



UniDive WG3

Subgroup - Multilingual Tool and Resource Documentation

Co-leaders:

A. Seza Dođruöz (Ghent University)

Maria Giagkou (Athena RC)

Teresa Lynn (Mohamed bin Zayed University of Artificial Intelligence)





Task Overview

Task 1

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

Task 2

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



Task 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 (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?


** (Activity will kick-off in Naples)

E.g Search for Albanian Tools - ELRA Catalogue

Browse Resources Information

Cart total View cart Register

albanian Search



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 € 

Screenshot

Clear All Filters

- Language
- Resource Type
- Media Type
- Availability
- Licence
- Restrictions of Use
- Linguality Type
- Language Variety

Resource Type:

- Corpus: 
- Lexical/Conceptual: 
- Tool/Service: 
- Language: 
- Description: 

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

CLARIN.SI repository / Search

Search

Selected Filters

Language : Albanian Clear All

[Advanced Search](#)

Limit your search

Author ▼

Subject ▼

Language (ISO) ▼

Type ▼

- corpus (3)
- lexicalConceptualResource (2)

Showing 1 through 5 out of 5 results

1 ⚙️

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

Browse

> All of the Repository ▼

My Account

Login

General Information

- Deposit
- Cite
- Submission Lifecycle
- FAQ
- About
- Help Desk

E.g Search for Albanian Tools - ELG Catalogue



Catalogue

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

Clear all filters

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

The screenshot shows the Hugging Face website interface. At the top, the browser address bar displays 'huggingface.co/models?search=albanian'. The Hugging Face logo and navigation menu are visible. The search bar contains the text 'albanian'. On the left sidebar, the 'Tasks' section is active, with a search filter 'Filter Tasks by name'. Below this, 'Multimodal' tasks are listed, including 'Text-to-Image', 'Image-to-Text', 'Image-to-Video', 'Text-to-Video', 'Visual Question Answering', 'Document Question Answering', 'Graph Machine Learning', and 'Text-to-3D'. 'Computer Vision' tasks include 'Depth Estimation', 'Image Classification', 'Object Detection', 'Image Segmentation', 'Image-to-Image', 'Unconditional Image Generation', 'Video Classification', 'Zero-Shot Image Classification', 'Mask Generation', and 'Zero-Shot Object Detection'. The main content area shows a list of models under the 'albanian' search filter. The models listed are:

- akdeniz27/mbert-base-albanian-cased-ner**
Token Classification • Updated Mar 22, 2023 • 17 downloads • 1 heart
- plattenschieber/albanian_gheg**
Updated Aug 10, 2021
- Kushtrim/bert-base-multilingual-cased-finetuned-albanian-ner**
Token Classification • Updated Nov 4, 2023 • 1 download
- Gerti/distilbert-base-multilingual-cased-finetuned-sentiment-albanian**
Text Classification • Updated Jul 29, 2023 • 4 downloads
- 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 1: Assessing the “discoverability” of NLP tools

Process

- A template will be provided with prompt questions
- Additional input also desired

Outcome

- 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 2: Tool Availability Analysis

