macocu

MaCoCu Corpora: Why Top-Level-Domain Crawling and Web Data Enrichment Matter

Nikola Ljubešić (Jožef Stefan Institute, Slovenia)

The MaCoCu crowd

HPLT & NLPL Winter School

Co-financed by the Connecting Europe Facility of the European Union

The MaCoCu project

Focus: get web corpora for languages with lack of data from EU members and candidate members (or interesting!). Two-year project: June 2021-June 2023

Monolingual and parallel corpora (with English) for following languages:

- ★ Batch 1 (May 2022): Bulgarian, Macedonian, Maltese, Slovenian, Turkish, Icelandic
- ★ Batch 2 (May 2023): Albanian, Bosnian, Croatian, Montenegrin, Serbian
- ★ Bonus (May 2023): Catalan, Greek, Ukrainian

Partners:

Universitat d'Alacant
Universidad de Alicante

rijksuniversiteit
groningen

Jožef Stefan Institute

prompsit

macocu

Use case - South Slavic languages

macocu

Use case - South Slavic languages

macocu

Overview

1. Top-level-domain crawling

   How are we obtaining our data?
   And why are we not using CommonCrawl?

2. Data enrichment

   We do not want to use data we know almost nothing about.
   Therefore we enrich our data with various metadata.

3. Modelling

   What are the effects of our quests above on the resulting language and translation models?

