UniDive WG3

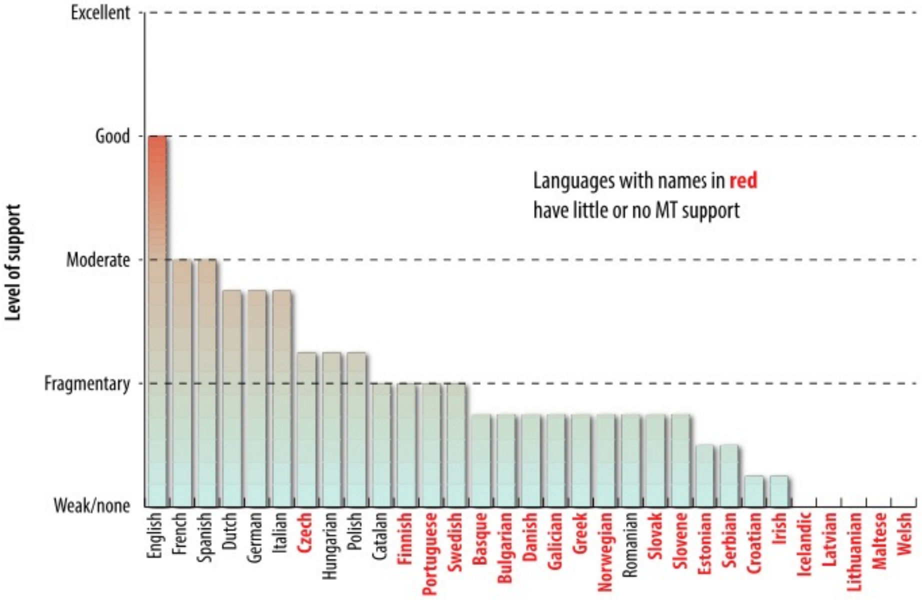
Subgroup - Multilingual Tool and Resource Documentation

Co-leaders:
A. Seza Doğruöz (Ghent University)
Maria Giagkou (Athena RC)
Teresa Lynn (Mohamed bin Zayed University of Artificial Intelligence)

UniDive

2012 META-NET White Paper Series: LT Support Levels

2012



Languages with names in **red** have little or no MT support

Source: META-NET White Paper Series: Europe's Languages in the Digital Age. Springer, Heidelberg, New York, Dordrecht, London, September 2012. Georg Rehm and Hans Uszkoreit (series editors)

“Language Equality” EP Resolution (2018)

EP Resolution “Language equality in the digital age”
P8_TA(2018)0332 – partially based on the STOA study

Voting (11 Sept. 2018): **592 yes** – 45 no

Selected Recommendations addressed by ELE:

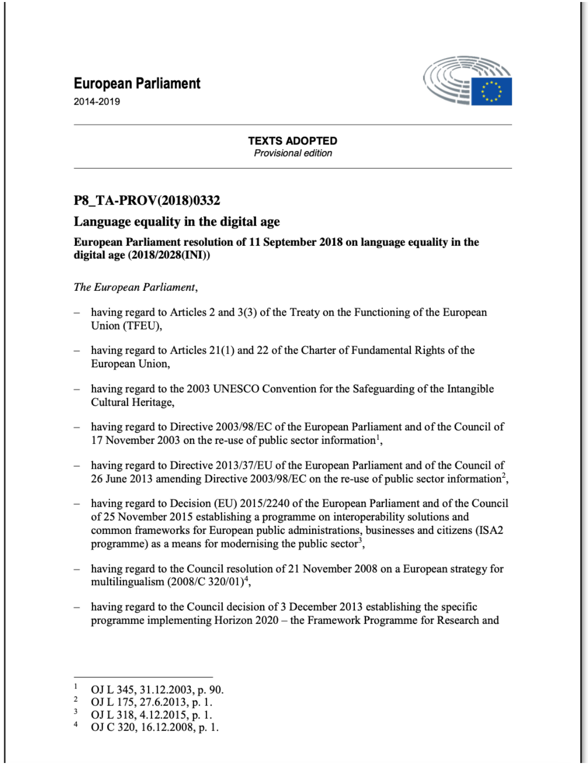
EUROPEAN LANGUAGE EQUALITY

25. Establish a large-scale, long-term coordinated funding programme for research, development and innovation in the field of language technologies, at European, national and regional levels, tailored specifically to Europe’s needs and demands

EUROPEAN LANGUAGE GRID

29. Create a European LT platform for sharing of services

27. Europe has to secure its leadership in language-centric AI



European Parliament
2014-2019

TEXTS ADOPTED
Provisional edition

P8_TA-PROV(2018)0332

Language equality in the digital age

European Parliament resolution of 11 September 2018 on language equality in the digital age (2018/2028(INI))

The European Parliament,

– having regard to Articles 2 and 3(3) of the Treaty on the Functioning of the European Union (TFEU),

– having regard to Articles 21(1) and 22 of the Charter of Fundamental Rights of the European Union,

– having regard to the 2003 UNESCO Convention for the Safeguarding of the Intangible Cultural Heritage,

– having regard to Directive 2003/98/EC of the European Parliament and of the Council of 17 November 2003 on the re-use of public sector information[1],

– having regard to Directive 2013/37/EU of the European Parliament and of the Council of 26 June 2013 amending Directive 2003/98/EC on the re-use of public sector information[2],

– having regard to Decision (EU) 2015/2240 of the European Parliament and of the Council of 25 November 2015 establishing a programme on interoperability solutions and common frameworks for European public administrations, businesses and citizens (ISA2 programme) as a means for modernising the public sector[3],

– having regard to the Council resolution of 21 November 2008 on a European strategy for multilingualism (2008/C 320/01)[4],

– having regard to the Council decision of 3 December 2013 establishing the specific programme implementing Horizon 2020 – the Framework Programme for Research and

[1] OJ L 345, 31.12.2003, p. 90.
[2] OJ L 175, 27.6.2013, p. 1.
[3] OJ L 318, 4.12.2015, p. 1.
[4] OJ C 320, 16.12.2008, p. 1.

European Language Equality Project. (2021-2022)

