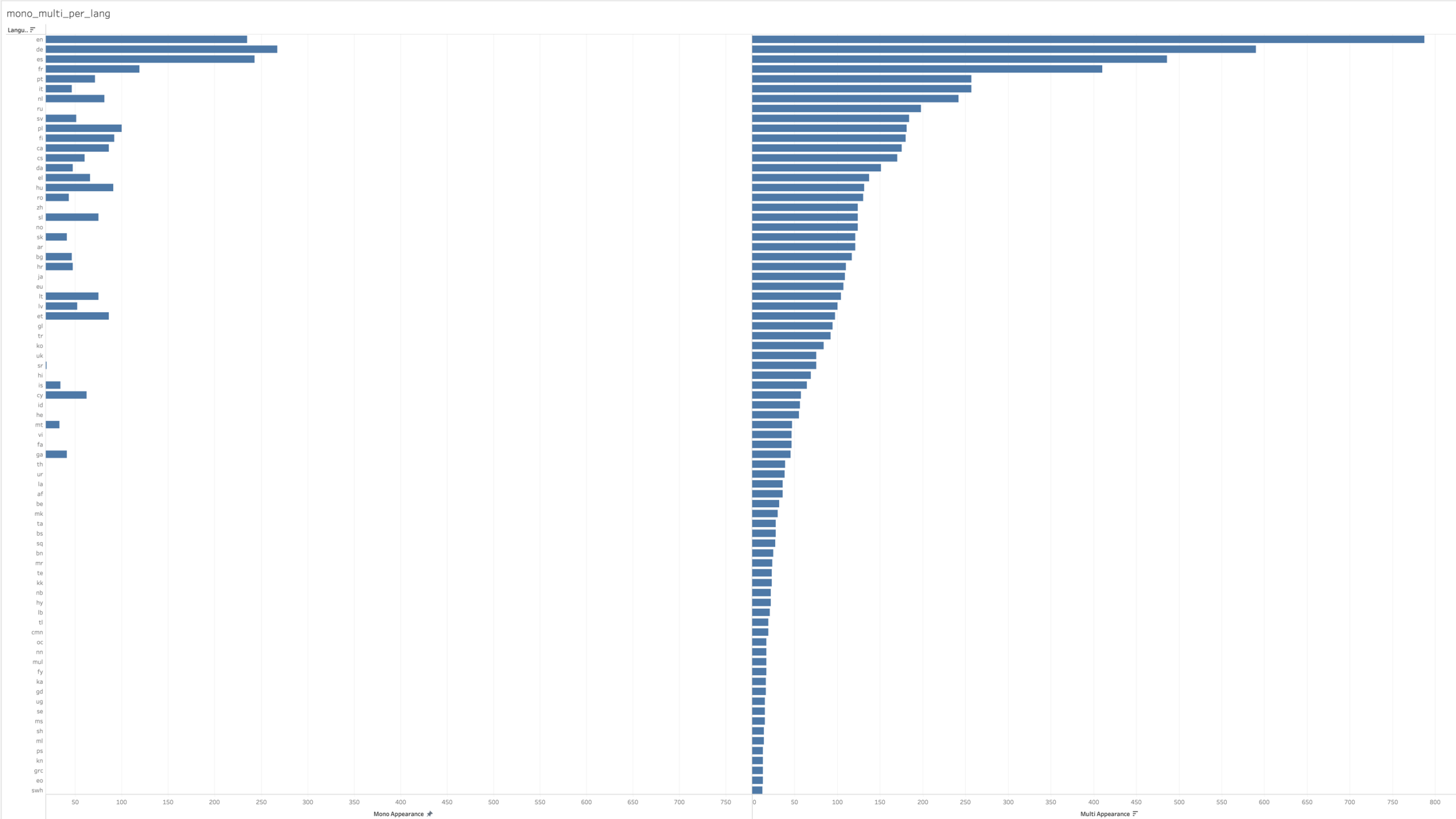
ELG Catalogue multilingual tools

Noémi Ligeti-Nagy

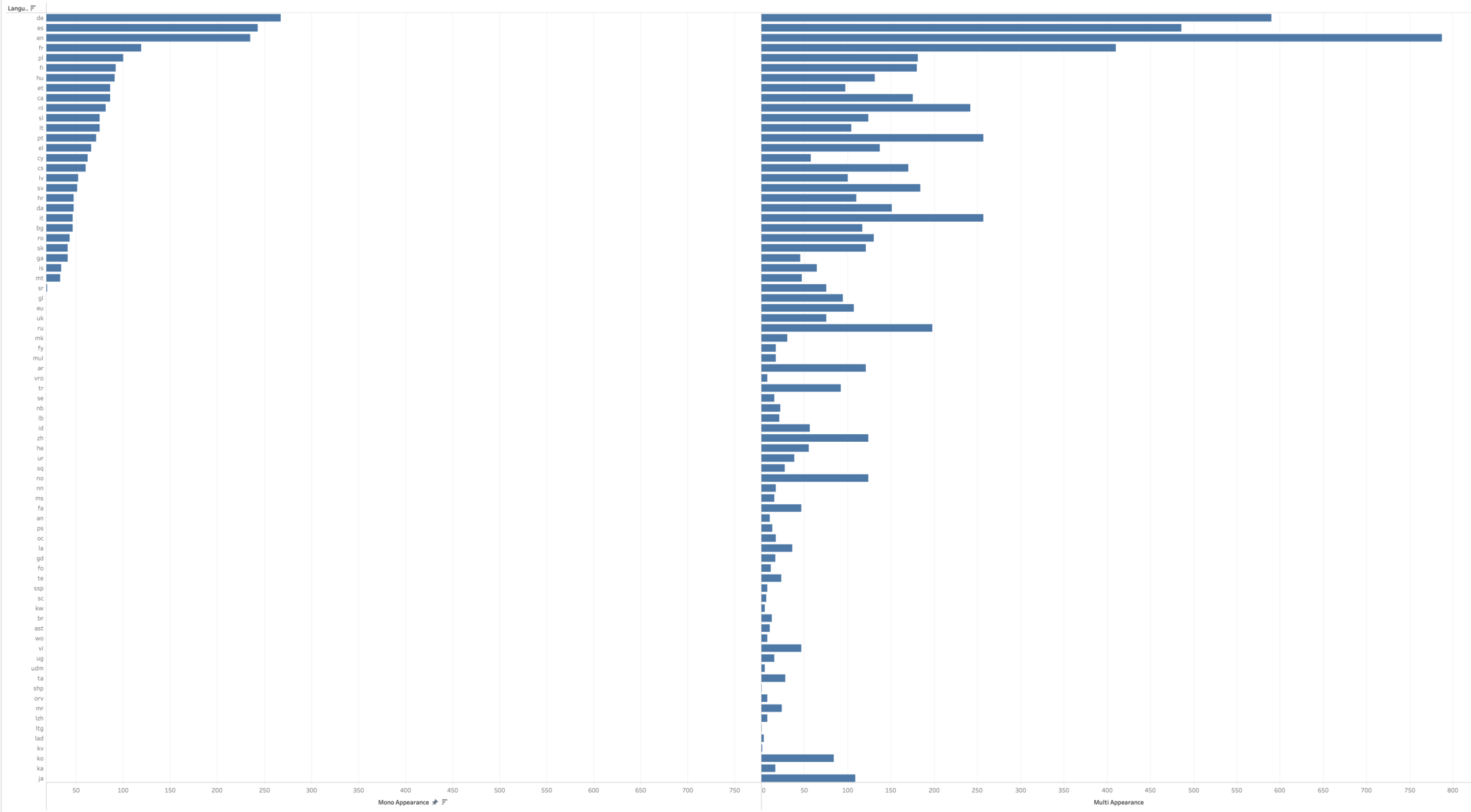
