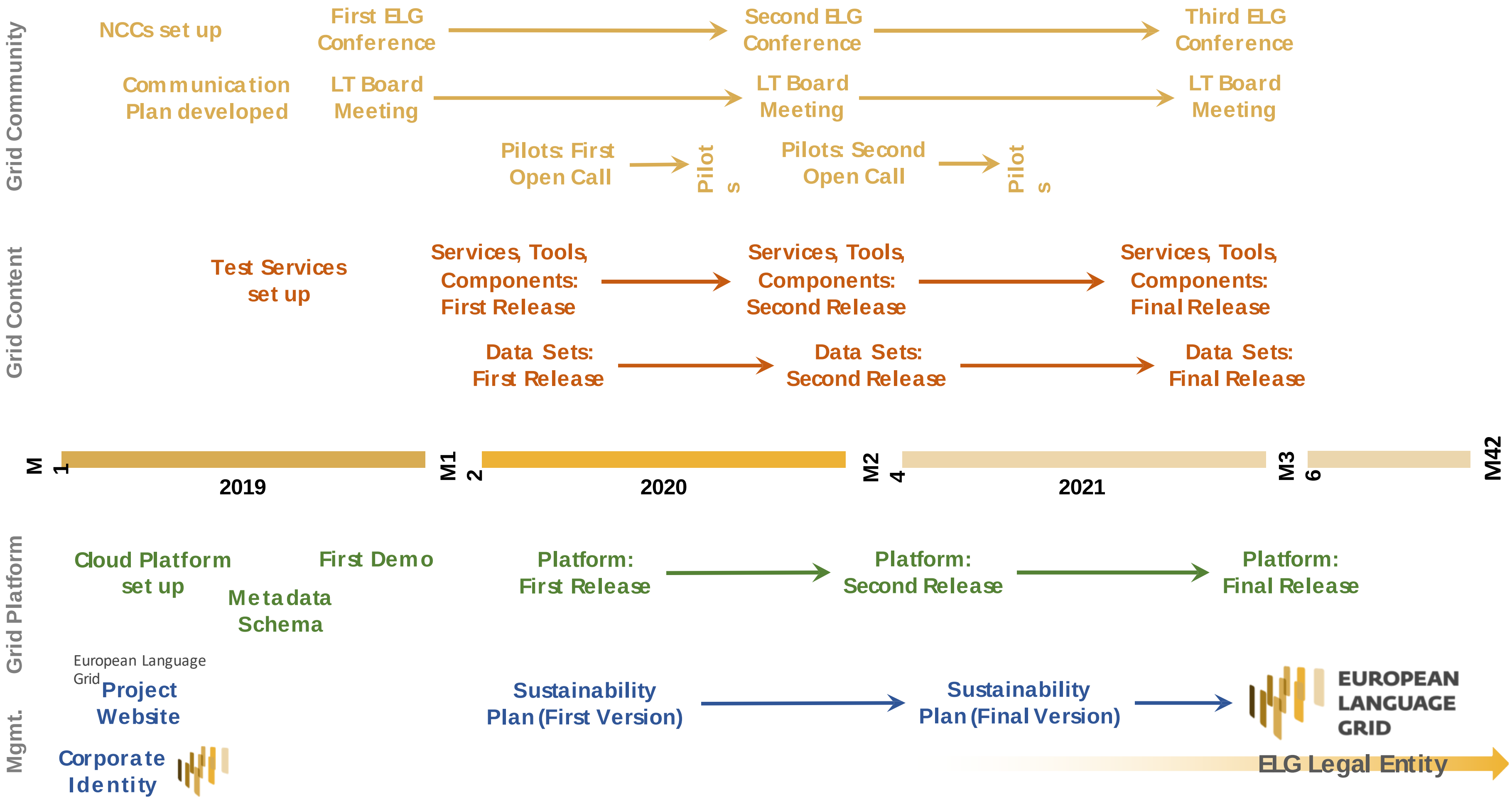
EUROPEAN LANGUAGE GRID

Kick-off meeting, 22/23 January 2019

European Language Grid

Objectives (Selection)

1. Establish the ELG as the primary Language Technology platform and market place in Europe to tackle the fragmentation of the European LT landscape.
2. ELG as a platform for commercial and non-commercial, industry-related LTs (functional and non-functional).
3. Enable the European LT community to upload services and data sets, to deploy them and to connect with, and make use of those resources made available by others.
4. Enable businesses to grow and benefit from scaling up.
5. Unleash enormous potential for innovation.



Grid Community

NCCs set up

First ELG Conference → Second ELG Conference → Third ELG Conference

Communication Plan developed

LT Board Meeting → LT Board Meeting → LT Board Meeting

Pilots: First Open Call → Pilots

Pilots: Second Open Call → Pilots

Grid Content

Test Services set up

Services, Tools, Components: First Release → Services, Tools, Components: Second Release → Services, Tools, Components: Final Release

Data Sets: First Release → Data Sets: Second Release → Data Sets: Final Release

M1 — 2019 — M12 — 2020 — M24 — 2021 — M36 — M42

Grid Platform

Cloud Platform set up

First Demo

Metadata Schema

Platform: First Release → Platform: Second Release → Platform: Final Release

Mgmt.

European Language Grid

Project Website

Corporate Identity

Sustainability Plan (First Version) → Sustainability Plan (Final Version) → EUROPEAN LANGUAGE GRID

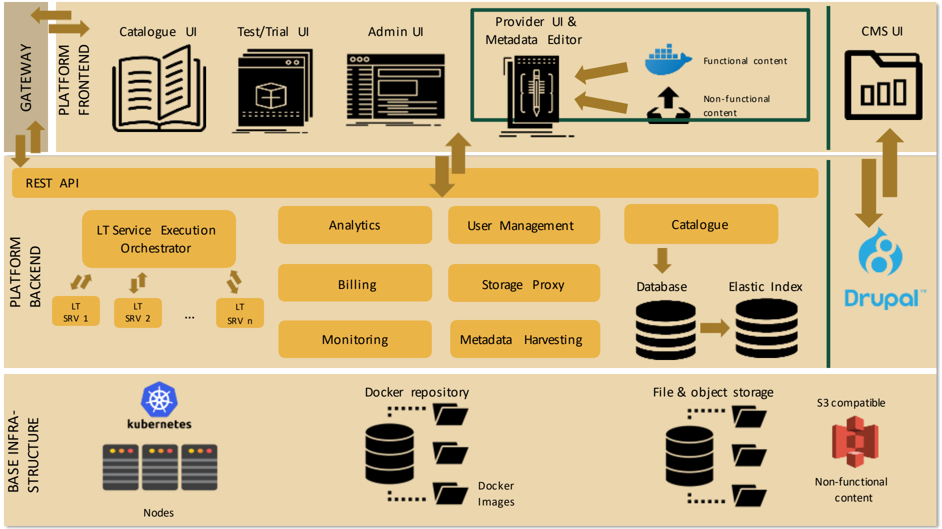
ELG Legal Entity

Current State of Play (December 2021)

EUROPEAN LANGUAGE GRID

Release 2 (Dec. 2021):

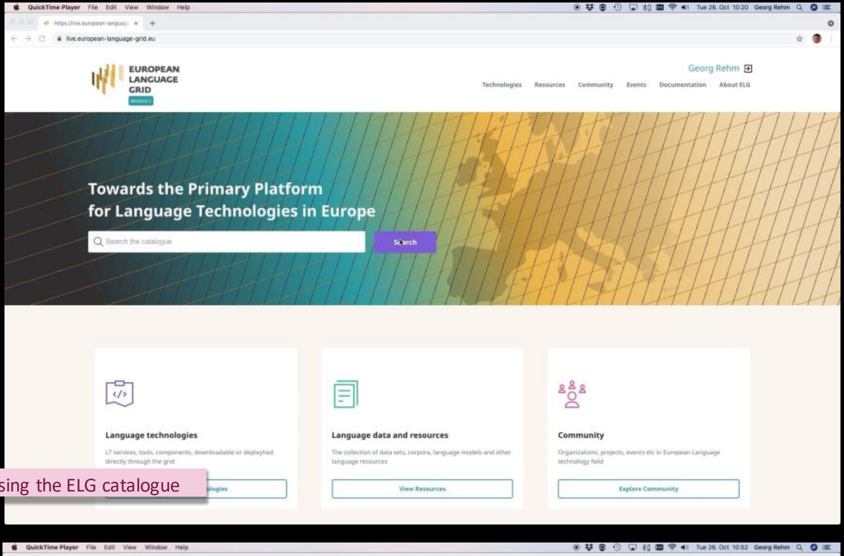
- 3626 corpora and data sets
- 477 functional services and tools
- 964 lexical/conceptual resources
- 27 lang. descriptions, grammars, models
- 1789 organisations (research organisations, companies)



Users can connect to the ELG cloud platform via ELG APIs, remote APIs, ELG GUI, Python SDK, download of containers or source code.

European Language Grid

- All layers built with robust scalable, reliable, widely used technologies
- Docker containers for all services and applications which comprise the ELG platform
- Kubernetes for container orchestration – ability to scale with the growing demand and supply of resources
- Laying the foundations for interoperable data and services spaces


QuickTime Player File Edit View Window Help
Tue 26. Oct 10:20 Georg Rehm
live.european-language-grid.eu
EUROPEAN LANGUAGE GRID
Georg Rehm
Technologies Resources Community Events Documentation About ELG
Search the catalogue
Search

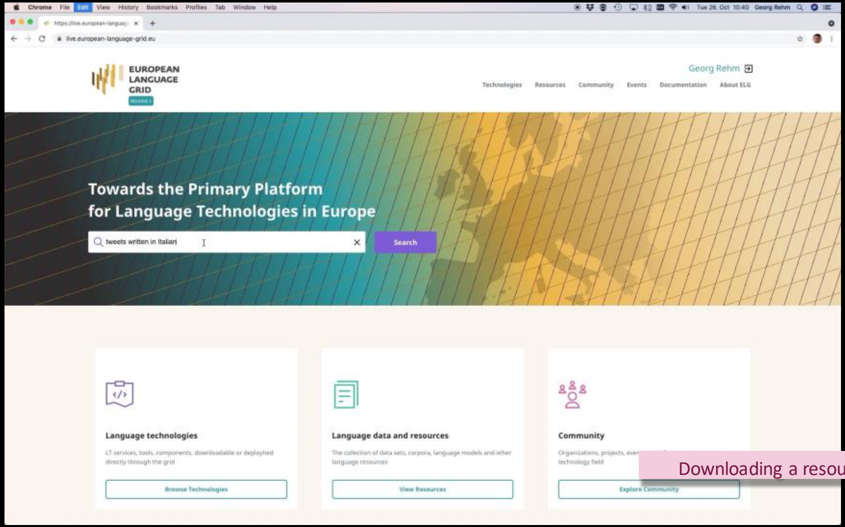
Towards the Primary Platform
for Language Technologies in Europe

Language technologies
LT services, tools, components, downloadable or deployed directly through the grid

Language data and resources
The collection of data sets, corpora, language models and other language resources
View Resources

Community
Organizations, projects, events etc in European Language technology field
Explore Community

Browsing the ELG catalogue


Chrome File Edit View History Bookmarks Profiles Tab Window Help
Tue 26. Oct 10:40 Georg Rehm
live.european-language-grid.eu
EUROPEAN LANGUAGE GRID
Georg Rehm
Technologies Resources Community Events Documentation About ELG
Search

Towards the Primary Platform
for Language Technologies in Europe

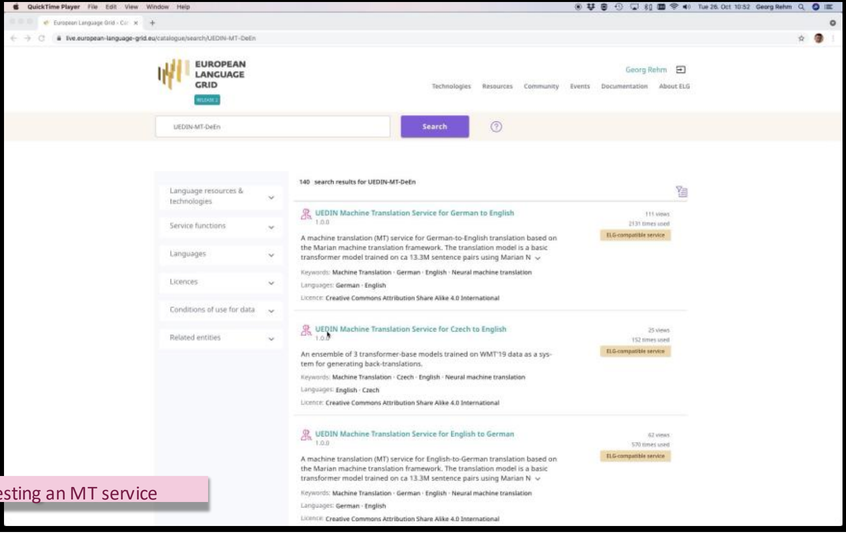
tweets written in Italian

Language technologies
LT services, tools, components, downloadable or deployed directly through the grid
Browse Technologies

Language data and resources
The collection of data sets, corpora, language models and other language resources
View Resources

Community
Organizations, projects, events etc in European Language technology field
Explore Community

Downloading a resource


QuickTime Player File Edit View Window Help
Tue 26. Oct 10:52 Georg Rehm
live.european-language-grid.eu/catalogue/search/UEDIN-MT-DeEn
EUROPEAN LANGUAGE GRID
Georg Rehm
Technologies Resources Community Events Documentation About ELG
Search
Language resources & technologies
Service functions
Languages
Licences
Conditions of use for data
Related entities

UEDIN-MT-DeEn

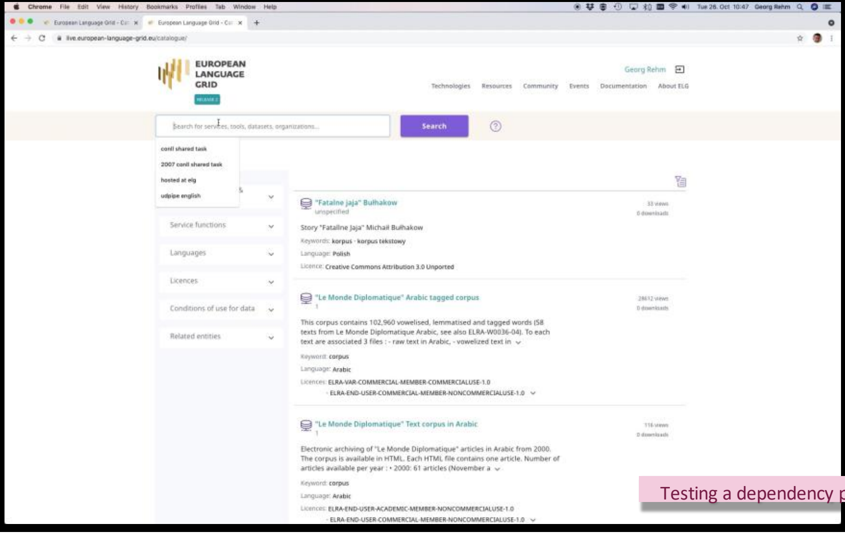
140 search results for UEDIN-MT-DeEn

UEDIN Machine Translation Service for German to English
1.0.0
111 views
2131 times used
ELG-compatible service
A machine translation (MT) service for German-to-English translation based on the Marian machine translation framework. The translation model is a basic transformer model trained on ca 13.3M sentence pairs using Marian N
Keywords: Machine Translation · German · English · Neural machine translation
Languages: German · English
Licence: Creative Commons Attribution Share Alike 4.0 International

UEDIN Machine Translation Service for Czech to English
1.0.0
25 views
152 times used
ELG-compatible service
An ensemble of 3 transformer-base models trained on WMT'19 data as a system for generating back-translations.
Keywords: Machine Translation · Czech · English · Neural machine translation
Languages: English · Czech
Licence: Creative Commons Attribution Share Alike 4.0 International

UEDIN Machine Translation Service for English to German
1.0.0
62 views
570 times used
ELG-compatible service
A machine translation (MT) service for English-to-German translation based on the Marian machine translation framework. The translation model is a basic transformer model trained on ca 13.3M sentence pairs using Marian N
Keywords: Machine Translation · German · English · Neural machine translation
Languages: German · English
Licence: Creative Commons Attribution Share Alike 4.0 International

Testing an MT service


Chrome File Edit View History Bookmarks Profiles Tab Window Help
Tue 26. Oct 10:47 Georg Rehm
European Language Grid - Ca
live.european-language-grid.eu/catalogue/
EUROPEAN LANGUAGE GRID
Georg Rehm
Technologies Resources Community Events Documentation About ELG
Search for services, tools, datasets, organizations...
Search
conll shared task
2007 conll shared task
hosted at elg
udpipe english
Service functions
Languages
Licences
Conditions of use for data
Related entities

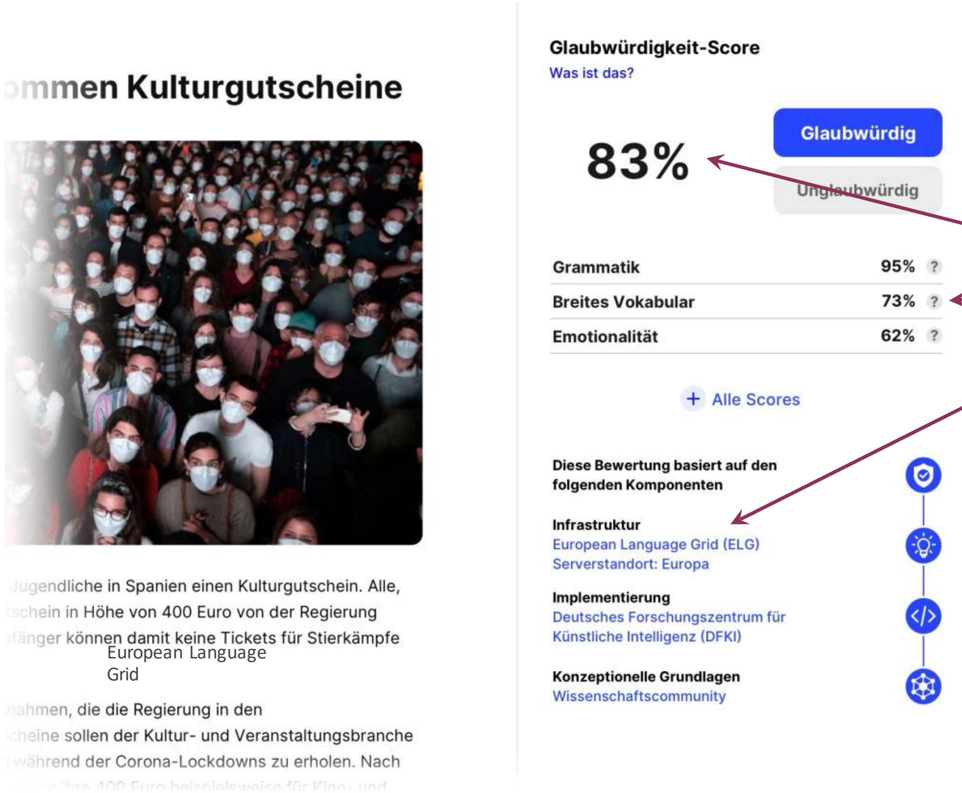
"Fatalne jaja" Bułhakow
unspecified
33 views
0 downloads
Story "Fatalne Jaja" Michaił Bułhakow
Keywords: korpus · korpus tekstowy
Language: Polish
Licence: Creative Commons Attribution 3.0 Unported

"Le Monde Diplomatique" Arabic tagged corpus
1
28612 views
0 downloads
This corpus contains 102,960 vowelised, lemmatised and tagged words (58 texts from Le Monde Diplomatique Arabic, see also ELRA-W0036-04). To each text are associated 3 files : - raw text in Arabic, - vowelized text in
Keyword: corpus
Language: Arabic
Licences: ELRA-VAR-COMMERCIAL-MEMBER-COMMERCIALUSE-1.0
· ELRA-END-USER-COMMERCIAL-MEMBER-NONCOMMERCIALUSE-1.0

"Le Monde Diplomatique" Text corpus in Arabic
1
116 views
0 downloads
Electronic archiving of "Le Monde Diplomatique" articles in Arabic from 2000. The corpus is available in HTML. Each HTML file contains one article. Number of articles available per year : • 2000: 61 articles (November a
Keyword: corpus
Language: Arabic
Licences: ELRA-END-USER-ACADEMIC-MEMBER-NONCOMMERCIALUSE-1.0
· ELRA-END-USER-COMMERCIAL-MEMBER-NONCOMMERCIALUSE-1.0

Testing a dependency parser

Using ELG Services – it's really very simple!



ommen Kulturgutscheine

Jugendliche in Spanien einen Kulturgutschein. Alle, tschein in Höhe von 400 Euro von der Regierung fänger können damit keine Tickets für Stierkämpfe

nahmen, die die Regierung in den cheine sollen der Kultur- und Veranstaltungsbranche während der Corona-Lockdowns zu erholen. Nach

Glaubwürdigkeit-Score

Was ist das?

83%

Glaubwürdig

Unglaubwürdig

Grammatik	95% ?
Breites Vokabular	73% ?
Emotionalität	62% ?

+ Alle Scores

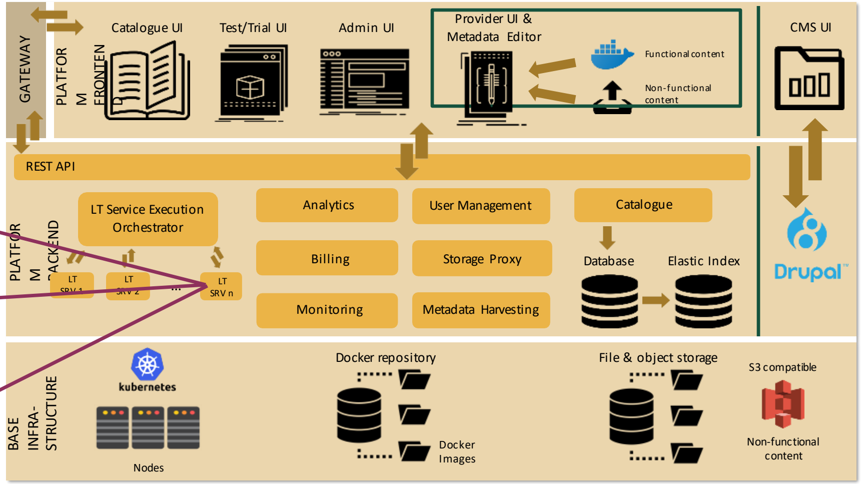
Diese Bewertung basiert auf den folgenden Komponenten

Infrastruktur
European Language Grid (ELG)
Serverstandort: Europa

Implementierung
Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI)

Konzeptionelle Grundlagen
Wissenschaftscommunity

European Language
Grid



- In the project PANQURA we perform automated credibility assessment of online content.
- The corresponding credibility assessment tool is hosted in ELG and can be called through the public ELG API.
- The service's response is exposed to the user – in this specific example clearly indicating the source of the service and also where the server is located, which is very important to users in the credibility assessment use case.