EUROPEAN
LANGUAGE
EQUALITY

Consortium: 52 partners from all over Europe

Coordinator: Dublin City University

Co-coordinator: DFKI

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

2022 LT Support Levels

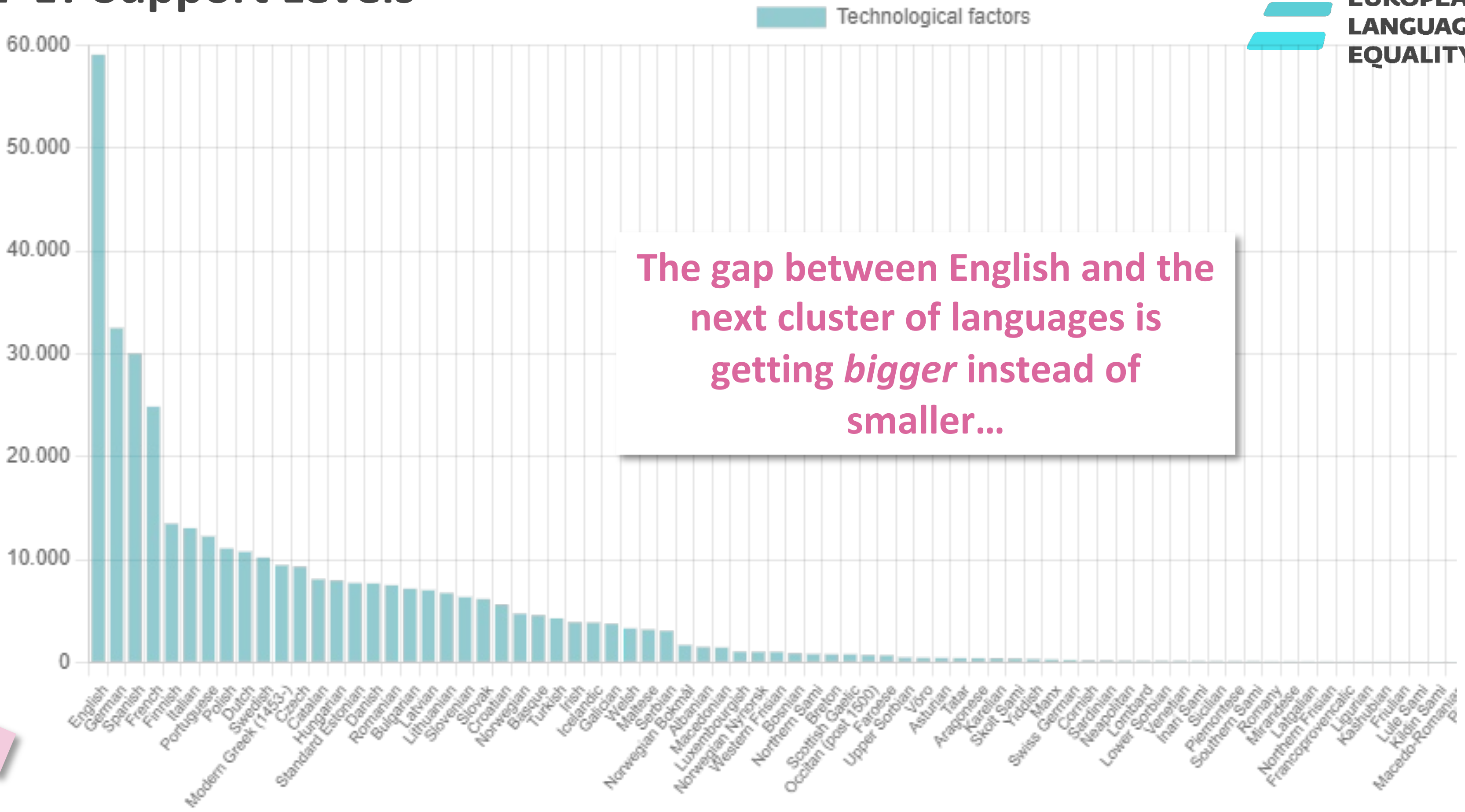
EUROPEAN
LANGUAGE
EQUALITY

Good support

Moderate support

Fragmentary support

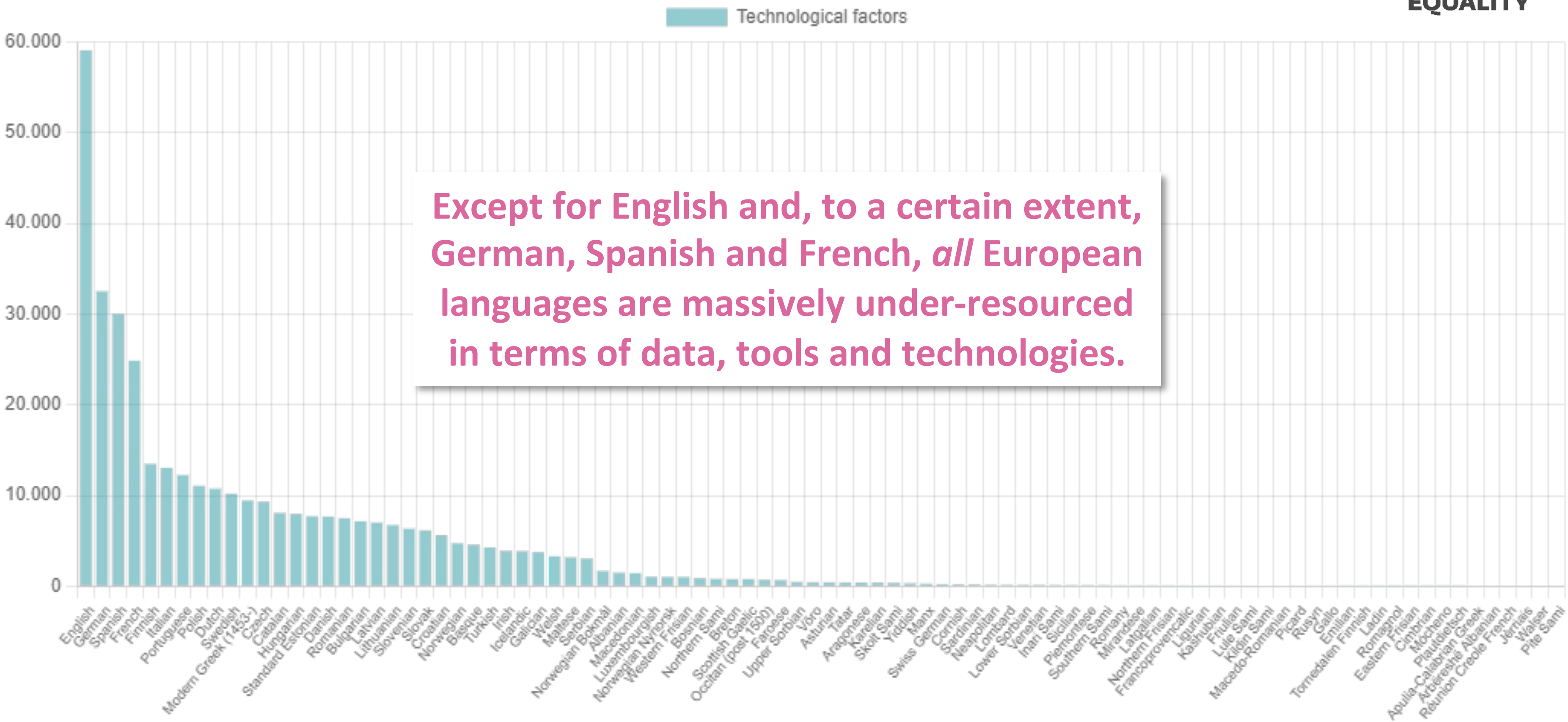
Weak or no support



The gap between English and the next cluster of languages is getting *bigger* instead of smaller…

