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)

Task Overview

Task 3.1.1

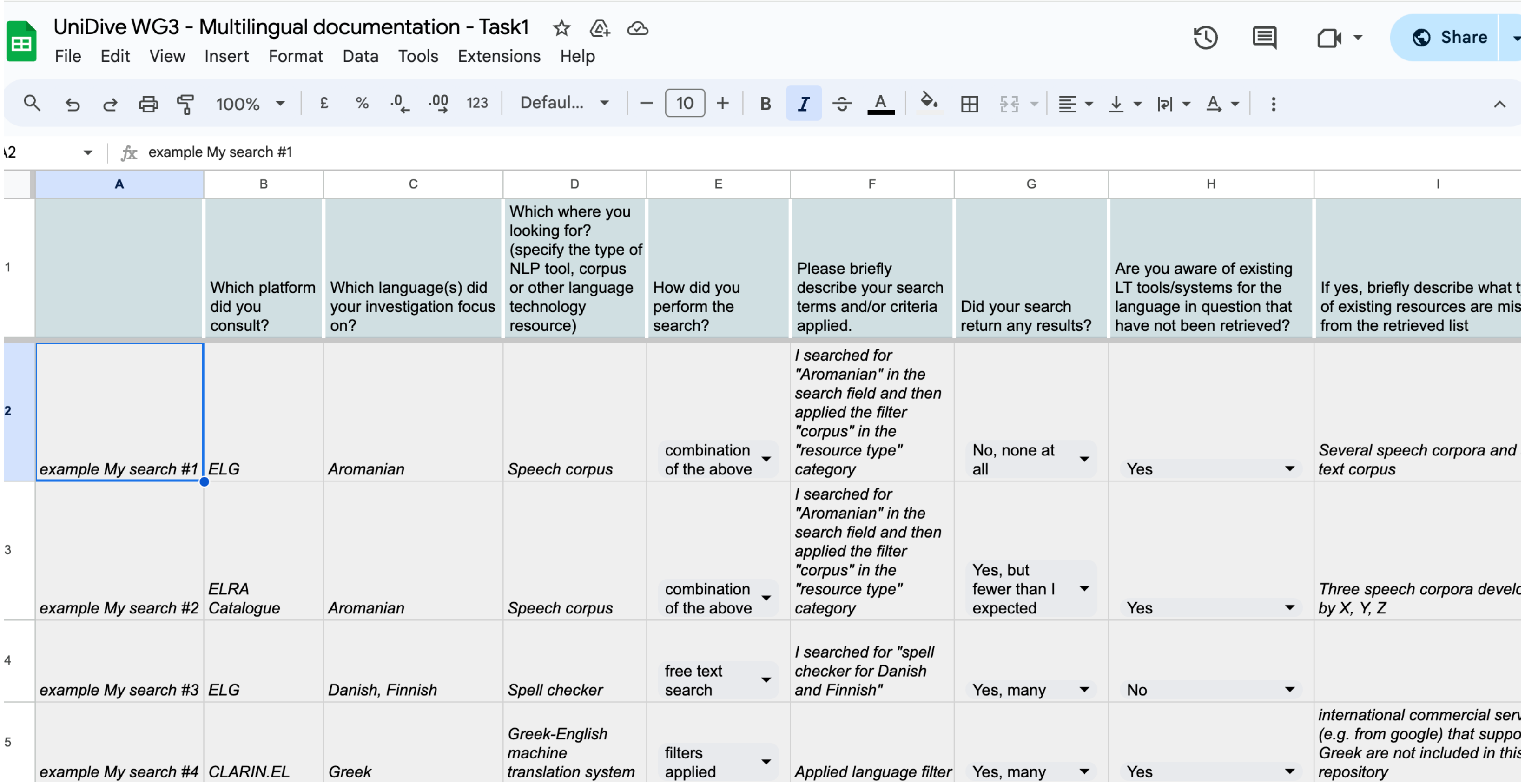
- Assess the “discoverability” of NLP tools and resources
- Who can participate?
  - Everyone 😃

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?

