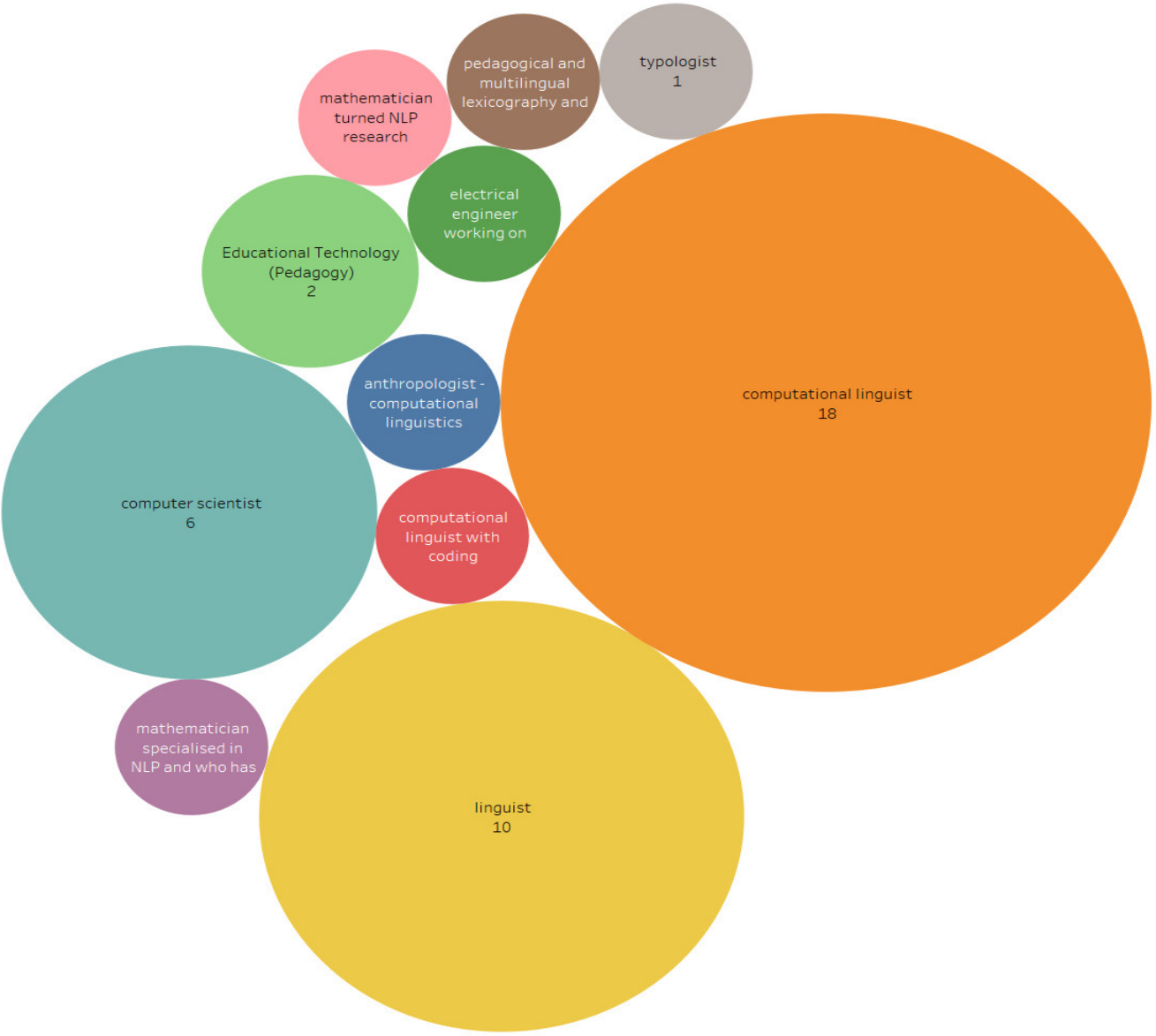
UniDive: WG3

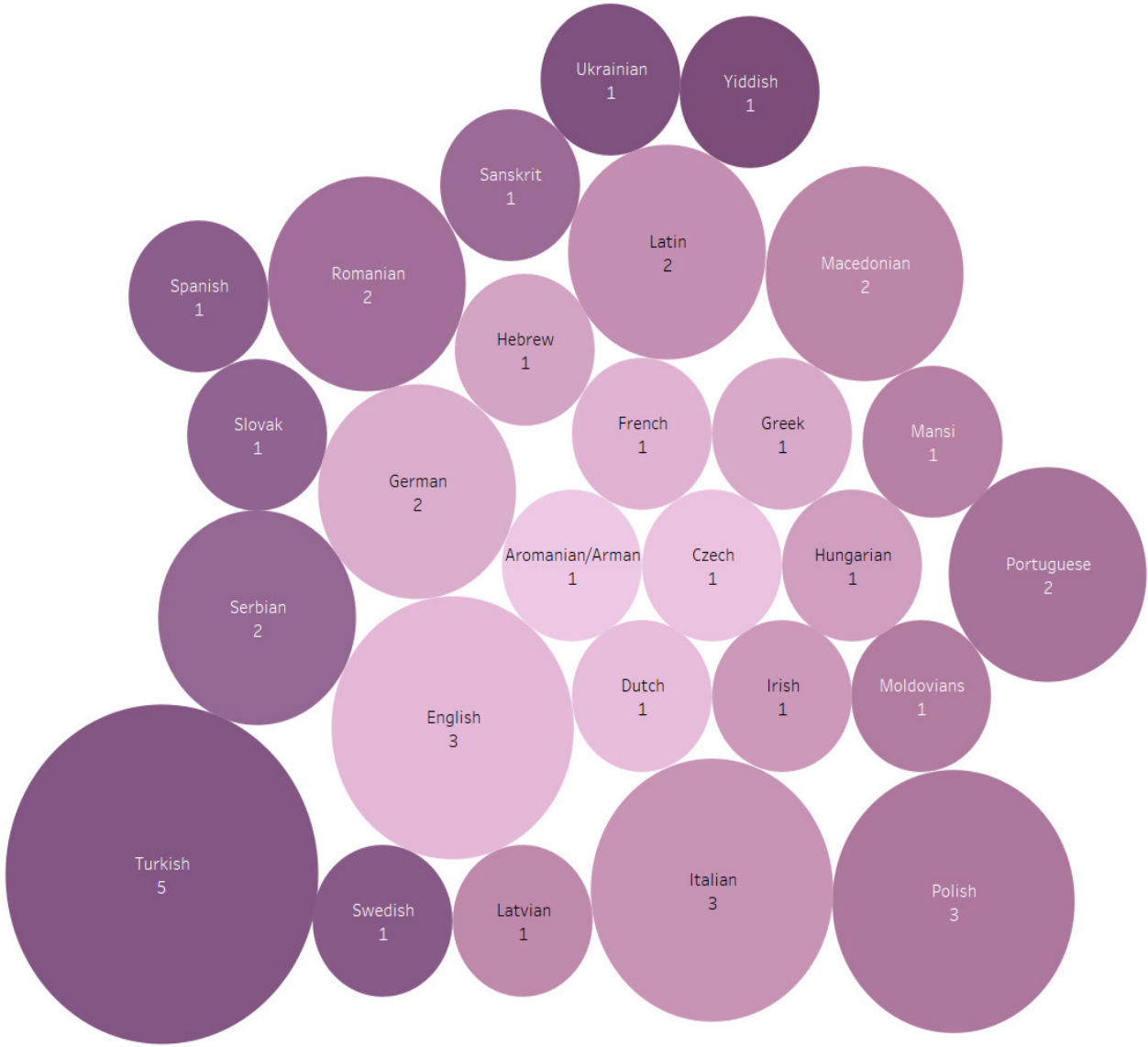
TASK 1

Learning about volunteers and their preferences on

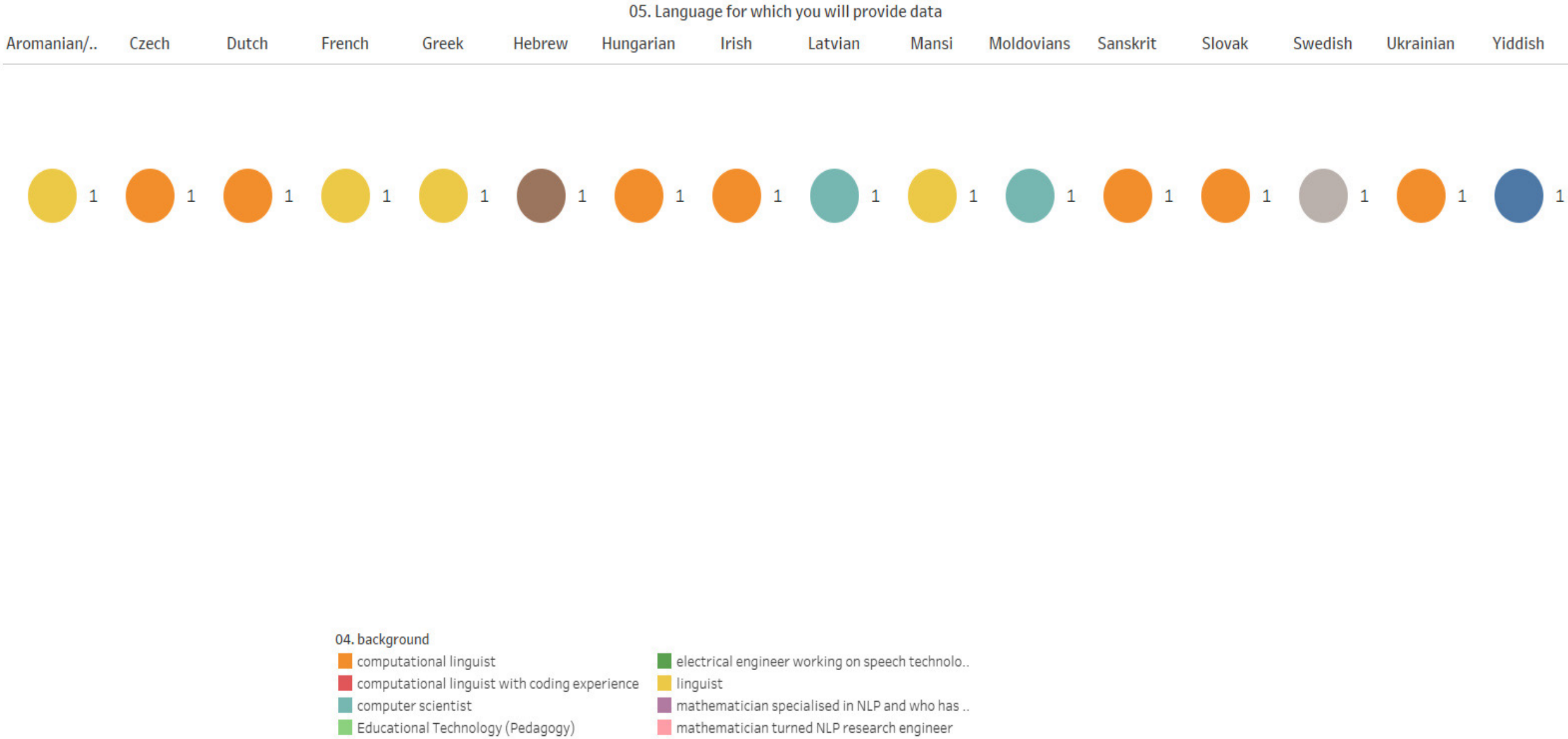
language technology platforms

Language specialists vs Languages





Type and number of experts for each language



Additional notes

Historical languages & language varieties:

*The question about rating the skills does not match perfectly with the nature of the **Latin** language, since it is a historical language with no native speakers. Thus, e.g. speaking is definitely less relevant than reading, which can be considered the main skill.*

