CA21167 - Universality, diversity and idiosyncrasy in language technology (UniDive)

UniDive: WG3

TASK 2

Learning about **ELG** and

data behind it

Number of resources per year

2020 - 92

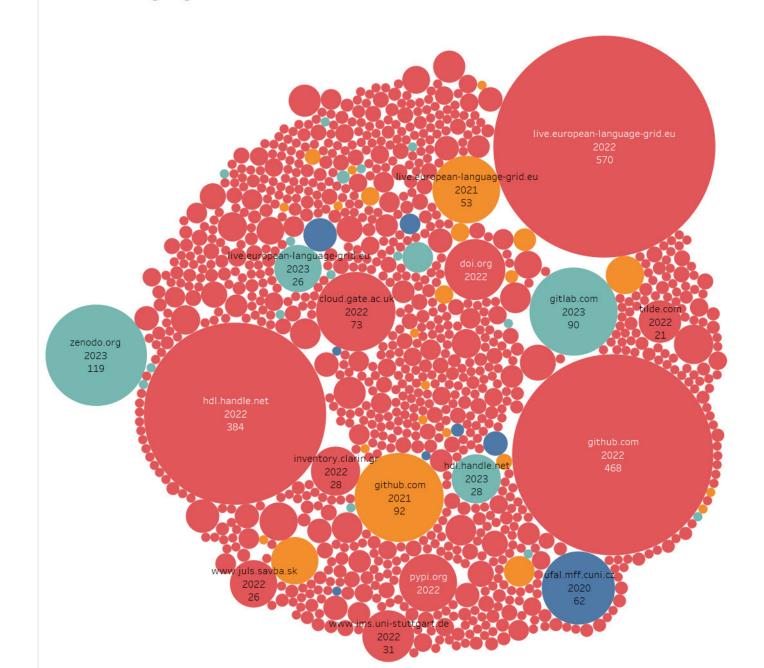
2021 - 237

2022 - 3 256

2023 - 290...

TOTAL - 3 875

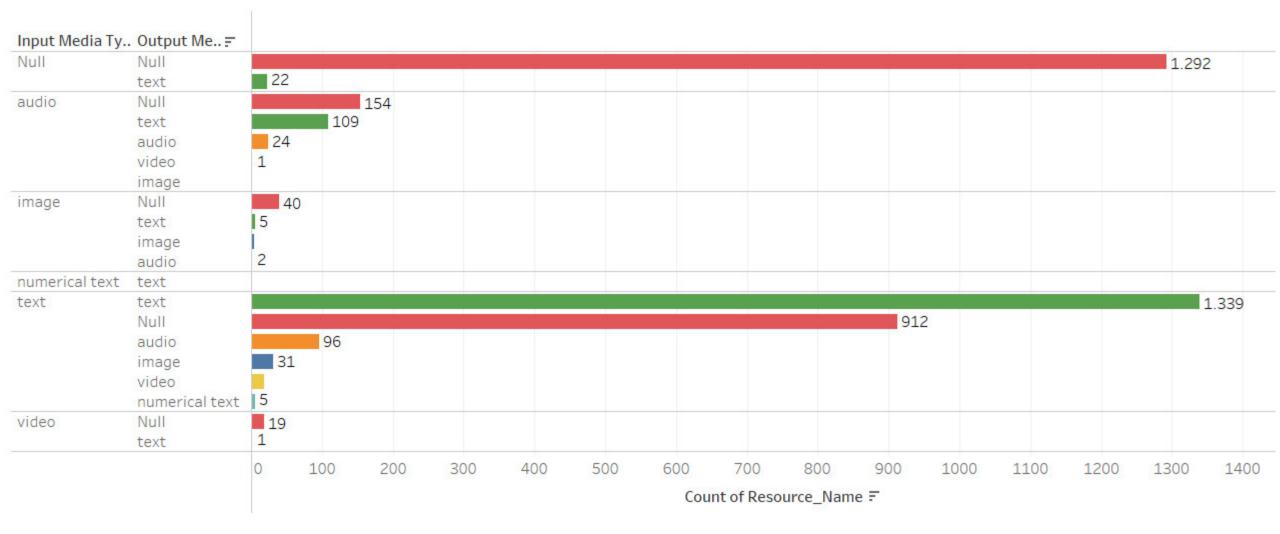
Number of resources added per **year** and its **Landing Page**