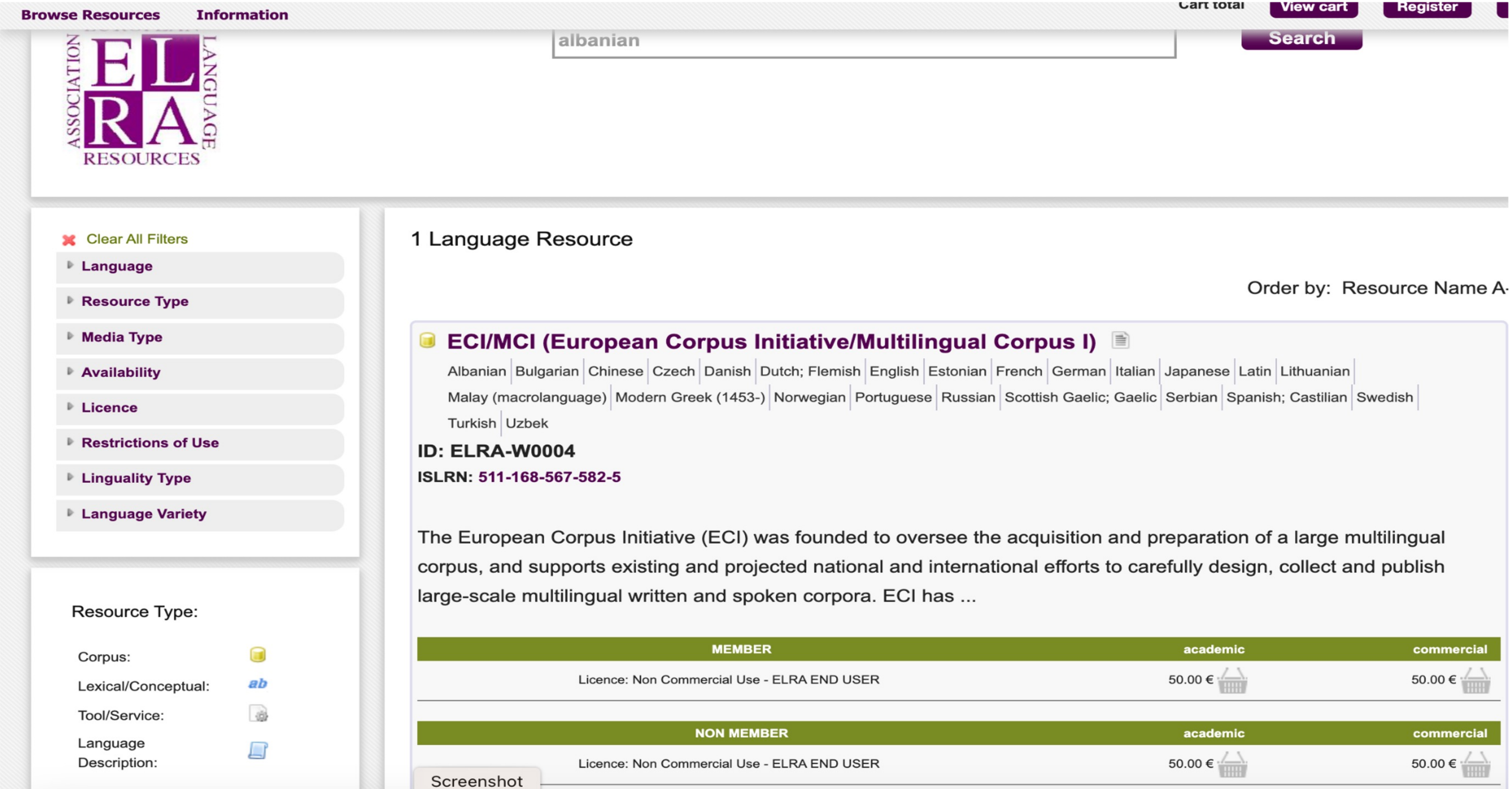


UniDive WG3 - Multilingual documentation - Task1

example My search #1

	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 develc 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



albanian

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 ...