***Latin** is not really a living language, though still in use in some milieus, so listening/speaking/writing skills are not applicable nor actually relevant*

*due to historic challenges, **Aromanian** language is not fully standardized and it is only in North Macedonia recognized as such. Other regional countries do not recognize it as language (Romania and Greece), but as historic linguistic variety (Albania) or just local dialect (Greece). Please consider this work as neutral approach, to be able to capture the language, without provoking any cultural misunderstandings, since it is highly sensitive matter in the region.*

***Portuguese** has different varieties, European, Brazilian and African varieties. Although all the varieties are low-resource, the later are even more. So it would be great to be able to work all the varieties.*

Language proficiency

05. Language ..	09.a. Readin..	09.b. Writin..	09.c. Listenin..	09.d. Speakin..	
Aromanian/Ar..	2	1	2	2	1
Czech	5	5	5	5	1
Dutch	5	5	5	5	1
English	3	3	3	3	1
	4	4	4	4	1
	5	5	5	5	1
French	5	5	5	5	1
German	3	2	1	2	1
	5	5	5	5	1
Greek	5	5	5	5	1
Hebrew	5	5	5	5	1
Hungarian	5	5	5	5	1
Irish	4	2	3	3	1
Italian	5	5	5	5	3
Latin	4	3	1	3	1
	5	3	1	1	1
Latvian	5	5	5	5	1
Macedonian	5	5	5	5	2
Mansi	5	4	4	3	1
Moldovians	5	5	5	5	1
Polish	5	5	5	5	3
Portuguese	5	5	5	5	2
Romanian	5	5	5	5	2
Sanskrit	3	2	2	2	1
Serbian	5	5	5	5	2
Slovak	5	5	5	5	1
Spanish	5	5	5	5	1
Swedish	3	3	3	1	1
Turkish	5	5	5	5	5
Ukrainian	5	5	5	5	1
Yiddish	1	1	1	1	1