1st October, 2024

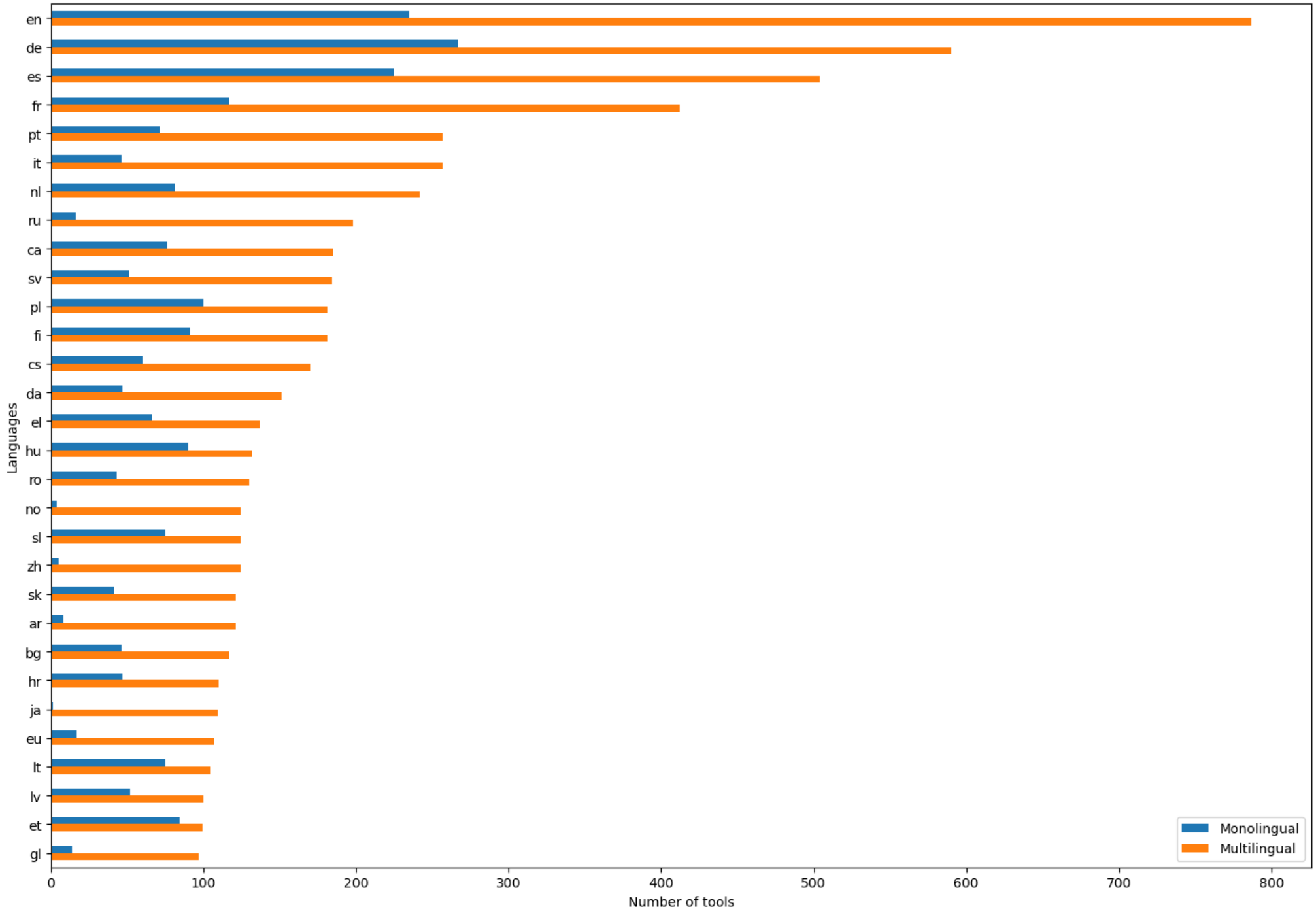
Mono- and multilingual tools for a given language