Resour	ce providers	for each l	angua	ge			Resou	rce providers	for each lang	Resource providers for each lang					
Inpu ∓	LanguageFullName									Inpu ∓	LanguageFullName	2000-1-1000-1	Inpu ∓	LanguageFullNa =	
en	English	CBC08 CB100	60 0		•					he	Hebrew	01000	sd	Sindhi	
de	German	(31913000) 0	0 0 0						•	fa	Persian	(000)	sn	Shona	
es	Spanish; Castilian	001001010000	•							vi	Vietnamese		st	Sotho, Southern	
fr	French	C000000 0		•						mk	Macedonian	(30) 0 00	wa	Walloon	• •
Null	Null	0.0000				•	•	•		ur	Urdu	(0000)	ay	Aymara	
pt	Portuguese	GH0303100	•							th	Thai	(000)	ce	Chechen	•
nl	Dutch; Flemish	GIO GEORGE				•				la	Latin	(00 00	ha	Hausa	
it	Italian	000000 0 0								af	Afrikaans	00 00 0	ig	Igbo	•
fi	Finnish	GBEED 000 0	0	•						be	Belarusian	000 0	kl	Kalaallisut; Greenla	
pl	Polish	60000 60 60	•							sq	Albanian	(3000	Ig	Ganda	•
ca	Catalan; Valencian	GD20010 050	• •							fy	Western Frisian	00	In	Lingala	•
SV	Swedish	000000000	0							lb	Luxembourgish; Letz.		nv	Navajo; Navaho	•
CS	Czech	6999000000				•				nb	Norwegian Bokmål		ny	Chichewa; Chewa; N.	
hu	Hungarian	G1910 0 G19		•						ta	Tamil	600 0	OS	Ossetian; Ossetic	•
ru	Russian	60 66600 0								bn	Bengali	000		Pali	
da	Danish	001010 00								mr	Marathi		pi	Samoan	
el	Greek, Modern (145	6016001610	0							te	Telugu		sm	Volapük	•
sl	Slovenian	(0000000 0	0							kk	Kazakh		VO	Abkhazian	
et	Estonian	00000 00000	0 0		•					hy	Armenian	(D 00	ab		
It	Lithuanian	GE000 00	•	60						bs	Bosnian	0.0	bm	Bambara	
ro	Romanian; Moldavia.	G0000 0 0 0								mul	Null		ee	Ewe Guarani	•
bg	Bulgarian	CO 000 COO 00								se	Northern Sami	600 60	gn	Hiri Motu	
hr	Croatian	C200 0000								nn	Norwegian Nynorsk	0 000	ho	Komi	•
sk	Slovak	GEG010130								OC	Occitan (post 1500)	(00	kv		
Iv	Latvian	GEOSCO - 00			•					cmn	Null		or	Oriya	•
ar	Arabic	G000000									Gaelic; Scottish Gael.		rw	Kinyarwanda	
zh	Chinese	(0000)								gd ms	Malay	(00 0	sg	Sango	•
eu	Basque	60000 O	0							tl	The state of the s	0.0	SU	Sundanese	
no	Norwegian	68100 0								ka	Tagalog Georgian	680	ti	Tigrinya	•
су	Welsh	G0800 0				0					Pushto; Pashto	0100	tn	Tswana	
gl	Galician	CiTI	US							ps	Uighur; Uyghur		to	Tonga (Tonga Island.	
	Galician; Gallegan	60880 00								ug		•	+-	Tonga (Nyasa)	
ja	Japanese	0.000								an br	Aragonese Breton		ts	Tsonga	
tr	Turkish	C30000								ml	Malayalam	• •	ty	Tahitian	
is	Icelandic	GDD • •	•								Null		ak	Akan	
uk	Ukrainian	600 0								vro	Esperanto		av	Avaric	
sr	Serbian	(30)00	0							eo	Faroese	60 0	bh	Bihari languages	
ga	Irish	CDO 0 CDO	•							TO	Null	•••	ie	Bihari	
ko	Korean	(00000)								grc	Kannada		ie	Interlingue; Occiden.	. 💌
mt	Maltese	COO CO O								kn sh	Serbo-Croatian		kg	Kongo	
hi	Hindi	600 00									Null	000	li	Limburgan; Limburg.	
id	Indonesian	(00)0								ast		•	SS	Siswant; Swati; Swa	
1	11 1	-								swh	Null		xh	Xhosa	•
		0 10	20	30		40 50 Count of Reso		70 80	90 100			0 10 2			0 10

Resour	ce providers f	for each langเ	ıage		Resou	rce providers	for each language	Resou	rce providers	for each lan		
Inpu ∓	LanguageFullNa 🗲						Inpu ∓	LanguageFullNa 🗲		Inpu ∓	LanguageFullNa 🗲	
en	English				• •		he	Hebrew		sd	Sindhi	
de	German					•	fa	Persian		sn	Shona	
es	Spanish; Castilian		•				vi	Vietnamese	227 7	st	Sotho, Southern	
fr	French		•				mk	Macedonian	*** ****	wa	Walloon	
Null	Null			•	•	•	ur	Urdu		ay	Aymara	
pt	Portuguese						th	Thai		ce	Chechen	
nl	Dutch; Flemish			•			la	Latin		ha	Hausa	
it	Italian						af	Afrikaans		ig	lgbo	
fi	Finnish		•				be	Belarusian		kl	Kalaallisut; Greenla	
pl	Polish		•				sq	Albanian		lg	Ganda	
ca	Catalan; Valencian						fy	Western Frisian		In	Lingala	
SV	Swedish						lb	Luxembourgish; Letz.		nv	Navajo; Navaho	
CS	Czech			•			nb	Norwegian Bokmål	77 7 7	ny	Chichewa; Chewa; N	٨
hu	Hungarian		•				ta	Tamil	•••	OS	Ossetian; Ossetic	• •
ru	Russian		•				bn	Bengali	•••	_ pi	Pali	
da	Danish						mr	Marathi		sm	Samoan	
el	Greek, Modern (145						te	Telugu		VO	Volapük	***
s	Slovenian		•				kk	Kazakh	* *	ab	Abkhazian	
et	Estonian		•				hy	Armenian		bm	Bambara	6)
It	Lithuanian		• •				bs	Bosnian	* 5	ee	Ewe	
ro	Romanian; Moldavia						mul	Null		gn	Guarani	<u>.</u>
bg	Bulgarian						se	Northern Sami	111 11	ho	Hiri Motu	
hr	Croatian	• • • • • •					nn	Norwegian Nynorsk		kv	Komi	•
sk	Slovak						OC	Occitan (post 1500)	•••	or	Oriya	
lv	Latvian	*****					cmn	Null	• •	rw	Kinyarwanda	•
ar	Arabic						gd	Gaelic; Scottish Gael.	. 1. 1. 1.	sg	Sango	
zh	Chinese						ms	Malay		su	Sundanese	
eu	Basque						tl	Tagalog		ti	Tigrinya	
no	Norwegian						ka	Georgian		tn	Tswana	•
су	Welsh		•				ps	Pushto; Pashto	••••	to	Tonga (Tonga Island.	
gl	Galician; Gallegan	····· ·· ·· CiTIUS					ug	Uighur; Uyghur			Tonga (Nyasa)	A
	Galician						an	Aragonese	••	ts	Tsonga	
ja	Japanese						br	Breton	•••	ty	Tahitian	A
tr	Turkish						ml	Malayalam	• •	ak	Akan	
is	Icelandic	• •					vro	Null		av	Avaric	<u></u>
uk	Ukrainian		•				eo	Esperanto		bh	Bihari languages	
sr	Serbian	•					fo	Faroese		- 3	Bihari	â.
ga	Irish						grc	Null	* * *	ie	Interlingue; Occiden.	•
ko	Korean						kn	Kannada		kg	Kongo	<u></u>
mt	Maltese						sh	Serbo-Croatian	••	_ li	Limburgan; Limburg	
hi	Hindi						ast	Null	***	SS	Siswant; Swati; Swa.	·
id	Indonesian						swh	Null		xh	Xhosa	
		0 10 20	30 4	10 50	60 70	80 90 100			0 10 20 3	0		0 10