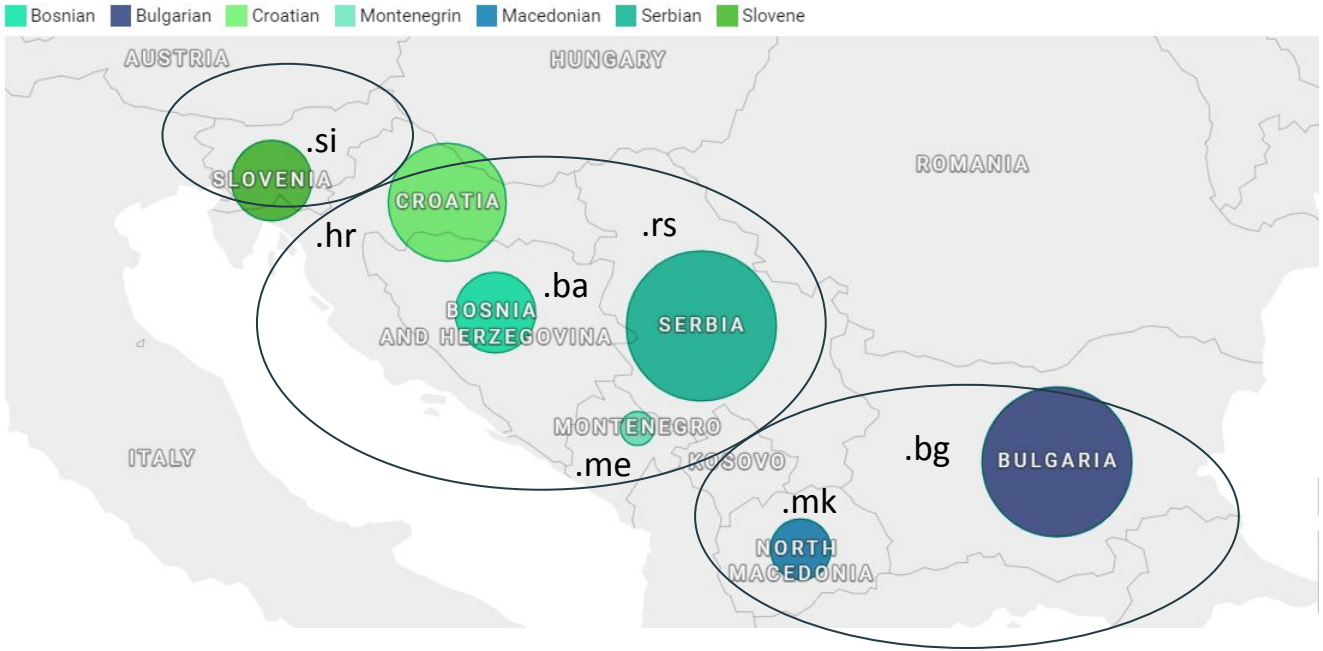
macocu

macocu

Top-level-domain crawling

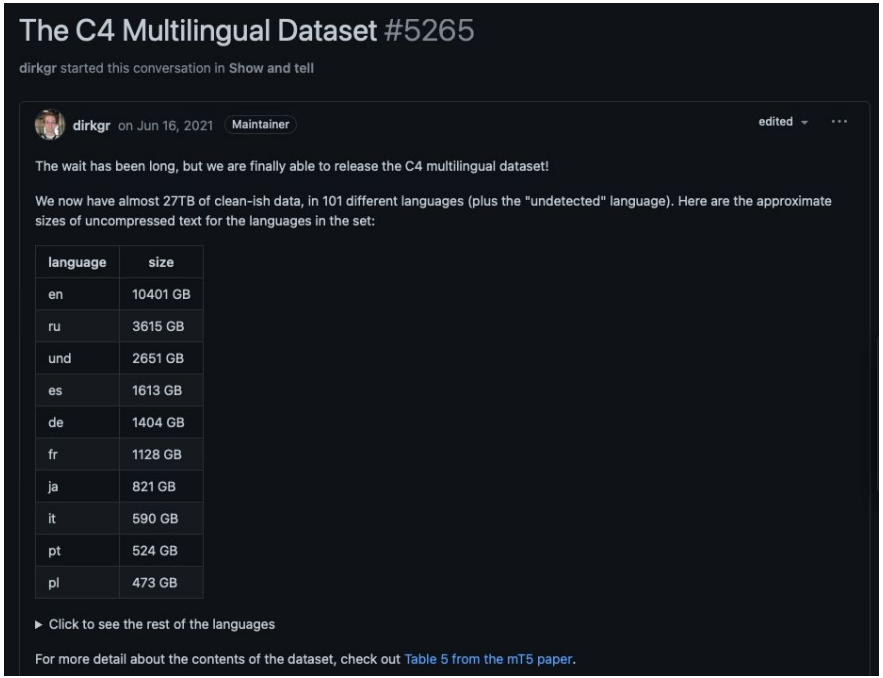
Co-financed by the Connecting Europe Facility of the European Union

What are top-level domains?



macocu

Existing corpora based on CommonCrawl



The C4 Multilingual Dataset #5265

dirkgr started this conversation in Show and tell

dirkgr on Jun 16, 2021 Maintainer edited

The wait has been long, but we are finally able to release the C4 multilingual dataset!

We now have almost 27TB of clean-ish data, in 101 different languages (plus the "undetected" language). Here are the approximate sizes of uncompressed text for the languages in the set:

language	size
en	10401 GB
ru	3615 GB
und	2651 GB
es	1613 GB
de	1404 GB
fr	1128 GB
ja	821 GB
it	590 GB
pt	524 GB
pl	473 GB

▸ Click to see the rest of the languages

For more detail about the contents of the dataset, check out Table 5 from the mT5 paper.