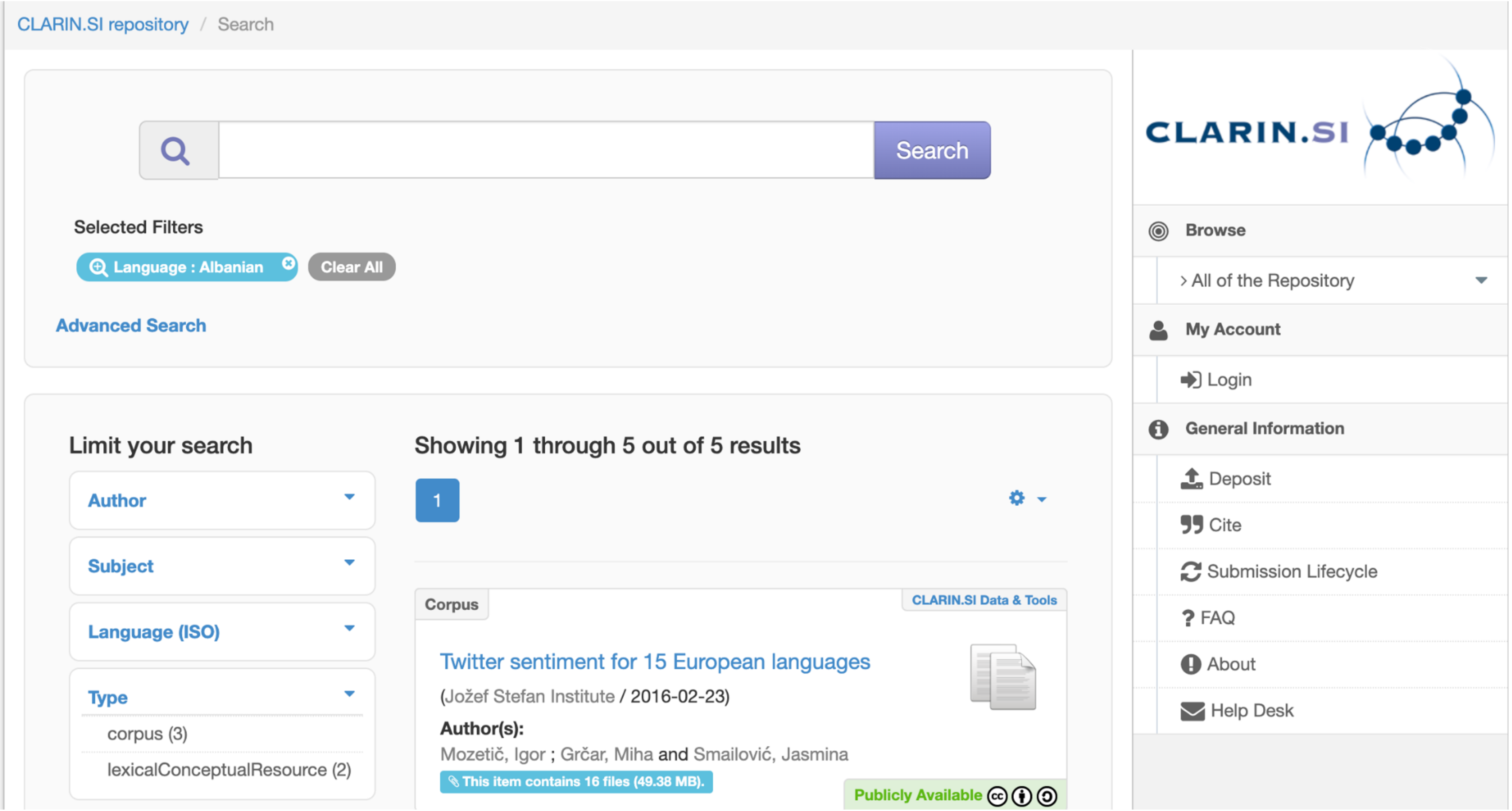
MEMBER	academic	commercial
Licence: Non Commercial Use - ELRA END USER	50.00 €	50.00 €

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

Resource Type:

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

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



CLARIN.SI repository / Search

Selected Filters

Language : Albanian

Limit your search

corpus (3)

lexicalConceptualResource (2)

Showing 1 through 5 out of 5 results

Corpus

CLARIN.SI Data & Tools

Twitter sentiment for 15 European languages

(Jožef Stefan Institute / 2016-02-23)