Output media type per Input media type



Different domains per Input media type





Different domains per Input media type

 Input Media Types

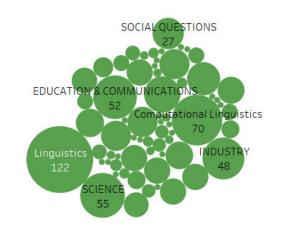
 Null
 audio
 image
 numerical text
 text
 video





Journalism & Publishing
News

Sign Language

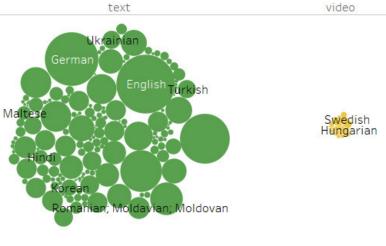




Different languages per Input media type Different languages per Input media type Input Media Types Input Media Types

		Input	t Media	Types				t Media 1	Media Types							
LanguageFullName	Null	audio	image	text F	video	Languag	geFullName N	ull audio	image	text ₹	video					
English	•	•>	*	O .		Siswant Swazi	t; Swati;									
German	•	•	•			Sango		*		**						
Spanish; Castilian	•	•				Pali				E						
French	•	•	*			Oriya										
Portuguese	•	*	•		(¥	Limburg Limburg	gan, ger; Limburgi									
Catalan; Valencian		•	18			Kongo				**						
Italian	•	•	6		(*	Komi				1						
Dutch; Flemish	•	Фт	•			Interling Occident				**						
Finnish	•	•	154			Hiri Mot	tu	*								
Russian	•	•		3	a.	Guarani		*		**						
Swedish	•		•	O \$.		Ewe										
Polish	•	•	•			Chechen	1	*		**						
Greek, Modern (1453-)	•	i w	lw.			Bambara	a	*								
Hungarian	•	•	×			Akan				**						
Estonian	•	14	is.	*		Abkhazi	an	*								
Czech	•	•	i.			Bihari la	anguages									
Danish	•	1.	14			Bihari			12							
Romanian;						Avaric										