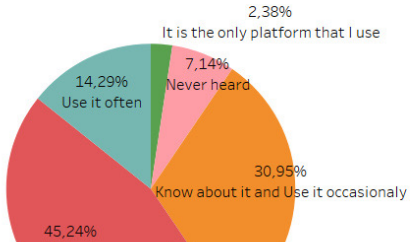
Language resource

platforms

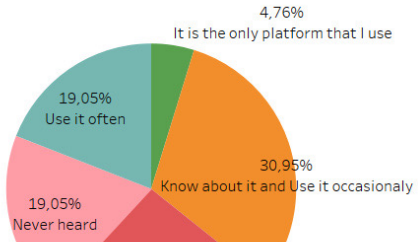
USAGE

Familiarity with Language platforms

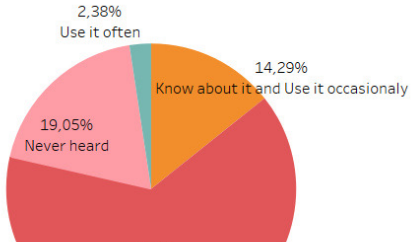
[CLARIN - Common LAnguage Resources and Technology INfrastructure]



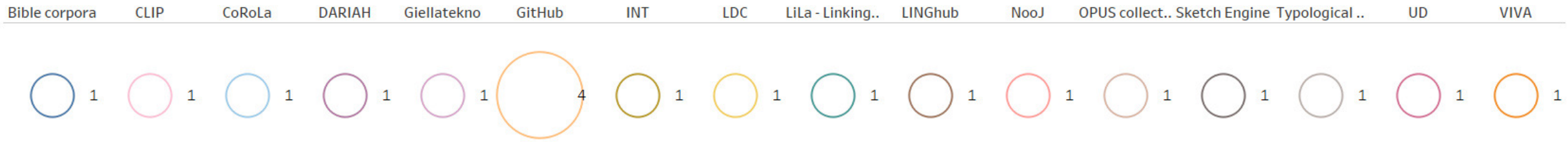
[Hugging Face]



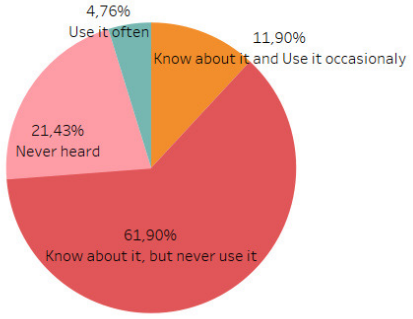
[ELRA - European Language Resource Association]



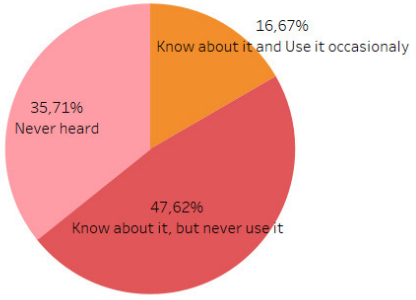
What other LT platforms do you know of?



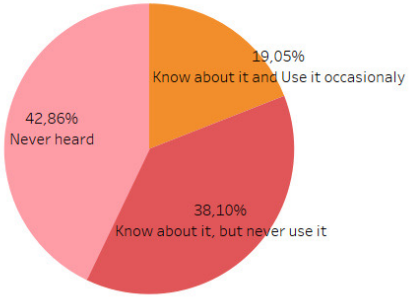
[ELG - European Language Grid]