Stakeholders and Users

Companies that

- … *develop* Language Technologies
- … *integrate* Language Technologies
- … *purchase* Language Technologies

Universities and research centres that

- … *develop* Language Technologies
- … *use* Language Technologies

META-FORUM 2019 (8/9 October) – Brussels, Belgium

Public administrations that *purchase* or *use* Language Technologies

Other organisations (e.g., NGOs) that *purchase* or *use* Language Technologies

Funding agencies that support the development of Language Technologies

Stakeholders and Users and Collaborators

EUROPEAN LANGUAGE EQUALITY

Prêt-à-LLOD

ELG Open Calls: 15 pilot projects

ICT-29b research and innovation projects

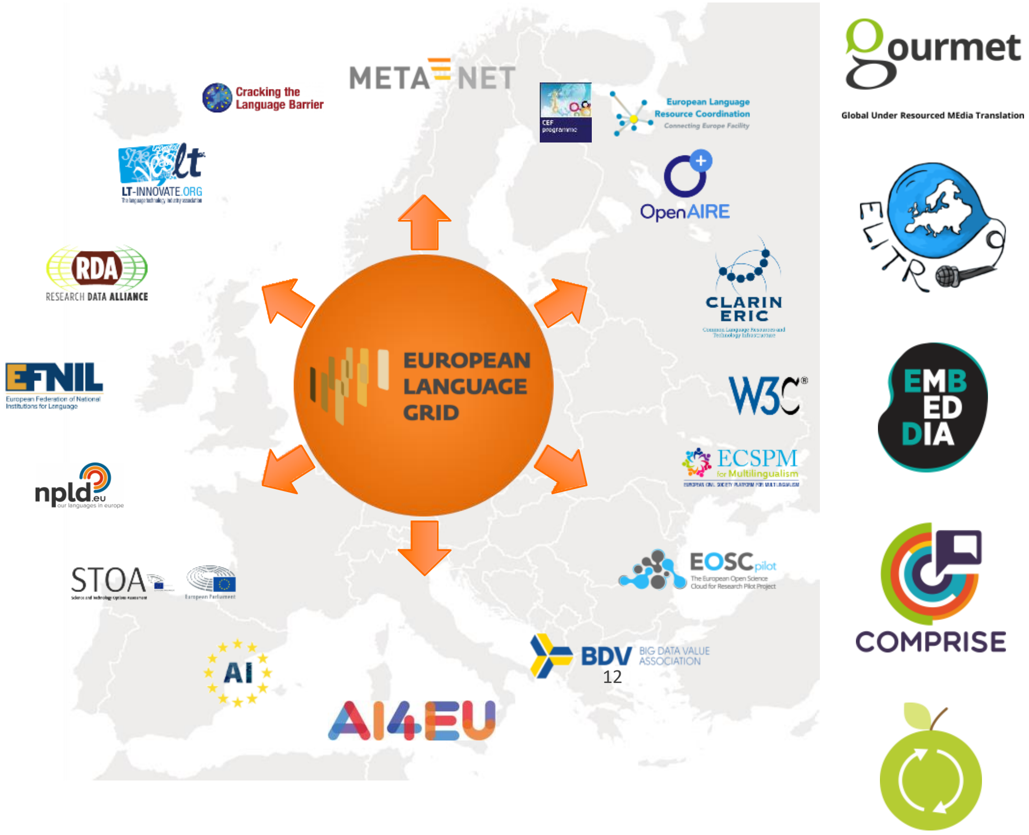
ELE – European Language Equality

AI4EU – European AI on demand platform

Other related projects: Elexis, Lynx, MeMAD, CEF INEA projects (MAPA, NTEU etc.), OpenGPT-X, NFDI4DataScience etc.

Core networks: META-NET, LT Innovate

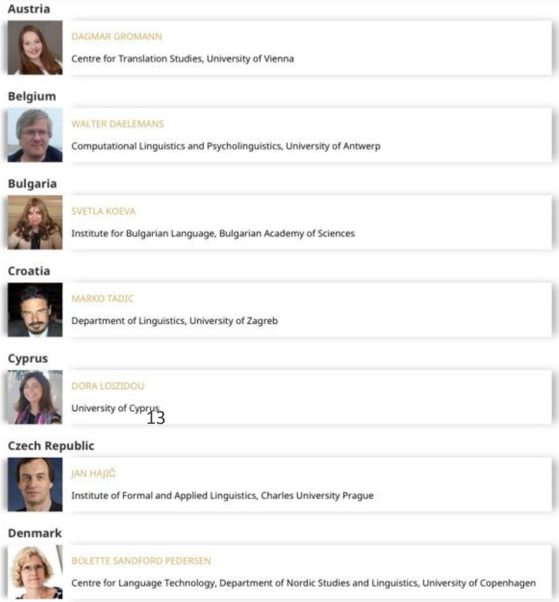
Additional initiatives: CLAIRE, AI PPP, ELRC, ECSPM, EFNIL, BDVA, NPLD, CLARIN, W3C, NGI, RDA, EOSC, OpenAIRE etc.


EUROPEAN LANGUAGE GRID
Cracking the Language Barrier
META-NET
CEF programme
European Language Resource Coordination
Connecting Europe Facility
OpenAIRE
CLARIN ERIC
W3C
ECSPM for Multilingualism
EOSC pilot
The European Open Science Cloud for Research Pilot Project
BDV BIG DATA VALUE ASSOCIATION
AI4EU
AI
STOA
European Parliament
npld.eu
our languages in europe
EFNIL
European Federation of National Institutions for Language
RDA
RESEARCH DATA ALLIANCE
LT-INNOVATE.ORG
Gourmet
Global Under Resourced MEdia Translation
ELITR
EMBEDDIA
COMPRISE

Community: NCCs and LTC

32 National Competence Centres (NCCs)

- Strong international network of national networks to broaden ELG’s reach, identify content for the ELG and interest companies in using the ELG.
- Main goal: *support the mission of the ELG project*.



Austria

DAGMAR GROMANN

Centre for Translation Studies, University of Vienna

Belgium

WALTER DAELEMANS

Computational Linguistics and Psycholinguistics, University of Antwerp

Bulgaria

SVETLA KOEVA

Institute for Bulgarian Language, Bulgarian Academy of Sciences

Croatia

MARKO TADIC

Department of Linguistics, University of Zagreb

Cyprus

DORA LOIZIDOU

University of Cyprus

Czech Republic

JAN HAJIČ

Institute of Formal and Applied Linguistics, Charles University Prague

Denmark

BOLETTE SANDFORD PEDERSEN

Centre for Language Technology, Department of Nordic Studies and Linguistics, University of Copenhagen

European LT Council (LTC) (work in progress)

- A pan-European body, in which strategic LT-related matters can be discussed and coordinated.
- Main goal: *represent and support European LT community*.

EUROPEAN LANGUAGE EQUALITY

1. **Austria:** Dagmar Gromann, Zentrum für Translationswissenschaft, Universität Wien
2. **Belgium:** Walter Daelemans, Comp. Ling. and Psycholing. Res. Centre (CLiPS), University of Antwerp
3. **Bulgaria:** Svetla Koeva, Bulgarian Academy of Sciences
4. **Croatia:** Marko Tadic, Inst. of Linguistics, Faculty of Hum. and Social Science, University of Zagreb
5. **Cyprus:** Dora Loizidou, University of Cyprus
6. **Czech Republic:** Jan Hajic, Inst. of Formal and Applied Linguistics, Charles University in Prague
7. **Denmark:** Bolette Sandford Pedersen, Centre for Lang. Technology, Department of Nordic Research, University of Copenhagen
8. **Estonia:** Kadri Vare, Department of Language, Estonian Ministry of Education and Resear
9. **Finland:** Krister Linden, Department of Modern Languages, University of Helsinki
10. **France:** François Yvon, CNRS-LIMSI

European Language Grid

11. **Germany:** Georg Rehm, Speech and Language Technology Lab, DFKI
12. **Greece:** Maria Gavriilidou, ILSP, R.C. "Athena"
13. **Hungary:** Tamás Várárdi, Research Institute for Linguistics, Hungarian Academy of Sciences
14. **Iceland:** Eirikur Rögnvaldsson, School of Humanities, University of Iceland
15. **Ireland:** Andy Way, ADAPT Centre and School of Computing, Dublin City University
16. **Italy:** Bernardo Magnini, Human Language Technology, Fondazione Bruno Kessler (FBK)
17. **Latvia:** Inguna Skadina, Institute of Mathematics and Computer Science, University of Latvia
18. **Lithuania:** Albina Auksoriūtė, Institute of the Lithuanian Language
19. **Luxembourg:** Dimitra Anastasiou, Luxembourg Institute of Science and Technology
20. **Malta:** Michael Rosner, Department Intelligent Computer Systems, University of Malta
21. **Netherlands:** Vincent Vandeghinste, Dutch Language Institute, Centre for Computational Linguistics, University of Leuven
22. **Norway:** Kristine Eide, The Language Council of Norway – Språkrådet
23. **Poland:** Maciej Ogrodniczuk, Institute of Computer Science, Polish Academy of Sciences
24. **Portugal:** António Branco, Department of Informatics, University of Lisbon
25. **Romania:** Dan Tufis, Research Institute for Artificial Intelligence, Romanian Academy of Sciences
26. **Serbia:** Cvetana Krstev, Faculty of Mathematics, Belgrade University (UBG)
27. **Slovakia:** Radovan Garabik, Ludovit Stur Institute of Linguistics, Slovak Academy of Sciences
28. **Slovenia:** Simon Krek, Jozef Stefan Institute
29. **Spain:** Marta Villegas, Barcelona Supercomputing
30. **Sweden:** Jens Edlund, Speech, Music & Hearing/Språkbanken Tal, KTH Royal Institute of Technology
31. **Switzerland:** Hervé Bourlard, Idiap Research Institute
32. **UK:** Kalina Bontcheva, Department of Computer Science, University of Sheffield

META-FORUM Conference Series

META-FORUM 2022 – June 08-10, Brussels, Belgium
Save the date!

META-FORUM 2021 – November 15-17, *virtual conference*
Using the European Language Grid

META-FORUM 2020 – December 01-03, *virtual conference*
Piloting the European Language Grid

META-FORUM 2019 – October 08/09, Brussels, Belgium
Introducing the European Language Grid

META-FORUM 2017 – November 13/14, Brussels, Belgium
Towards a Human Language Project

META-FORUM 2016 – July 04/05, Lisbon, Portugal
Beyond Multilingual Europe

META-FORUM 2015 – April 27, Riga, Latvia
Technologies for the Multilingual Digital Single Market

META-FORUM 2013 – September 19/20, Berlin, Germany
Connecting Europe for New Horizons

META-FORUM 2012 – June 20/21, Brussels, Belgium
A Strategy for Multilingual Europe

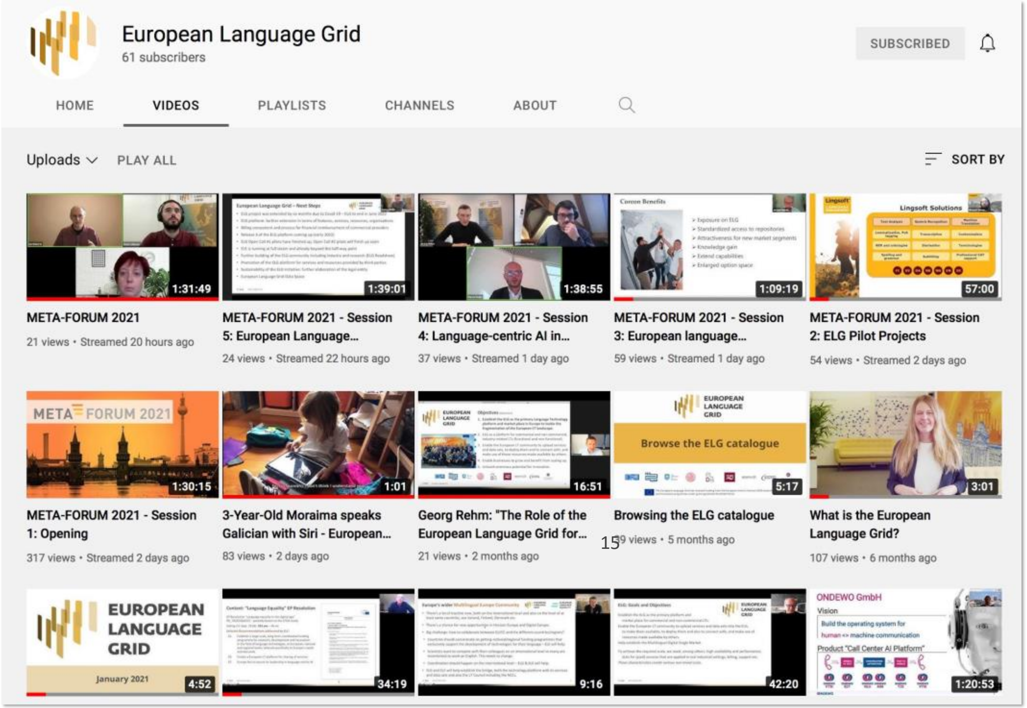
European Language Grid

META-FORUM 2011 – June 27/28, Budapest, Hungary
Solutions for Multilingual Europe

META-FORUM 2010 – November 17/18, Brussels, Belgium
Challenges for Multilingual Europe

All META-FORUM 2020 and META-FORUM 2021 sessions are available on the European Language Grid YouTube channel:

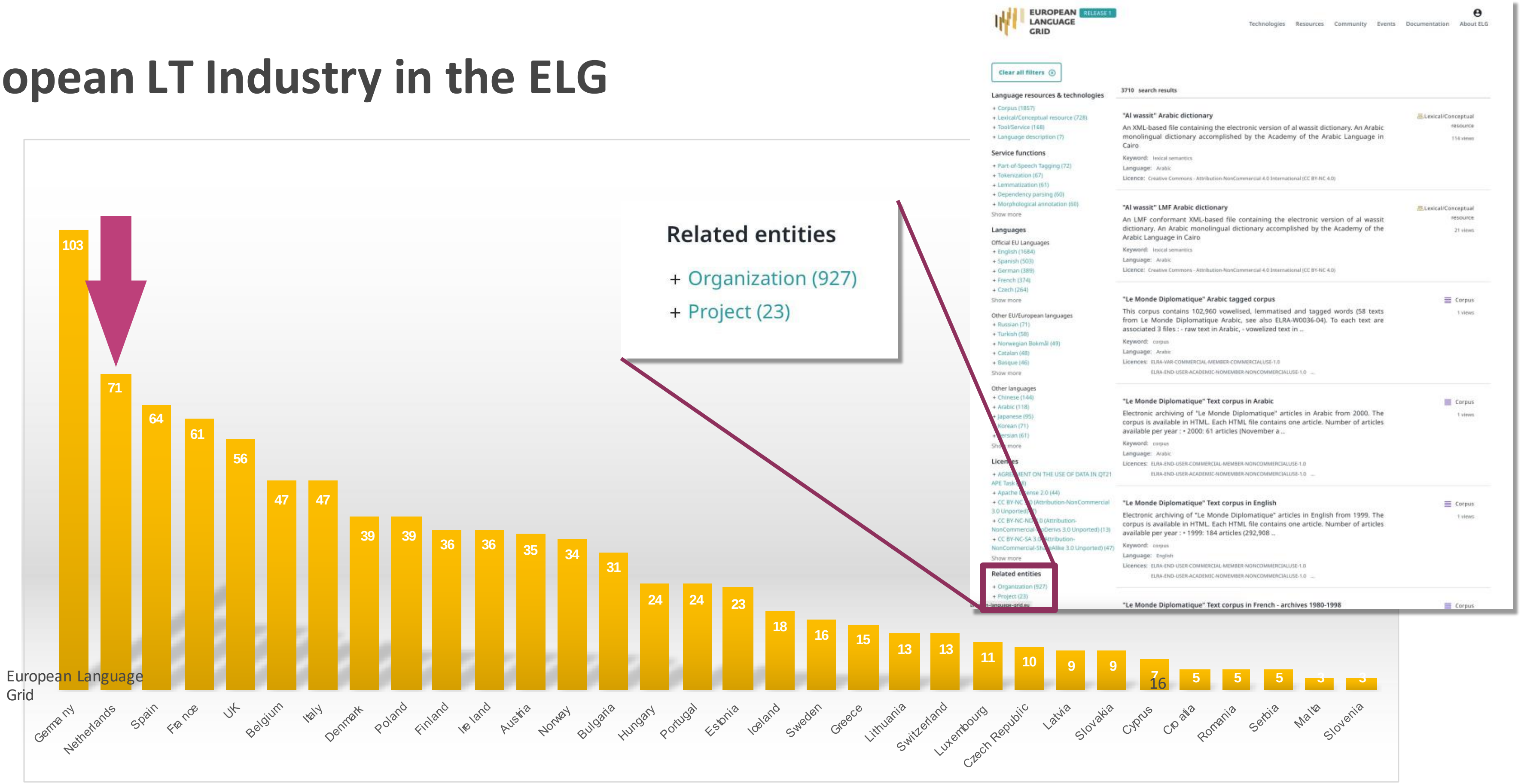
https://www.youtube.com/channel/UCarEHmsWT2JslcvvWkbhL4A



European Language Grid
61 subscribers

- 1:31:49 META-FORUM 2021 — 21 views • Streamed 20 hours ago
- 1:39:01 META-FORUM 2021 - Session 5: European Language... — 24 views • Streamed 22 hours ago
- 1:38:55 META-FORUM 2021 - Session 4: Language-centric AI in... — 37 views • Streamed 1 day ago
- 1:09:19 META-FORUM 2021 - Session 3: European language... — 59 views • Streamed 1 day ago
- 57:00 META-FORUM 2021 - Session 2: ELG Pilot Projects — 54 views • Streamed 2 days ago
- 1:30:15 META-FORUM 2021 - Session 1: Opening — 317 views • Streamed 2 days ago
- 1:01 3-Year-Old Moraima speaks Galician with Siri - European... — 83 views • 2 days ago
- 16:51 Georg Rehm: "The Role of the European Language Grid for... — 21 views • 2 months ago
- 5:17 Browsing the ELG catalogue — 9 views • 5 months ago
- 3:01 What is the European Language Grid? — 107 views • 6 months ago
- 4:52
- 34:19
- 9:16
- 42:20
- 1:20:53

European LT Industry in the ELG



Related entities

+ Organization (927)

+ Project (23)

Language resources & technologies

+ Corpus (1857)
+ Lexical/Conceptual resource (728)
+ Tool/Service (168)
+ Language description (7)

Service functions

+ Part-of-Speech Tagging (72)
+ Tokenization (67)
+ Lemmatization (61)
+ Dependency parsing (60)
+ Morphological annotation (60)

Show more

Languages

Official EU Languages

+ English (1684)
+ Spanish (503)
+ German (389)
+ French (374)
+ Czech (264)

Show more

Other EU/European languages

+ Russian (71)
+ Turkish (58)
+ Norwegian Bokmål (49)
+ Catalan (48)
+ Basque (46)

Show more

Other languages

+ Chinese (144)
+ Arabic (118)
+ Japanese (95)
+ Korean (71)
+ ersian (61)

Sho more

Licen es

+ AGRE MENT ON THE USE OF DATA IN QT21 APE Task
+ Apache L ense 2.0 (44)
+ CC BY-NC (Attribution-NonCommercial 3.0 Unported)
+ CC BY-NC-ND .0 (Attribution-NonCommercial- oDerivs 3.0 Unported) (13)
+ CC BY-NC-SA 3.0 Attribution-NonCommercial-Sh eAlike 3.0 Unported) (47)

Show more

Related entities

+ Organization (927)
+ Project (23)

n-language-grid.eu

3710 search results

"Al wassit" Arabic dictionary — Lexical/Conceptual resource — 114 views

An XML-based file containing the electronic version of al wassit dictionary. An Arabic monolingual dictionary accomplished by the Academy of the Arabic Language in Cairo

Keyword: lexical semantics
Language: Arabic
Licence: Creative Commons - Attribution-NonCommercial 4.0 International (CC BY-NC 4.0)

"Al wassit" LMF Arabic dictionary — Lexical/Conceptual resource — 21 views

An LMF conformant XML-based file containing the electronic version of al wassit dictionary. An Arabic monolingual dictionary accomplished by the Academy of the Arabic Language in Cairo

Keyword: lexical semantics
Language: Arabic
Licence: Creative Commons - Attribution-NonCommercial 4.0 International (CC BY-NC 4.0)

"Le Monde Diplomatique" Arabic tagged corpus — Corpus — 1 views

This corpus contains 102,960 vowelised, lemmatised and tagged words (58 texts from Le Monde Diplomatique Arabic, see also ELRA-W0036-04). To each text are associated 3 files : - raw text in Arabic, - vowelized text in ...

Keyword: corpus
Language: Arabic
Licences: ELRA-VAR-COMMERCIAL-MEMBER-COMMERCIALUSE-1.0
ELRA-END-USER-ACADEMIC-NOMEMBER-NONCOMMERCIALUSE-1.0 ...

"Le Monde Diplomatique" Text corpus in Arabic — Corpus — 1 views

Electronic archiving of "Le Monde Diplomatique" articles in Arabic from 2000. The corpus is available in HTML. Each HTML file contains one article. Number of articles available per year : • 2000: 61 articles (November a ...

Keyword: corpus
Language: Arabic
Licences: ELRA-END-USER-COMMERCIAL-MEMBER-NONCOMMERCIALUSE-1.0
ELRA-END-USER-ACADEMIC-NOMEMBER-NONCOMMERCIALUSE-1.0 ...

"Le Monde Diplomatique" Text corpus in English — Corpus — 1 views

Electronic archiving of "Le Monde Diplomatique" articles in English from 1999. The corpus is available in HTML. Each HTML file contains one article. Number of articles available per year : • 1999: 184 articles (292,908 ...

Keyword: corpus
Language: English
Licences: ELRA-END-USER-COMMERCIAL-MEMBER-NONCOMMERCIALUSE-1.0
ELRA-END-USER-ACADEMIC-NOMEMBER-NONCOMMERCIALUSE-1.0 ...