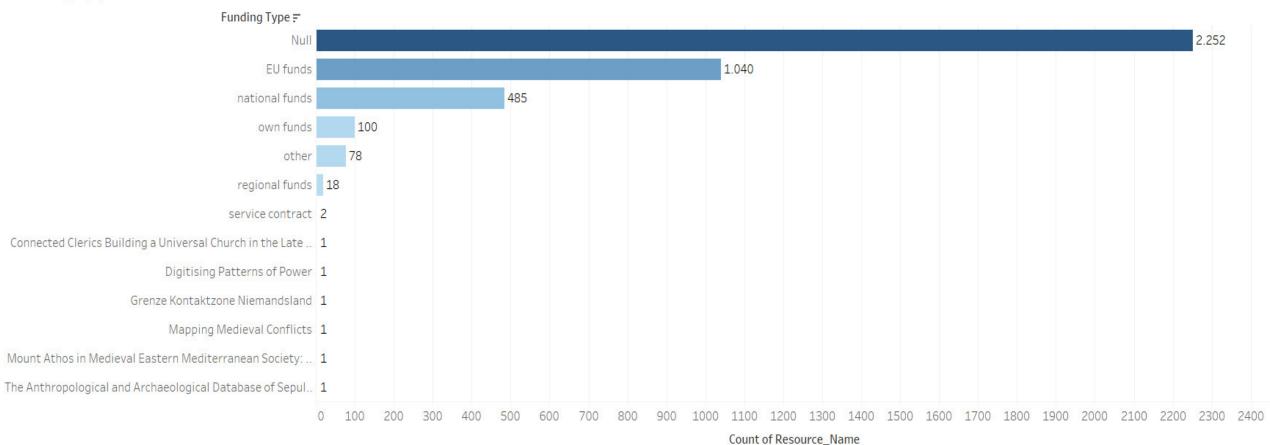
Moldavian; Moldovan



											O	utput La	nguages	/ Langu	ageFullN	lame (La	nguages	1)									
		Null	bg	ca	CS	су	da	de	el	en	es	et	eu	fi	fr	hr	hu	it	It	lv	nl	pl	pt	ro	ru	sl	SV
1 4	LanguagoEu	Mull	Bulga	Catal	Czoch	Wolch	Danish	Corm	Grook	English	Cnani	Eston	Pacauo	Einnich	Eronch	Croati	Luna	Italian	Lithu	Latvian	Dutch	Dolich	Dortu	Domo	Russi	Clava	Swodi
l å	LanguageFu Bulgarian	IVUII	69	Catal	CZECII	Weisii	Dallisli	deriii	Greek	English	Spaili	ESCOII	Dasque	FIIIIISII	French	Croati	nung	Italiali	Littiu	Latviaii	Dutti	PUIISII	Portu	KUIIIa	Kussi	Slove	Sweur
bg	Catalan; Val		09	178						75	85																
ca	Czech	65		1/0	103					82	0.0																
CS	Welsh	05			103	69				02																	
cy	Danish	82				09	69			67																	
	German	486					03	212		185	111				113			82			69		73				
е	Greek, Mod	65							91		111				110			02			03		7.5				
	English	398		68	74			168	31	441	181	<u> </u>		73	165			90			72	65	100		68		
	Spanish; Ca	206		79				105		201	402				124			84			68		101				
e	Estonian											90															
eu	Basque												66														
eu	Finnish	111								80				107													
	French	213						113		180	135				220			87			68		85				
h	Croatian															73											
h	Hungarian	77															98										
	Italian	119		2. A		2. 4		82		105	92			2	88	,		108			65	20 J	74			0	
It	Lithuanian	65																	71								
lv	Latvian																			75							
nl	Dutch; Flem	168						69		86	66				68			65			86						
Null	Null	365																									
pl	Polish	84								73												128					
pt	Portuguese	129						75		112	104				82			74					129				
ro	Romanian;									,														69			
ru	Russian	90						, , , , , , , , , , , , , , , , , , ,		77															73		
sl	Slovenian																									93	
SV	Swedish	97								70																	81