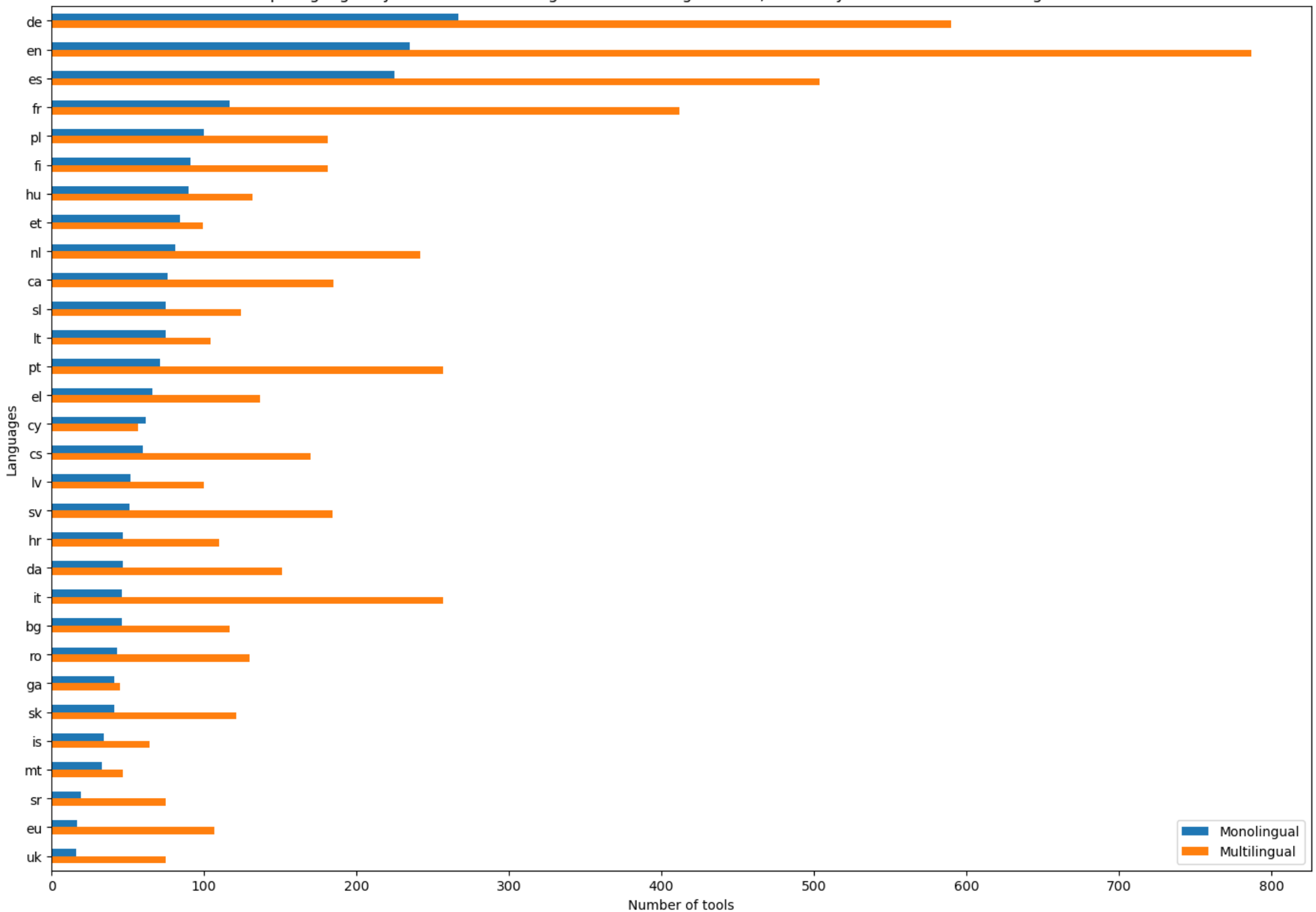
- 1st: ordered by the number of multiling. tools
- 2nd: ordered by the number of monoling. tools


mono_multi_per_lang
Langu..
en
de
es
fr
pt
it
nl
ru
sv
pl
fi
ca
cs
da
el
hu
ro
zh
sl
no
sk
ar
bg
hr
ja
eu
lt
lv
et
gl
tr
ko
uk
sr
hi
is
cy
id
he
mt
vi
fa
ga
th
ur
la
af
be
mk
ta
bs
sq
bn
mr
te
kk
nb
hy
lb
tl
cmn
oc
nn
mul
fy
ka
gd
ug
se
ms
sh
ml
ps
kn
grc
eo
swh
50 100 150 200 250 300 350 400 450 500 550 600 650 700 750
Mono Appearance
0 50 100 150 200 250 300 350 400 450 500 550 600 650 700 750 800
Multi Appearance

mono_multi_per_lang




Top languages by number of monolingual and multilingual tools, sorted by the number of multilingual tools
Languages
en
de
es
fr
pt
it
nl
ru
ca
sv
pl
fi
cs
da
el
hu
ro
no
sl
zh
sk
ar
bg
hr
ja
eu
lt
lv
et
gl
0
100
200
300
400
500
600
700
800
Number of tools
Monolingual
Multilingual


Top languages by number of monolingual and multilingual tools, sorted by the number of monolingual tools
Languages
de
en
es
fr
pl
fi
hu
et
nl
ca
sl
lt
pt
el
cy
cs
lv
sv
hr
da
it
bg
ro
ga
sk
is
mt
sr
eu
uk
0
100
200
300
400
500
600
700
800
Number of tools
Monolingual
Multilingual

Mono- and multilingual tools for a given language

- 1st: ordered by the number of multiling. tools
- 2nd: ordered by the number of monoling. tools
- Conclusion:
  - If a lang. is well-resourced in monoling. tools, it is well-resourced in multiling. tools as well (en-de-es being the top 3)
  - But not vice versa
  - **Many languages: many multi, zero mono**

Languages	Monolingual Tools	Multilingual Tools
cy	62	57
ltg	1	0
shp	1	0

Base NLP tasks and their coverage for languages in mono- or multiling. tools

- just some random collection, not a scientifically proven list
- 1st table: a summary
- 2nd, bar charts: a summary with an aggregated number
- 3rd: focus on some European languages.