"Le Monde Diplomatique" Text corpus in French - archives 1980-1998 — Corpus

European Language Grid: Sustainable OperaTonal Model & Legal EnTty

- ELG is supposed to be a long-term, sustainable initiative – a legal entity is needed.
- The technical and operational requirements – high availability and performance, SLAs, billing, support etc. – create non-trivial costs: hosting; bandwidth; ELG team; legal; etc.
- We've identified several ways and approaches of covering the costs on a long-term basis.
- Establish consensus for a sustainable operational model.
- Options: a) for-profit or b) *not-for-profit company*, c) association, d) foundation
- Business Model Canvas prepared
- First set of approx. 20 products defined
- Next up: market validation of the product ideas
- Q1/Q2 2022: Establish legal entity

Open Calls for Pilot Projects

Two open calls for pilot projects

- **Open Call #1:** 03/04 2020
- **Open Call #2:** 10/11 2020

Pilot projects shall

- **Type A:** broaden ELG's portfolio or
- **Type B:** demonstrate usefulness of ELG

Up to €200,000 per project

Approx. €2,000,000 FSTP in total

Available in Open Call #1: 1.3M€

Available in Open Call #2: 585k€

Project duration: **9-12 months**

Eligibility: **SMEs, research organisations**

Open Call #1 – StaTsTcs

110 project proposals evaluated

- **Type A:** 79 – **Type B:** 31 proposals

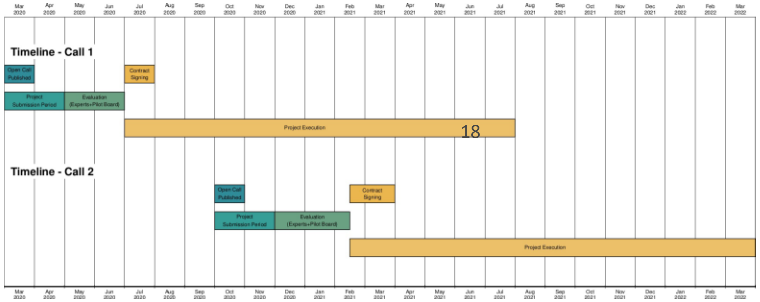
Total amount requested: 16.9M€

□ **10 projects selected**

Open Call #2 – Statistics

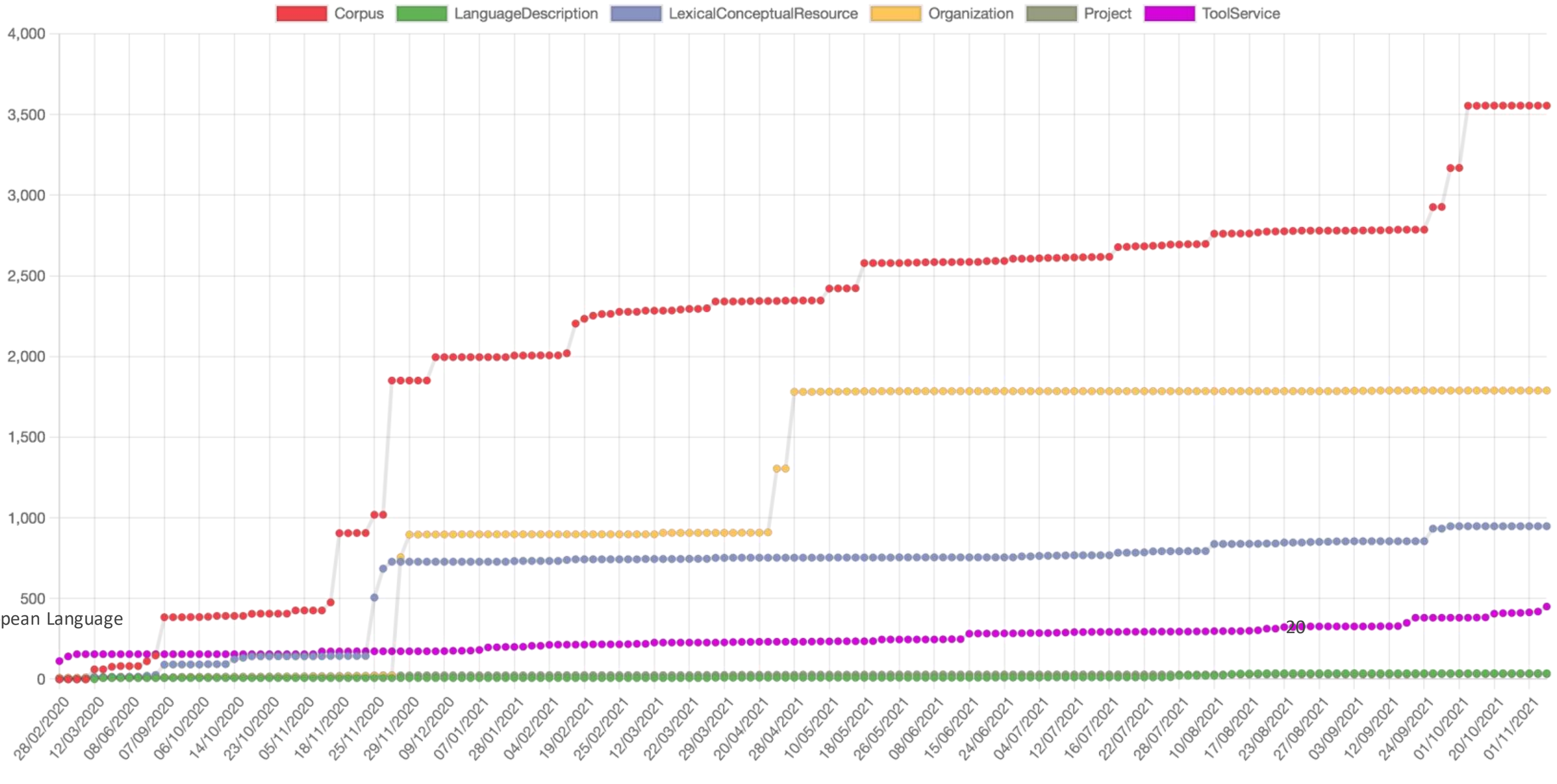
106 project proposals evaluated

□ **5 projects selected**



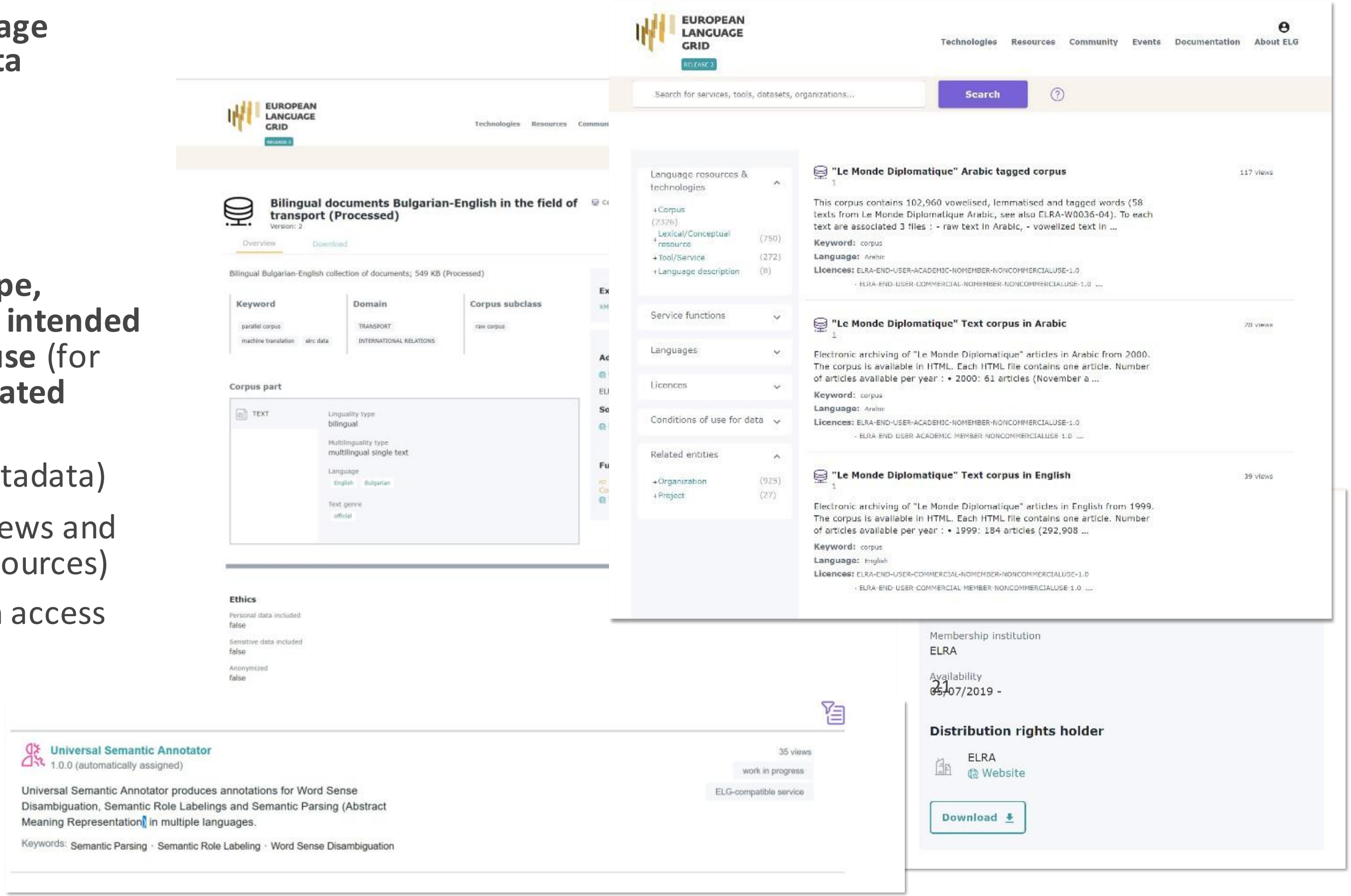
Open Call #1	Fondazione Bruno Kessler	European Clinical Case Corpus ✓	IT	139,370€
	Lingsoft, Inc.	Lingsoft Solutions as Distributable Container ✓	FI	140,625€
	Coreon GmbH	MKS as Linguistic Linked Open Data ✓	DE	167,375€
	Elhuyar Fundazioa	Basque-speaking smart speaker based on Mycroft A ✓	ES	117,117€
	Universita' Degli Studi di Torino	Italian EVALITA Benchmark Linguistic Resources, NLP Services and Too ✓	IT	126,125€
	University of Helsinki	Open Translation Models, Tools and Service ✓	FI	154,636€
	University of Vienna	Extracting Terminological Concept Systems from Natural Language Tex ✓	AT	132,977€
	University of Turku	Textual paraphrase dataset for deep language modellin ✓	FI	166,085€
	Weber Consulting KG	Virtual Personal Assistant Prototyp ✓	AT	87,445€
	FZI Research Center for IT	Streaming Language Processing in Manufacturin ✓	DE	132,160€
Open Call #2	Institute for Bulgarian Language	Multilingual Image Corpus 2021	BG	110,960€
	EDIA BV	CEFR Labelling and Assessment Services	NL	137,560€
	University of West Bohemia	Motion Capture 3D Sign Language Resources	CZ	85,421€
	Sapienza University of Rome	Universal Semantic Annotator	IT	113,228€
	Sign Time GmbH	Sign language explanations for terms in a text	AT	137,227€

Number of Resources and Resource Types over Time



On the data consumer side

- Consumers can **search and browse** the ELG Catalogue
  - for **different types of language processing services** and **data**
  - related **projects** and **organisations** in Europe
  - using simple and advanced free text search
  - using **facets for resource type, language, service function, intended application, conditions of use** (for data resources), **license, related entities**
- View detailed information (metadata)
- View statistics of number of views and downloads (for ELG hosted resources)
- Download data (depending on access conditions)
- Export metadata records
- Check what is forthcoming in terms of data and services



EUROPEAN LANGUAGE GRID

Technologies Resources Community Events Documentation About ELG

Search for services, tools, datasets, organizations...

Search

Language resources & technologies
- Corpus (2326)
- Lexical/Conceptual resource (750)
- Tool/Service (272)
- Language description (8)

Service functions

Languages

Licences

Conditions of use for data

Related entities
- Organization (925)
- Project (27)

"Le Monde Diplomatique" Arabic tagged corpus — 117 views

This corpus contains 102,960 vowelised, lemmatised and tagged words (58 texts from Le Monde Diplomatique Arabic, see also ELRA-W0036-04). To each text are associated 3 files : - raw text in Arabic, - vowelized text in ...

Keyword: corpus
Language: Arabic
Licences: ELRA-END-USER-ACADEMIC-NOMEMBER-NONCOMMERCIALUSE-1.0 · ELRA-END-USER-COMMERCIAL-NOMEMBER-NONCOMMERCIALUSE-1.0 ...

"Le Monde Diplomatique" Text corpus in Arabic — 28 views

Electronic archiving of "Le Monde Diplomatique" articles in Arabic from 2000. The corpus is available in HTML. Each HTML file contains one article. Number of articles available per year : • 2000: 61 articles (November a ...

Keyword: corpus
Language: Arabic
Licences: ELRA-END-USER-ACADEMIC-NOMEMBER-NONCOMMERCIALUSE-1.0 · ELRA-END-USER-ACADEMIC-MEMBER-NONCOMMERCIALUSE-1.0 ...

"Le Monde Diplomatique" Text corpus in English — 39 views

Electronic archiving of "Le Monde Diplomatique" articles in English from 1999. The corpus is available in HTML. Each HTML file contains one article. Number of articles available per year : • 1999: 184 articles (292,908 ...

Keyword: corpus
Language: English
Licences: ELRA-END-USER-COMMERCIAL-NOMEMBER-NONCOMMERCIALUSE-1.0 · ELRA-END-USER-COMMERCIAL-MEMBER-NONCOMMERCIALUSE-1.0 ...

Bilingual documents Bulgarian-English in the field of transport (Processed)
Version: 2

Overview · Download

Bilingual Bulgarian-English collection of documents; 549 KB (Processed)

Keyword: parallel corpus · machine translation · elrc data
Domain: TRANSPORT · INTERNATIONAL RELATIONS
Corpus subclass: raw corpus

Corpus part
TEXT
Linguality type: bilingual
Multilinguality type: multilingual single text
Language: English · Bulgarian
Text genre: official

Ethics
Personal data included: false
Sensitive data included: false
Anonymized: false

Membership institution: ELRA
Availability: 05/07/2019 -

Distribution rights holder
ELRA
Website

Download

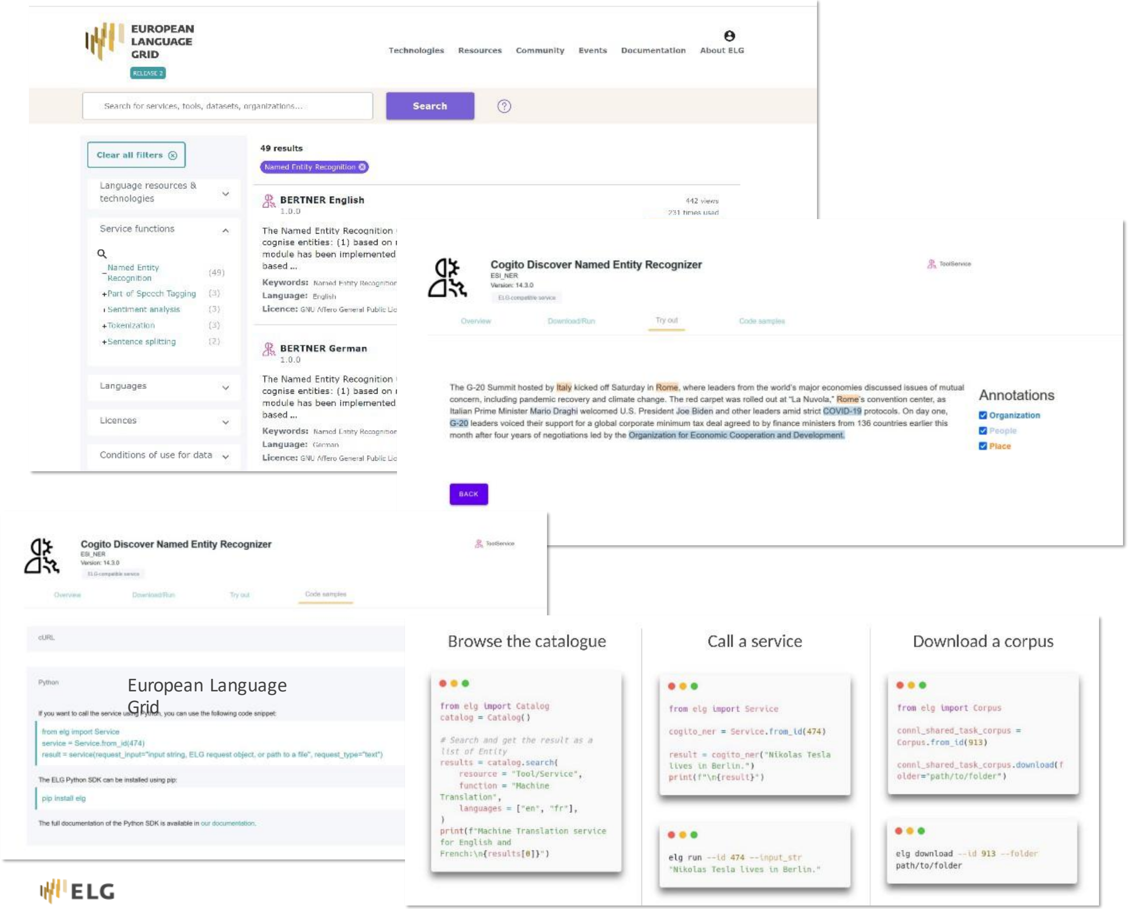
Universal Semantic Annotator — 35 views
1.0.0 (automatically assigned)

work in progress · ELG-compatible service

Universal Semantic Annotator produces annotations for Word Sense Disambiguation, Semantic Role Labelings and Semantic Parsing (Abstract Meaning Representation) in multiple languages.

Keywords: Semantic Parsing · Semantic Role Labeling · Word Sense Disambiguation

On the services consumer side



EUROPEAN LANGUAGE GRID RELEASE 2

Technologies Resources Community Events Documentation About ELG

Search for services, tools, datasets, organizations...

Search

Clear all filters ⊗

Language resources & technologies

Service functions

– Named Entity Recognition	(49)
+ Part of Speech Tagging	(3)
+ Sentiment analysis	(3)
+ Tokenization	(3)
+ Sentence splitting	(2)

Languages

Licences

Conditions of use for data

49 results

Named Entity Recognition ⊗

BERTNER English
1.0.0
442 views
231 times used

The Named Entity Recognition
cognise entities: (1) based on
module has been implemented
based ...

Keywords: Named Entity Recognition
Language: English
Licence: GNU Affero General Public Lic

BERTNER German
1.0.0

The Named Entity Recognition
cognise entities: (1) based on
module has been implemented
based ...

Keywords: Named Entity Recognition
Language: German
Licence: GNU Affero General Public Lic

Cogito Discover Named Entity Recognizer
ESI_NER
Version: 14.3.0
ELG-compatible service
ToolService

Overview Download/Run Try out Code samples

The G-20 Summit hosted by Italy kicked off Saturday in Rome, where leaders from the world's major economies discussed issues of mutual concern, including pandemic recovery and climate change. The red carpet was rolled out at "La Nuvola," Rome's convention center, as Italian Prime Minister Mario Draghi welcomed U.S. President Joe Biden and other leaders amid strict COVID-19 protocols. On day one, G-20 leaders voiced their support for a global corporate minimum tax deal agreed to by finance ministers from 136 countries earlier this month after four years of negotiations led by the Organization for Economic Cooperation and Development.

Annotations
☑ Organization
☑ People
☑ Place

BACK

Cogito Discover Named Entity Recognizer
ESI_NER
Version: 14.3.0
ELG-compatible service
ToolService

Overview Download/Run Try out Code samples

cURL

Python

If you want to call the service using Python, you can use the following code snippet:

```
from elg import Service
service = Service.from_id(474)
result = service(request_input="input string, ELG request object, or path to a file", request_type="text")
```

The ELG Python SDK can be installed using pip:

```
pip install elg
```

The full documentation of the Python SDK is available in our documentation.

Browse the catalogue

```
from elg import Catalog
catalog = Catalog()

# Search and get the result as a
list of Entity
results = catalog.search(
    resource = "Tool/Service",
    function = "Machine
Translation",
    languages = ["en", "fr"],
)
print(f"Machine Translation service
for English and
French:\n{results[0]}")
```

Call a service

```
from elg import Service

cogito_ner = Service.from_id(474)

result = cogito_ner("Nikolas Tesla
lives in Berlin.")
print(f"\n{result}")
```

```
elg run --id 474 --input_str
"Nikolas Tesla lives in Berlin."
```

Download a corpus

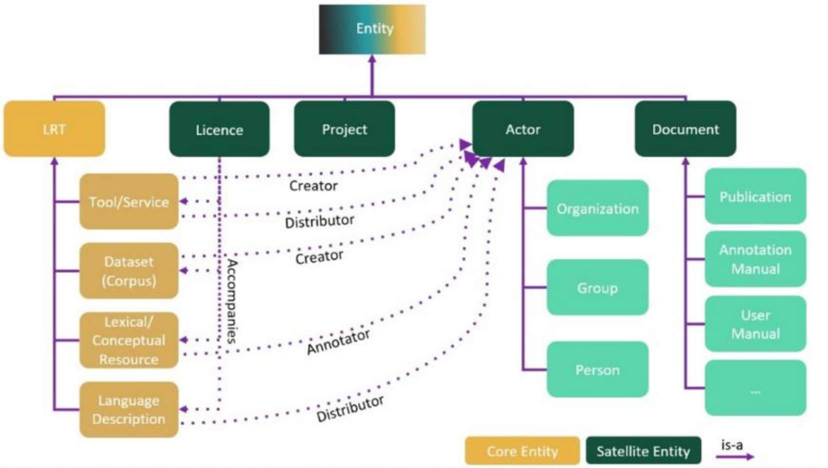
```
from elg import Corpus

conll_shared_task_corpus =
Corpus.from_id(913)

conll_shared_task_corpus.download(f
older="path/to/folder")
```

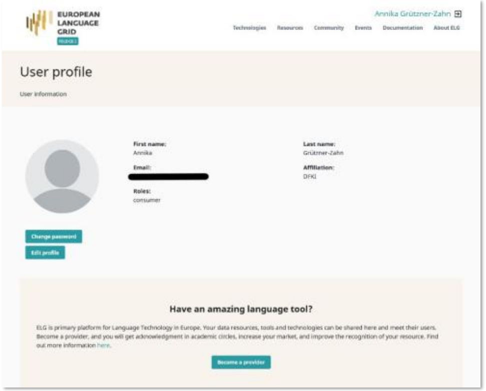
```
elg download --id 913 --folder
path/to/folder
```

- Consumers can **try out and test** language processing services
  - registration/authentication is required
  - daily quotas apply
- **Call a service** from the command line directly (via its common REST API) and integrate it in their own workflows
- Current APIs support
  - **machine translation**
  - **information extraction**
  - **text classification**
  - **speech recognition**
  - **speech synthesis**
- **Python-based API** for accessing the ELG catalogue, searching and directly fetching datasets to feed them into, e.g., their model training pipeline, running services, combining services in a pipeline, etc.

On the provider side

- LT community members
  - can **contribute** by providing formal descriptions compliant with a dedicated metadata schema
    - describe **organisations** (parents or divisions) & **projects**
    - describe & provide access (remote or integrated in ELG) **LRTs**
    - **linked with each other**
  - must register and get authenticated as providers





User profile

User information

First name:
Annika

Last name:
Grützner-Zahn

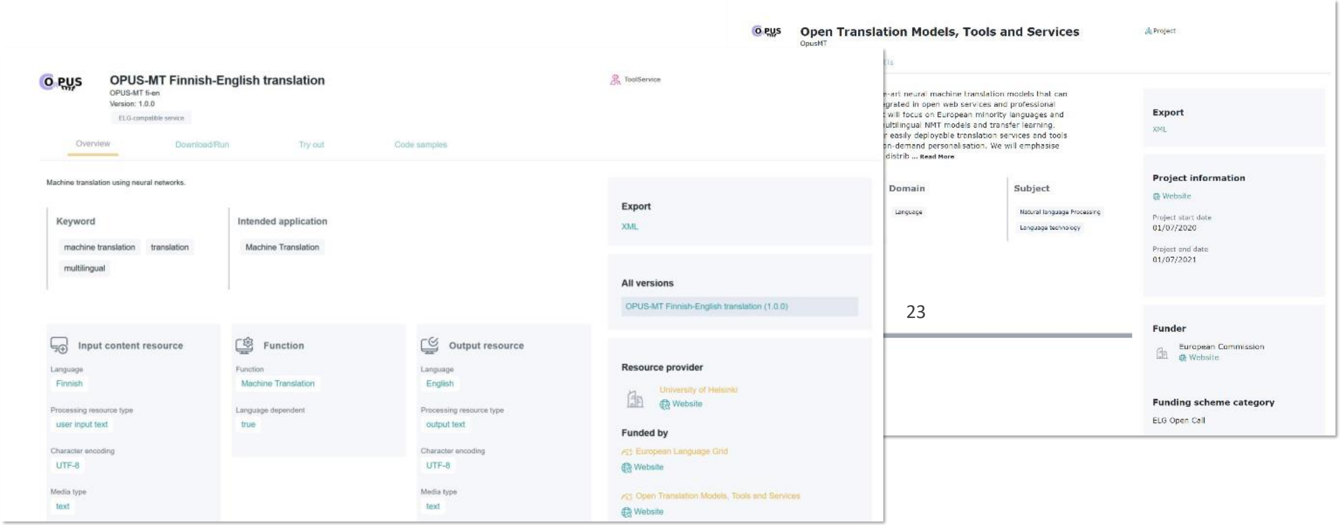
Email:
[redacted]

Affiliation:
DFKI

Roles:
consumer

Have an amazing language tool?

ELG is primary platform for Language Technology in Europe. Your data resources, tools and technologies can be shared here and meet their users. Become a provider, and you will get acknowledgment in academic circles, increase your market, and improve the recognition of your resource. Find out more information here.