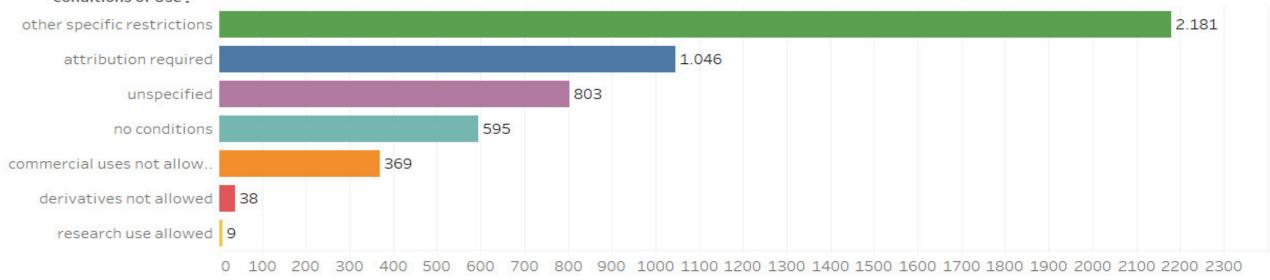
FUNDING

Funding type



Licences



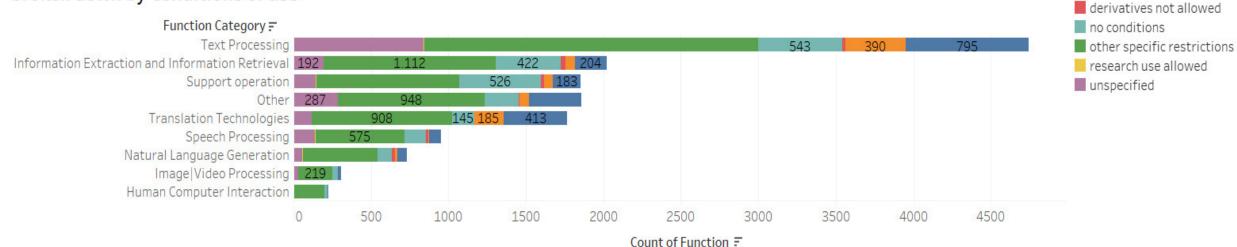


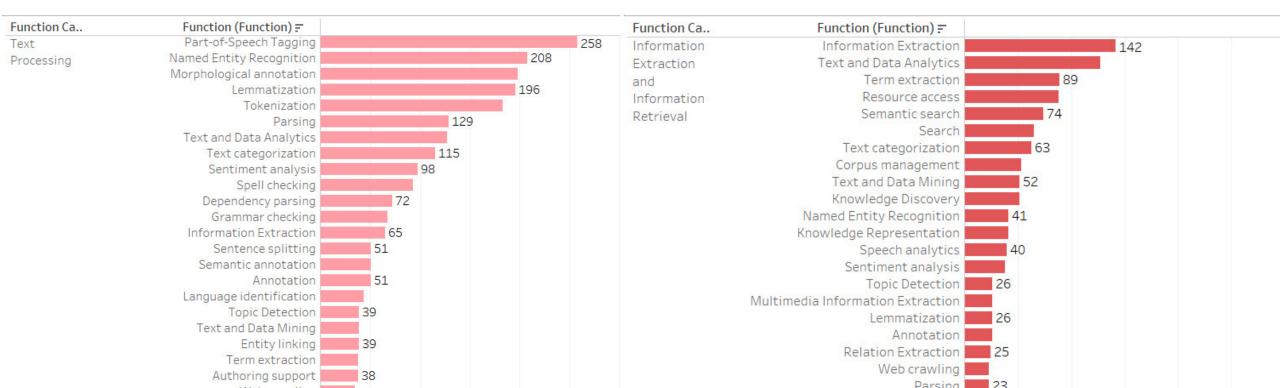
Count of Resource_Name ₹



Function Category COLLECTIONS

broken down by *Conditions of use*





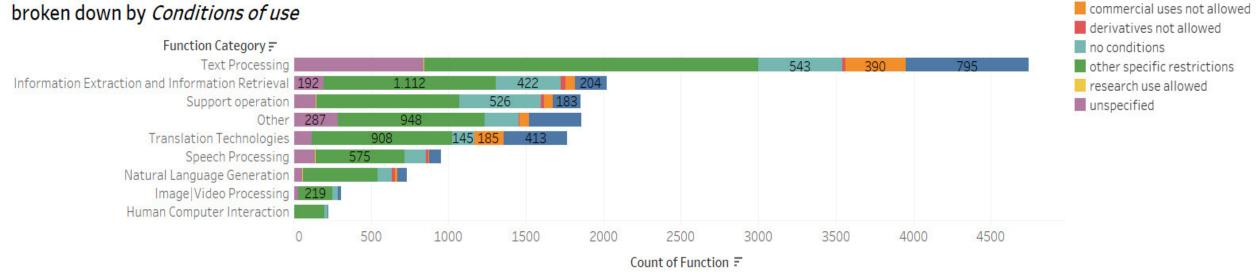
Conditions of Use