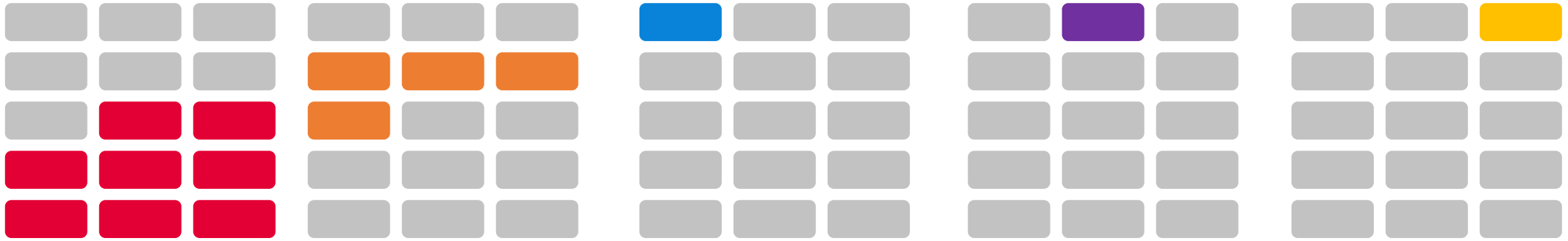
[META-SHARE]



[ELRC-SHARE - European Language Resource Coordination repository]



Reasons for not using available language resources



53%

Available resources are **outside** my research **domain**

26%

It would require too much time or additional knowledge to adjust the **format** of available resources to my research needs

7%

I **don't trust** the resources that I have not prepared myself.

7%

Available resources have too many **legal dilemmas**.

7%

Available resources are **outside** my research **domain**

It would require too much time or additional knowledge to adjust the **format** of available resources to my research needs

Available resources have too many **legal dilemmas**.

Language platforms

FEATURES

Important features on a language technology platform

SPECIALIST	CORPUS **DESCRIPTION**	TYPE OF **ANNOTATION**	**LICENCE**	CORPUS **MEDIA TYPE** *[written, spoken, multimodal…]*	CORPUS **SIZE**	CORPUS **DOMAIN**	LANG. **NAME**	CORPUS **TYPE** *[PDF, TXT, CSV…]*	**AUTHOR**SHIP	CORPUS **NAME**	CORPUS RECORDING **PROVENANCE**	LANG. **FAMILY**
Anthropologist												
Computational Linguist												
Computer Scientist												
Pedagogist												
Speech Technologiest												
Linguist												
Mathematician												
Lexicographer												
Typologist												

Important features on a language technology platform

Specialist	additional metadata you consider important
Computational Linguist	▪ direct link to data or service ▪ language code (ISO 639, maybe also Glottolog); ▪ dialect, register, historical language stage (may or may not be covered by "corpus domain"); ▪ is it code-switching? ▪ is it parallel corpus? is it a non-corpus (lexicon, ontology etc.) ▪ manually reviewed vs. created/annotated automatically ▪ document the use case of the data ▪ if the dataset has been annotated/modified and reuploaded ▪ script - if it has been transliterated, or if the language can be written in multiple scripts ▪ taxonomic morphosyntactic tags for all parts of speech
Computer Scientist	▪ Level of quality of the data (with a clear indication of the estimation method) ▪ Research question or task to which the data/tools can contribute ▪ Source data, in case a data set adds annotations to a pre-existing corpus ▪ OCR quality, annotators skills, selection criteria ▪ Old - Modern language identifier. At least from which century is resourced collected. ▪ Structural annotations: sentence tags, paragraph tags.
Speech Technologiest	▪ if it's a speech corpus then number and sex of speakers, amount of data available for each speaker, facilities used for the recording
Linguist	▪ for corpora in which several domains etc exist it is important for the user to know to which domain etc. each sentence/document belongs to
Mathematician	▪ not exactly meta data but, if available, annotation guideline should be included.
Pedagogist	▪ linguistic inclusion of gender

Expectations from a language technology platform

SPECIALIST	Good documentation	Data quality	Licence clearly stated	Taxon. Of resources	Taxon. Of tools	Taxon. of research questions	Resource download	Dynamic use of resourc.	Unified access to resources & tools	Easy to use	Good filtering / search tools	Single place to rule them all *(to annotate at all levels)*	Examples of use
Anthropologist			1/1								1/1		
Computational Linguist	4/19	4/19	2/19	2/19	2/19		2/19		2/19	7/19	1/19	1/19	
Computer Scientist	1/6	1/6	1/6	1/6	1/6	1/6	1/6	1/6	1/6			1/19	
Pedagogist										1/2			
Speech Technologiest							1/1						
Linguist	2/10		1/10						1/10	3/10		4/10	1/10
Mathematician			1/2							2/2	1/2		
Lexicographer													
Typologist													

UniDive: WG3

TASK 1

Learning about volunteers and their preferences on

language technology platforms