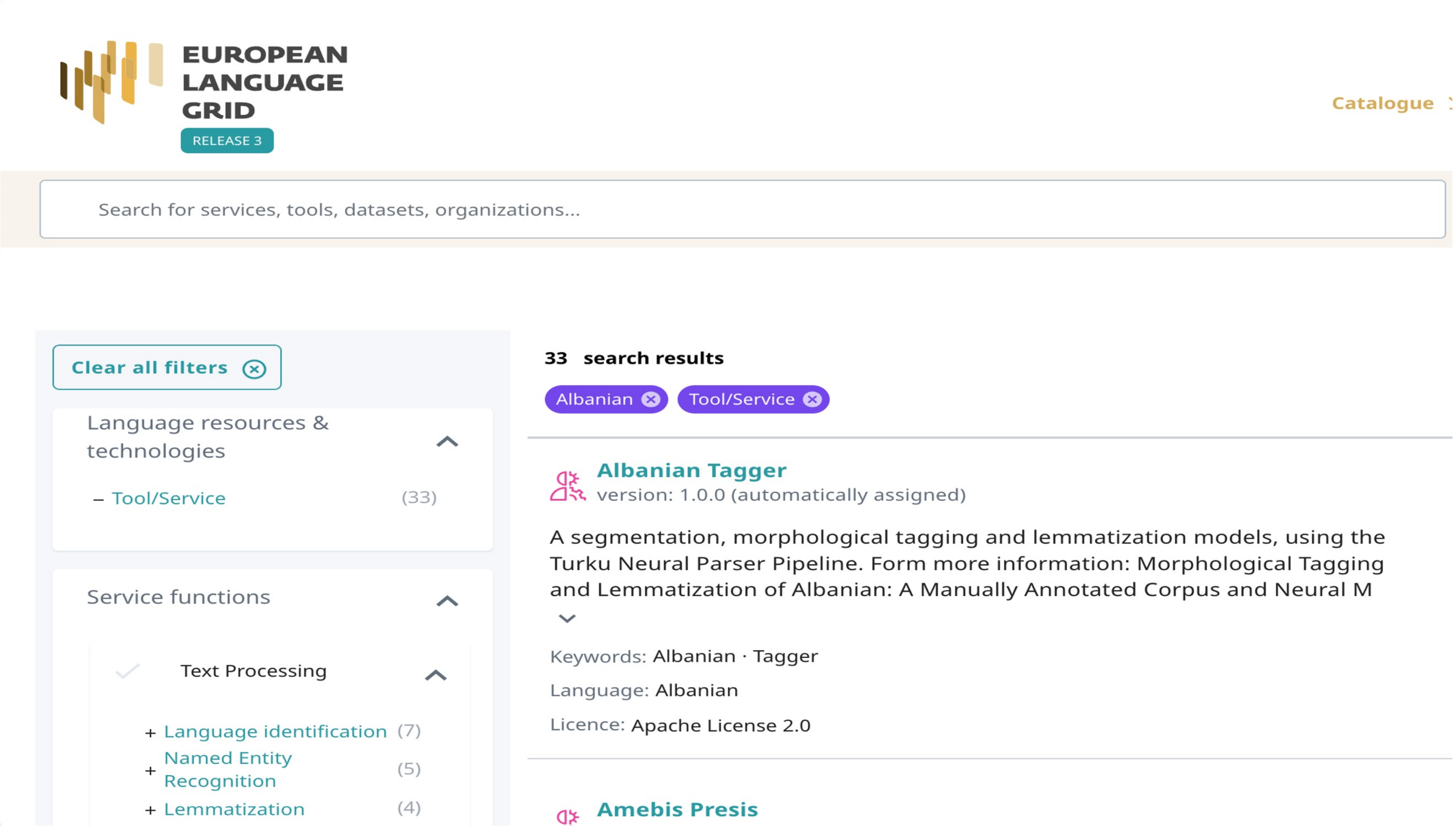
Author(s):

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



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

33 search results

Albanian ⊗ Tool/Service ⊗

Language resources & technologies

– Tool/Service (33)

Service functions

Text Processing

+ Language identification (7)

+ Named Entity Recognition (5)

+ Lemmatization (4)