attribution required

commercial uses not allowed

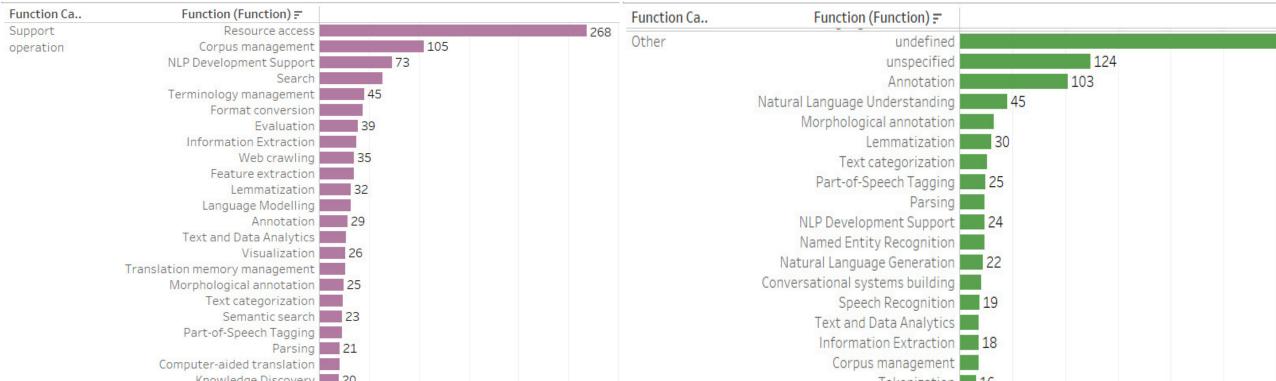
Function Category COLLECTIONS

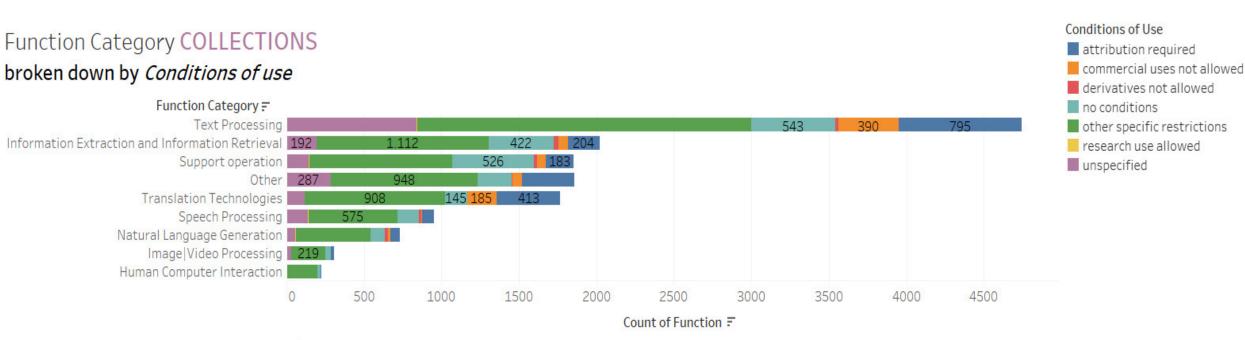
broken down by *Conditions of use*

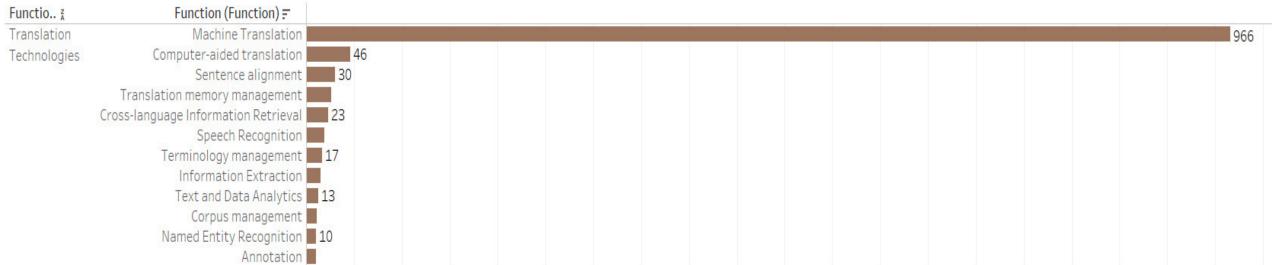


Conditions of Use

attribution required

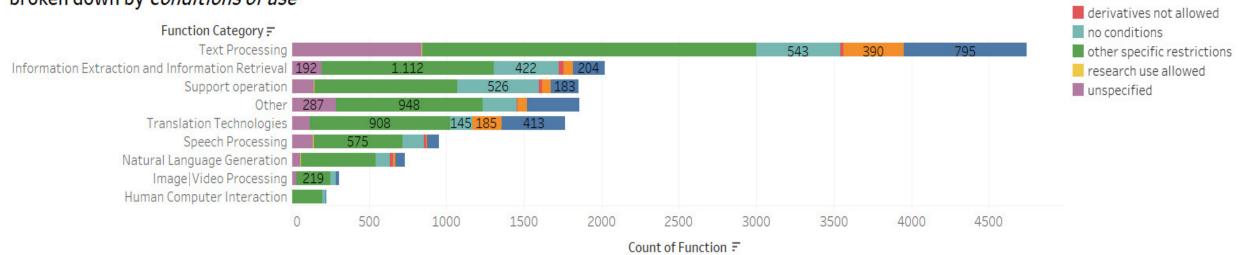