2022

https://live.european-language-grid.eu/catalogue/dashboard

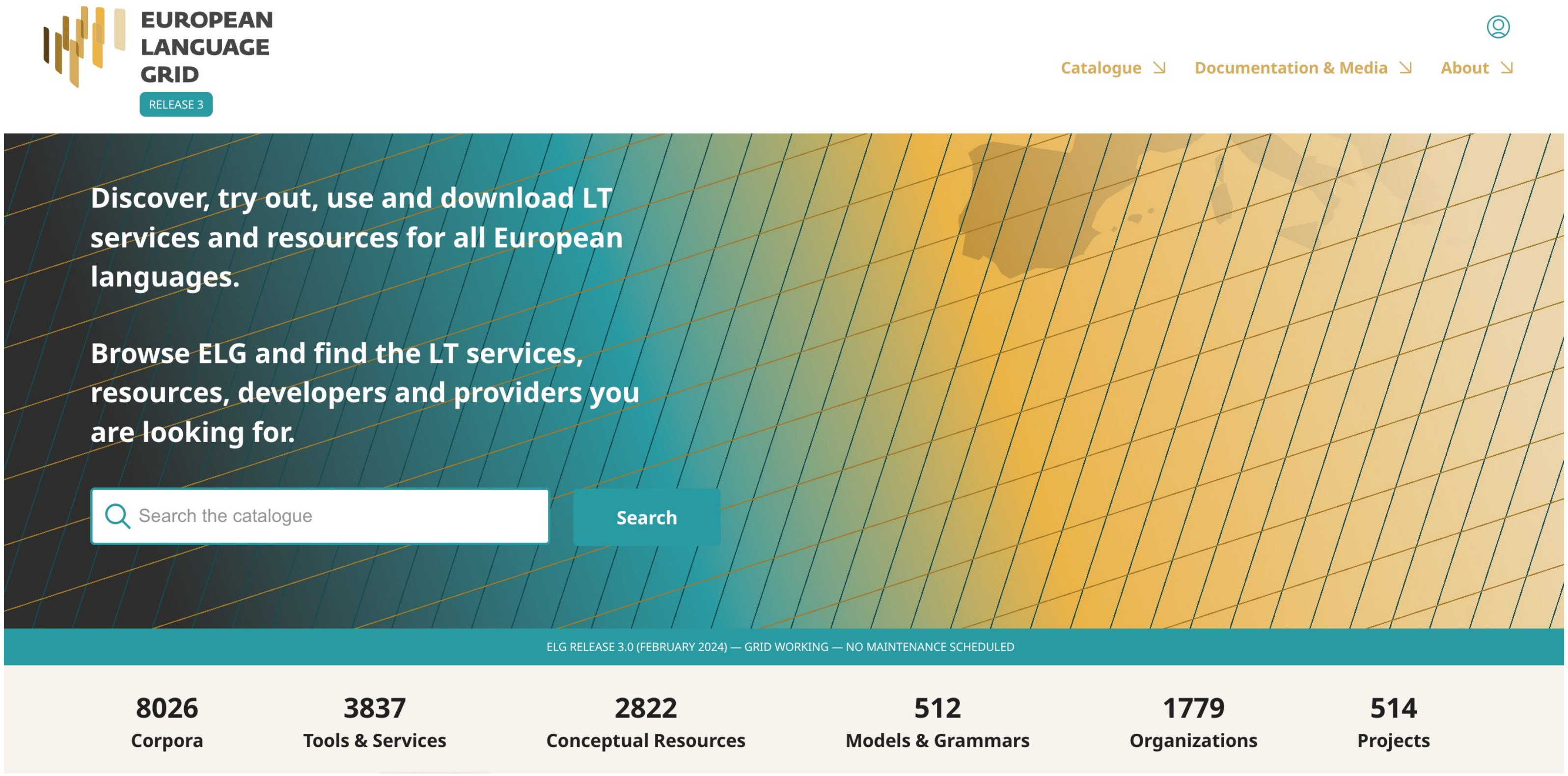


Except for English and, to a certain extent, German, Spanish and French, *all* European languages are massively under-resourced in terms of data, tools and technologies.

https://live.european-language-grid.eu/catalogue/dashboard

European Language Grid – “the yellow pages” of LT

EUROPEAN LANGUAGE EQUALITY



Discover, try out, use and download LT services and resources for all European languages.

Browse ELG and find the LT services, resources, developers and providers you are looking for.

ELG RELEASE 3.0 (FEBRUARY 2024) — GRID WORKING — NO MAINTENANCE SCHEDULED

8026	3837	2822	512	1779	514
Corpora	Tools & Services	Conceptual Resources	Models & Grammars	Organizations	Projects

https://live.european-language-grid.eu/

European Language Grid – harvests other platforms

EUROPEAN LANGUAGE RESOURCES ASSOCIATION
ELRA

European Language Resource Coordination
Connecting Europe Facility
ELRC-SHARE

AI4EU

META-SHARE

zenodo

EUROPEAN
LANGUAGE
GRID

LINDAT
CLARIAH-CZ

Hugging Face

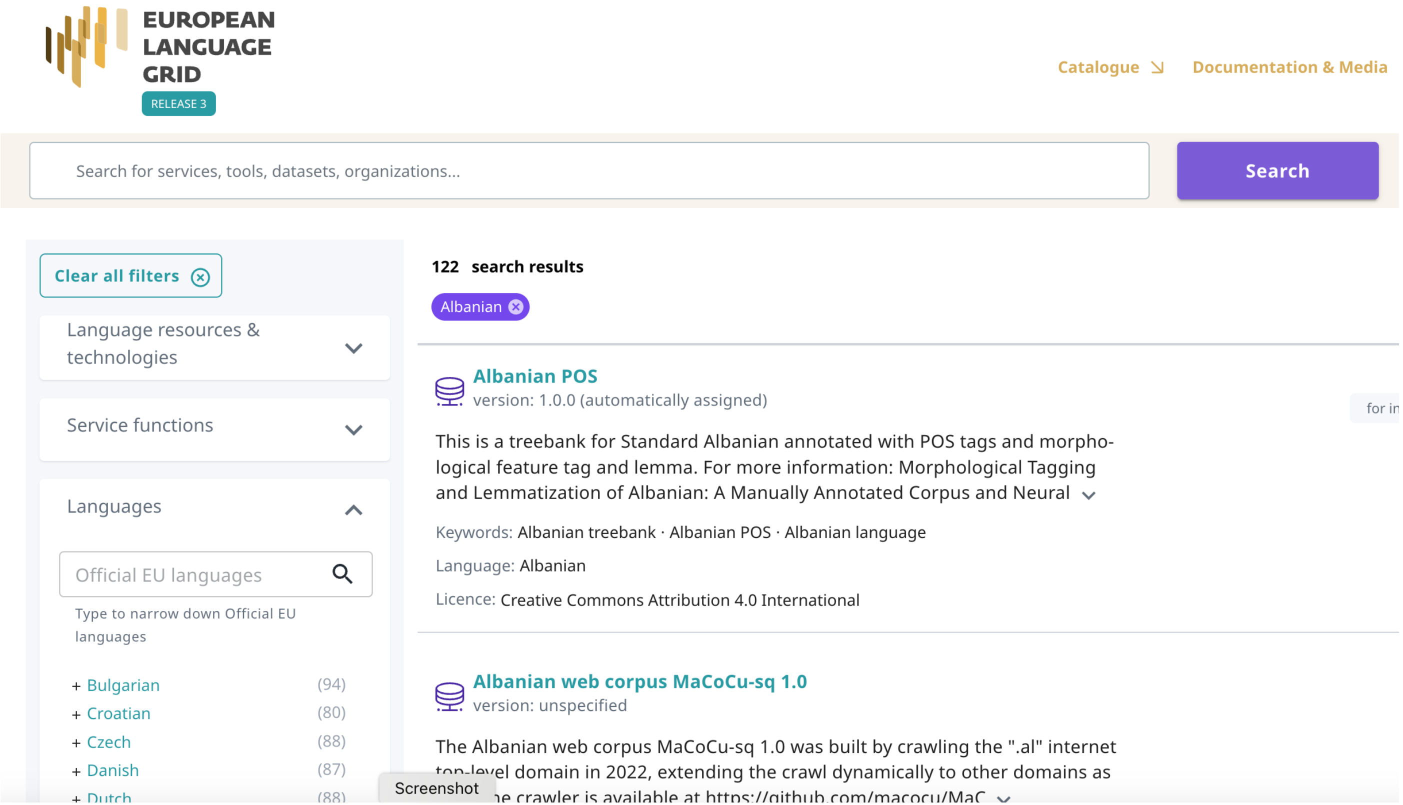
CLARIN.SI

CLARIN

https://live.european-language-grid.eu/

European Language Grid – catalogue search

EUROPEAN LANGUAGE EQUALITY



EUROPEAN LANGUAGE GRID
RELEASE 3

Catalogue ↘ Documentation & Media

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

Search

Clear all filters ⊗

Language resources & technologies

Service functions

Languages

Official EU languages

Type to narrow down Official EU languages

+ Bulgarian	(94)
+ Croatian	(80)
+ Czech	(88)
+ Danish	(87)
+ Dutch	(88)

122 search results

Albanian ⊗

Albanian POS
version: 1.0.0 (automatically assigned)

for in

This is a treebank for Standard Albanian annotated with POS tags and morphological feature tag and lemma. For more information: Morphological Tagging and Lemmatization of Albanian: A Manually Annotated Corpus and Neural

Keywords: Albanian treebank · Albanian POS · Albanian language

Language: Albanian

Licence: Creative Commons Attribution 4.0 International

Albanian web corpus MaCoCu-sq 1.0
version: unspecified

The Albanian web corpus MaCoCu-sq 1.0 was built by crawling the ".al" internet top-level domain in 2022, extending the crawl dynamically to other domains as he crawler is available at https://github.com/macocu/MaC

Screenshot

https://live.european-language-grid.eu/

Task Overview

Task 3.1.1

- Assess the “discoverability” of NLP tools and resources
- Who can participate?
    - Everyone 😃

