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 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

- Kicked-off in Naples
- Template provided with instructions for semi-guided searches
  - 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 it easily? 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

UniDive WG3 - Multilingual documentation - Task3.1.1

File Edit View Insert Format Data Tools Extensions Help

Share

Menus 100% Defaul... 10

A6

	A	B	C	D	E	F	G	H	I	J	K	L
1		Which platform did you consult?	Which language(s) did your investigation focus on?	What 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 type of existing resources are missing from the retrieved list	What metadata (if any) is missing that you consider critical for your search?	What kind of difficulties did you face in this search effort?	Further comments and feedback you share
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 a text corpus*	*NA*	*I couldn't find any speech corpora*	*I'm disappointed th corpora are not on to help contribute t catalogue further w data*
3	*example My search #2*	*Portulan Clarin https://portulanclarin.net/workbench/*	*Portuguese*	*written corpus*	free text search	*I searched for "inocente" to retrieve instances in context*	Yes, but fewer than I expected	No	*Three speech corpora developed by X, Y, Z*	*- geographic variance for Aromanian was not specified -homepage URL for original resource missing*	*A Captcha was requested at each query and this makes the search slower*	
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	*NA*	*The Danish spell-checker's meta-data was missing terms-of-use information*	*None*	
5	*example My search #4*	*CLARIN.EL*	*Greek*	*Greek-English machine translation system*	filters applied	*Applied language filter*	Yes, many	Yes	*international commercial services (e.g. from google) that support Greek are not included in this repository*	*None*		*I would prefer to be full list of open-sou commercial tools*
6												
7												
8												
9												
10												
11	...											
12												
13												
14												
15												
16												
17												
18												

1. READ ME Guide | 2. Meta-data examples | 3. Your input

Observations

- 9 participants
- 68 search logs
- Repositories/platforms consulted:
  - ELG (27)
  - Clarin (24)
  - ELRA Catalogue (8)
  - LDC catalogue (6)
  - Hugging face (2)
  - https://corpus-analysis.com/ (1)

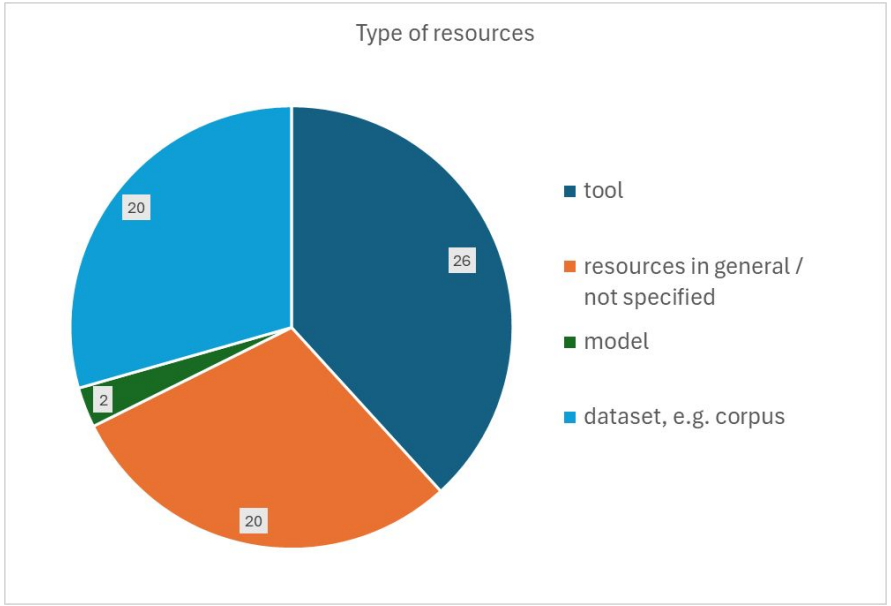
→ Only platforms provided as examples were consulted - one exception

→ Perhaps we have been too prescriptive in our guidelines?

Observations

- Languages searched for:
  - Ancient Egyptian
  - Ancient Italian
  - English
  - Florentine
  - Old Florentine
  - Old Italian
  - Polish
  - Portuguese
  - Serbian
  - Swedish
  - Turkish

- Resources searched for:



→ the repositories suggested are not appropriate for diachronic linguistic studies (very few tools and resources for extinct languages)

→ 7 logs (by one respondent) were attempts to actually run a NLP service offered through ELG and CLARIN → out of scope for this task

Selected feedback

→ Models reported as missing

→ Some tools retrieved are outdated

→ Metadata issues reported:

- ❏ geographic variants not always available (e.g. Brazilian vs European Portuguese)
- ❏ missing time coverage to facilitate diachronic studies
- ❏ limited metadata specific to language models
- ❏ keywords not specific enough
- ❏ level of annotation

→ Many irrelevant results, low precision scores especially for free text searches

→ Respondents expected that tools would run natively on platform, not just described

Thank you for your attention!

Questions?

UniDive