Open Translation Models, Tools and Services
OpusMT

e-art neural machine translation models that can
egrated in open web services and professional
t will focus on European minority languages and
ultilingual NMT models and transfer learning.
r easily deployable translation services and tools
on-demand personalisation. We will emphasise
distrib ... Read More

Domain
Language

Subject
Natural language Processing
Language technology

Export
XML

Project information
Website
Project start date
01/07/2020
Project end date
01/07/2021

Funder
European Commission
Website

Funding scheme category
ELG Open Call

OPUS-MT Finnish-English translation
OPUS-MT fi-en
Version: 1.0.0
ELG-compatible service

Machine translation using neural networks.

Keyword
machine translation
translation
multilingual

Intended application
Machine Translation

Input content resource
Language
Finnish
Processing resource type
user input text
Character encoding
UTF-8
Media type
text

Function
Function
Machine Translation
Language dependent
true

Output resource
Language
English
Processing resource type
output text
Character encoding
UTF-8
Media type
text

Export
XML

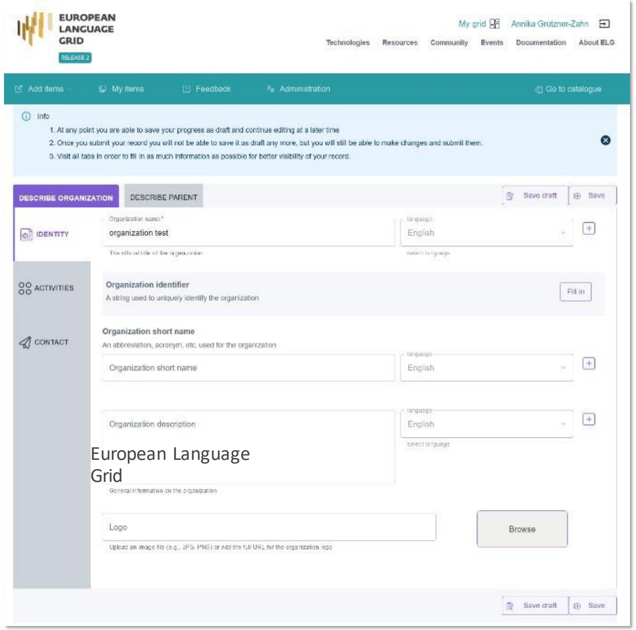
All versions
OPUS-MT Finnish-English translation (1.0.0)

Resource provider
University of Helsinki
Website

Funded by
European Language Grid
Website
Open Translation Models, Tools and Services
Website

European Language Grid: Provider’s grid

- Providers can create new items
  - by validating and uploading a schema compliant metadata record (single or batch)
  - using an **interactive metadata editor**



Info

1. At any point you are able to save your progress as draft and continue editing at a later time
2. Once you submit your record you will not be able to save it as draft any more, but you will still be able to make changes and submit them.
3. Visit all tabs in order to fill in as much information as possible for better visibility of your record.

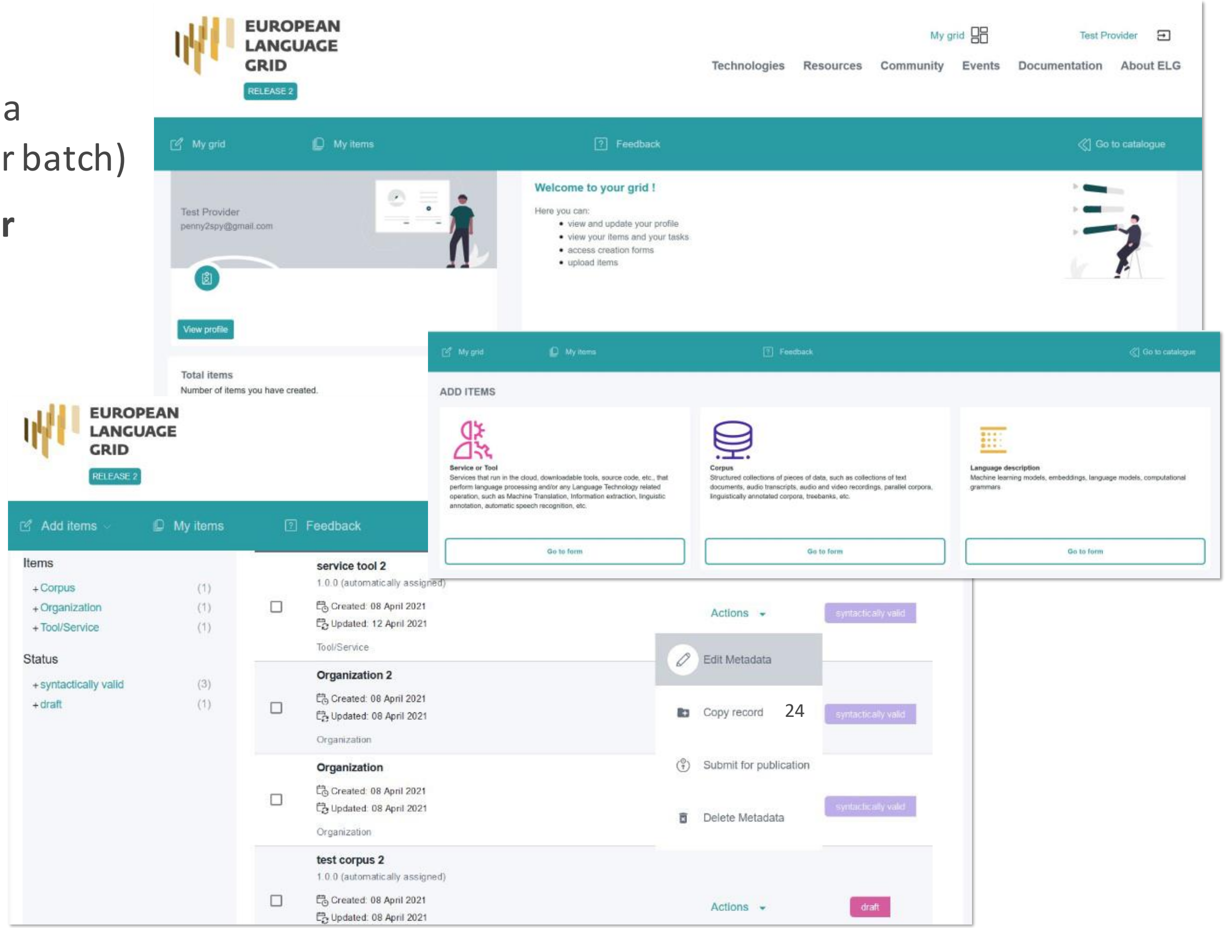
Organization name*: organization test — language: English
The official title of the organization — select language

Organization identifier
A string used to uniquely identify the organization

Organization short name
An abbreviation, acronym, etc. used for the organization
Organization short name — language: English

Organization description — language: English — select language
European Language Grid
General information on the organization

Logo
Upload an image file (e.g., JPG, PNG) or add the full URL for the organization logo



Test Provider
penny2spy@gmail.com

Welcome to your grid !
Here you can:
- view and update your profile
- view your items and your tasks
- access creation forms
- upload items

Total items
Number of items you have created.

ADD ITEMS

Service or Tool
Services that run in the cloud, downloadable tools, source code, etc., that perform language processing and/or any Language Technology related operation, such as Machine Translation, Information extraction, linguistic annotation, automatic speech recognition, etc.

Corpus
Structured collections of pieces of data, such as collections of text documents, audio transcripts, audio and video recordings, parallel corpora, linguistically annotated corpora, treebanks, etc.

Language description
Machine learning models, embeddings, language models, computational grammars

Items
+ Corpus (1)
+ Organization (1)
+ Tool/Service (1)

Status
+ syntactically valid (3)
+ draft (1)

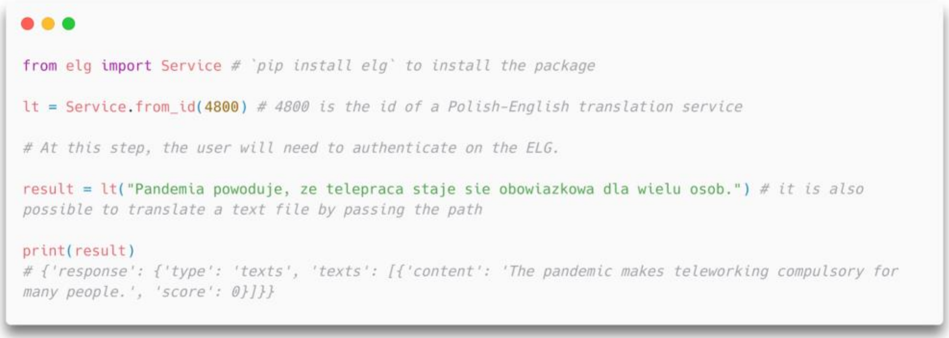
□ **service tool 2**
1.0.0 (automatically assigned)
Created: 08 April 2021
Updated: 12 April 2021
Tool/Service
syntactically valid

□ **Organization 2**
Created: 08 April 2021
Updated: 08 April 2021
Organization
syntactically valid

□ **Organization**
Created: 08 April 2021
Updated: 08 April 2021
Organization
syntactically valid

□ **test corpus 2**
1.0.0 (automatically assigned)
Created: 08 April 2021
Updated: 08 April 2021
draft

ELG Python SDK – Overview

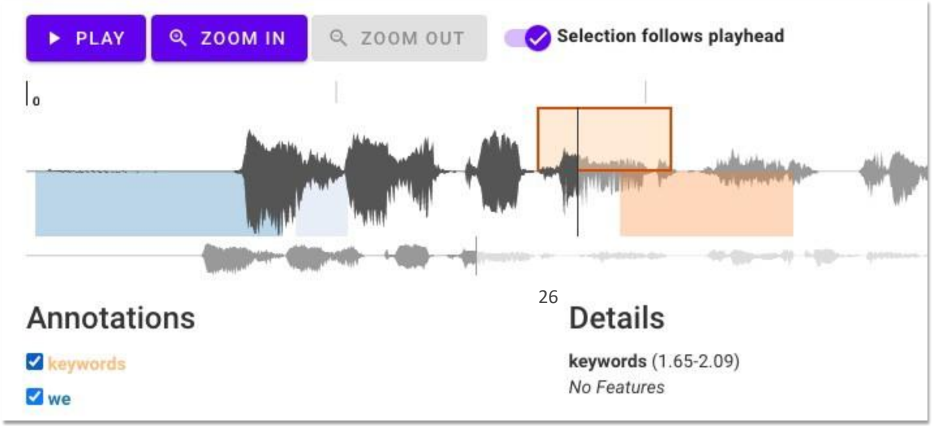


```
from elg import Service # `pip install elg` to install the package

lt = Service.from_id(4800) # 4800 is the id of a Polish-English translation service

# At this step, the user will need to authenticate on the ELG.

result = lt("Pandemia powoduje, ze telepraca staje sie obowiazkowa dla wielu osob.") # it is also
possible to translate a text file by passing the path

print(result)
# {'response': {'type': 'texts', 'texts': [{'content': 'The pandemic makes teleworking compulsory for
many people.', 'score': 0}]}}
```

- Brings main ELG functionalities to Python.
- Integrate ELG services into pipelines.
- Repository available on Gitlab: https://gitlab.com/european-language-grid/platform/python-client
- Currently available functionalities:
  - Search the ELG catalogue
  - Download ELG resources
  - Call ELG services
  - Compare multiple ELG services on various inputs (benchmarking)
- Coming up:
  - Contribute resources/services to ELG using Python through automated deployment
  - Enable running ELG services locally (by running Docker images locally)
  - Improve search experience
  - Many additional features …

LT Services: From the consortium

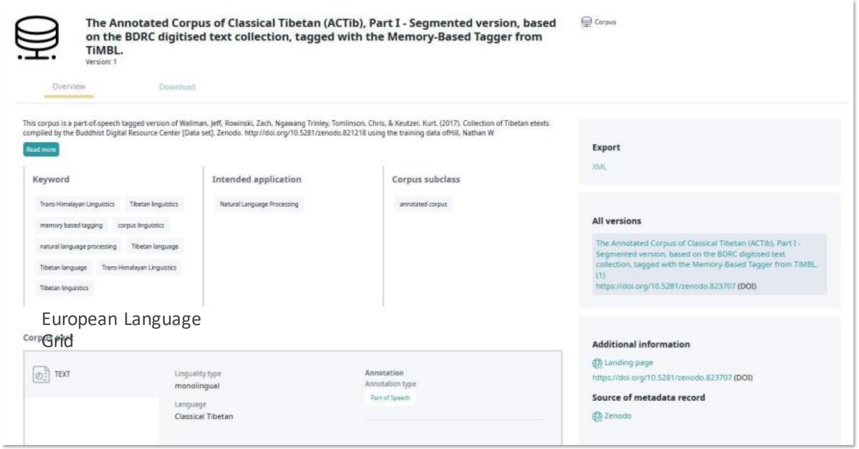
- ELG Release 1 in April 2020 finalized APIs for major classes of services (ASR, IE, MT, TTS)
- Concentrated on a subset of EU languages – Czech, English, French, German, Greek, Latvian, Spanish (native languages of the ELG consortium)
  - 9 ASR services, approx. 150 distinct services for IE & Text Analytics, 24 MT, 2 TTS
- ELG Release 2 (early 2021) added support for more EU and related languages
  - 8 further ASR services, approx. 300 IE & Text Analytics, 24 MT, 17 TTS
- … and APIs for additional service types
  - 12 services doing keyword spotting in audio
- Release 3 (2022) will add more languages beyond the EU (13 ASR, approx. 150 IE, 9 MT), and new service types including OCR
- More services from pilot projects and others



Language Resources: Metadata Description and Harvesting

ELG has imported metadata from many other repositories:

- ELRA Catalogue, ELRC-SHARE, ELRA-SHARE-LRs, LINDAT/CLARIAH-CZ, CLARIN PL repository, CLARIN Slovenia repository, META-SHARE-DFKI, META-SHARE-ELDA and META-SHARE-ILSP, Hugging Face, Quantum Stat, Zenodo
- Some metadata can be imported automatically, some need manual correction



The Annotated Corpus of Classical Tibetan (ACTib), Part I - Segmented version, based on the BDRC digitised text collection, tagged with the Memory-Based Tagger from TiMBL.

Version: 1

This corpus is a part-of-speech tagged version of Wallman, Jeff, Rowinski, Zach, Ngawang Trinley, Tomlinson, Chris, & Keutzer, Kurt. (2017). Collection of Tibetan etexts compiled by the Buddhist Digital Resource Center [Data set]. Zenodo. http://doi.org/10.5281/zenodo.821218 using the training data ofHill, Nathan W

Keyword

Trans-Himalayan Linguistics, Tibetan linguistics, memory based tagging, corpus linguistics, natural language processing, Tibetan language, Tibetan language, Trans-Himalayan Linguistics, Tibetan linguistics

Intended application

Natural Language Processing

Corpus subclass

annotated corpus

Corpus part

TEXT

Linguality type
monolingual

Language
Classical Tibetan

Annotation

Annotation type
Part of Speech

Export

XML

All versions

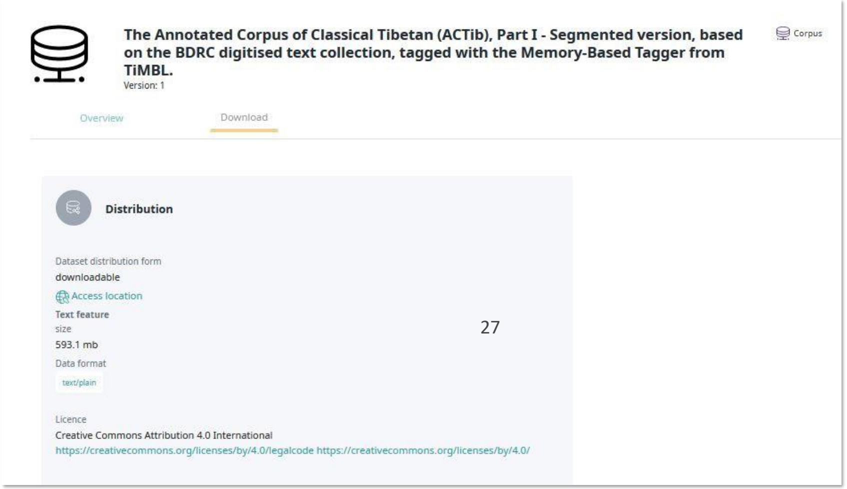
The Annotated Corpus of Classical Tibetan (ACTib), Part I - Segmented version, based on the BDRC digitised text collection, tagged with the Memory-Based Tagger from TiMBL. (1)
https://doi.org/10.5281/zenodo.823707 (DOI)

Additional information

Landing page
https://doi.org/10.5281/zenodo.823707 (DOI)

Source of metadata record

Zenodo



The Annotated Corpus of Classical Tibetan (ACTib), Part I - Segmented version, based on the BDRC digitised text collection, tagged with the Memory-Based Tagger from TiMBL.

Version: 1

Distribution

Dataset distribution form
downloadable

Access location

Text feature
size
593.1 mb

Data format
text/plain

Licence
Creative Commons Attribution 4.0 International
https://creativecommons.org/licenses/by/4.0/legalcode https://creativecommons.org/licenses/by/4.0/

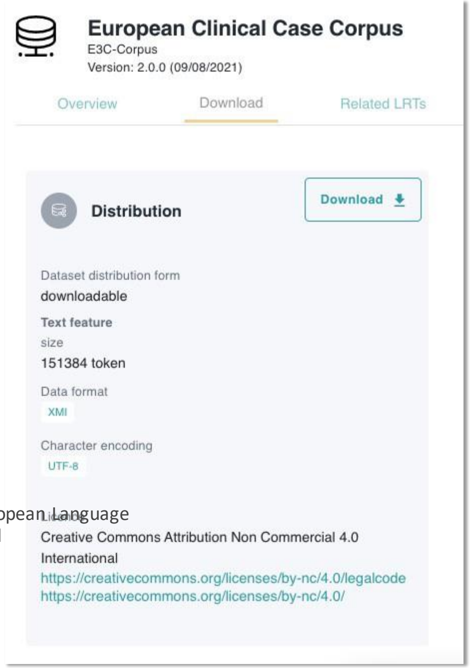
Language Resources in ELG (only those that are publicly visible)

Repository	Corpora	Lexical/Conceptual Resources	Models & Computational grammars	Total
ELRA	635	545	–	1180
ELRC-SHARE	1249	50	–	1299
META-SHARE	52	12	7	71
ELRA-SHARE-LRs	92	33	1	126
LINDAT/CLARIAH-CZ	274	79	–	353
CLARIN.SI	140	78	–	218
CLARIN.PL	239	15	–	254
Quantum Stat	255	5	–	260
Zenodo	137	99	24	*260
Tudatalib	1	–	–	1
HuggingFace	385	–	–	385
Others	72	28	5	105
TOTAL	3531	944	37	4512

*approx. 400 ingested resources are still under validation

Language Resources: Hosted Datasets

Data resources can be hosted in ELG for direct download



European Clinical Case Corpus
E3C-Corpus
Version: 2.0.0 (09/08/2021)

Distribution

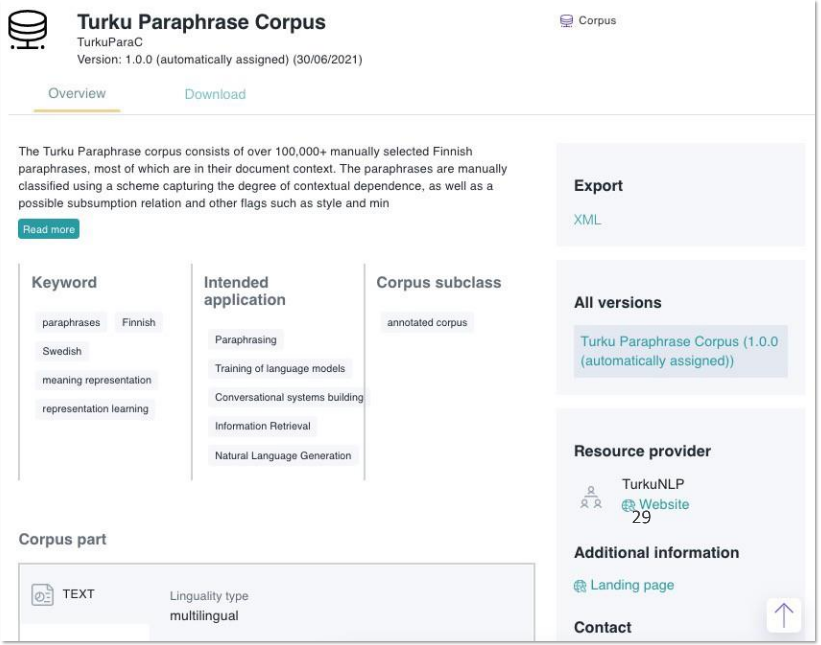
Dataset distribution form
downloadable

Text feature
size
151384 token

Data format
XMI

Character encoding
UTF-8

Licence
Creative Commons Attribution Non Commercial 4.0 International
https://creativecommons.org/licenses/by-nc/4.0/legalcode
https://creativecommons.org/licenses/by-nc/4.0/



Turku Paraphrase Corpus
TurkuParaC
Version: 1.0.0 (automatically assigned) (30/06/2021)

The Turku Paraphrase corpus consists of over 100,000+ manually selected Finnish paraphrases, most of which are in their document context. The paraphrases are manually classified using a scheme capturing the degree of contextual dependence, as well as a possible subsumption relation and other flags such as style and min

Keyword
paraphrases
Finnish
Swedish
meaning representation
representation learning

Intended application
Paraphrasing
Training of language models
Conversational systems building
Information Retrieval
Natural Language Generation

Corpus subclass
annotated corpus

Corpus part
TEXT
Linguality type
multilingual

Export
XML

All versions
Turku Paraphrase Corpus (1.0.0 (automatically assigned))

Resource provider
TurkuNLP
Website

Additional information
Landing page

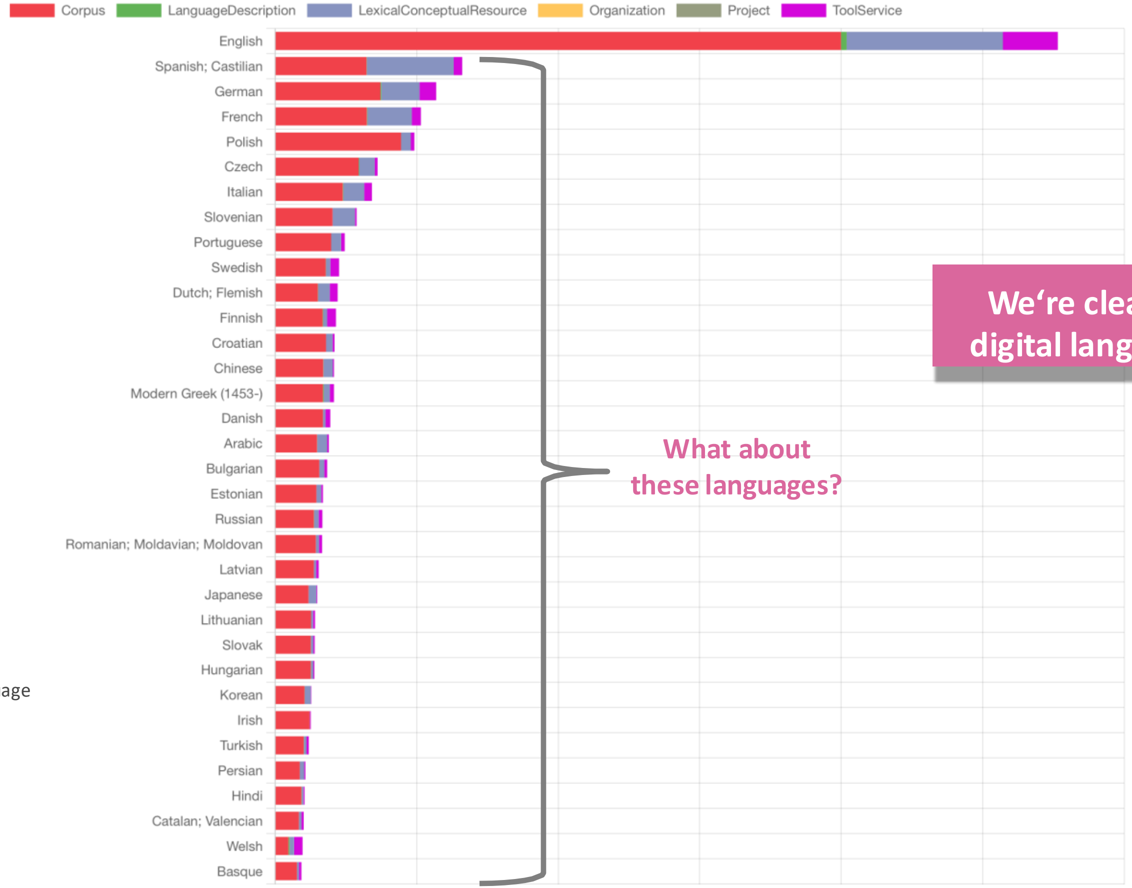
Contact

ELG in the wider LT and AI Ecosystem

- ELG is **building bridges** to exiscng plalorms and infrastructures
  - Mainly in terms of metadata-based descripcons of resources
  - Based on **open protocols** (OAI-PMH), or **APIs** offered by the plalorm or infrastructure providers
- ELG is also used for and within ELE
  - Using a mixture of automacc and collaboracve populacon of the ELG catalogue

European Language Resource Coordination
Connecting Europe Facility
ELRC-SHARE
EUROPEAN LANGUAGE RESOURCES ASSOCIATION
ELRA
META SHARE
AI4EU
EUROPEAN LANGUAGE GRID
LINDAT
CLARIAH-CZ
zenodo
CLARIN.SI
Hugging Face
CLARIN
EUROPEAN LANGUAGE EQUALITY

Languages of the Resources



What about these languages?