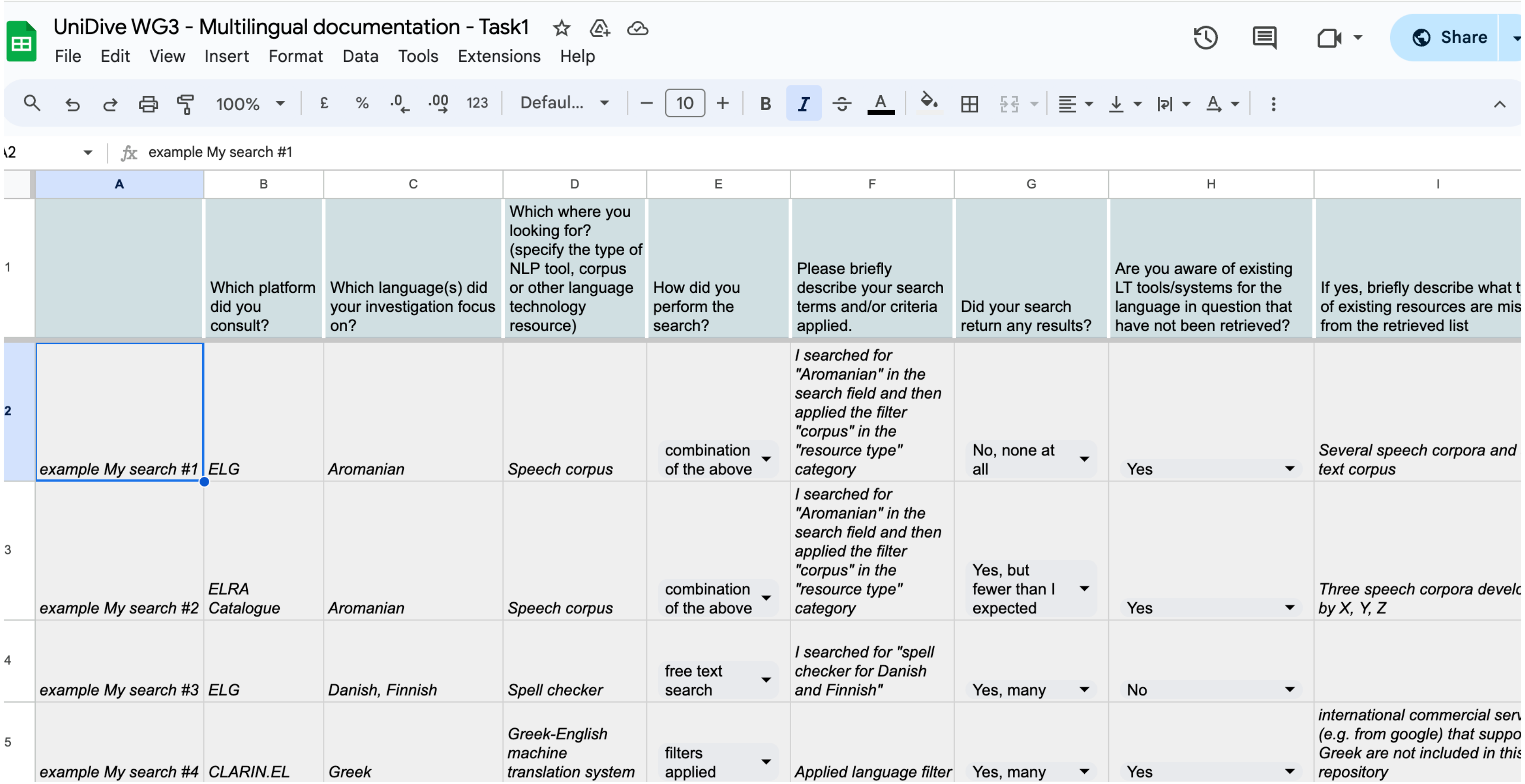
Task 3.1.2

- Analyse the NLP tool availability in the ELG catalogue
- Who can participate?
    - Excel or Tableau enthusiasts
    - Those with skills in data visualisation

Task 3.1.1: Assessing the “discoverability” of NLP tools

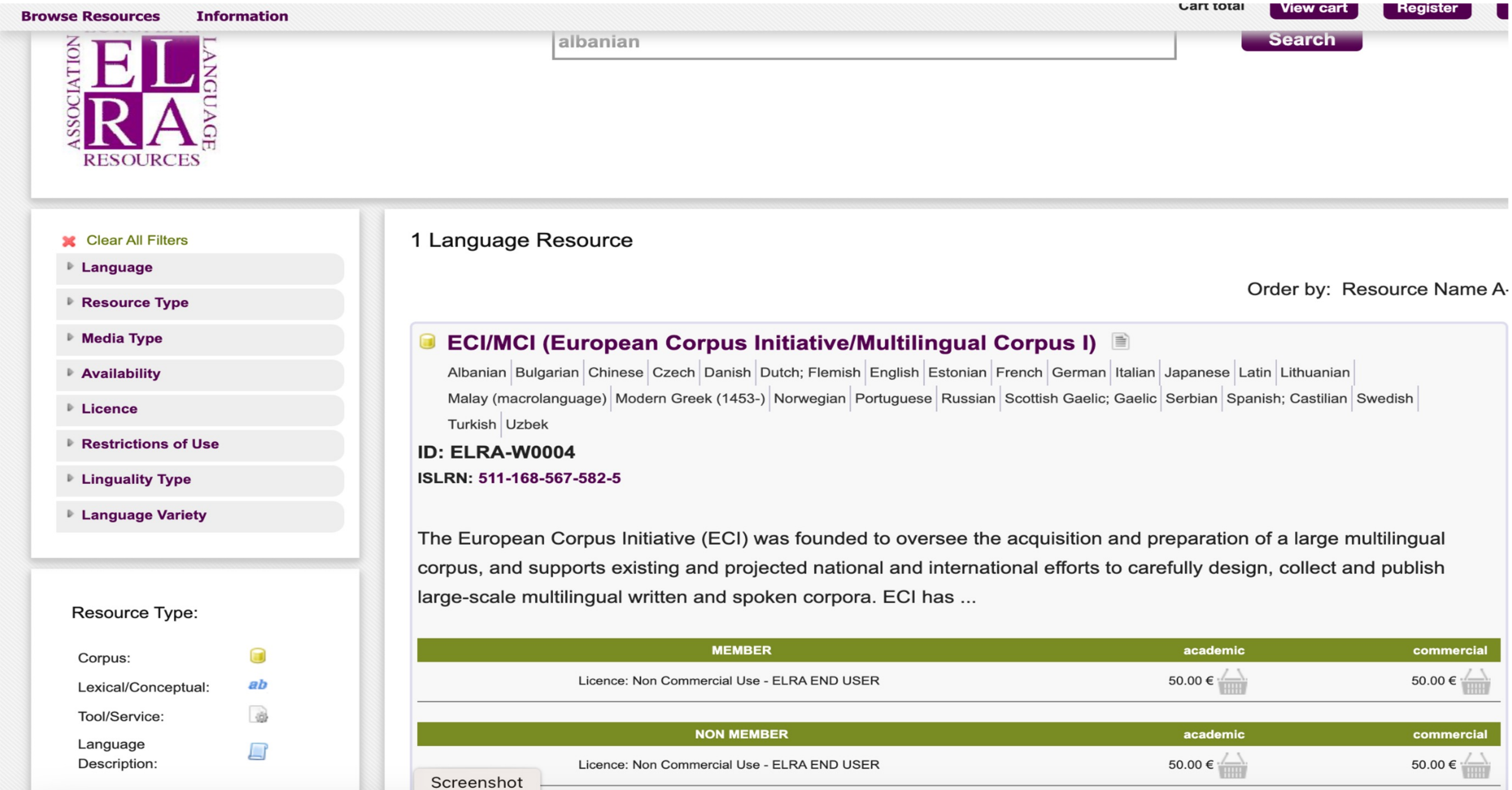
- Choose your language(s) and NLP task(s) of interest
- Search for the relevant tools across a number of platforms
- Report on the discoverability of desired tool/resource
  (Could you find easily it or not? What challenges?)
- Report on the metadata information available (was it sufficient and accurate?)
- What metadata do you recommend should be provided for a similar search?
- Is there a tool/ resource you are aware of that you can’t find on these platforms?