base_NLP_tasks

Input Langu..	Language T..	In / Out of Base_NLP / Function In: coref	Dependency parsing	Hyphenation	Lemmatization	Machine Translation	Morphological analysis	Morphological disamb..	Morphosyntactic tagg..	Named Entity Recogni..	NER	Noun phrase chunking	Parsing	Part-of-Speech Tagging	Semantic annotation	Sentence splitting	Tokenization	Word Sense Disambig..
Null	Mono				4	6				3	1	1	4	4	3		4	3
ady	Multi				1								1	1				
af	Mono		1		1									1			1	
	Multi		2		4	12				3			1	4		3	4	1
afb	Multi				1								1	1				
ak	Multi				1								1	1				
am	Multi					3												
ame	Multi				1								1	1				
an	Multi					6												
ang	Multi				1								1	1				
ar	Mono		1		1					1				1			1	
	Multi	1	1		9	27	1			9			7	9	4	3	5	1
arl	Multi																	1
arn	Multi				1								1	1				
arz	Multi				1								1	1				
as	Multi					2												
ase	Mono					1												
ast	Mono												1	1				
	Multi	1			1	3				1			2	2	2		1	1
ay	Multi				1								1	1				
az	Multi				1	3							1	1				
azg	Multi				1								1	1				
ba	Multi				1	1							1	1				
be	Mono		1		1									1			1	
	Multi				2	16							1	2			1	
bg	Mono		1		1	26				1			1	2		1	1	
	Multi	2	2		10	41				15			9	14	6	5	11	1
bho	Multi				1									1				
bn	Multi				2	8				3			3	2		1	1	1
bo	Multi				1	1							1	1				
br	Mono					2												
	Multi		1		2	3				1			1	3		1	1	1
bs	Multi		1		1	7				2				1		1	1	1
bua	Multi		1		1									1		1	1	
ca	Mono		1		5	1				1			6	3			2	
	Multi	4	2		10	39				11			10	19	7	4	13	4
cbk	Multi					1												
ceb	Multi				1	1							1	1				
ckb	Multi				1	3							1	1				
cly	Multi				1								1	1				
cmn	Multi														1	1		
cni	Multi				1								1	1				
cnr	Multi									1								
co	Multi		1		1	1				1				1		1	1	1
cop	Mono		1		1									1			1	
	Multi				1									1			1	
cpa	Multi				1								1	1				
crh	Multi				1								1	1				
cs	Mono		1		1	29				2				2			1	
	Multi	3	2		12	45				11			8	13	3	4	10	1
csb	Multi				1								1	1				
ctp	Multi				1								1	1				
cu	Mono		1		1									1			1	
	Multi		1		3								1	3		1	2	
cy	Mono		1		2	10	1			1				6	1			
	Multi	1	1		2	17				4			3	5	4	1	2	2
czn	Multi				1								1	1				
da	Mono		2		4	25	1			6				6		1	2	
	Multi	2	2		13	43				13			6	12	4	4	14	2
de	Mono	2	2	1	12	40	3			12			6	13	9	6	9	1
	Multi	5	3		33	83	1			50			34	40	20	10	30	12
dsb	Multi				1	2							1	1				
dv	Multi					1												
el	Mono		2		2	25				7			2	6		2	3	
	Multi	2	2		12	30				12			8	14	4	3	13	1
en	Mono		2		2	99	2			21		1	2	9	2	3	7	3
	Multi	7	3	1	42	133	1			60			47	53	25	14	40	13
enh	Multi				1		1											
eo	Multi		1		1	5				1				1		1	1	1
es	Mono		1		12	32				7			8	8	3	1	8	2
	Multi	6	3		30	76				46			21	43	18	8	31	9
et	Mono	1	3		5	27	1		3	2			2	3		2	3	
	Multi	2	2		9	35				11			5	11	2	4	10	1
eu	Mono		1		4	2				1				3			2	
	Multi	2	2		11	21				12			8	13	2	3	11	2
fa	Mono		1		1					1				1			1	
	Multi		2		6	11				5			4	6	1	3	6	1
fi	Mono		2	2	6	33	2		1	11			5	11		1	4	1