Function Category COLLECTIONS

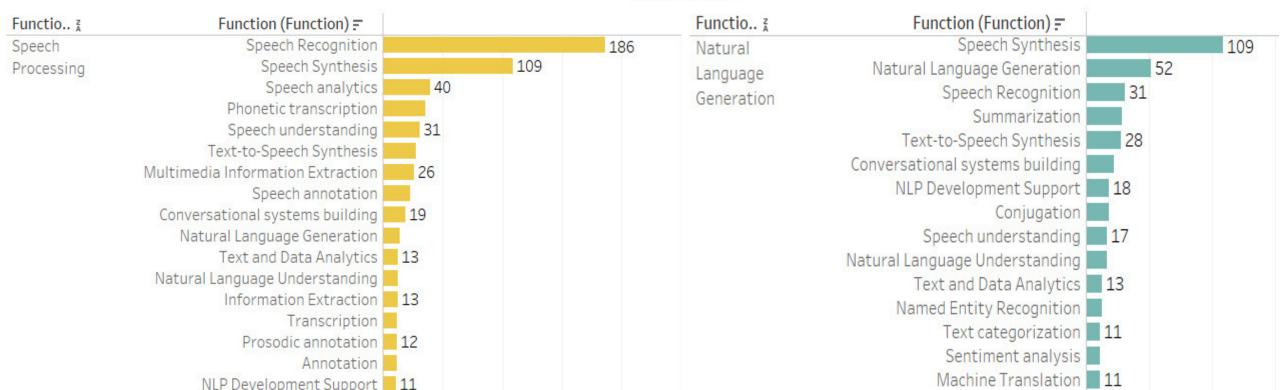
broken down by Conditions of use

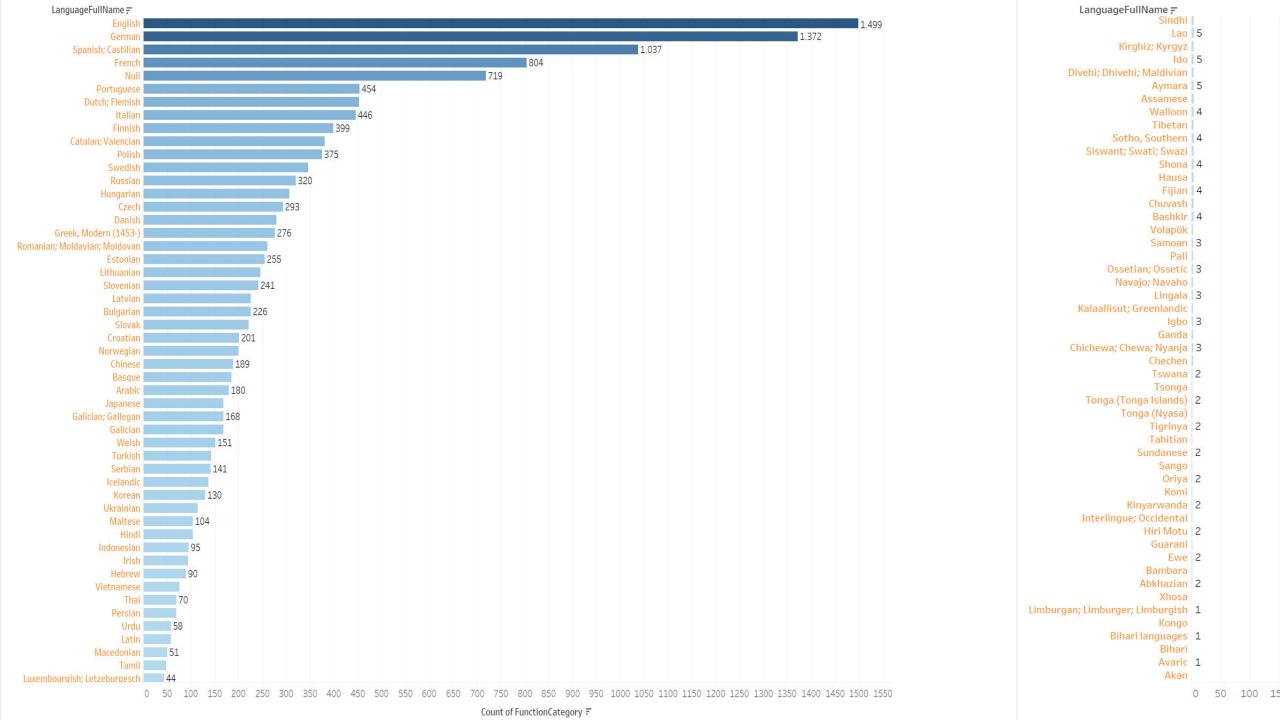


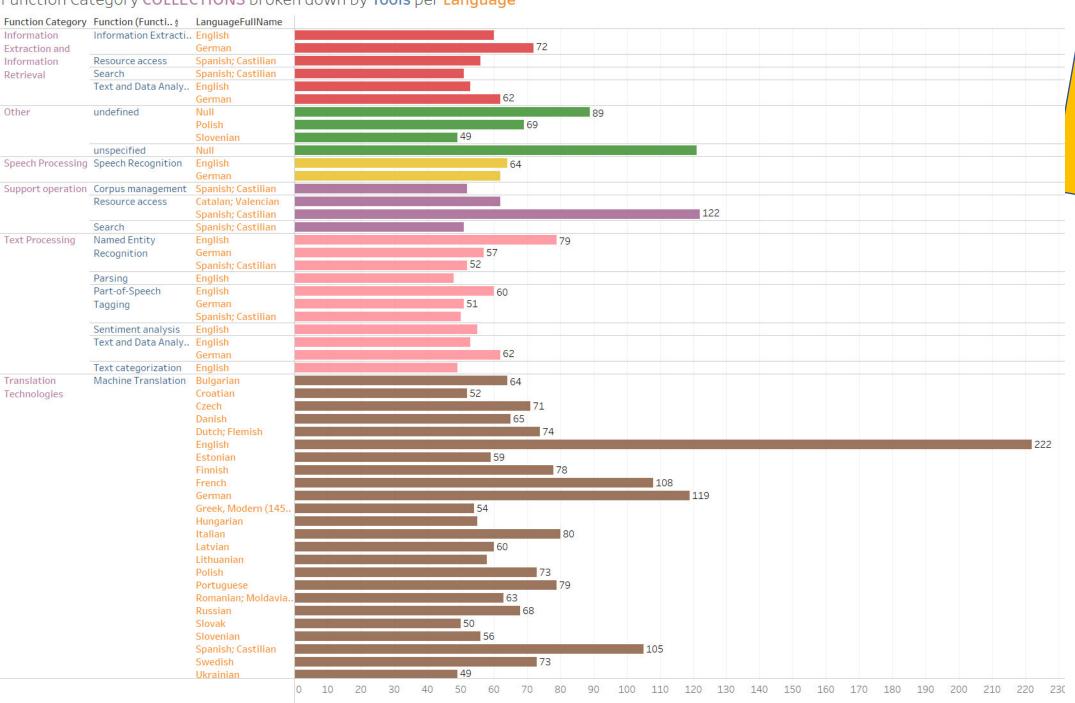
Conditions of Use

attribution required

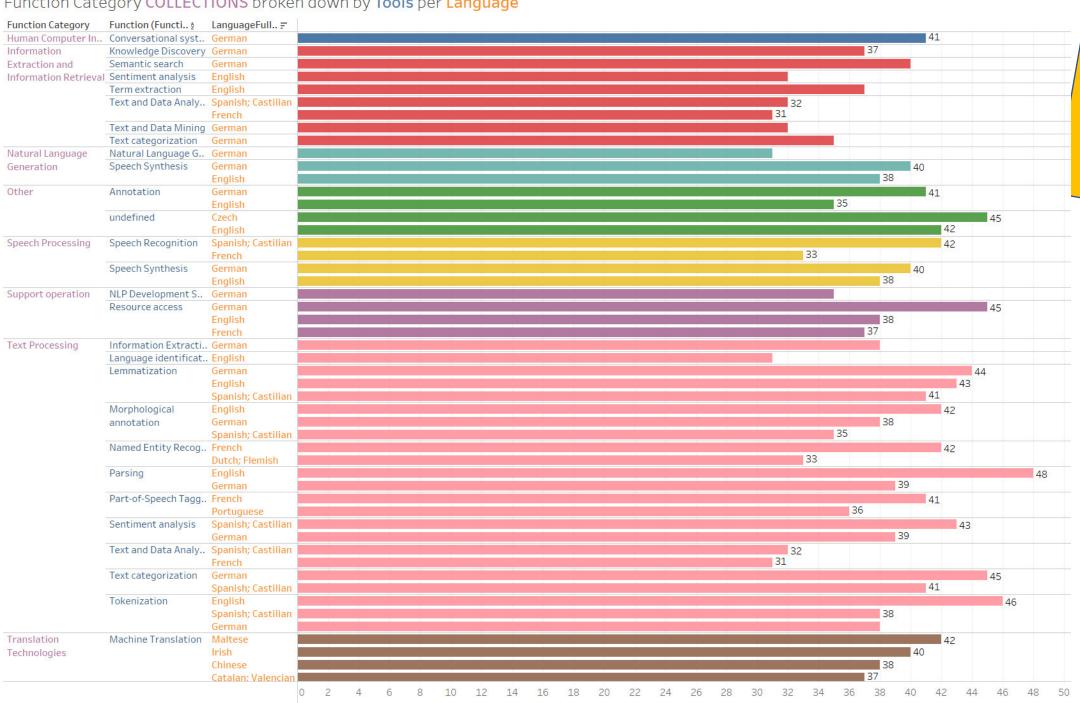
commercial uses not allowed

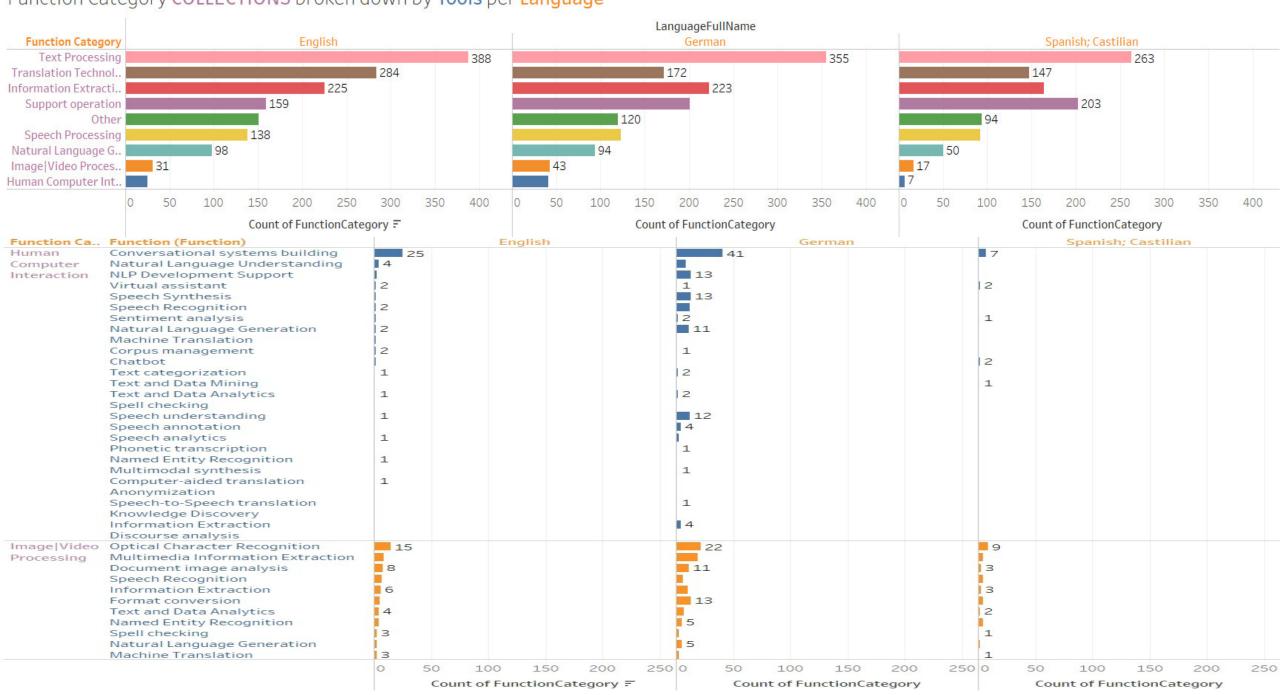