UniDive



	A	B	C	D	E	F	G	H	I
1		Which platform did you consult?	Which language(s) did your investigation focus on?	Which where you looking for? (specify the type of NLP tool, corpus or other language technology resource)	How did you perform the search?	Please briefly describe your search terms and/or criteria applied.	Did your search return any results?	Are you aware of existing LT tools/systems for the language in question that have not been retrieved?	If yes, briefly describe what t of existing resources are mis from the retrieved list
2	*example My search #1*	*ELG*	*Aromanian*	*Speech corpus*	combination of the above	*I searched for "Aromanian" in the search field and then applied the filter "corpus" in the "resource type" category*	No, none at all	Yes	*Several speech corpora and text corpus*
3	*example My search #2*	*ELRA Catalogue*	*Aromanian*	*Speech corpus*	combination of the above	*I searched for "Aromanian" in the search field and then applied the filter "corpus" in the "resource type" category*	Yes, but fewer than I expected	Yes	*Three speech corpora develo by X, Y, Z*
4	*example My search #3*	*ELG*	*Danish, Finnish*	*Spell checker*	free text search	*I searched for "spell checker for Danish and Finnish"*	Yes, many	No	
5	*example My search #4*	*CLARIN.EL*	*Greek*	*Greek-English machine translation system*	filters applied	*Applied language filter*	Yes, many	Yes	*international commercial serv (e.g. from google) that suppo Greek are not included in this repository*

E.g Search for Albanian Tools - ELRA Catalogue



Resource Type:

Corpus:	🟨
Lexical/Conceptual:	ab
Tool/Service:	⚙
Language Description:	📘

1 Language Resource

Order by: Resource Name A-

ECI/MCI (European Corpus Initiative/Multilingual Corpus I)

Albanian | Bulgarian | Chinese | Czech | Danish | Dutch; Flemish | English | Estonian | French | German | Italian | Japanese | Latin | Lithuanian | Malay (macrolanguage) | Modern Greek (1453-) | Norwegian | Portuguese | Russian | Scottish Gaelic; Gaelic | Serbian | Spanish; Castilian | Swedish | Turkish | Uzbek

ID: ELRA-W0004

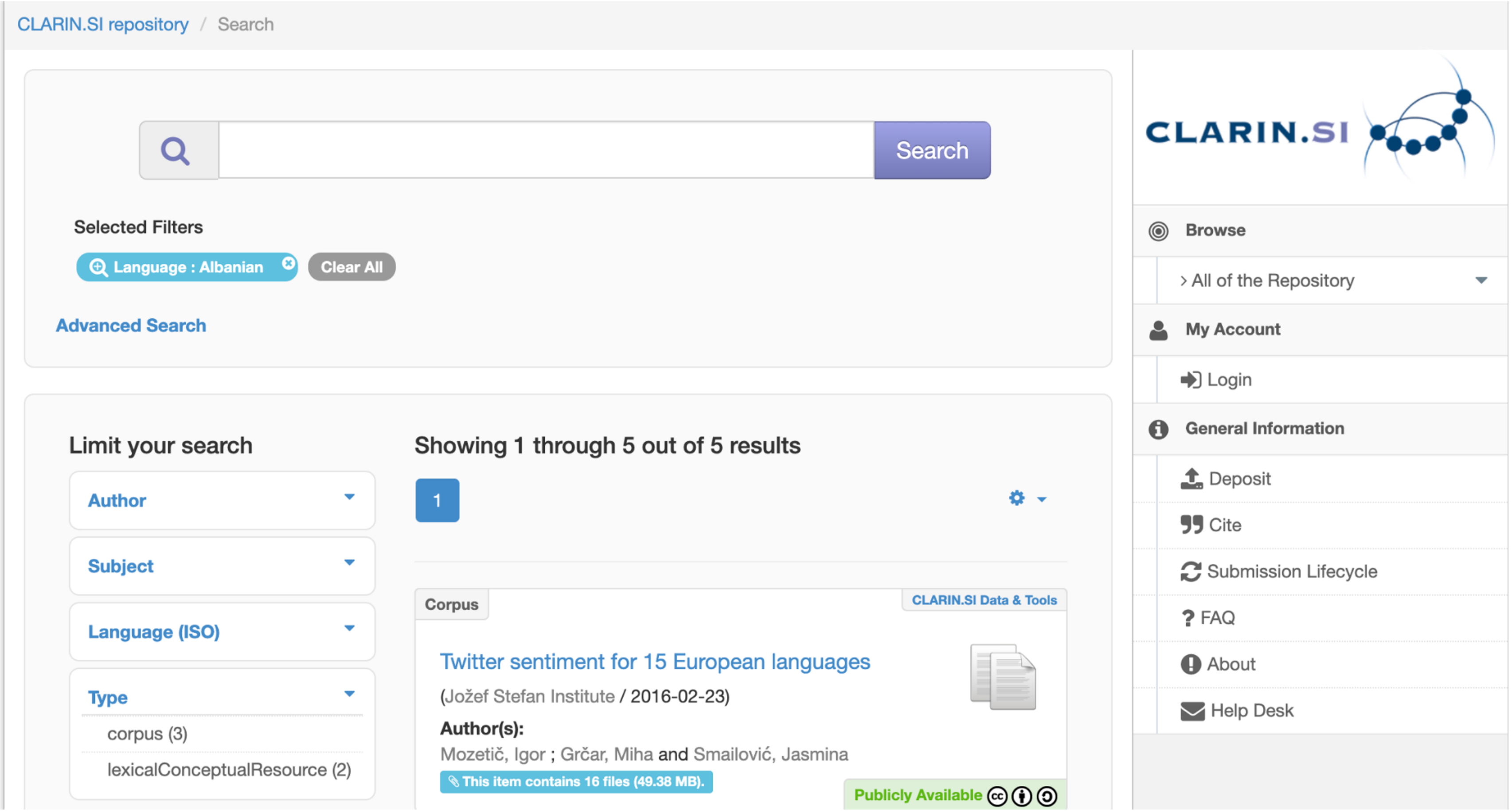
ISLRN: 511-168-567-582-5

The European Corpus Initiative (ECI) was founded to oversee the acquisition and preparation of a large multilingual corpus, and supports existing and projected national and international efforts to carefully design, collect and publish large-scale multilingual written and spoken corpora. ECI has ...

MEMBER	academic	commercial
Licence: Non Commercial Use - ELRA END USER	50.00 €	50.00 €

NON MEMBER	academic	commercial
Licence: Non Commercial Use - ELRA END USER	50.00 €	50.00 €

E.g Search for Albanian Tools - CLARIN-SI Catalogue



Selected Filters

Language : Albanian

Limit your search

Type
- corpus (3)
- lexicalConceptualResource (2)

Showing 1 through 5 out of 5 results

Twitter sentiment for 15 European languages

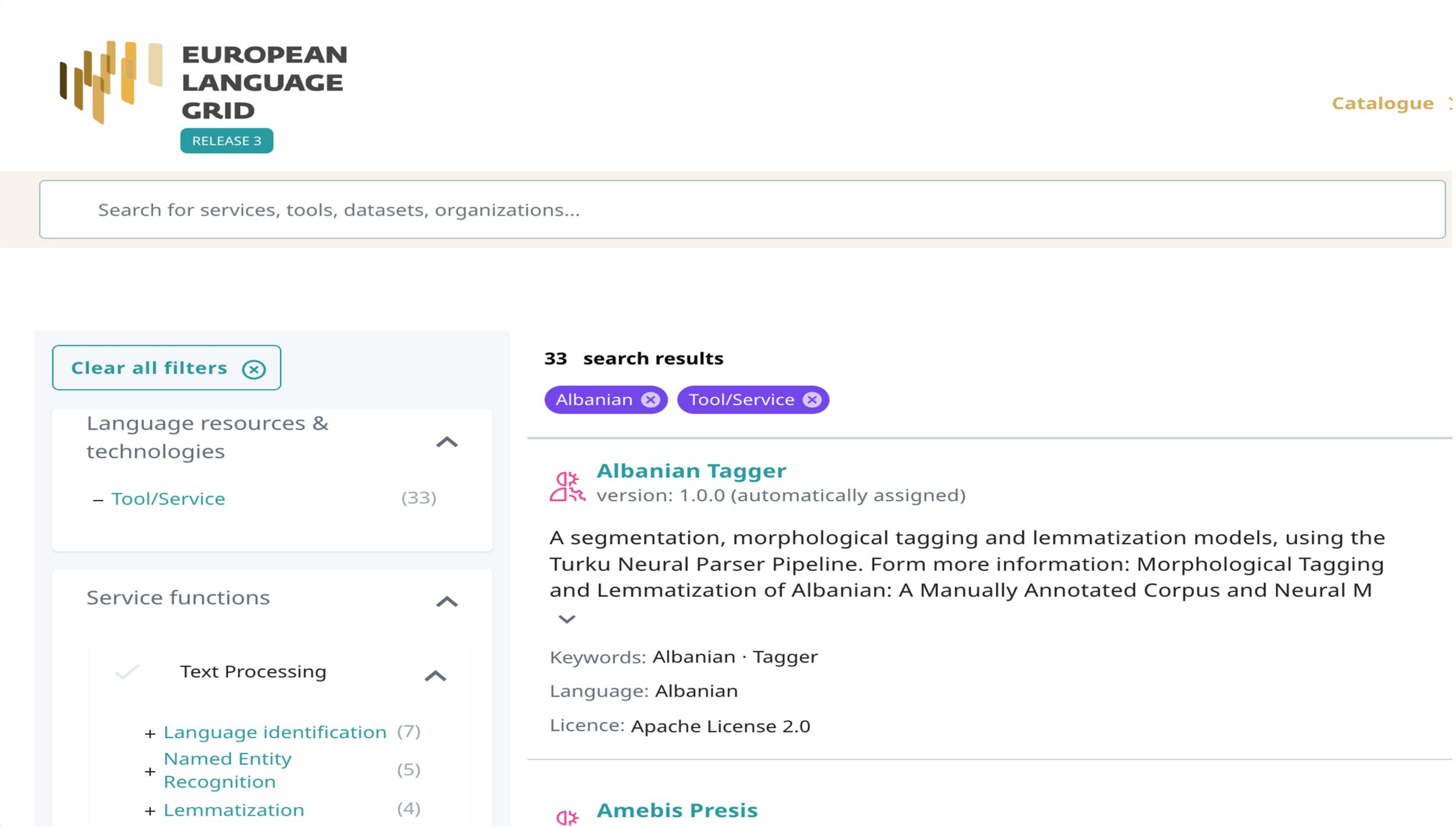
(Jožef Stefan Institute / 2016-02-23)