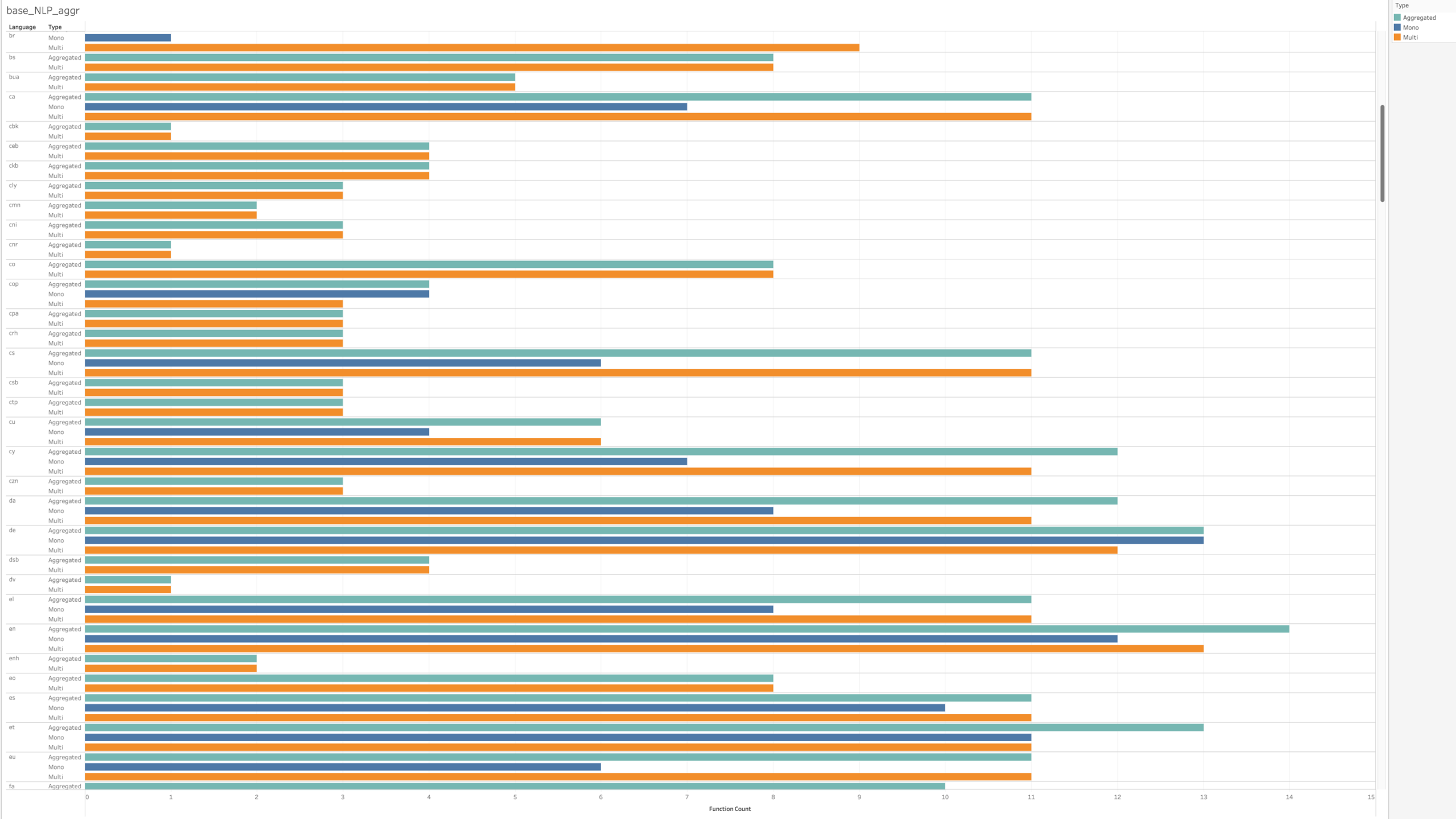
Base_NLP

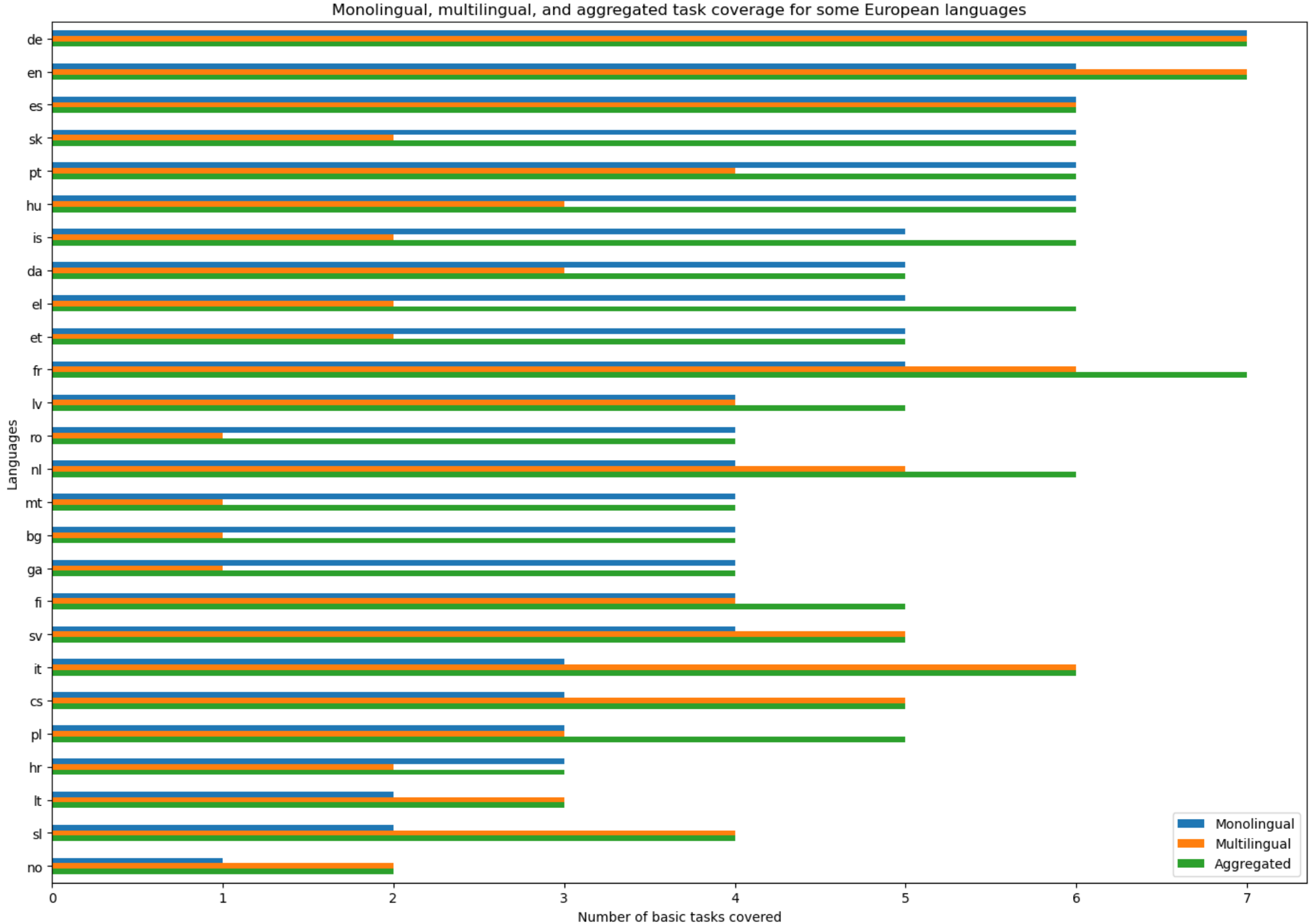
- (All)
- Chunking
- coref
- Dependency parsing
- Hyphenation
- Lemmatization
- Machine Translation
- Morphological analysis
- Morphological disambiguation
- Morphosyntactic tagging
- Named Entity Recognition
- NER
- Noun phrase chunking
- Parsing
- Part-of-Speech Tagging
- Semantic annotation
- Sentence splitting
- Tokenization
- Word Sense Disambiguation

Function

- (All)
- ~~Accentuation~~
- ~~Age-bracket detection~~
- ~~Alignment~~
- ~~Annotation~~
- ~~Annotation of argumentation~~
- ~~Annotation of compounds~~
- ~~Annotation of document structure~~
- ~~Annotation of measurements~~
- ~~Annotation of modalities~~
- ~~Annotation of multi-word units~~
- ~~Annotation of suprasegmental features~~
- ~~Annotation of textual entailment~~
- ~~Anonymization~~
- ~~API~~
- ~~API Connector~~
- ~~App development~~
- ~~Argument detection~~
- ~~Argument mining~~
- ~~Article extraction~~
- ~~Article search~~
- ~~Audio processing~~
- ~~Audio segmentation~~
- ~~Author age identification~~
- ~~Author classification~~
- ~~Author profiling~~
- ~~Authoring support~~
- ~~Authorship attribution~~
- ~~Automatic analysis~~
- ~~Automatic Speech Recognition~~
- ~~Automatic subtitling~~
- ~~Avatar synthesis~~
- ~~Behavior Rules~~
- ~~Braille screen review~~
- ~~Caching~~
- ~~Calculation of multilingual sentence embe..~~
- ~~Capitalization~~
- ~~Case disambiguation~~
- ~~Chatbot~~
- Chunking
- ~~Clustering~~
- ~~Co-occurrence detection~~
- ~~Collocations extraction~~
- ~~Comparing automatic labelling tools~~
- ~~Comparing collocations of two words~~
- ~~Compound splitting~~
- ~~Computational argumentation~~
- ~~Computer-aided translation~~
- ~~Conceptual analysis~~
- ~~Concordance search~~
- ~~Concordancing~~
- ~~Conjugation~~
- ~~Conversational systems building~~
- coref
- ~~Coreference annotation~~
- ~~Corpus analysis~~
- ~~Corpus management~~
- ~~Corpus recording~~
- ~~Corpus viewing~~
- ~~Cross-language Information Retrieval~~
- ~~Data collection~~
- ~~Data labelling~~
- ~~Data splitting~~
- ~~Deaf Support~~
- ~~Dehyphenation~~

base_NLP_aggr




Monolingual, multilingual, and aggregated task coverage for some European languages
Languages
de
en
es
sk
pt
hu
is
da
el
et
fr
lv
ro
nl
mt
bg
ga
fi
sv
it
cs
pl
hr
lt
sl
no
0
1
2
3
4
5
6
7
Number of basic tasks covered
Monolingual
Multilingual
Aggregated