We‘re clearly very far away from digital language equality in Europe.

EUROPEAN
LANGUAGE
EQUALITY

Consortium: 52 partners from all over Europe

Coordinator: ADAPT Centre (Dublin City University)

Co-Coordinator: DFKI

Objective: *development of a strategic research, innovation and deployment agenda to achieve digital language equality in Europe by 2030*

EUROPEAN
LANGUAGE
GRID

Runtime: 18 months – ELE & ELG will both finish in June 2022

Start on **1 January 2021**

PP/PA (*not* Horizon 2020) – Pilot Project/Preparatory Action

Budget: 1,8M€ (Coordination and Support Action)

Nov. 2021: more than halfway through …

Current State of Play in the Project

EUROPEAN LANGUAGE EQUALITY

- **Digital Language Equality:** concept defined, dozens of *technological* and *contextual factors* longlisted, currently working on the selection of factors and the DLE formula
- **Surveys** conducted, now in the analysis phase – **Citizen survey** currently prepared for launch
- **Update of META-NET White Papers**, especially regarding the overview and qualitative analysis of technology support for Europe's languages – plus: *dynamic and fully empirical dashboard*
- **Collection of LRs/LTs:** with the whole ELE consortium we collected 6000+ additional records
- **ELG:** these 6000+ records will be included in ELG, arriving at approx. *11,000 resources* in total
- **Dynamic DLE computation:** based on these 11,000 metadata records we can compute and visualise the DLE metric for all European languages *through a dedicated dashboard in ELG*
- **Continuously:** this approach enables us to monitor the development of DLE in Europe over time
- **ELG:** Making a new resource available through ELG will increase the DLE metric for that language
- Almost all analyses to be ready in approx. April 2022 – ELE book to be published in mid 2022

Resource Types

		01/12/2020 01/12/2021
Corpora	1851	3626
Tools/Services	173	477
Language Descriptions	7	34
Lexical Conceptual Resources	728	964
Total	**2759**	**5101**
Additional resources collected by ELE	–	approx. 6000
Total (online soon)	**–**	**11101**

European Language Grid: An Overview

Making Metadata Fit for Next Generation Language Technology Platforms: The Metadata Schema of the European Language Grid

The European Language Technology Landscape in 2020: Language-Centric and Human-Centric AI for Cross-Cultural Communication in Multilingual Europe

European Language Grid: A Joint Platform for the European Language Technology Community

Towards an Interoperable Ecosystem of AI and LT Platforms: A Roadmap for the Implementation of Different Levels of Interoperability

LREC 2020 Workshop
Language Resources and Evaluation Conference
11–16 May 2020

IWLTP 2020 – 1st International Workshop on Language Technology Platforms

PROCEEDINGS

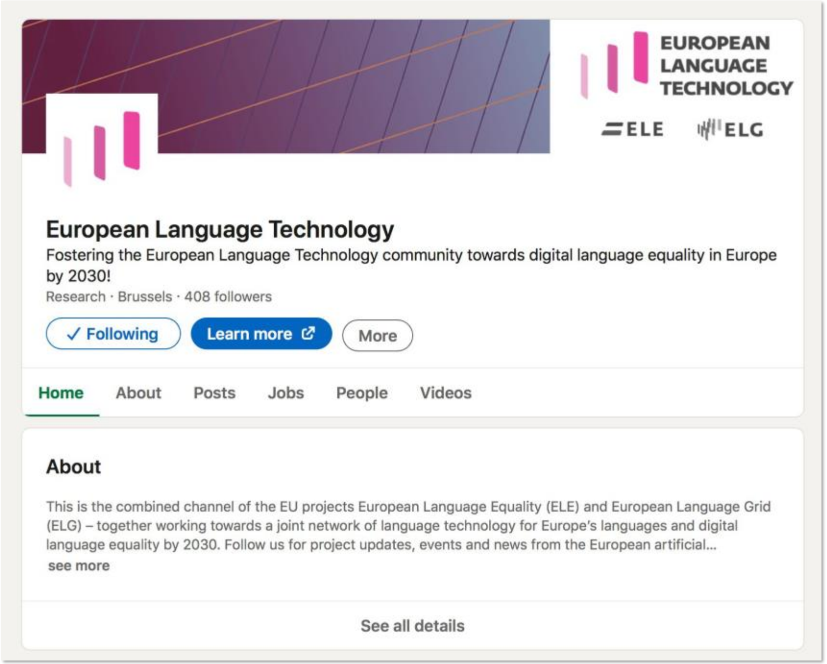
Edited by:
Georg Rehm, Kalina Bontcheva, Khalid Choukri,
Jan Hajicˇ, Stelios Piperidis, Andrejs Vasiļjevs

European Language Grid

Coming up:

- European Language Grid book (2022)
- European Language Equality book (2022)

Plus: ELG Online Documentation



EUROPEAN LANGUAGE TECHNOLOGY

ELE ELG

European Language Technology

Fostering the European Language Technology community towards digital language equality in Europe by 2030!

Research · Brussels · 408 followers

About

This is the combined channel of the EU projects European Language Equality (ELE) and European Language Grid (ELG) – together working towards a joint network of language technology for Europe's languages and digital language equality by 2030. Follow us for project updates, events and news from the European artificial...

see more

See all details

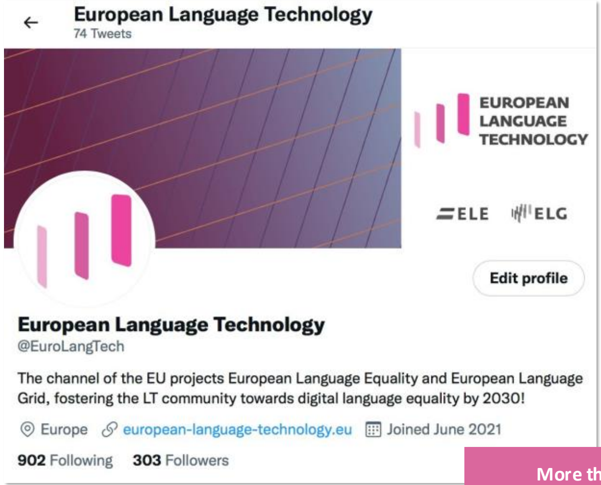
https://www.linkedin.com/company/european-language-technology

https://twitter.com/EuroLangTech

https://www.european-language-technology.eu

Subscribe to our newsletter

More than 2200 subscribers already!



European Language Technology

74 Tweets

EUROPEAN LANGUAGE TECHNOLOGY

ELE ELG

European Language Technology

@EuroLangTech

The channel of the EU projects European Language Equality and European Language Grid, fostering the LT community towards digital language equality by 2030!

Europe · european-language-technology.eu · Joined June 2021

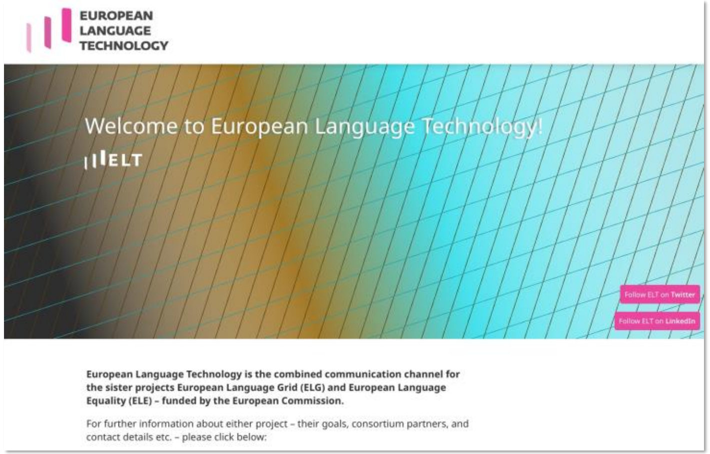
902 Following 303 Followers

More than 14k visitors on and more than 65k impressions already!

EUROPEAN LANGUAGE TECHNOLOGY

EUROPEAN LANGUAGE GRID

EUROPEAN LANGUAGE EQUALITY



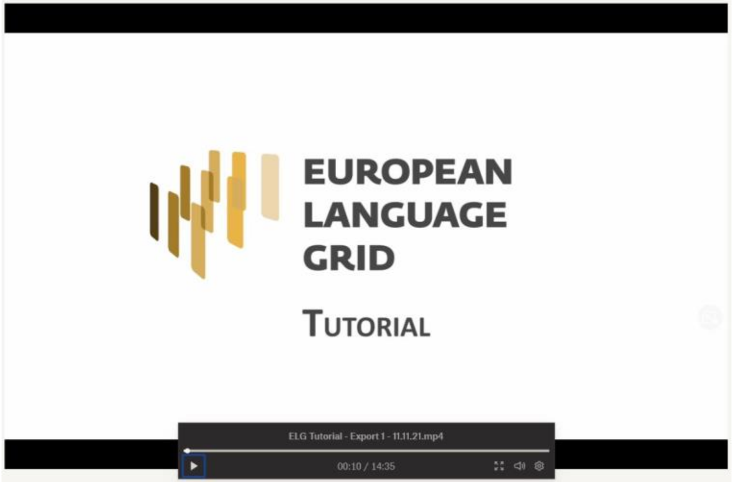
EUROPEAN LANGUAGE TECHNOLOGY

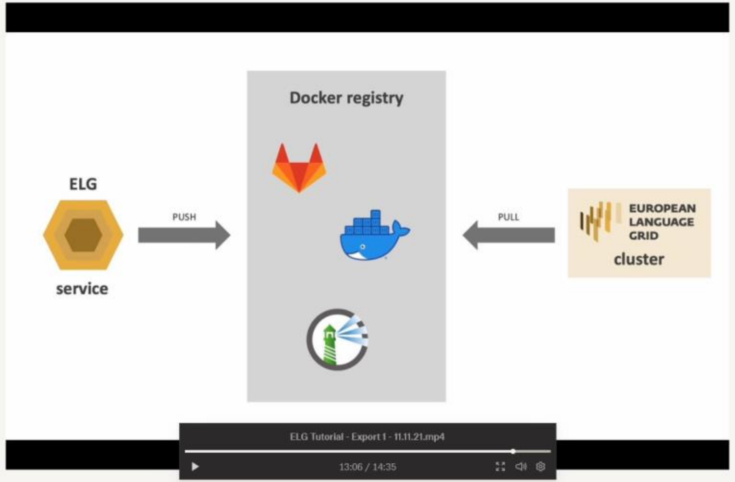
Welcome to European Language Technology!

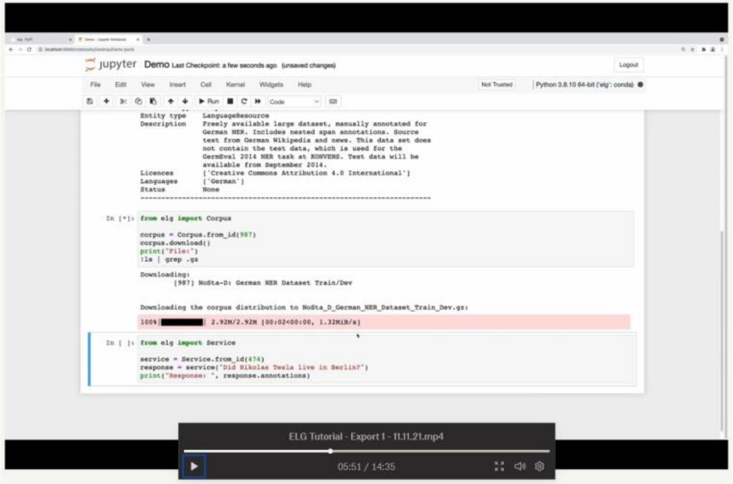
ELT

European Language Technology is the combined communication channel for the sister projects European Language Grid (ELG) and European Language Equality (ELE) – funded by the European Commission.

For further information about either project – their goals, consortium partners, and contact details etc. – please click below:





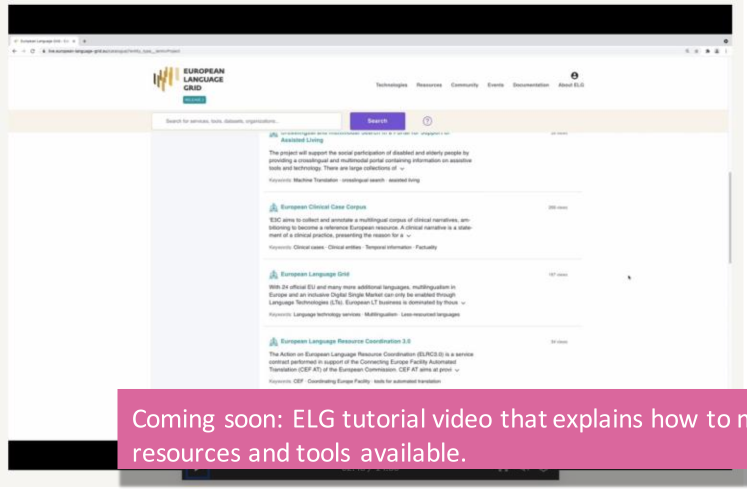

jupyter
Demo Last Checkpoint: a few seconds ago (unsaved changes)
Logout
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Not Trusted
Python 3.8.10 64-bit ('elg': conda)
Run
Code
ELG Tutorial - Export 1 - 11.11.21.mp4
05:51 / 14:35

```
Entity type    LanguageResource
Description    Freely available large dataset, manually annotated for
               German NER. Includes nested span annotations. Source
               text from German Wikipedia and news. This data set does
               not contain the test data, which is used for the
               GermEval 2014 NER task at KONVENS. Test data will be
               available from September 2014.
Licences       ['Creative Commons Attribution 4.0 International']
Languages      ['German']
Status         None
------------------------------------------------------------------------
```

```
In [*]: from elg import Corpus

        corpus = Corpus.from_id(987)
        corpus.download()
        print("File:")
        !ls | grep .gz
```

```
Downloading:
        [987] NoSta-D: German NER Dataset Train/Dev


Downloading the corpus distribution to NoSta_D_German_NER_Dataset_Train_Dev.gz:
100%|██████████| 2.92M/2.92M [00:02<00:00, 1.32MiB/s]
```

```
In [ ]: from elg import Service

        service = Service.from_id(474)
        response = service("Did Nikolas Tesla live in Berlin?")
        print("Response: ", response.annotations)
```



Assisted Living

The project will support the social participation of disabled and elderly people by providing a crosslingual and multimodal portal containing information on assistive tools and technology. There are large collections of

Keywords: Machine Translation · crosslingual search · assisted living

European Clinical Case Corpus

'E3C aims to collect and annotate a multilingual corpus of clinical narratives, ambitioning to become a reference European resource. A clinical narrative is a statement of a clinical practice, presenting the reason for a

Keywords: Clinical cases · Clinical entities · Temporal information · Factuality

European Language Grid

With 24 official EU and many more additional languages, multilingualism in Europe and an inclusive Digital Single Market can only be enabled through Language Technologies (LTs). European LT business is dominated by thous

Keywords: Language technology services · Multilingualism · Less-resourced languages

European Language Resource Coordination 3.0

The Action on European Language Resource Coordination (ELRC3.0) is a service contract performed in support of the Connecting Europe Facility Automated Translation (CEF AT) of the European Commission. CEF AT aims at provi

Keywords: CEF · Coordinating Europe Facility · tools for automated translation

Coming soon: ELG tutorial video that explains how to make resources and tools available.

Summary and Next Steps

- **Establish ELG as the primary platform and marketplace for Language Technology in Europe.**
- An initiative ***from*** the European LT community ***for*** the European LT community.
- European LT landscape is **highly fragmented**: ELG aims to provide just the right **umbrella platform**.
- Global market size by 2025 is enormous: we want the European LT community to be a **key player**.
- We want to **increase the visibility and reach of all members of the European LT landscape**.
- ELG is a long-term initiative: we will establish a **legal entity** for sustainability, which will operate and maintain the technology platform for the whole LT community as a joint marketplace.
- Contribute to **Digital Language Equality** in Europe by giving all our languages one virtual home and umbrella platform that collects **all** services and resources (**ELE**).
- **Next up:** attach more **data repositories**; include ELG in **various infrastructures** (e.g., NFDI, GAIA-X); **raise awareness** of LT for Europe; organise **META-FORUM 2022**; develop **ELG Release 3** (early 2022); establish the **ELG legal entity** (2022); contribute to the **Language Data Space** (2022/2023).

European Language Grid

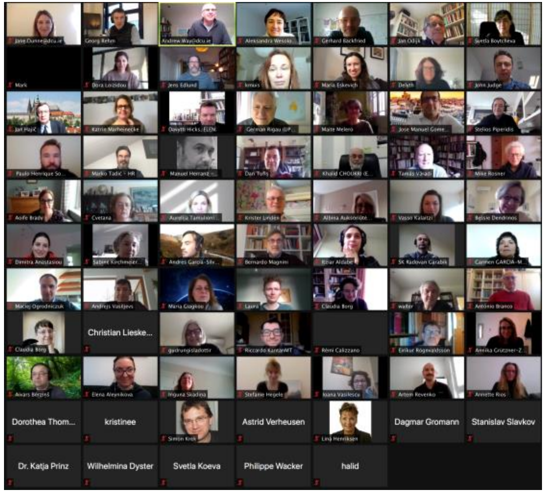
With many thanks to all colleagues in the ELG and ELE teams and consortia!

EUROPEAN LANGUAGE TECHNOLOGY

EUROPEAN LANGUAGE GRID

EUROPEAN LANGUAGE EQUALITY

Thank you!



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

Prof. Dr. Georg Rehm (DFKI GmbH, Germany)
Coordinator ELG, Co-coordinator ELE

03-12-2021 ELG NCC Workshop Netherlands
http://www.european-language-grid.eu
http://www.european-language-equality.eu

14.00 uur	Welkom	Vincent Vandeghinste (INT)
14.05 uur	Introducing the European Language Grid	Georg Rehm (DFKI)
14.35 uur	Demonstratie van de ELG, met focus op het Nederlands	Vincent Vandeghinste (INT)
15.00 uur	Vraag- en antwoordsessie en discussie	
15.10 uur	Pauze	
15.25 uur	Taalmaterialen voor het Nederlands	Bob Boelhouwer (INT)
15.35 uur	Presentaties van verschillende bedrijven (3) ● EDIA (Mark Breuker) ● Telecats (Arjan van Hessen) ● Y.digital (Ian Fitzpatrick)	Oele Koornwinder (NOTaS)
16.25 uur	Vraag- en antwoordsessie en discussie	

/instituut voor de
Nederlandse taal/

EUROPEAN
LANGUAGE
GRID

Demonstratie van de European Language Grid

met focus op (het) Nederland(s)

Vincent Vandeghinste

Waarom organiseert het Instituut voor de Nederlandse Taal deze workshop?

National Competence Centre Nederland

De European Language Grid heeft een sterk netwerk van 32 National Competence Centres (NCCs). De NCCs fungeren als lokale en nationale brug naar het ELG consortium en de European Language Grid.

Vincent Vandeghinste is NCC lead voor Nederland.

Als u verdere vragen heeft:

vincent.vandeghinste@ivdnt.org

INT en internationale infrastructuur

- CLARIN
  - Technisch centrum voor Nederland en Vlaanderen
    - tools en services
  - nationale coördinatie voor CLARIN-België
    - consortiumpartner in CLARIN-VL
    - consortiumpartner in CLARIAH-VL
  - K-Dutch: CLARIN kenniscentrum voor het Nederlands
- Consortium partner in European Language Equality project
- national anchor point voor ELRC voor Nederland

CLARIN

EUROPEAN
LANGUAGE
EQUALITY

ELG vs CLARIN

The ELG develops and deploys a **scalable cloud platform**, providing, in an **easy-to-integrate** way, access to hundreds of **commercial** and non-commercial **Language Technologies** for **all European languages**, including running tools and services as well as data sets and resources.

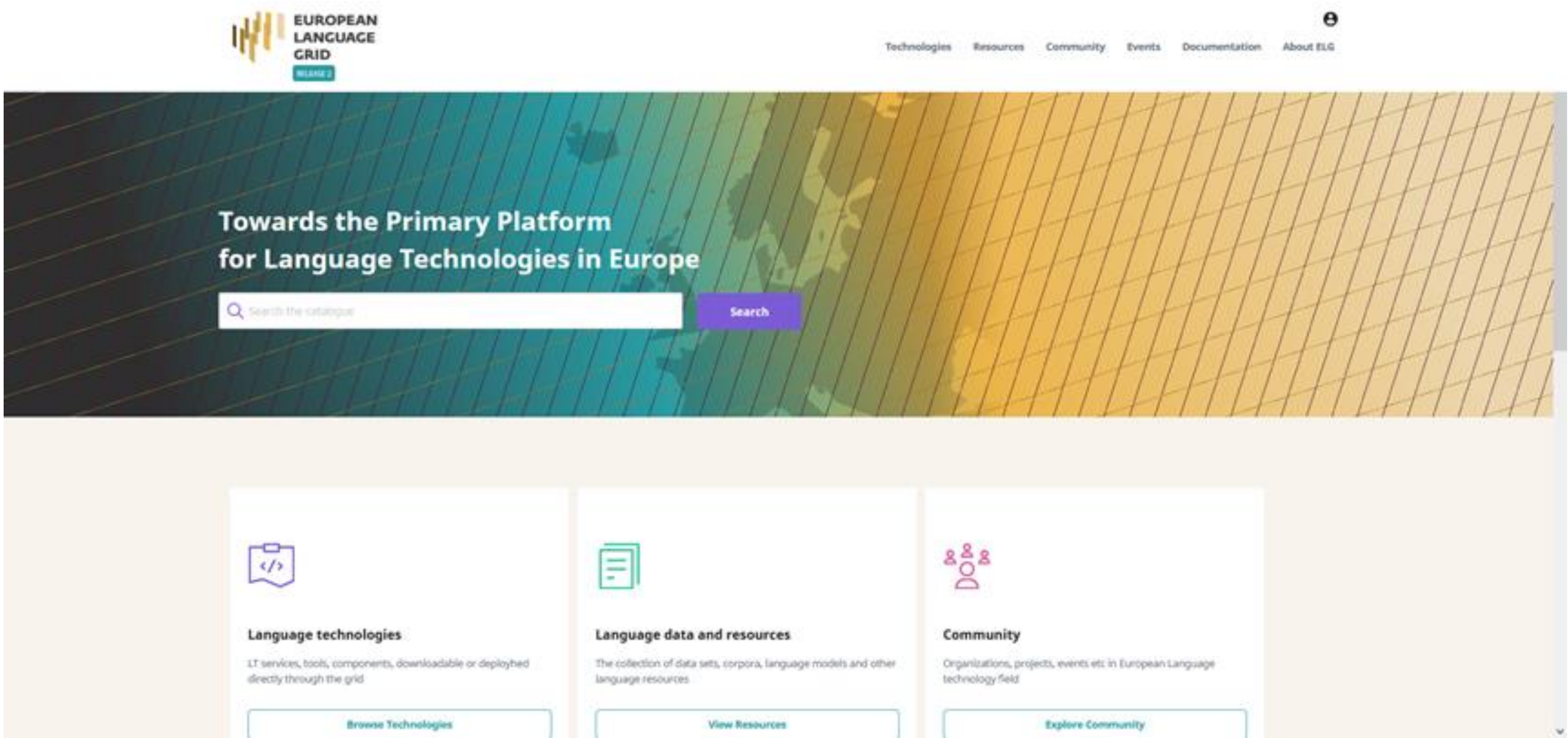
It is a research infrastructure that was initiated from the vision that **all digital language resources and tools** from all over Europe and beyond are accessible through a **single sign-on** online environment for the **support of researchers** in the humanities and social sciences.