Author(s):
Mozetič, Igor ; Grčar, Miha and Smailović, Jasmina

This item contains 16 files (49.38 MB).

Publicly Available

E.g Search for Albanian Tools - ELG Catalogue



Language resources & technologies

– Tool/Service (33)

Service functions

Text Processing

+ Language identification (7)
+ Named Entity Recognition (5)
+ Lemmatization (4)

33 search results

Albanian ⊗ Tool/Service ⊗

Albanian Tagger
version: 1.0.0 (automatically assigned)

A segmentation, morphological tagging and lemmatization models, using the Turku Neural Parser Pipeline. Form more information: Morphological Tagging and Lemmatization of Albanian: A Manually Annotated Corpus and Neural M

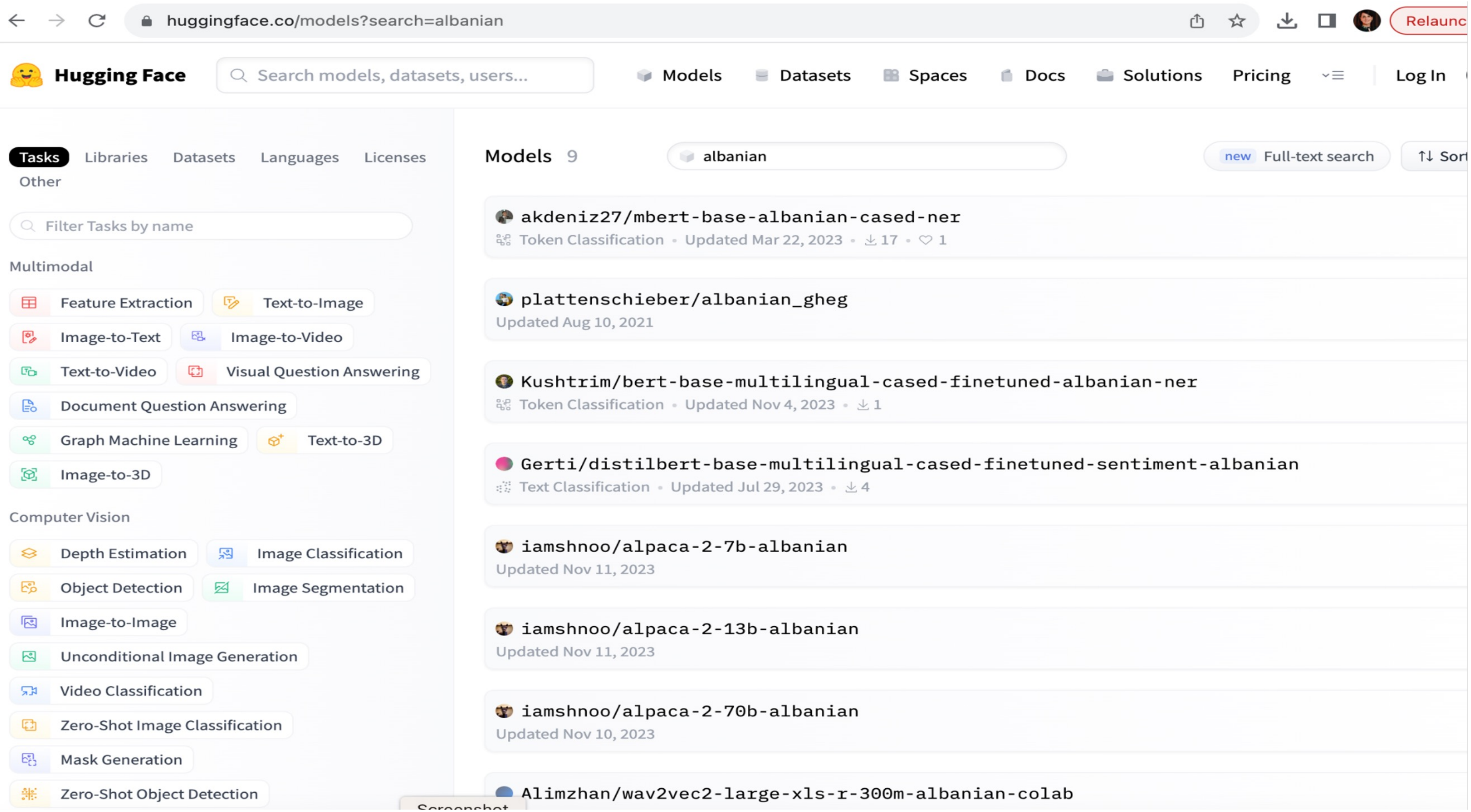
Keywords: Albanian · Tagger

Language: Albanian

Licence: Apache License 2.0

Amebis Presis

E.g Search for Albanian Tools - Hugging Face



Models 9

albanian

akdeniz27/mbert-base-albanian-cased-ner
Token Classification • Updated Mar 22, 2023 • 17 • 1

plattenschieber/albanian_gheg
Updated Aug 10, 2021

Kushtrim/bert-base-multilingual-cased-finetuned-albanian-ner
Token Classification • Updated Nov 4, 2023 • 1

Gerti/distilbert-base-multilingual-cased-finetuned-sentiment-albanian
Text Classification • Updated Jul 29, 2023 • 4

iamshnoo/alpaca-2-7b-albanian
Updated Nov 11, 2023

iamshnoo/alpaca-2-13b-albanian
Updated Nov 11, 2023

iamshnoo/alpaca-2-70b-albanian
Updated Nov 10, 2023

Alimzhan/wav2vec2-large-xls-r-300m-albanian-colab

Task 3.1.1: Assessing the “discoverability” of NLP tools

Process

- A template will be provided with suggested queries
- Additional input also desired