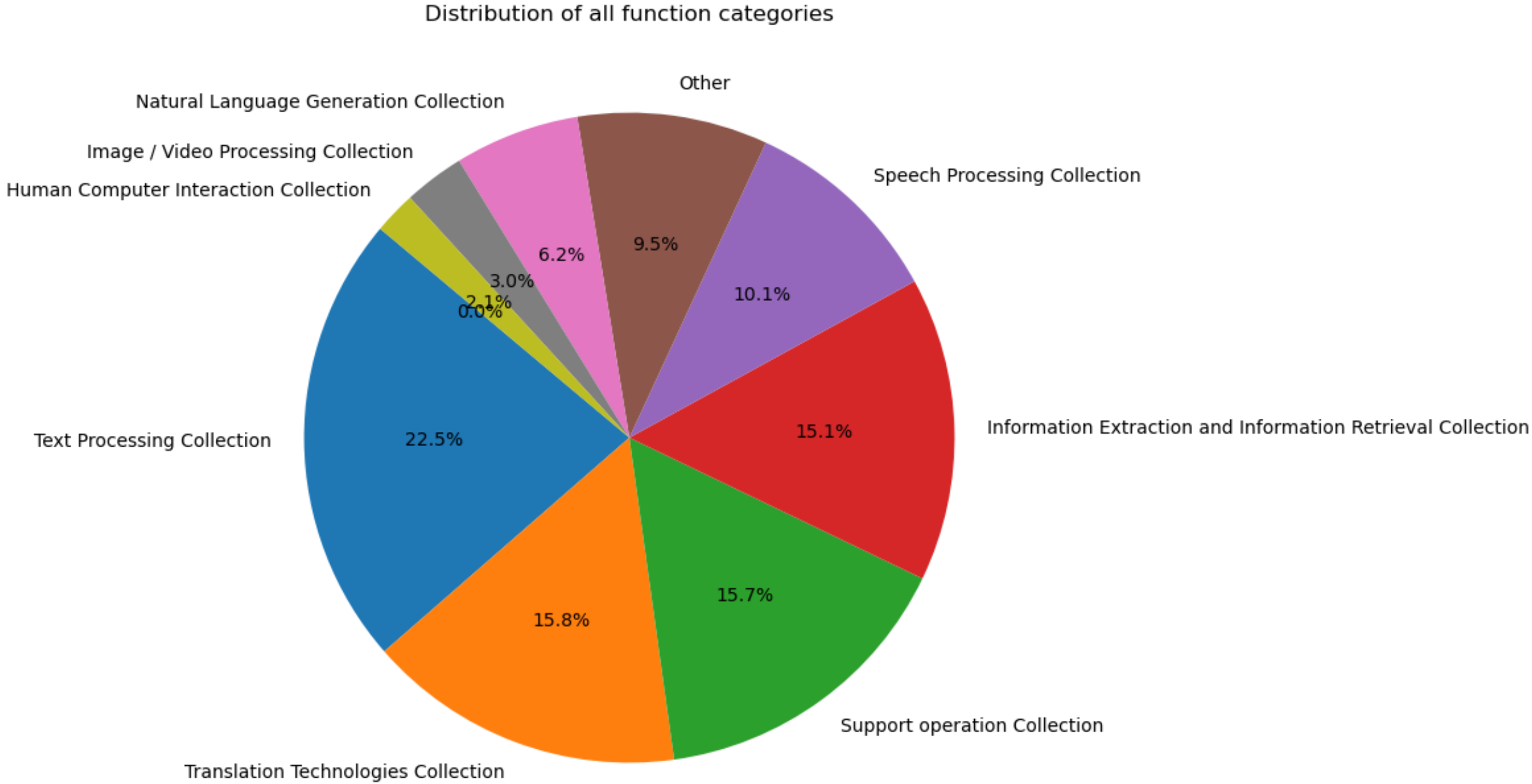
Base NLP tasks and their coverage for languages in mono- or multiling. tools

- just some random collection, not a scientifically proven list
- 1st table: a summary
- 2nd, bar charts: a summary with an aggregated number
- 3rd: focus on some European languages.

No big surprises here (aggregated means the number of base NLP tasks for a given language covered by either mono- or multiling. tools). If aggregated is higher than multilingual, it means that there is at least one task that is only covered for a language in a monoling. tool.