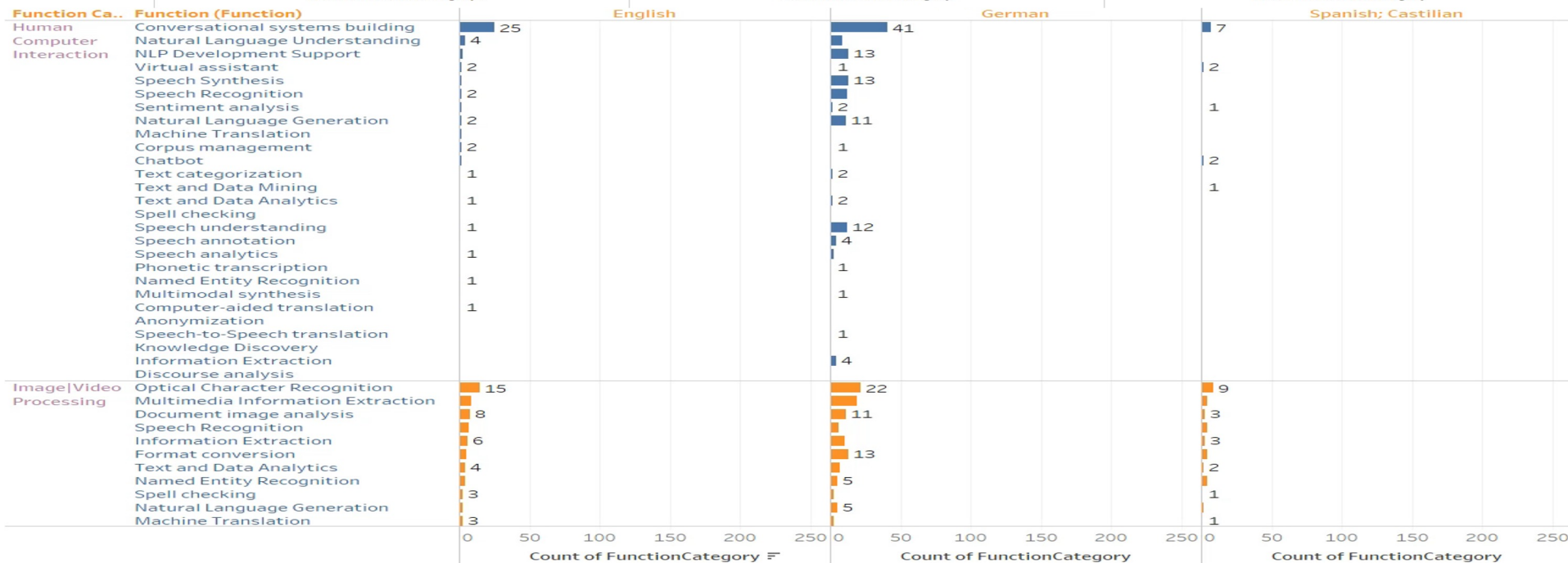
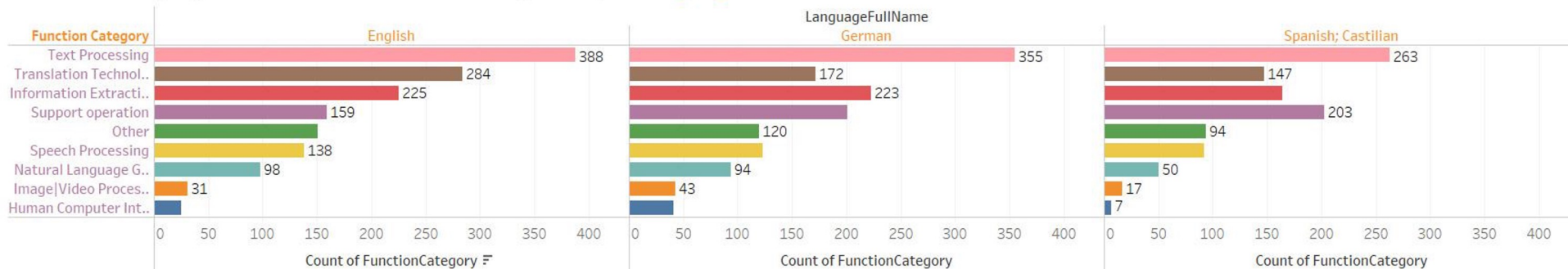
- Seeking volunteers with strong Excel/ Tableau skills
- Analysis required on ELG catalogue export - similar to Kristina's report as starting point
- Prompts below can be the start of investigation - let's see what else emerges:
 1. The tools that certain languages are missing (e.g. Irish doesn't have 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

01 x ✓ fx Licences

	B	G	J	K	O	R
1	Resource Name	Function	Input Media	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 md myv s j sw h		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/en
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	Г„Гӑni 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 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)



Thank you for your attention!

Questions?