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

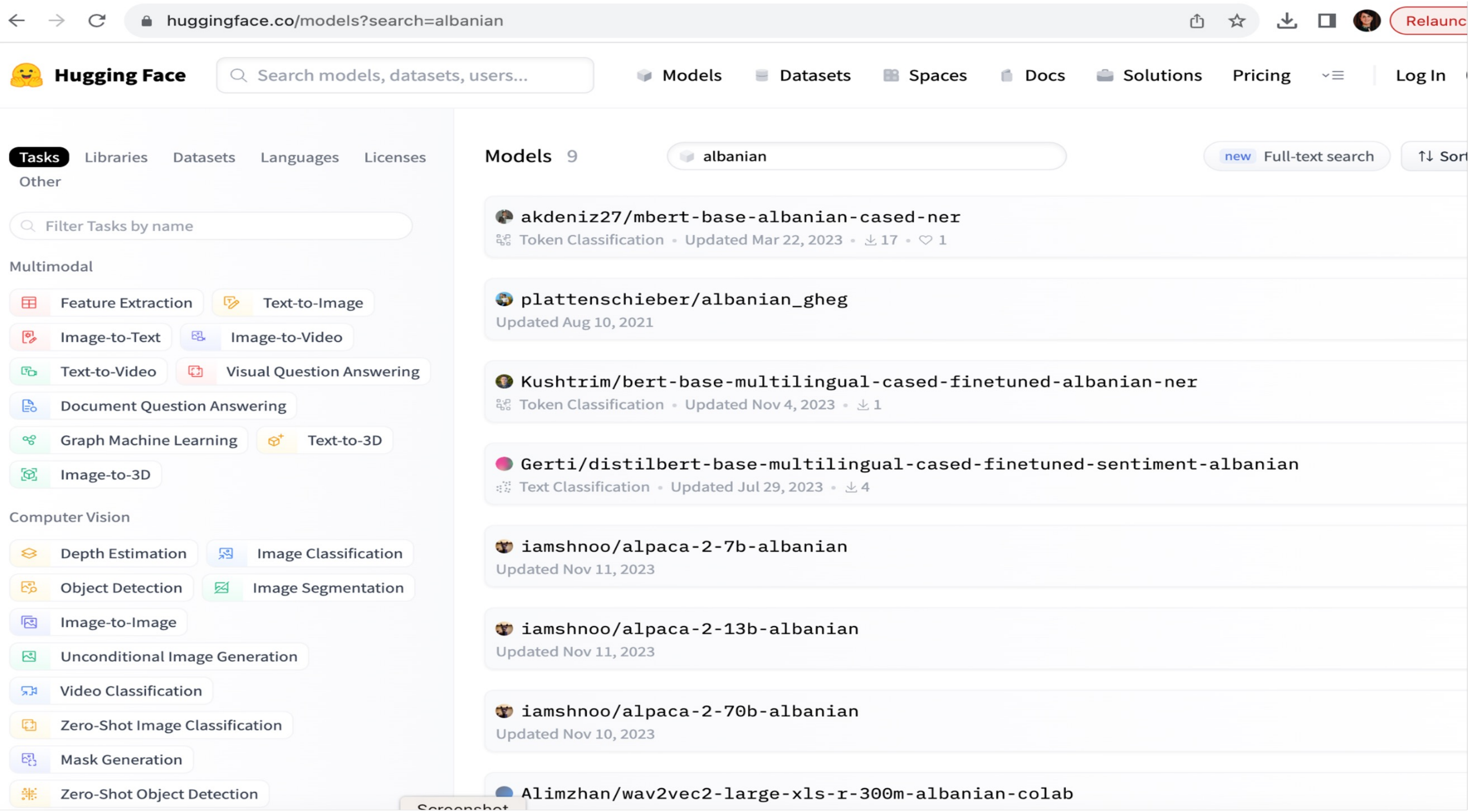
Keywords: Albanian · Tagger

Language: Albanian

Licence: Apache License 2.0

Amebis Presis

E.g Search for Albanian Tools - Hugging Face



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

Responses

- **20 responses**
- **16 countries**
- **22 languages**
- **range: 2-60 entries**
- **average: 6* entries**

Czechia
Denmark
Finland
Georgia
Germany
Hungary
Italy
Moldova
Poland
Serbia
Slovenia
Spain
Sweden
Turkey
Ukraine
United Kingdom

Albania
Armenia
Belgium
Bosnia and Herzegovina
Bulgaria
Croatia
Estonia
France
Greece
Ireland
Israel
Latvia
Lithuania
Malta
Netherlands
North Macedonia
Norway
Portugal
Romania
Slovakia
Switzerland