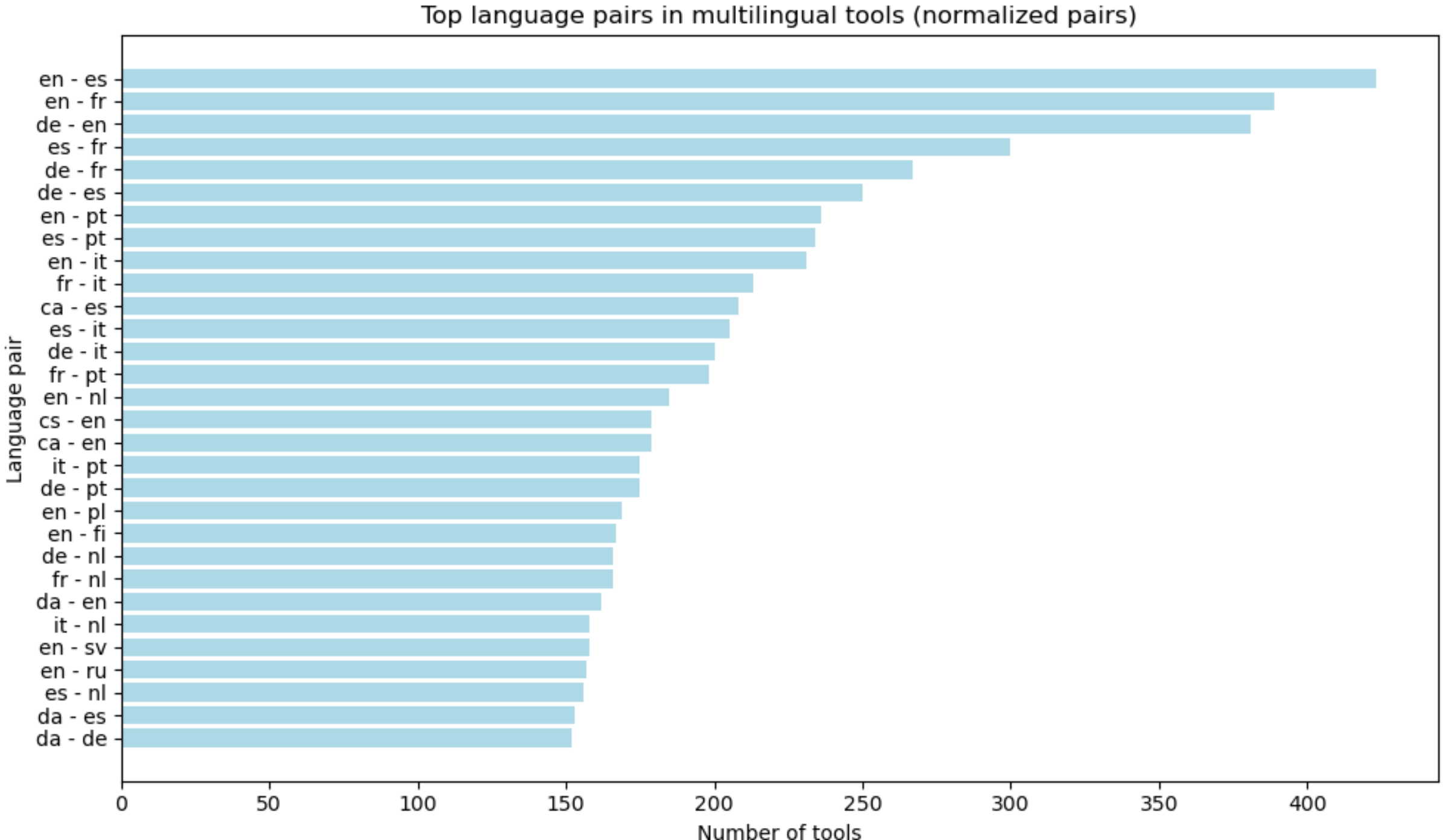
Function categories in multilingual tools

- The pie chart illustrates the **distribution of the top 10 most frequent function categories** in the dataset.
- "Others" aggregates the less frequent categories.
- Each slice of the pie represents a category's proportion relative to the total occurrences of function categories.


Distribution of all function categories
Other
Natural Language Generation Collection
Image / Video Processing Collection
Human Computer Interaction Collection
Speech Processing Collection
Information Extraction and Information Retrieval Collection
Text Processing Collection
Support operation Collection
Translation Technologies Collection
9.5%
6.2%
3.0%
2.1%
0.0%
10.1%
15.1%
22.5%
15.7%
15.8%

Top language pairs in multilingual tools

- top language pairs used in multilingual tools
- pairs like en-es and es-en are treated as one
- the most commonly supported cross-lingual combinations