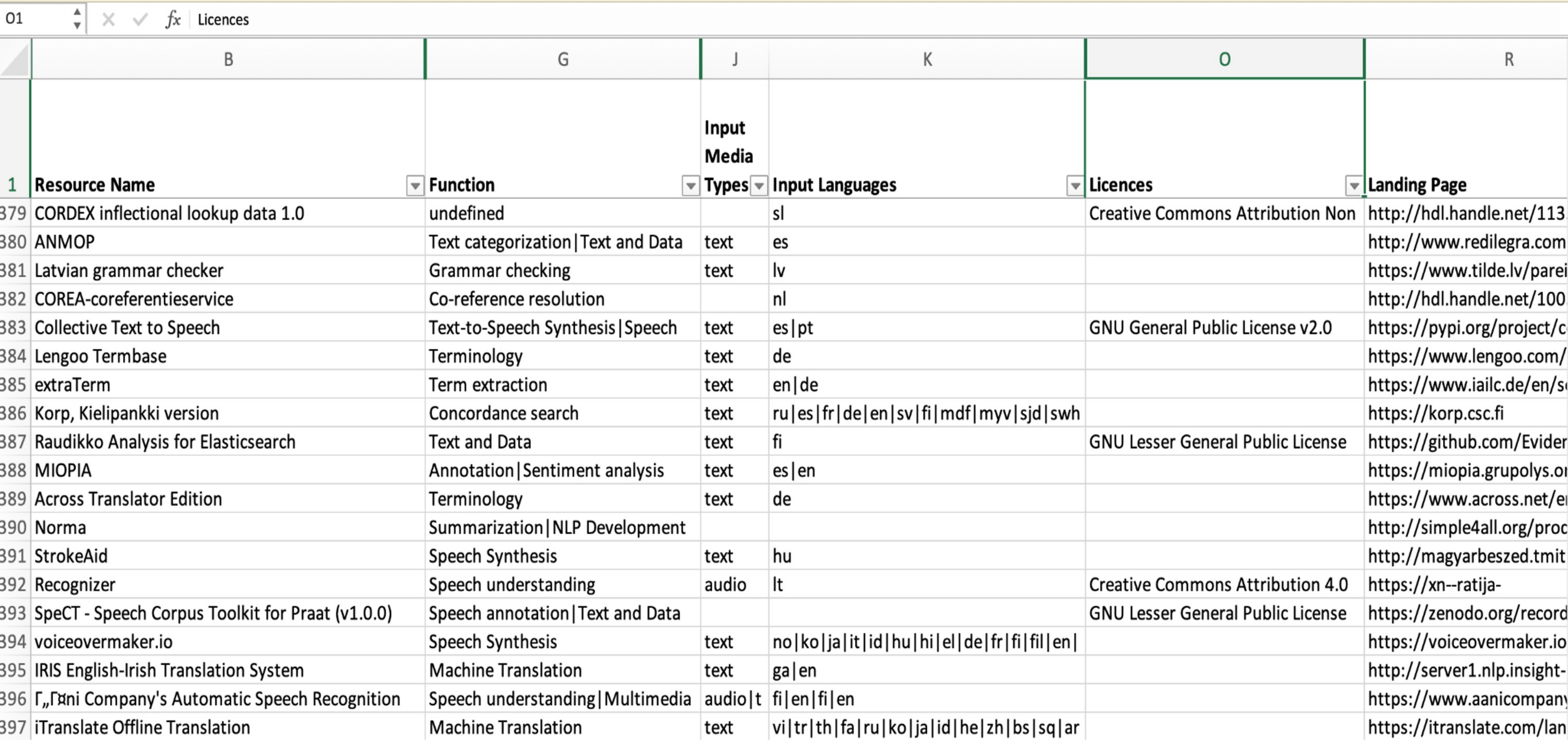
Outcomes

- An increased awareness and understanding of language technology platforms
- Insight into limitations of current schemas
- Honed research skills in searching for NLP tools/ resources
- Recommendations for improving discoverability of tools/ resources

Task 3.1.2: Tool Availability Analysis

- Seeking volunteers with strong Excel/ Tableau skills
- Analysis required on ELG catalogue export – (large excel doc, ~3900 entries)
- Prompts below can be the start of investigation - let's see what else emerges:
    1. The tools for which certain languages are missing (e.g. Irish doesn't have NER, Sentiment Analyser, etc)
    2. The multilingual tool types that are lacking across languages (e.g. NER is only available for X, Y, Z languages)
    3. Which languages tend to be left out of "multilingual" tools?

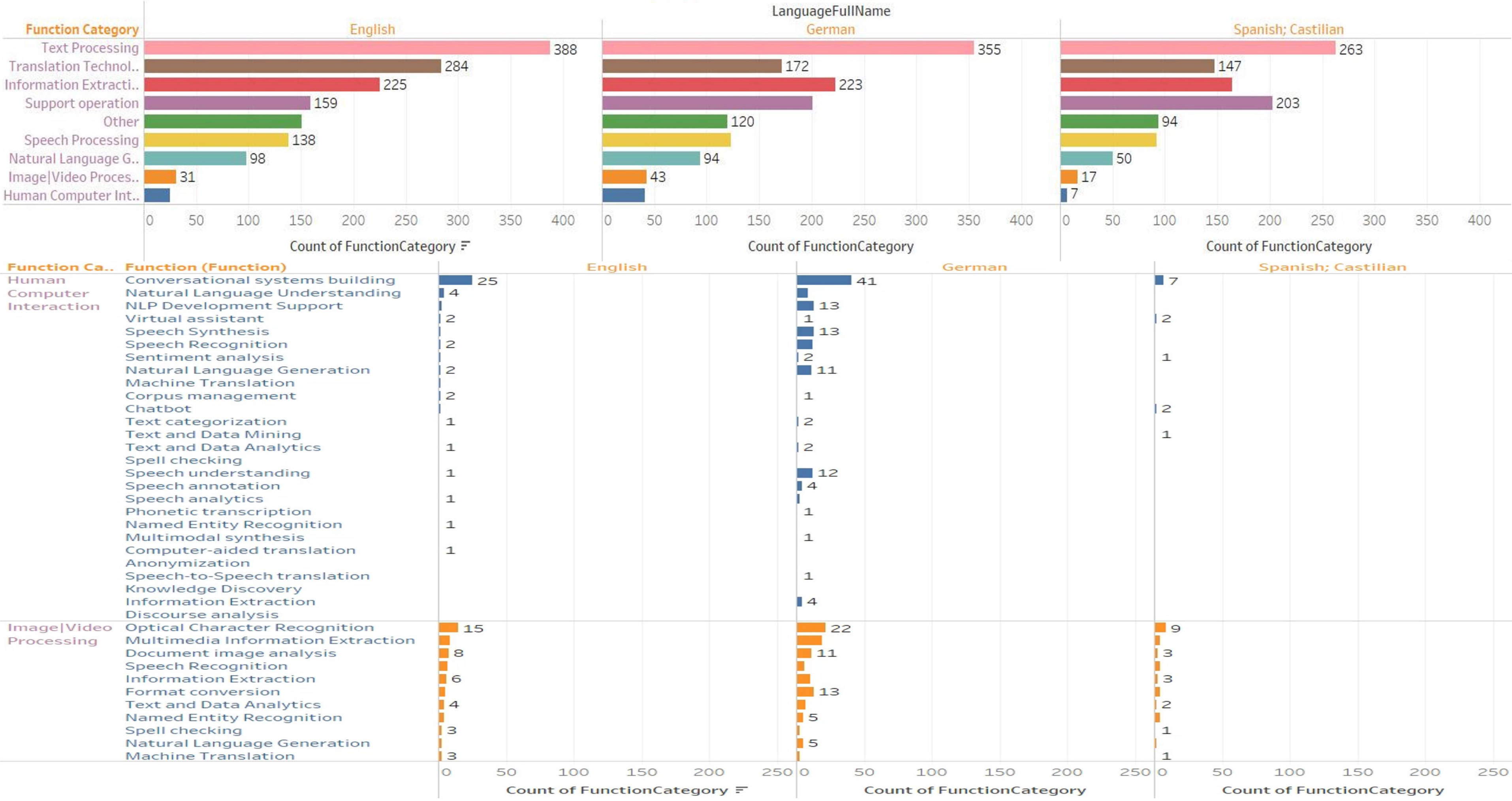
ELG Catalogue Export



O1 fx Licences

	B	G	J	K	O	R
1	**Resource Name**	**Function**	**Input Media Types**	**Input Languages**	**Licences**	**Landing Page**
379	CORDEX inflectional lookup data 1.0	undefined		sl	Creative Commons Attribution Non	http://hdl.handle.net/113
380	ANMOP	Text categorization|Text and Data	text	es		http://www.redilegra.com
381	Latvian grammar checker	Grammar checking	text	lv		https://www.tilde.lv/parei
382	COREA-coreferentieservice	Co-reference resolution		nl		http://hdl.handle.net/100
383	Collective Text to Speech	Text-to-Speech Synthesis|Speech	text	es|pt	GNU General Public License v2.0	https://pypi.org/project/c
384	Lengoo Termbase	Terminology	text	de		https://www.lengoo.com/
385	extraTerm	Term extraction	text	en|de		https://www.iailc.de/en/s
386	Korp, Kielipankki version	Concordance search	text	ru|es|fr|de|en|sv|fi|mdf|myv|sjd|swh		https://korp.csc.fi
387	Raudikko Analysis for Elasticsearch	Text and Data	text	fi	GNU Lesser General Public License	https://github.com/Evider
388	MIOPIA	Annotation|Sentiment analysis	text	es|en		https://miopia.grupolys.o
389	Across Translator Edition	Terminology	text	de		https://www.across.net/e
390	Norma	Summarization|NLP Development				http://simple4all.org/proc
391	StrokeAid	Speech Synthesis	text	hu		http://magyarbeszed.tmit
392	Recognizer	Speech understanding	audio	lt	Creative Commons Attribution 4.0	https://xn--ratija-
393	SpeCT - Speech Corpus Toolkit for Praat (v1.0.0)	Speech annotation|Text and Data			GNU Lesser General Public License	https://zenodo.org/record
394	voiceovermaker.io	Speech Synthesis	text	no|ko|ja|it|id|hu|hi|el|de|fr|fi|fil|en|		https://voiceovermaker.io
395	IRIS English-Irish Translation System	Machine Translation	text	ga|en		http://server1.nlp.insight-
396	Γ„Γ¤ni Company's Automatic Speech Recognition	Speech understanding|Multimedia	audio|t	fi|en|fi|en		https://www.aanicompany
397	iTranslate Offline Translation	Machine Translation	text	vi|tr|th|fa|ru|ko|ja|id|he|zh|bs|sq|ar		https://itranslate.com/lan