ELG vs Nederlandse AI voor het Nederlands (NAIN)

NAIN is een use case van de Nederlandse AI Coalitie

Het doel van het project is **spraaktechnologie** beschikbaar te maken voor iedereen die Nederlands spreekt en daarvoor niet afhankelijk te zijn van de willekeur van grote buitenlandse commerciële partijen. De ambitie is de krachten te bundelen en **als Nederland zelf** een grote verbeterslag maken in spraaktechnologie, met name omdat het **verzamelen en transcriberen van relevant trainingsmateriaal** niet voor iedere individuele Nederlandse organisatie haalbaar is.

Wat is er al beschikbaar in de ELG voor het Nederlands



Towards the Primary Platform
for Language Technologies in Europe

Language technologies

LT services, tools, components, downloadable or deployhed directly through the grid

Language data and resources

The collection of data sets, corpora, language models and other language resources

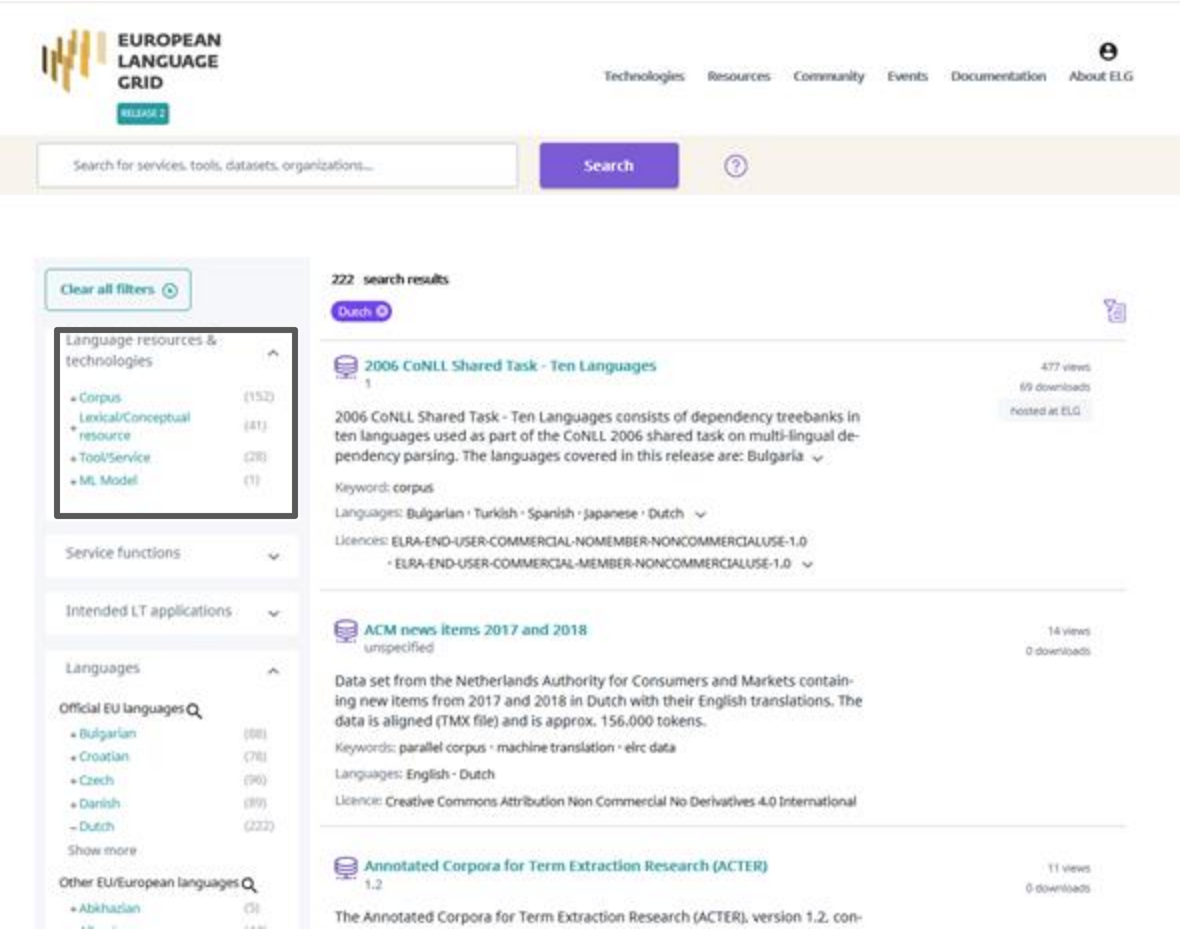
Community

Organizations, projects, events etc in European Language technology field

ELG voor het Nederlands

Taal: Nederlands: 222 zoekresultaten (26.11.2021)

- ○ corpora: 152
- ○ lexical resources: 41
- ○ tools / services: 28
- ○ machine learning model: 1



EUROPEAN LANGUAGE GRID

RELEASE 2

Technologies Resources Community Events Documentation About ELG

Search for services, tools, datasets, organizations...

Search

Clear all filters

Language resources & technologies

+ Corpus	(152)
+ Lexical/Conceptual resource	(41)
+ Tool/Service	(28)
+ ML Model	(1)

Service functions

Intended LT applications

Languages

Official EU languages

+ Bulgarian	(88)
+ Croatian	(78)
+ Czech	(96)
+ Danish	(89)
– Dutch	(222)

Show more

Other EU/European languages

+ Abkhazian	(3)

222 search results

Dutch

2006 CoNLL Shared Task - Ten Languages

1

477 views

69 downloads

hosted at ELG

2006 CoNLL Shared Task - Ten Languages consists of dependency treebanks in ten languages used as part of the CoNLL 2006 shared task on multi-lingual dependency parsing. The languages covered in this release are: Bulgaria

Keyword: corpus

Languages: Bulgarian · Turkish · Spanish · Japanese · Dutch

Licences: ELRA-END-USER-COMMERCIAL-NOMEMBER-NONCOMMERCIALUSE-1.0 · ELRA-END-USER-COMMERCIAL-MEMBER-NONCOMMERCIALUSE-1.0

ACM news items 2017 and 2018

unspecified

14 views

0 downloads

Data set from the Netherlands Authority for Consumers and Markets containing new items from 2017 and 2018 in Dutch with their English translations. The data is aligned (TMX file) and is approx. 156.000 tokens.

Keywords: parallel corpus · machine translation · elrc data

Languages: English · Dutch

Licence: Creative Commons Attribution Non Commercial No Derivatives 4.0 International

Annotated Corpora for Term Extraction Research (ACTER)

1.2

11 views

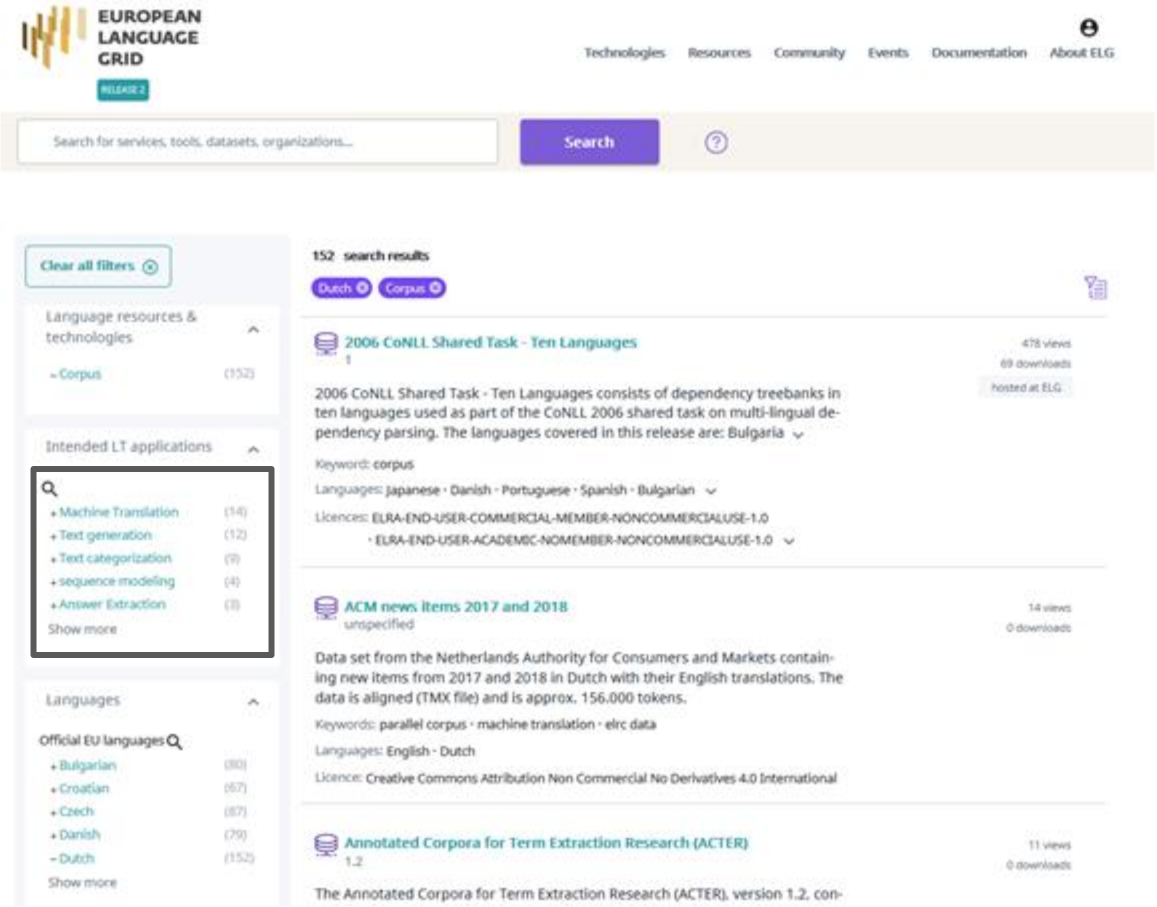
0 downloads

The Annotated Corpora for Term Extraction Research (ACTER), version 1.2, con-

Corpora

Intended applications:

- Machine Translation: 14
- Text generation: 12
- Text categorization: 9
- Sequence modeling: 4
- Answer extraction: 3
- ...



2006 CoNLL Shared Task - Ten Languages
1

478 views
69 downloads
hosted at ELG

2006 CoNLL Shared Task - Ten Languages consists of dependency treebanks in ten languages used as part of the CoNLL 2006 shared task on multi-lingual dependency parsing. The languages covered in this release are: Bulgaria

Keyword: corpus

Languages: Japanese · Danish · Portuguese · Spanish · Bulgarian

Licences: ELRA-END-USER-COMMERCIAL-MEMBER-NONCOMMERCIALUSE-1.0 · ELRA-END-USER-ACADEMIC-NOMEMBER-NONCOMMERCIALUSE-1.0

ACM news items 2017 and 2018
unspecified

14 views
0 downloads

Data set from the Netherlands Authority for Consumers and Markets containing new items from 2017 and 2018 in Dutch with their English translations. The data is aligned (TMX file) and is approx. 156.000 tokens.

Keywords: parallel corpus · machine translation · elrc data

Languages: English · Dutch

Licence: Creative Commons Attribution Non Commercial No Derivatives 4.0 International

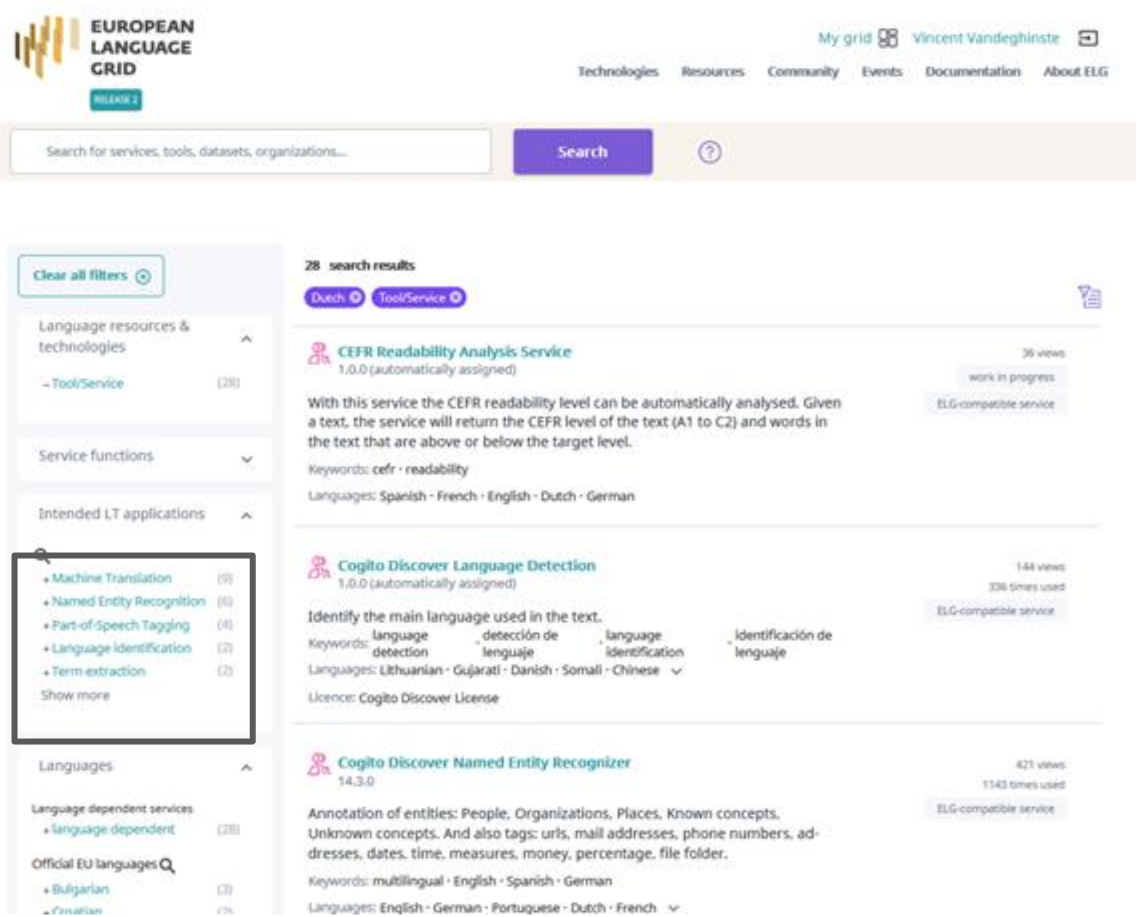
Annotated Corpora for Term Extraction Research (ACTER)
1.2

11 views
0 downloads

The Annotated Corpora for Term Extraction Research (ACTER), version 1.2, con-

Tools voor het Nederlands

- Automatische vertaling: 9
- Named Entity Recognition: 6
- POS tagging: 4
- Taalidentificatie: 2
- Termextractie: 2
- ...



28 search results

CEFR Readability Analysis Service
1.0.0 (automatically assigned)

36 views
work in progress
ELG-compatible service

With this service the CEFR readability level can be automatically analysed. Given a text, the service will return the CEFR level of the text (A1 to C2) and words in the text that are above or below the target level.

Keywords: cefr · readability

Languages: Spanish · French · English · Dutch · German

Cogito Discover Language Detection
1.0.0 (automatically assigned)

144 views
336 times used
ELG-compatible service

Identify the main language used in the text.

Keywords: language detection · detección de lenguaje · language identification · identificación de lenguaje

Languages: Lithuanian · Gujarati · Danish · Somali · Chinese

Licence: Cogito Discover License

Cogito Discover Named Entity Recognizer
14.3.0

421 views
1143 times used
ELG-compatible service

Annotation of entities: People, Organizations, Places, Known concepts, Unknown concepts. And also tags: urls, mail addresses, phone numbers, addresses, dates, time, measures, money, percentage, file folder.

Keywords: multilingual · English · Spanish · German

Languages: English · German · Portuguese · Dutch · French

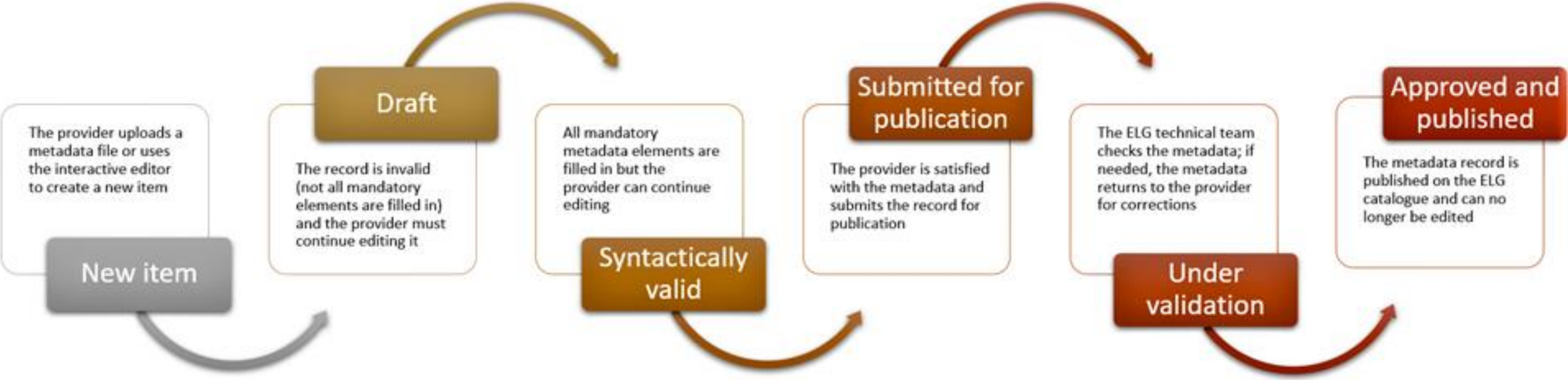
Tools voor het Nederlands -- uitproberen

- Taalidentificatie https://live.european-language-grid.eu/catalogue/tool-service/473
- Parsing https://live.european-language-grid.eu/catalogue/tool-service/451
- Spraakherkenning: https://live.european-language-grid.eu/catalogue/tool-service/8154
- Keyword spotting: https://live.european-language-grid.eu/catalogue/tool-service/8166
- Summarization / keyword extraction: https://live.european-language-grid.eu/catalogue/tool-service/478
- Named Entity Recognition: https://live.european-language-grid.eu/catalogue/tool-service/8134

Hoe kan je zelf tools en data beschikbaar maken?

Registreer als provider: https://european-language-grid.readthedocs.io/en/stable/all/3_Contributing/Contributing.html

Publicatiecyclus:



New item

The provider uploads a metadata file or uses the interactive editor to create a new item

Draft

The record is invalid (not all mandatory elements are filled in) and the provider must continue editing it

Syntactically valid

All mandatory metadata elements are filled in but the provider can continue editing

Submitted for publication

The provider is satisfied with the metadata and submits the record for publication

Under validation

The ELG technical team checks the metadata; if needed, the metadata returns to the provider for corrections

Approved and published

The metadata record is published on the ELG catalogue and can no longer be edited

Een corpus / dataset bijdragen

https://european-language-grid.readthedocs.io/en/stable/all/3_Contributing/Corpus.html

- Gehost door ELG: maak er een zip file van
- Beschrijf het corpus (metadata) volgens metadata schema
- Registreer het corpus bij ELG
  - uploaden metadata
  - interactieve editor
- Dien in voor publicatie

Een ELG-compatibele service bijdragen

https://european-language-grid.readthedocs.io/en/stable/all/3_Contributing/Service.html

Momenteel ondersteunt ELG de integratie van tools/diensten in deze categorieën:

- **Informatie-extractie** (IE) : Diensten die tekst annoteren met metadata over specifieke segmenten, zoals **Named Entity Recognition** (NER), de taak om personen, locaties en organisaties uit een bepaalde tekst te extraheren.
- **Tekstclassificatie** (TC) : Diensten die tekst classificeren volgens een eindige reeks klassen, b.v. Tekstcategorisatie, de taak om tekst te categoriseren in (meestal gelabelde) georganiseerde categorieën.
- **Machinevertaling** (MT) : Diensten die tekst vertalen naar een andere taal, mogelijk met aanvullende metadata die aan elk segment zijn gekoppeld (zin, woordgroep, enz.).
- **Automatische spraakherkenning** (ASR): services die audio als invoer gebruiken en tekst (bijvoorbeeld een transcriptie) als uitvoer produceren, mogelijk met metagegevens die aan elk segment zijn gekoppeld.
- **Text-to-Speech** synthese (TTS) : Diensten die tekst omzetten naar audio

Een ELG-compatibele service bijdragen

1. Dockerize je service

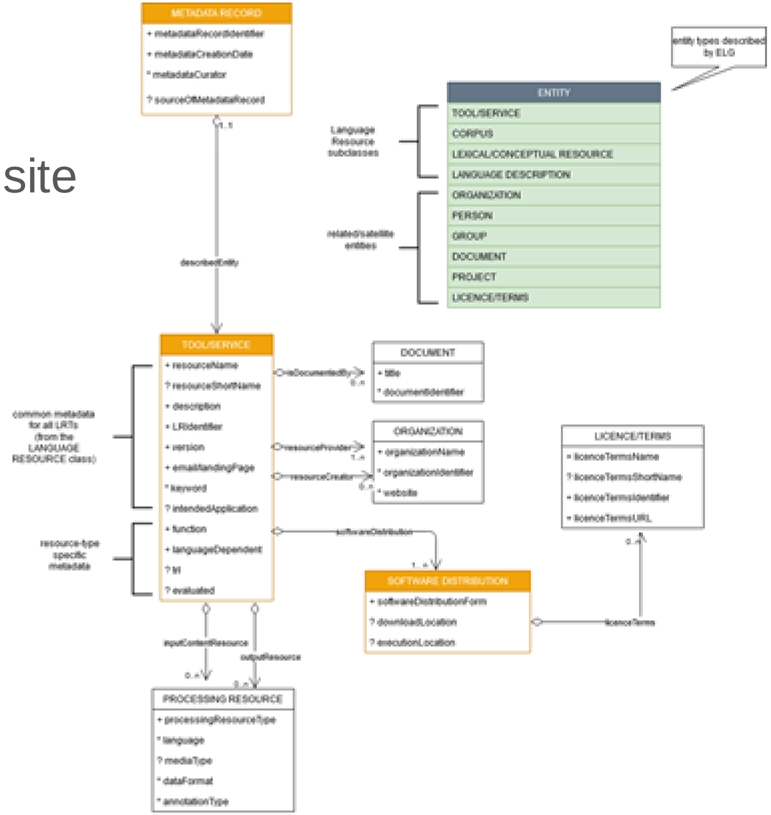
   voorbeelden zijn beschikbaar op de website