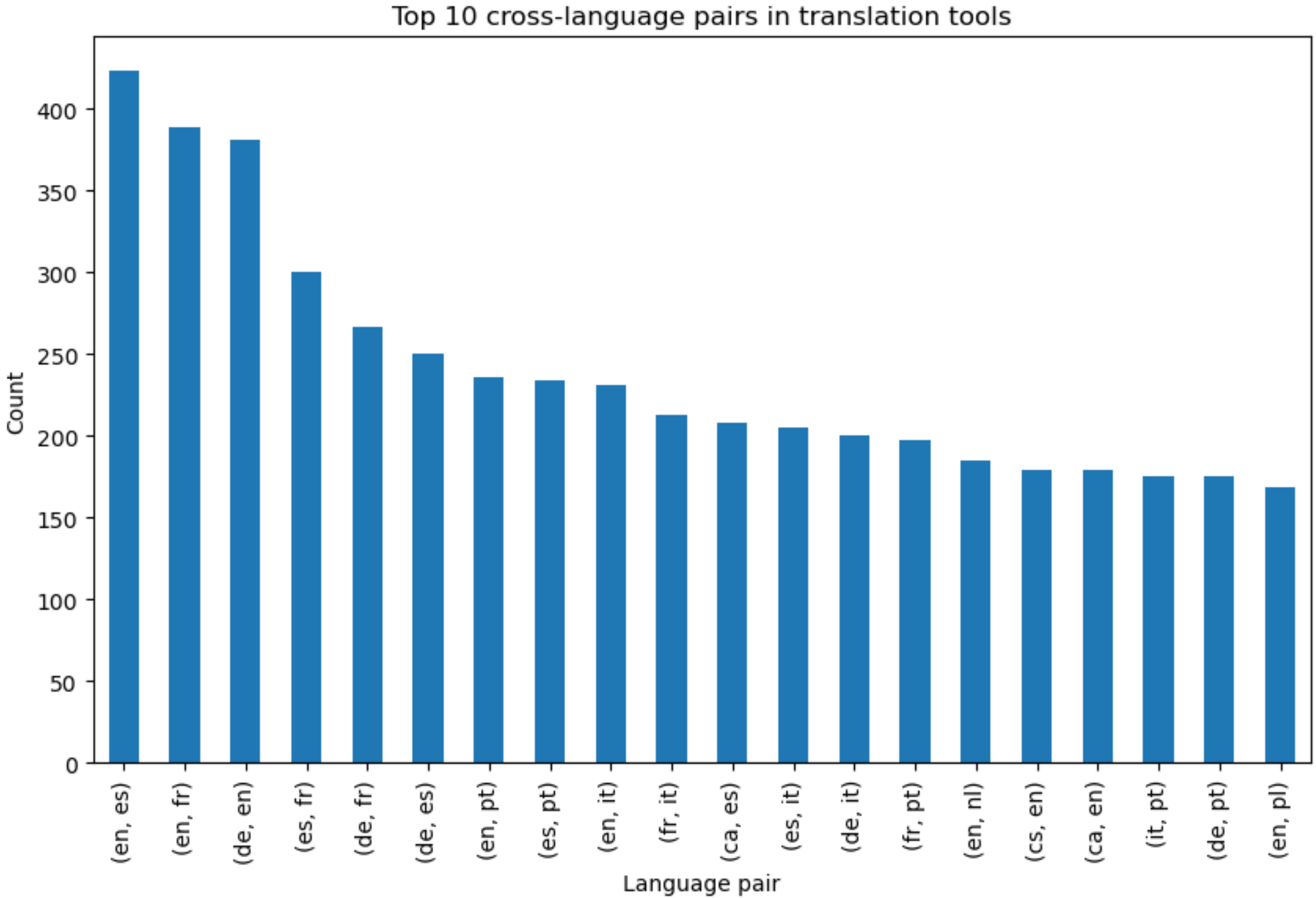

Top language pairs in multilingual tools (normalized pairs)
Language pair
en - es
en - fr
de - en
es - fr
de - fr
de - es
en - pt
es - pt
en - it
fr - it
ca - es
es - it
de - it
fr - pt
en - nl
cs - en
ca - en
it - pt
de - pt
en - pl
en - fi
de - nl
fr - nl
da - en
it - nl
en - sv
en - ru
es - nl
da - es
da - de
0
50
100
150
200
250
300
350
400
Number of tools

Top language pairs in multilingual tools

- top language pairs used in multilingual tools
- pairs like en-es and es-en are treated as one
- the most commonly supported cross-lingual combinations
- those involving English and Spanish dominate the landscape

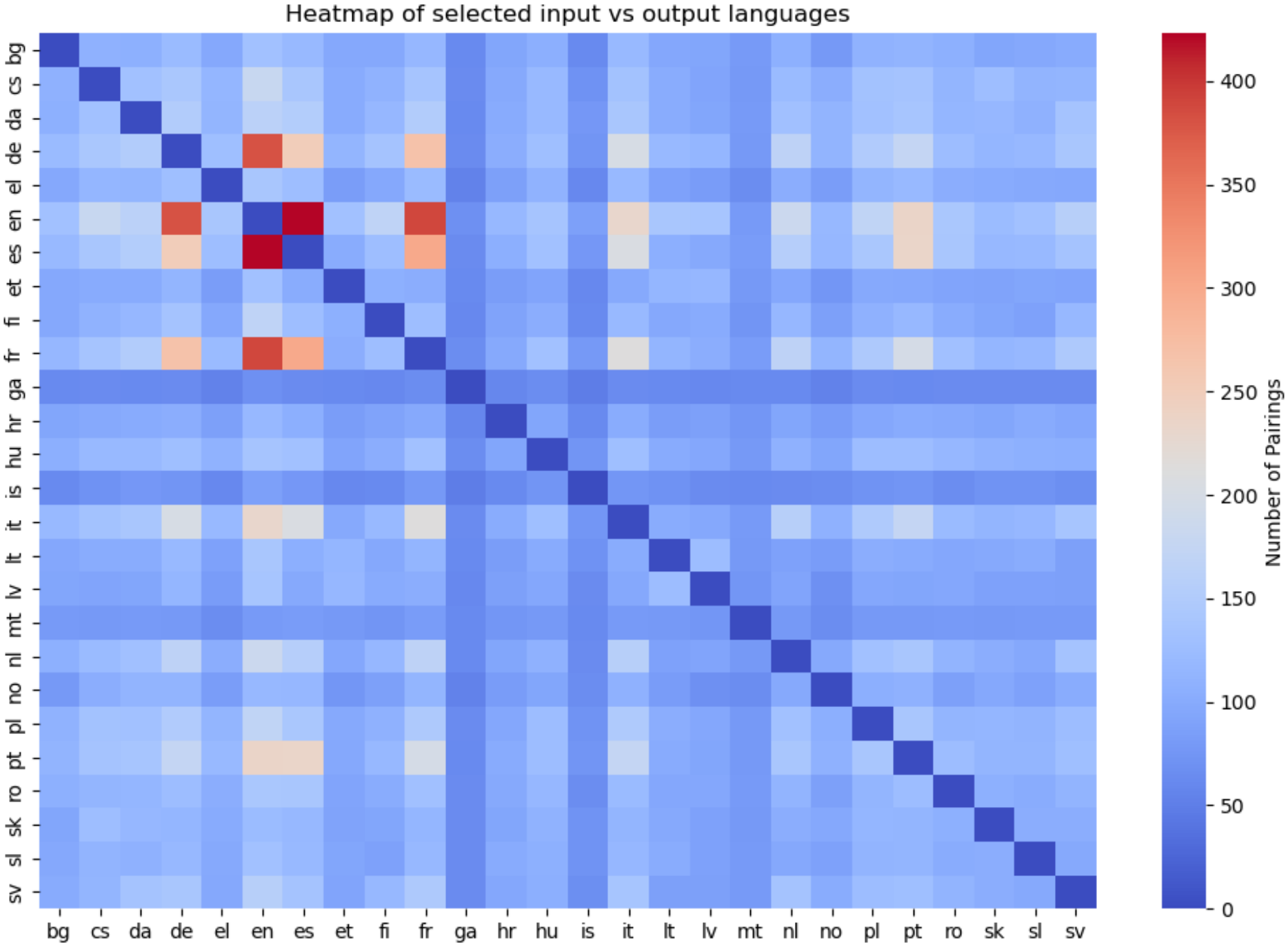
Top language pairs for one task: translation

- English-Spanish is the most frequently supported pair, followed by English-French and German-English
- Many romance languages


Top 10 cross-language pairs in translation tools
Count
0
50
100
150
200
250
300
350
400
(en, es)
(en, fr)
(de, en)
(es, fr)
(de, fr)
(de, es)
(en, pt)
(es, pt)
(en, it)
(fr, it)
(ca, es)
(es, it)
(de, it)
(fr, pt)
(en, nl)
(cs, en)
(ca, en)
(it, pt)
(de, pt)
(en, pl)
Language pair

Heatmap of language pairs

- All multilingual tools
- Focus on some European languages


Heatmap of selected input vs output languages
bg cs da de el en es et fi fr ga hr hu is it lt lv mt nl no pl pt ro sk sl sv
Number of Pairings
400
350
300
250
200
150
100
50
0