Function Category COLLECTIONS broken down by Tools per Language

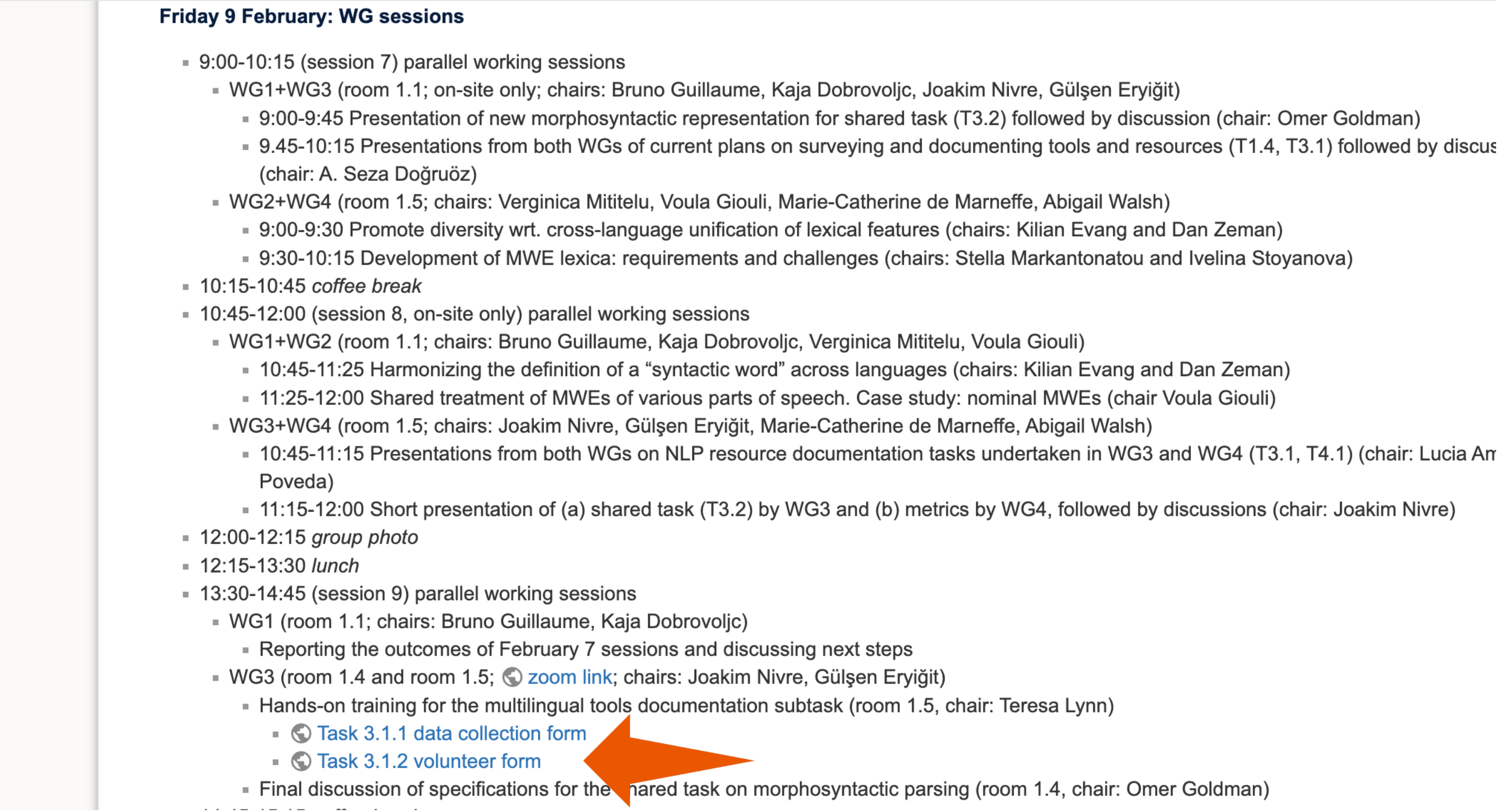


Task 3.1.2: Tool Availability Analysis

Expected Outcomes

- A better insight into current NLP tool availability
- A better insight into existing gaps and digital language inequality
- A basis for improved reporting on language support or tool availability status (visually/ written reports)

Task 3.1.2: Tool Availability Analysis – sign up!



Friday 9 February: WG sessions

- 9:00-10:15 (session 7) parallel working sessions
  - WG1+WG3 (room 1.1; on-site only; chairs: Bruno Guillaume, Kaja Dobrovoljc, Joakim Nivre, Gülşen Eryiğit)
    - 9:00-9:45 Presentation of new morphosyntactic representation for shared task (T3.2) followed by discussion (chair: Omer Goldman)
    - 9.45-10:15 Presentations from both WGs of current plans on surveying and documenting tools and resources (T1.4, T3.1) followed by discus (chair: A. Seza Doğruöz)
  - WG2+WG4 (room 1.5; chairs: Verginica Mititelu, Voula Giouli, Marie-Catherine de Marneffe, Abigail Walsh)
    - 9:00-9:30 Promote diversity wrt. cross-language unification of lexical features (chairs: Kilian Evang and Dan Zeman)
    - 9:30-10:15 Development of MWE lexica: requirements and challenges (chairs: Stella Markantonatou and Ivelina Stoyanova)
- 10:15-10:45 *coffee break*
- 10:45-12:00 (session 8, on-site only) parallel working sessions
  - WG1+WG2 (room 1.1; chairs: Bruno Guillaume, Kaja Dobrovoljc, Verginica Mititelu, Voula Giouli)
    - 10:45-11:25 Harmonizing the definition of a “syntactic word” across languages (chairs: Kilian Evang and Dan Zeman)
    - 11:25-12:00 Shared treatment of MWEs of various parts of speech. Case study: nominal MWEs (chair Voula Giouli)
  - WG3+WG4 (room 1.5; chairs: Joakim Nivre, Gülşen Eryiğit, Marie-Catherine de Marneffe, Abigail Walsh)
    - 10:45-11:15 Presentations from both WGs on NLP resource documentation tasks undertaken in WG3 and WG4 (T3.1, T4.1) (chair: Lucia An Poveda)
    - 11:15-12:00 Short presentation of (a) shared task (T3.2) by WG3 and (b) metrics by WG4, followed by discussions (chair: Joakim Nivre)
- 12:00-12:15 *group photo*
- 12:15-13:30 *lunch*
- 13:30-14:45 (session 9) parallel working sessions
  - WG1 (room 1.1; chairs: Bruno Guillaume, Kaja Dobrovoljc)
    - Reporting the outcomes of February 7 sessions and discussing next steps
  - WG3 (room 1.4 and room 1.5; zoom link; chairs: Joakim Nivre, Gülşen Eryiğit)
    - Hands-on training for the multilingual tools documentation subtask (room 1.5, chair: Teresa Lynn)
      - Task 3.1.1 data collection form
      - Task 3.1.2 volunteer form
    - Final discussion of specifications for the hared task on morphosyntactic parsing (room 1.4, chair: Omer Goldman)

UniDive

Thank you for your attention!

Go raibh maith agaibh!

Questions?

Ceisteanna?

UniDive