1. Beschrijf de service

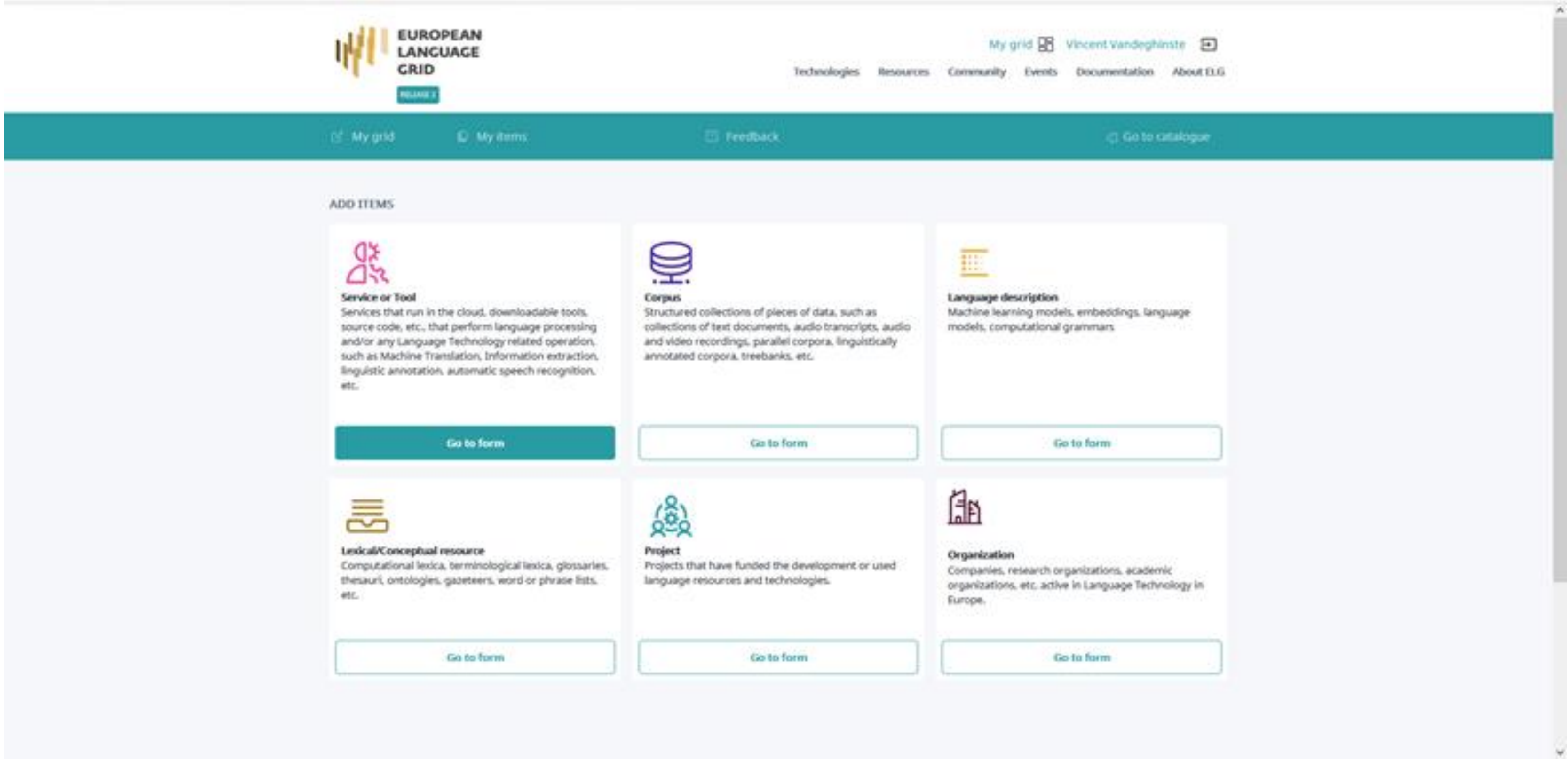
   volgens metadata schema

1. Registreer de service bij ELG
   a. via interactieve ELG editor
   b. het uploaden van een metadatabestand

2. Dien de service in bij ELG voor publicatie



Demonstratie van de *Interactive Editor*



ADD ITEMS

Service or Tool
Services that run in the cloud, downloadable tools, source code, etc., that perform language processing and/or any Language Technology related operation, such as Machine Translation, Information extraction, linguistic annotation, automatic speech recognition, etc.

Go to form

Corpus
Structured collections of pieces of data, such as collections of text documents, audio transcripts, audio and video recordings, parallel corpora, linguistically annotated corpora, treebanks, etc.

Go to form

Language description
Machine learning models, embeddings, language models, computational grammars

Go to form

Lexical/Conceptual resource
Computational lexica, terminological lexica, glossaries, thesauri, ontologies, gazeteers, word or phrase lists, etc.

Go to form

Project
Projects that have funded the development or used language resources and technologies.

Go to form

Organization
Companies, research organizations, academic organizations, etc. active in Language Technology in Europe.

Go to form

Conclusie

- Als je tools of datasets zoekt ivm taal- of spraaktechnologie, kijk dan eens in de ELG
- Ook voor het Nederlands is al heel wat beschikbaar
- Als je zelf tools of datasets ontwikkelt of ontwikkeld hebt, kan je die publiek bekendmaken, al dan niet met demo
- Contacteer je NCC-lead vincent.vandeghinste@ivdnt.org als iets niet helemaal duidelijk is

15u00 Vragen & Antwoorden

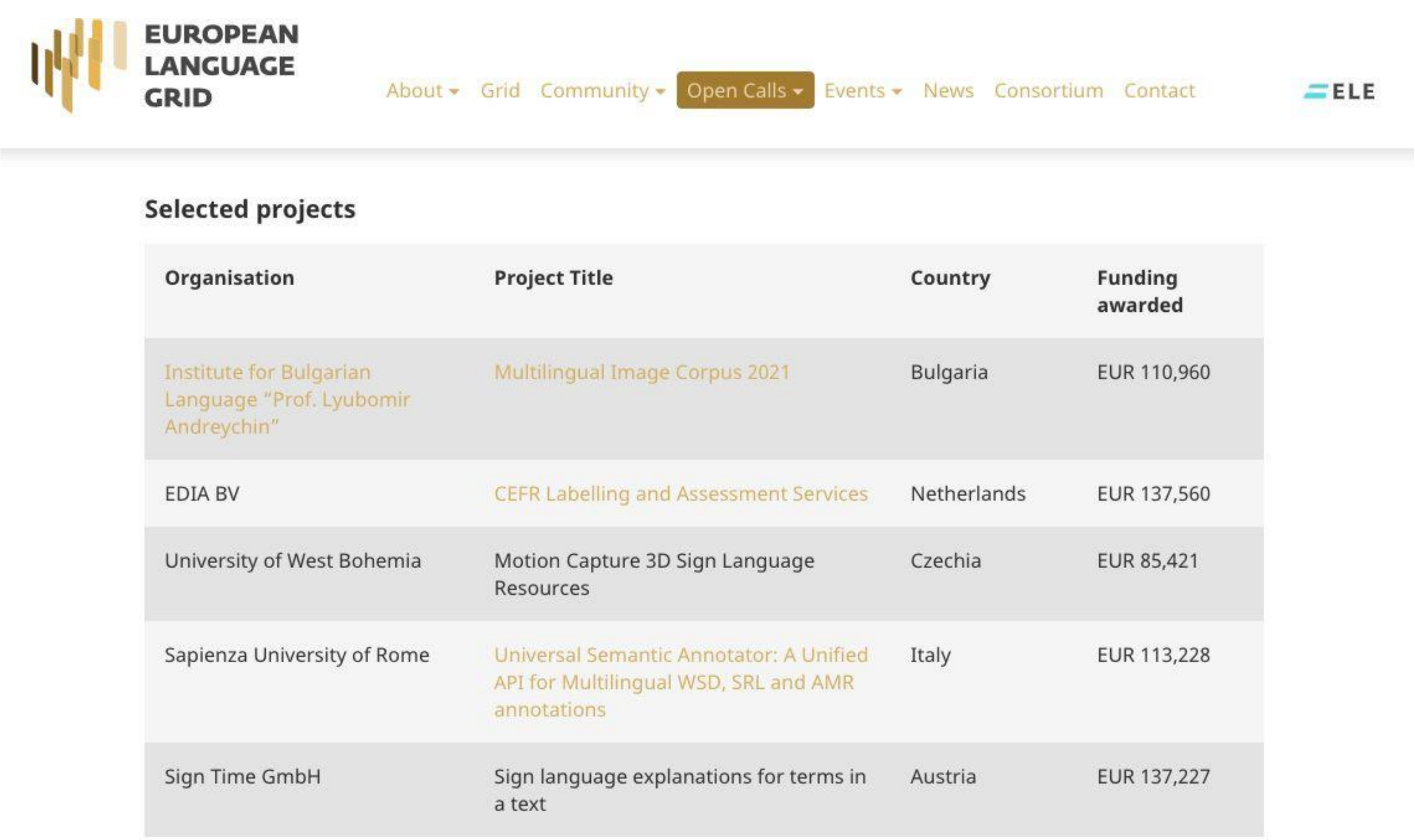
15u10 Pauze

14.00 uur	Welkom	Vincent Vandeghinste (INT)
14.05 uur	Introducing the European Language Grid	Georg Rehm (DFKI)
14.35 uur	Demonstratie van de ELG, met focus op het Nederlands	Vincent Vandeghinste (INT)
15.00 uur	Vraag- en antwoordsessie en discussie	
15.10 uur	Pauze	
15.25 uur	Taalmaterialen voor het Nederlands	Bob Boelhouwer (INT)
15.35 uur	Presentaties van verschillende bedrijven (3) ● EDIA (Mark Breuker) ● Telecats (Arjan van Hessen) ● Y.digital (Ian Fitzpatrick)	Oele Koornwinder (NOTaS)
16.25 uur	Vraag- en antwoordsessie en discussie	

CEFR Readability Labelling and Assessment Services

European Language Grid pilot project

Presenter: Mark Breuker (EDIA)



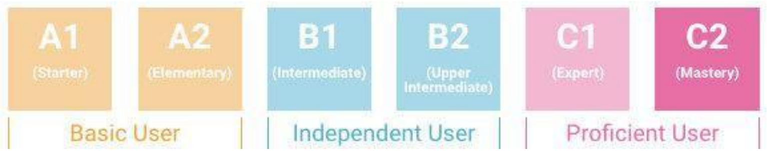
Selected projects

Organisation	Project Title	Country	Funding awarded
Institute for Bulgarian Language "Prof. Lyubomir Andreychin"	Multilingual Image Corpus 2021	Bulgaria	EUR 110,960
EDIA BV	CEFR Labelling and Assessment Services	Netherlands	EUR 137,560
University of West Bohemia	Motion Capture 3D Sign Language Resources	Czechia	EUR 85,421
Sapienza University of Rome	Universal Semantic Annotator: A Unified API for Multilingual WSD, SRL and AMR annotations	Italy	EUR 113,228
Sign Time GmbH	Sign language explanations for terms in a text	Austria	EUR 137,227

https://www.european-language-grid.eu/open-calls/open-call-2/

Common European Framework of Reference (CEFR)

The CEFR describes foreign language proficiency at six levels:



- to establish learning and teaching objectives
- to review curricula
- to design teaching materials
- to provide a basis for recognising language qualifications thus facilitating educational and occupational mobility.

https://www.coe.int/en/web/common-european-framework-reference-languages/home

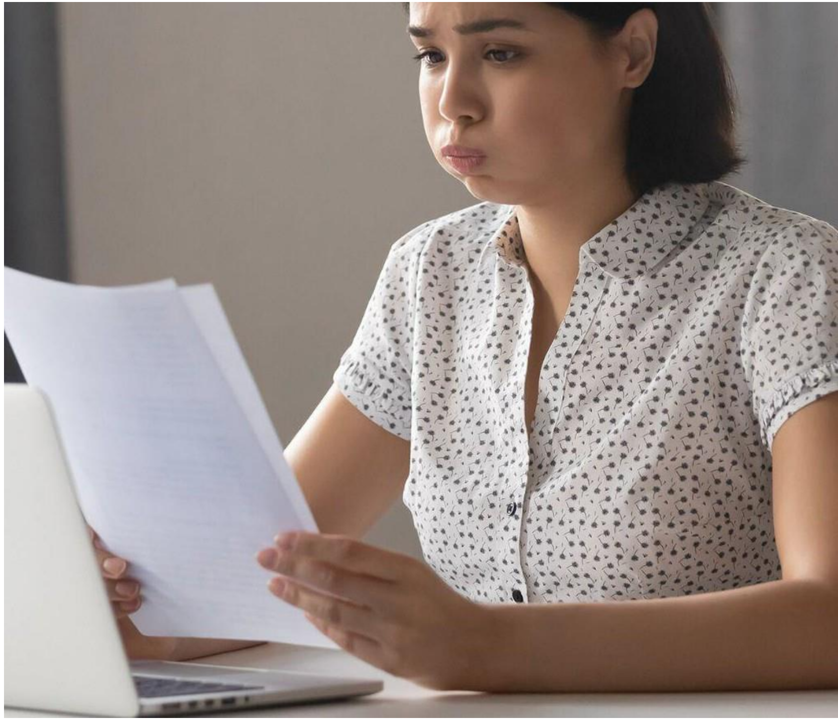
Project objectives

Problem

Manual assessment of reading materials on CEFR readability levels depends on human experts: inconsistent and not scalable.

Our solution

Develop a set of tools, datasets and services (infrastructure) to enable automatic classification of CEFR reading difficulty



EDIA

Applications

Personalised learning

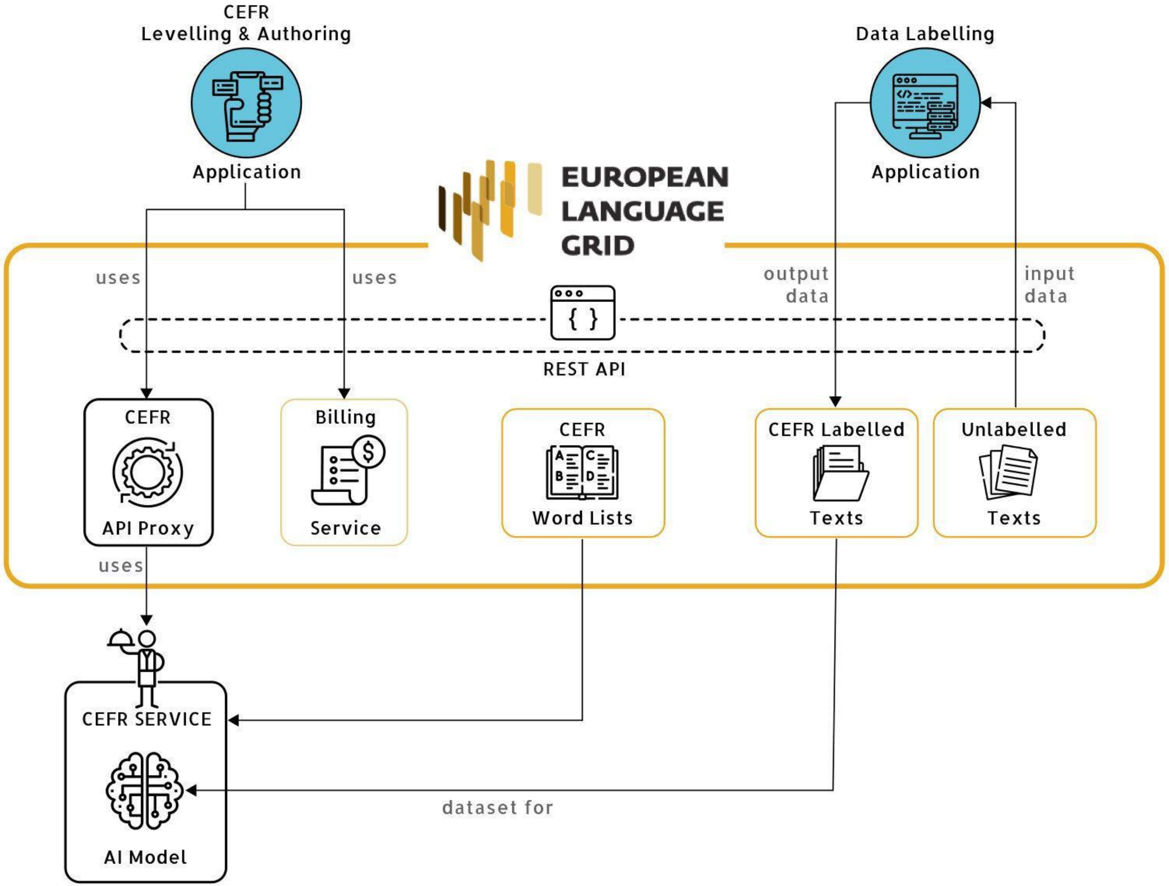
Teachers want to use authentic and current content in class to boost student engagement.

Automatic CEFR classification supports easy discovery and assessment of texts that match individual student's reading proficiency.

Inclusiveness in society

News and information from governmental organisations is often too difficult to read for people with low literacy skills

Automatic CEFR classification helps to identify difficult texts and make them more readable.



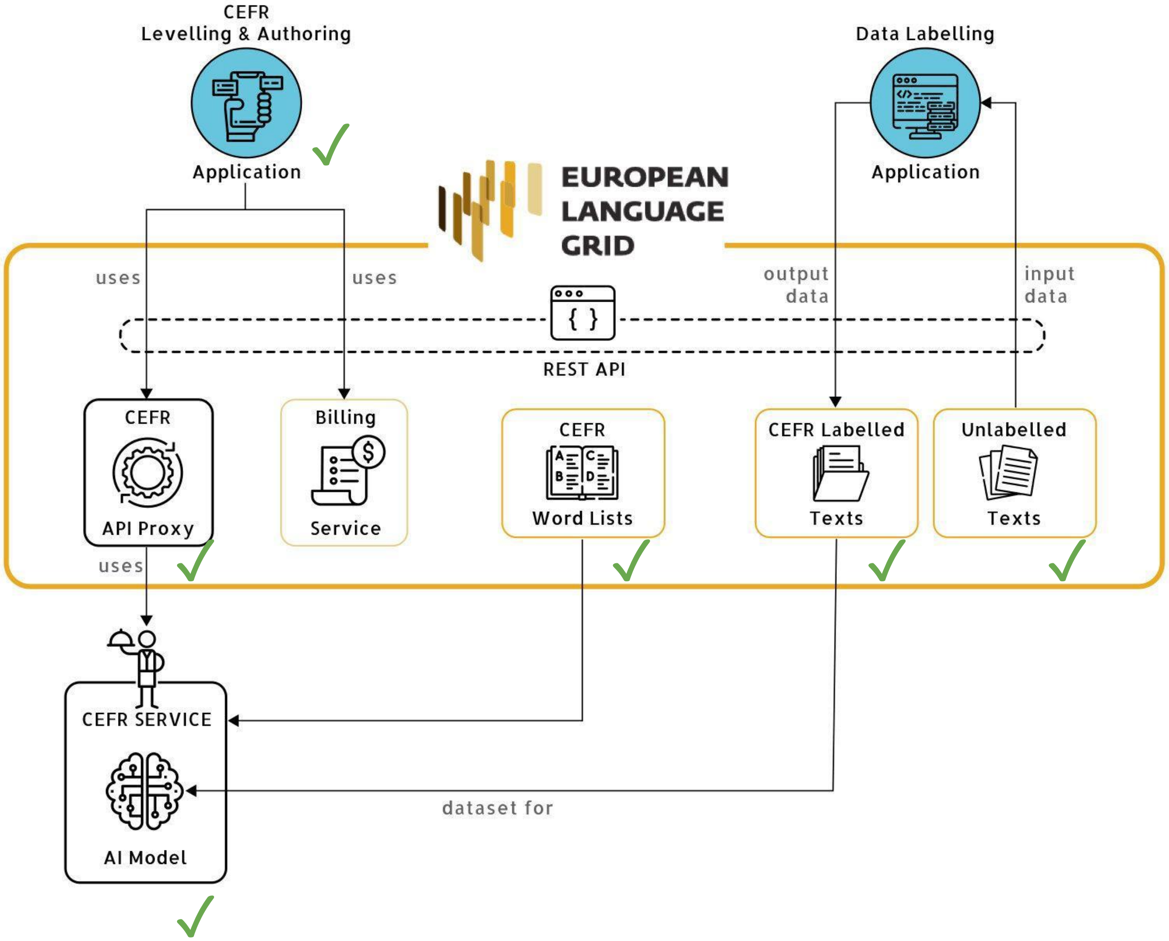
Approach

1. Collect unlabelled texts
2. Label texts on CEFR level using human experts (teachers)
3. Train AI model on dataset
4. Expose model as service (REST API) on ELG
5. Integrate API in end-user applications

Languages

- English, German (available)
- Dutch, Spanish (in development)

EDIA

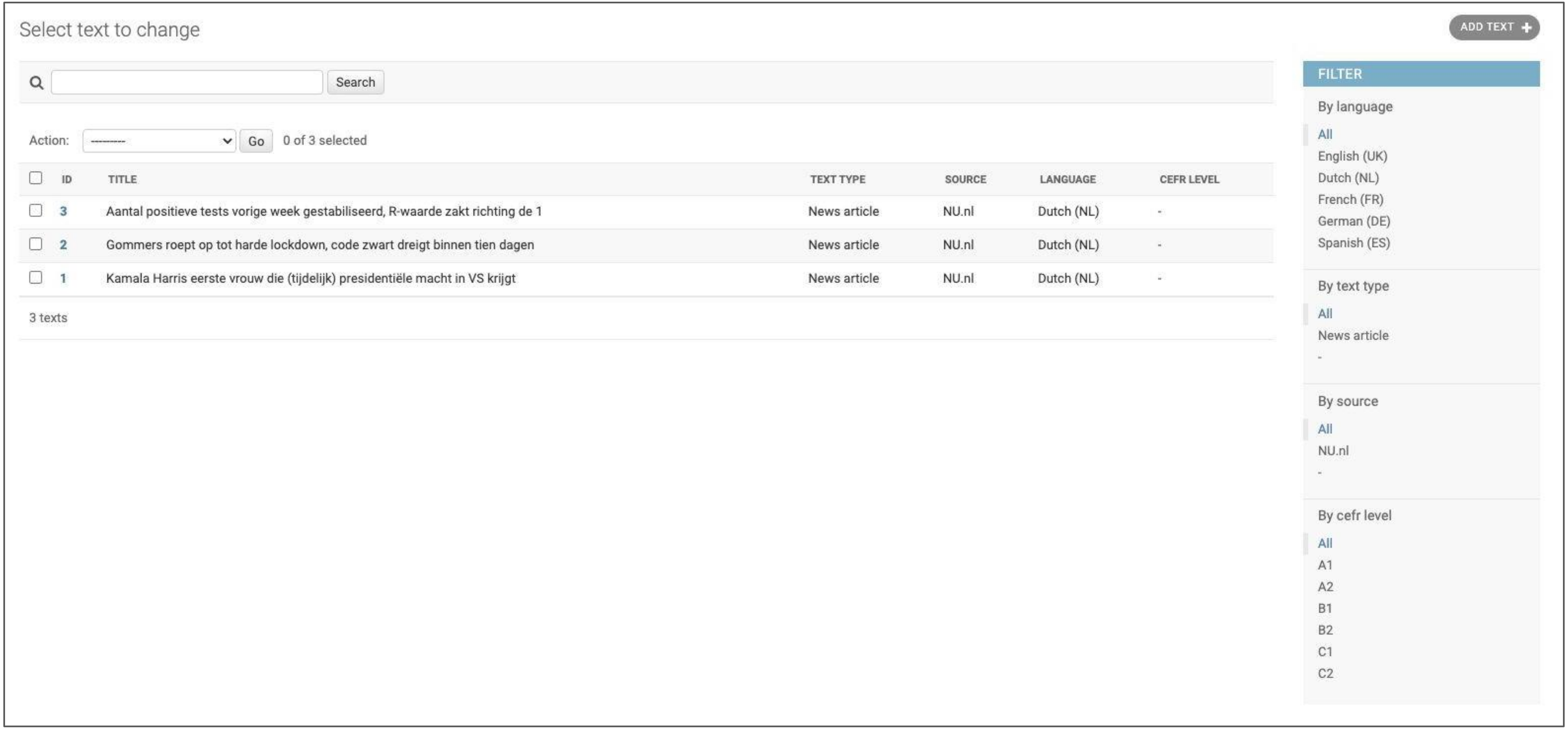


Approach

1. Collect unlabelled texts
2. Label texts on CEFR level using human experts (teachers)
3. Train AI model on dataset
4. Expose model as service (REST API) on ELG
5. Integrate API in end-user applications