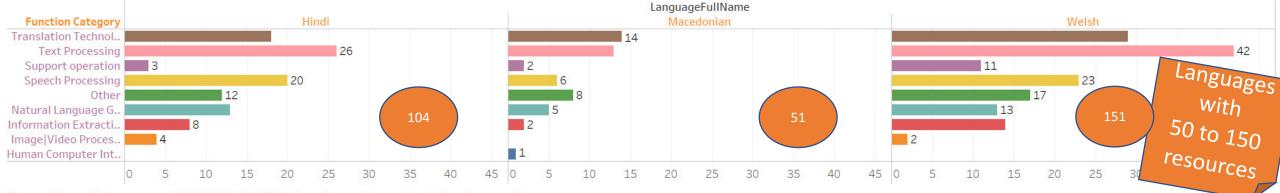
Language With 49 and more resources for a Tool



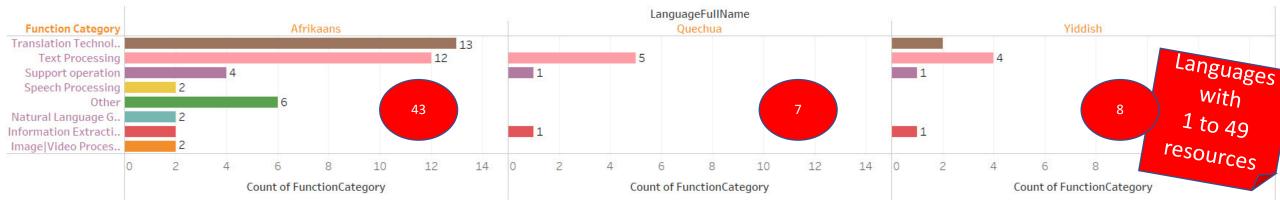




Function Category COLLECTIONS broken down by Tools per Language



Function Category COLLECTIONS broken down by Tools per Language





TASK 2

Learning about **ELG** and

data behind it