Language Coverage

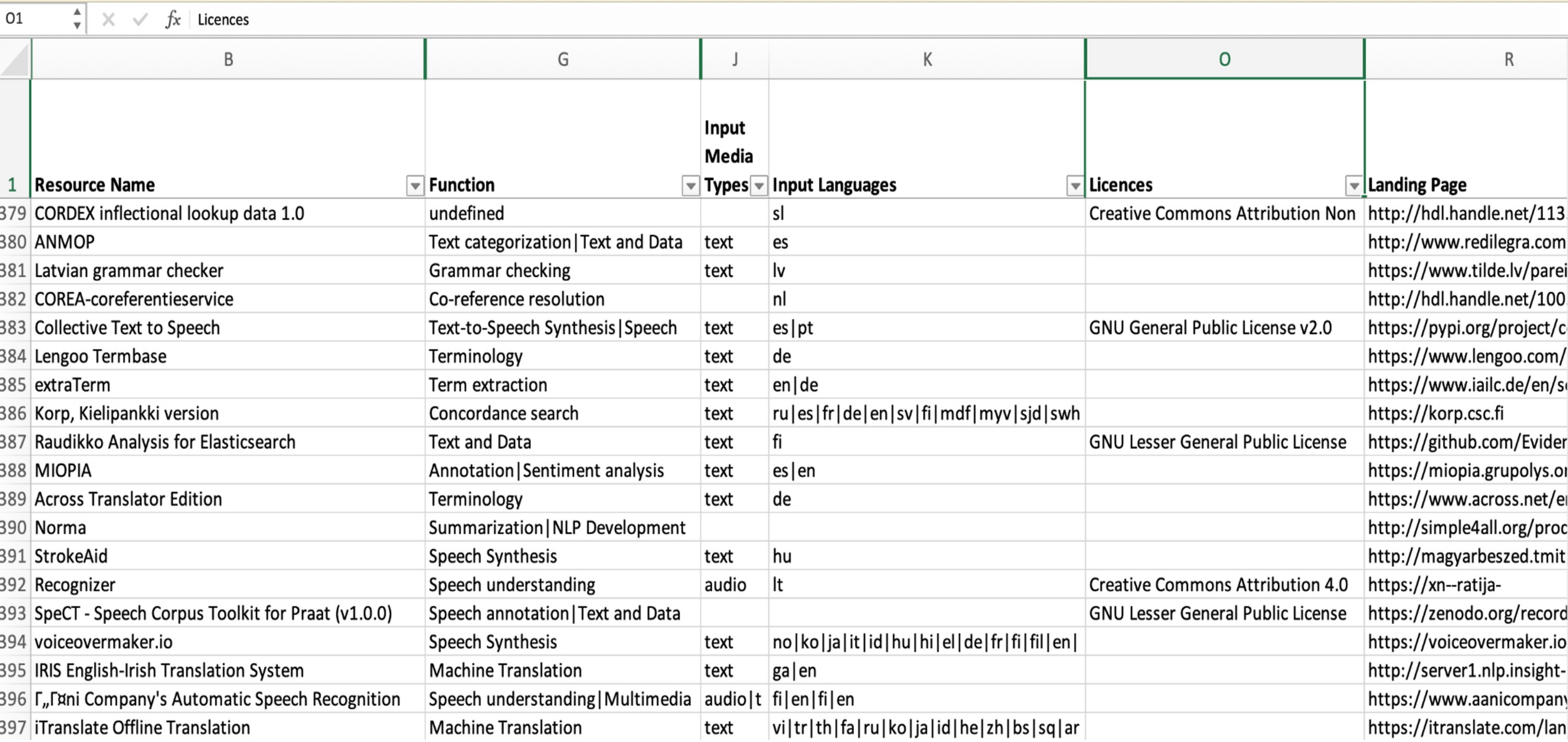
- Ancient Egyptian
- Czech
- Danish, Frisian
- Georgian, Megrelian, Svan
- Hungarian
- Mansi
- Moldovan, Romanian
- Old Italian, Old Florentine
- Polish
- Portuguese
- Serbian
- Slovak
- Slovenian, (spoken) English
- Swedish
- Turkish
- Ukrainian

- Platforms: ELG, ELRA, CLARIN, Portulan (CLARIN), Hugging Face
- *Are you aware of existing LT tools/systems for the language in question that have not been retrieved?*
  - Predominantly YES
- Please add your entries to commonly used portals ☺

Task 3.1.2: Tool Availability Analysis

- Seeking volunteers with strong Excel/ Tableau skills
- Analysis required on ELG catalogue export
- Ideas below can be the start of investigation - let's see what else emerges:
    1. The tools that 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.or
389	Across Translator Edition	Terminology	text	de		https://www.across.net/er
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

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)

Data Visualisation Report:

Dr Ivelina Stoyanova
Department of Computational Linguistics
Institute for Bulgarian Language