Languages

- English, German (available)
- Dutch, Spanish (in development)



Select text to change

ADD TEXT +

Search

Action: --------- Go 0 of 3 selected

□	ID	TITLE	TEXT TYPE	SOURCE	LANGUAGE	CEFR LEVEL
□	3	Aantal positieve tests vorige week gestabiliseerd, R-waarde zakt richting de 1	News article	NU.nl	Dutch (NL)	-
□	2	Gommers roept op tot harde lockdown, code zwart dreigt binnen tien dagen	News article	NU.nl	Dutch (NL)	-
□	1	Kamala Harris eerste vrouw die (tijdelijk) presidentiële macht in VS krijgt	News article	NU.nl	Dutch (NL)	-

3 texts

FILTER

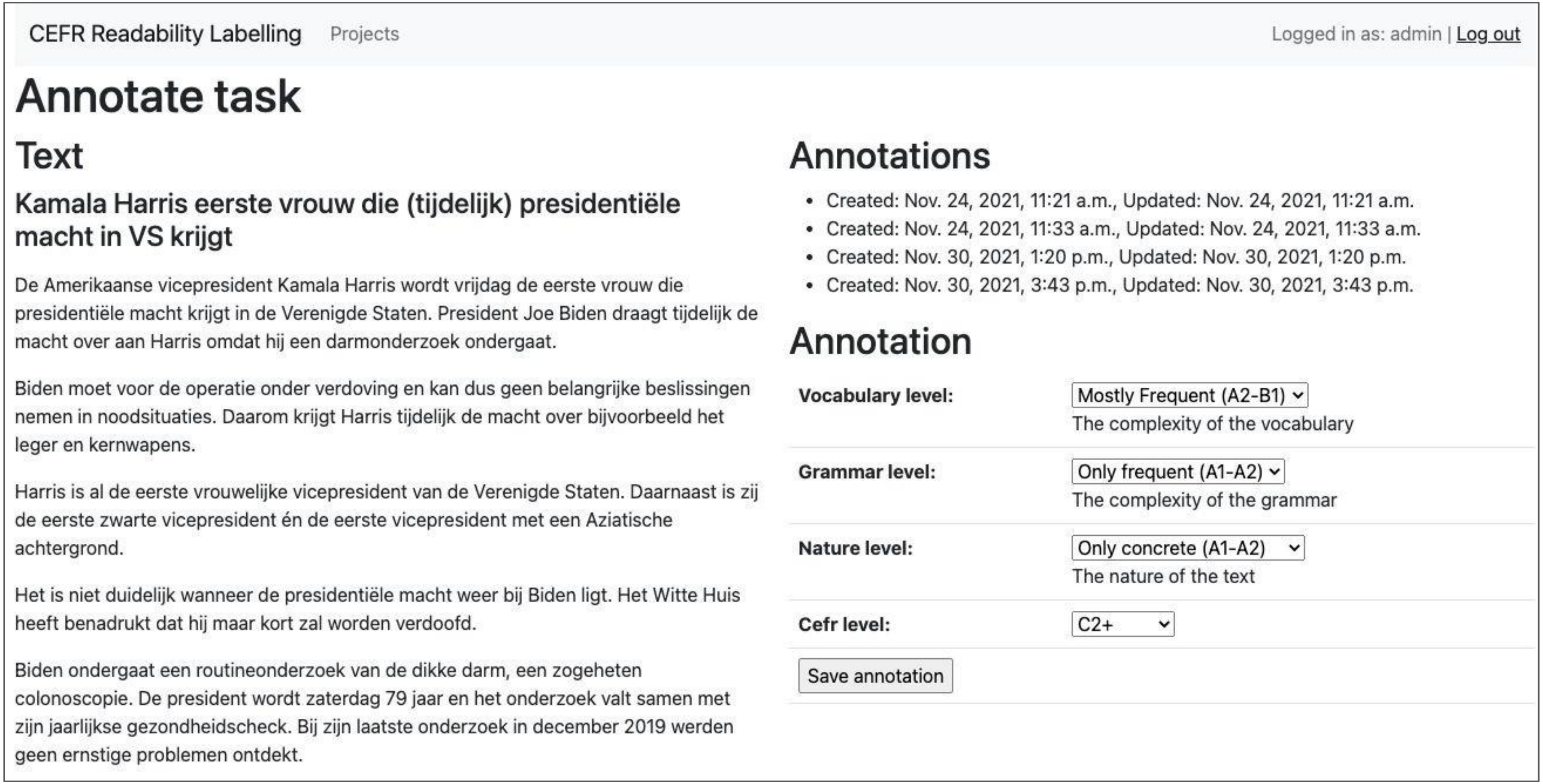
By language
- All
- English (UK)
- Dutch (NL)
- French (FR)
- German (DE)
- Spanish (ES)

By text type
- All
- News article
- -

By source
- All
- NU.nl
- -

By cefr level
- All
- A1
- A2
- B1
- B2
- C1
- C2

Step 1: Collect unlabelled texts



Annotate task

Text

Kamala Harris eerste vrouw die (tijdelijk) presidentiële macht in VS krijgt

De Amerikaanse vicepresident Kamala Harris wordt vrijdag de eerste vrouw die presidentiële macht krijgt in de Verenigde Staten. President Joe Biden draagt tijdelijk de macht over aan Harris omdat hij een darmonderzoek ondergaat.

Biden moet voor de operatie onder verdoving en kan dus geen belangrijke beslissingen nemen in noodsituaties. Daarom krijgt Harris tijdelijk de macht over bijvoorbeeld het leger en kernwapens.

Harris is al de eerste vrouwelijke vicepresident van de Verenigde Staten. Daarnaast is zij de eerste zwarte vicepresident én de eerste vicepresident met een Aziatische achtergrond.

Het is niet duidelijk wanneer de presidentiële macht weer bij Biden ligt. Het Witte Huis heeft benadrukt dat hij maar kort zal worden verdoofd.

Biden ondergaat een routineonderzoek van de dikke darm, een zogeheten colonoscopie. De president wordt zaterdag 79 jaar en het onderzoek valt samen met zijn jaarlijkse gezondheidscheck. Bij zijn laatste onderzoek in december 2019 werden geen ernstige problemen ontdekt.

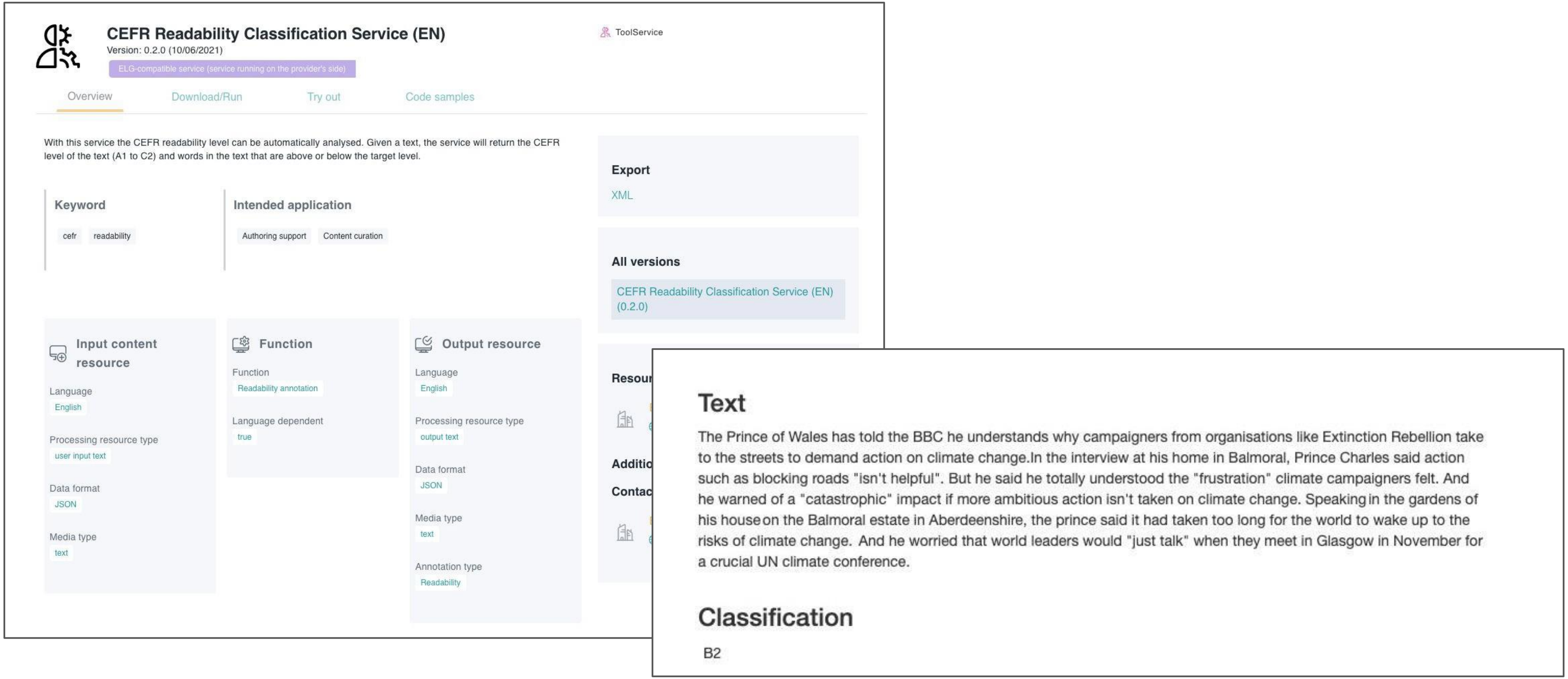
Annotations

- Created: Nov. 24, 2021, 11:21 a.m., Updated: Nov. 24, 2021, 11:21 a.m.
- Created: Nov. 24, 2021, 11:33 a.m., Updated: Nov. 24, 2021, 11:33 a.m.
- Created: Nov. 30, 2021, 1:20 p.m., Updated: Nov. 30, 2021, 1:20 p.m.
- Created: Nov. 30, 2021, 3:43 p.m., Updated: Nov. 30, 2021, 3:43 p.m.

Annotation

Vocabulary level:	Mostly Frequent (A2-B1) The complexity of the vocabulary
Grammar level:	Only frequent (A1-A2) The complexity of the grammar
Nature level:	Only concrete (A1-A2) The nature of the text
Cefr level:	C2+

Step 2: Label texts on CEFR level using human annotators



CEFR Readability Classification Service (EN)

Version: 0.2.0 (10/06/2021)

ELG-compatible service (service running on the provider's side)

ToolService

With this service the CEFR readability level can be automatically analysed. Given a text, the service will return the CEFR level of the text (A1 to C2) and words in the text that are above or below the target level.

Keyword

cefr readability

Intended application

Authoring support Content curation

Input content resource

Language: English

Processing resource type: user input text

Data format: JSON

Media type: text

Function

Function: Readability annotation

Language dependent: true

Output resource

Language: English

Processing resource type: output text

Data format: JSON

Media type: text

Annotation type: Readability

Export

XML

All versions

CEFR Readability Classification Service (EN) (0.2.0)

Resour

Additio

Contac

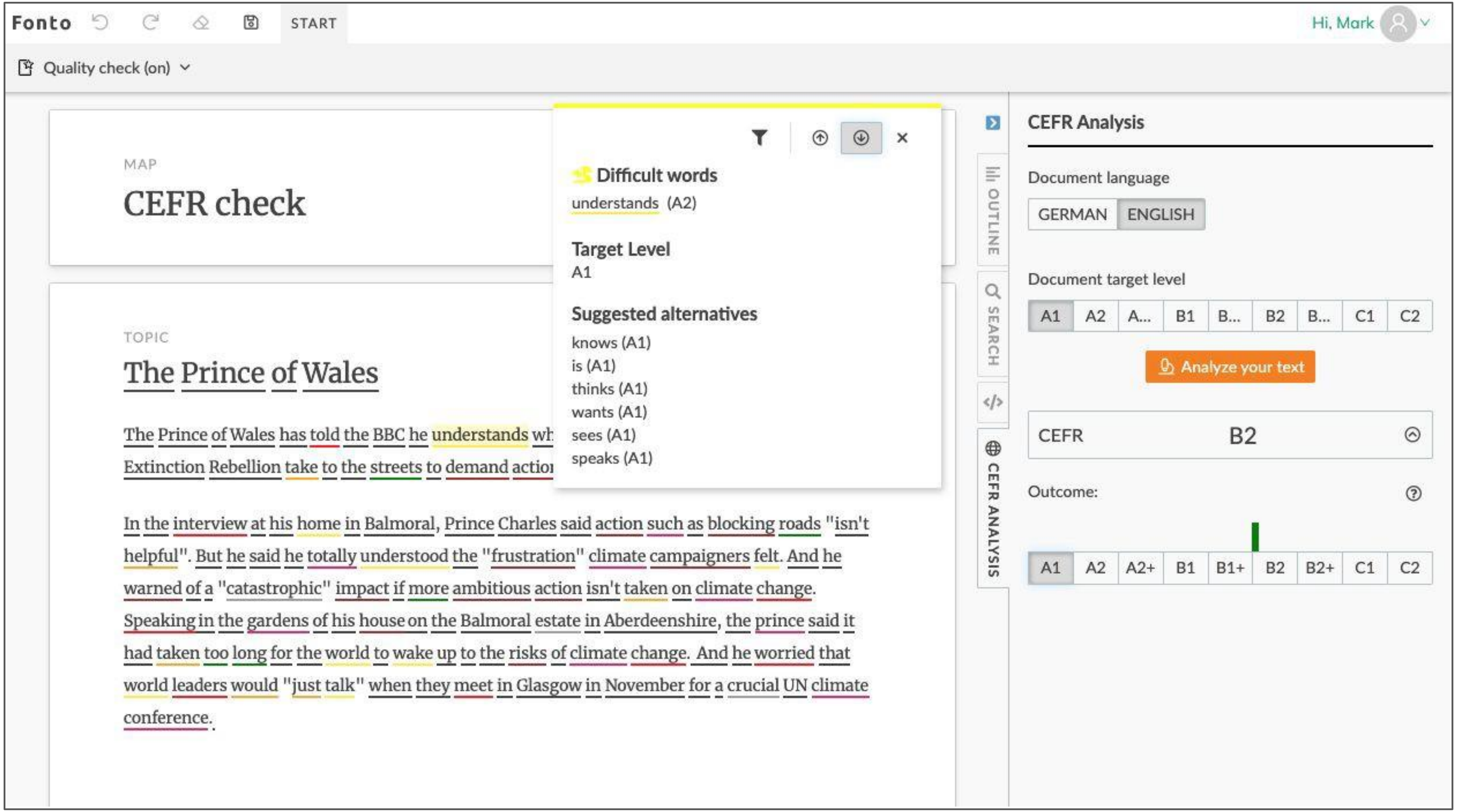
Text

The Prince of Wales has told the BBC he understands why campaigners from organisations like Extinction Rebellion take to the streets to demand action on climate change.In the interview at his home in Balmoral, Prince Charles said action such as blocking roads "isn't helpful". But he said he totally understood the "frustration" climate campaigners felt. And he warned of a "catastrophic" impact if more ambitious action isn't taken on climate change. Speaking in the gardens of his house on the Balmoral estate in Aberdeenshire, the prince said it had taken too long for the world to wake up to the risks of climate change. And he worried that world leaders would "just talk" when they meet in Glasgow in November for a crucial UN climate conference.

Classification

B2

Step 4: Expose AI models as services (REST API) on ELG



MAP

CEFR check

TOPIC

The Prince of Wales

The Prince of Wales has told the BBC he understands wh

Extinction Rebellion take to the streets to demand actio

In the interview at his home in Balmoral, Prince Charles said action such as blocking roads "isn't helpful". But he said he totally understood the "frustration" climate campaigners felt. And he warned of a "catastrophic" impact if more ambitious action isn't taken on climate change. Speaking in the gardens of his house on the Balmoral estate in Aberdeenshire, the prince said it had taken too long for the world to wake up to the risks of climate change. And he worried that world leaders would "just talk" when they meet in Glasgow in November for a crucial UN climate conference.

Difficult words

understands (A2)

Target Level

A1

Suggested alternatives

knows (A1)
is (A1)
thinks (A1)
wants (A1)
sees (A1)
speaks (A1)

CEFR Analysis

Document language

GERMAN [✓ ENGLISH]

Document target level

[✓ A1] A2 A... B1 B... B2 B... C1 C2

Analyze your text

CEFR B2

Outcome:

[✓ A1] A2 A2+ B1 B1+ B2 B2+ C1 C2

Step 5: Integrate services into end-user applications

Why ELG

1. Opportunity for extending our language coverage
2. Additional go-to-market channel
   a. Exposure and discoverability (marketplace)
   b. Lower adoption barrier for external developers (e.g. using ELG SDK)

Experiences so far

- Required metadata for publishing resources and services quite extensive / complex
- Very supportive ELG technical team
  - Help understand metadata scheme
  - Build and deploy services (by example)
  - Incorporated feedback into ELG platform
- Standardization and improvement of API design (both ways)
- Commercial services (eg customer registration, metering, billing) not yet available

14.00 uur	Welkom	Vincent Vandeghinste (INT)
14.05 uur	Introducing the European Language Grid	Georg Rehm (DFKI)
14.35 uur	Demonstratie van de ELG, met focus op het Nederlands	Vincent Vandeghinste (INT)
15.00 uur	Vraag- en antwoordsessie en discussie	
15.10 uur	Pauze	
15.25 uur	Taalmaterialen voor het Nederlands	Bob Boelhouwer (INT)
15.35 uur	Presentaties van verschillende bedrijven (3) ● EDIA (Mark Breuker) ● Telecats (Arjan van Hessen) ● Y.digital (Ian Fitzpatrick)	Oele Koornwinder (NOTaS)
16.25 uur	Vraag- en antwoordsessie en discussie	

Telecats

THE SEARCH FOR SUITABLE DATA

About Telecats

- An innovative “customer contact provider” with HLT (language and speech technology) and AI (artificial intelligence) that empower customer service.
- Part of Webhelp (France) since 1-1-2020
- Automation of “as much as possible/desirable” calls

About Telecats

We **facilitate the contact centre** by identifying the customer, classifying the customer’s question and routing the call.

We **support employees** by delivering all relevant information to them and showing them suggestions for answers.

We **assist the customer** intelligently routing the call to the most suitable employee or by offering self-service

We **identify trends, opportunities and problems** based on conversation characteristics and choice of words.

Speech Routing

Routing your customer contact with speech recognition

- Speech recognition provides customers with the opportunity to ask their question using their own words.
- 90% of all questions can be properly classified by using speech recognition which means fewer questions have to be put through to another advisor.
- Routing with speech recognition and CTI can result in an average time gain of 45 seconds for each conversation.

Recording & Transcription

Insight into your customer contact with speech technology

- Why do your customers call?
- What are the conversations about?
- What are customers (not) satisfied about?

This is valuable information for the entire organization. Because of recording and transcription this information becomes available. Phone calls are recorded in high quality and are transcribed, sometimes in real time. When added to the Call-Detail-Record this information can be the basis for extensive analysis.

Old situation

- Recording (telephony) data
- Recognising the spoken data
- Asking customer experts to label the data
- Train the machine with recognition results AND the label(s)

What we miss

- There is more information in the speech than we can extract now. Tone of voice, hesitations, etc.
- We do not know anything about the background of an individual caller.
- We cannot extract “emotion”.

Pro’s and Con’s

- Works well as long as people speak “normally”
- You need a lot of data, time (and money)
- Once trained, it is stable

  but changing the content needs a lot of work
- It is AI, but…

What we need

- We need a lot of labelled, spoken (and eventually transcribed) data, handling all kind of company-customer dialogs.
- We need, besides HQ ASR, additional algorithms like emotion-recognition, meaning extraction and identification algorithms.

What we need

- We need the data from various sources (telephony, broadband, with/without background noise, etc.)
- We are NOT interested in the person-related content, but modern GDPR…
- We need this data for Dutch, English, French and other languages

What can we do?

- Collect, recognise, and label our own data
- Use data from “other parties” but GDPR…
- Train with other data without access to the speech and transcripts (via a data vault)
- Nothing…

Thanks

AND… QUESTIONS?

EUROPEAN LANGUAGE GRID

14.00 uur	Welkom	Vincent Vandeghinste (INT)
14.05 uur	Introducing the European Language Grid	Georg Rehm (DFKI)
14.35 uur	Demonstratie van de ELG, met focus op het Nederlands	Vincent Vandeghinste (INT)
15.00 uur	Vraag- en antwoordsessie en discussie	
15.10 uur	Pauze	
15.25 uur	Taalmaterialen voor het Nederlands	Bob Boelhouwer (INT)
15.35 uur	Presentaties van verschillende bedrijven (3) ● EDIA (Mark Breuker) ● Telecats (Arjan van Hessen) ● Y.digital (Ian Fitzpatrick)	Oele Koornwinder (NOTaS)
16.25 uur	Vraag- en antwoordsessie en discussie	

/instituut voor
de Nederlandse
taal/

Taalmaterialen

Dutch Language Resources Service

Bob Boelhouwer
Bob.boelhouwer@ivdnt.org

Wat kunnen we bieden?

- Hergebruik van taalmaterialen
- Archivering
- Permanente verwijzingen (incl. PID’s)
- Onderhoud
- Ondersteuning

Geschiedenis van Taalmaterialen

- Opgericht in 2004 door de NTU, als onderdeel van STEVIN; een groot onderzoeks-programma.
- Oorspronkelijke naam: TST-Centrale, uitgevoerd door het INT (toen INL).
- STEVIN data is nog steeds de kern van de catalogus, maar veel andere organisaties, commercieel and niet-commercieel hebben data bijgedragen.

Voor wie is het bedoeld?

- Academici (CLARIN B-Centre)
- Commerciële ondernemingen
- Algemene publiek

Een paar statistieken

- 129 producten in de catalogus
- Mirror voor een aantal Zuid-Afrikaanse talen

Aantal downloads in 2020:

- 1360 downloads niet-commercieel
- 45 commercieel

EUROPEAN LANGUAGE GRID

14.00 uur	Welkom	Vincent Vandeghinste (INT)
14.05 uur	Introducing the European Language Grid	Georg Rehm (DFKI)
14.35 uur	Demonstratie van de ELG, met focus op het Nederlands	Vincent Vandeghinste (INT)
15.00 uur	Vraag- en antwoordsessie en discussie	
15.10 uur	Pauze	
15.25 uur	Taalmaterialen voor het Nederlands	Bob Boelhouwer (INT)
15.35 uur	Presentaties van verschillende bedrijven (3) ● EDIA (Mark Breuker) ● Telecats (Arjan van Hessen) ● Y.digital (Ian Fitzpatrick)	Oele Koornwinder (NOTaS)
16.25 uur	Vraag- en antwoordsessie en discussie	