mC4

Based on all pre-2021 CC dumps

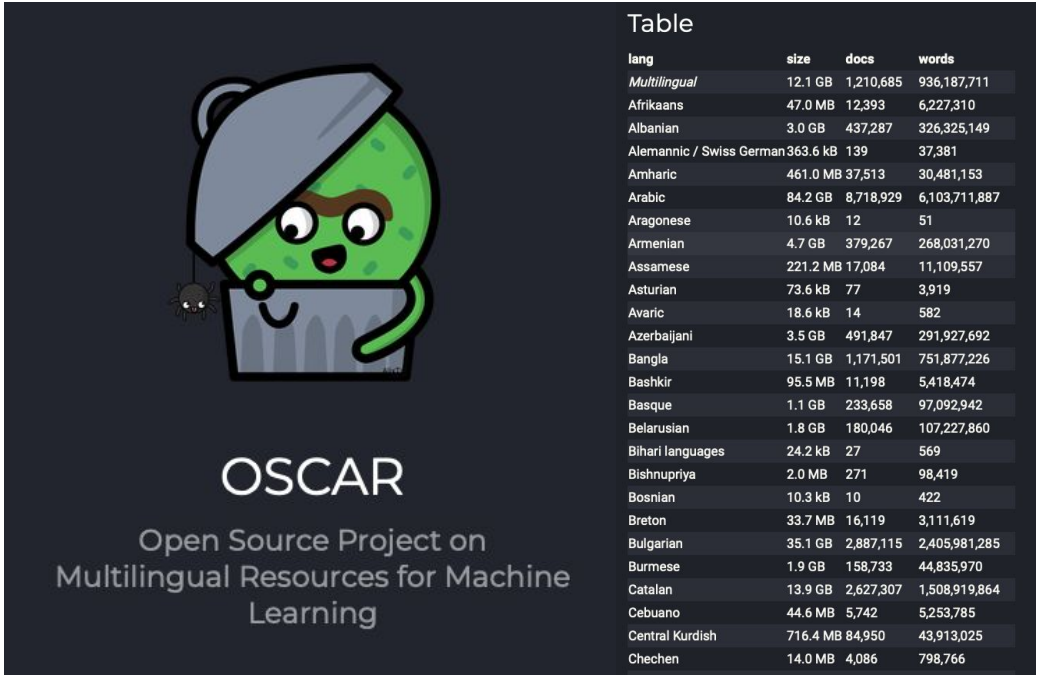
Uses CLD3 for language identification

Huge!, but also noisy

Blind spots - almost no Latin BCMS present

macocu

Existing corpora based on CommonCrawl



Table

lang	size	docs	words
Multilingual	12.1 GB	1,210,685	936,187,711
Afrikaans	47.0 MB	12,393	6,227,310
Albanian	3.0 GB	437,287	326,325,149
Alemannic / Swiss German	363.6 kB	139	37,381
Amharic	461.0 MB	37,513	30,481,153
Arabic	84.2 GB	8,718,929	6,103,711,887
Aragonese	10.6 kB	12	51
Armenian	4.7 GB	379,267	268,031,270
Assamese	221.2 MB	17,084	11,109,557
Asturian	73.6 kB	77	3,919
Avaric	18.6 kB	14	582
Azerbaijani	3.5 GB	491,847	291,927,692
Bangla	15.1 GB	1,171,501	751,877,226
Bashkir	95.5 MB	11,198	5,418,474
Basque	1.1 GB	233,658	97,092,942
Belarusian	1.8 GB	180,046	107,227,860
Bihari languages	24.2 kB	27	569
Bishnupriya	2.0 MB	271	98,419
Bosnian	10.3 kB	10	422
Breton	33.7 MB	16,119	3,111,619
Bulgarian	35.1 GB	2,887,115	2,405,981,285
Burmese	1.9 GB	158,733	44,835,970
Catalan	13.9 GB	2,627,307	1,508,919,864
Cebuano	44.6 MB	5,742	5,253,785
Central Kurdish	716.4 MB	84,950	43,913,025
Chechen	14.0 MB	4,086	798,766

OSCAR

Based on a single CC snapshot

Cleaner than mC4 - Caswell et al. (2021) report, when sampling by language sample, 13% of OSCAR data have issues, while it is 28% for mC4

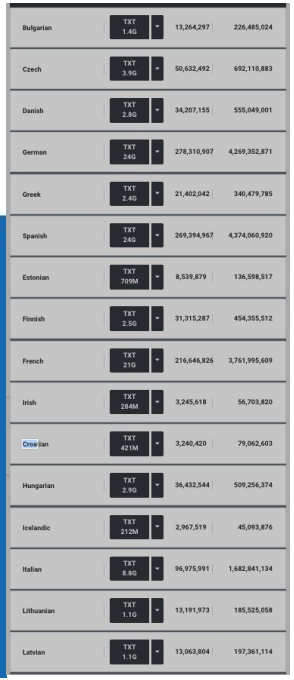
macocu

Existing corpora based on CommonCrawl

ParaCrawl

Broader/Continued Web-Scale Provision of Parallel Corpora for European Languages

Learn More



Bulgarian	TXT 1.4G	13,264,297	226,485,024
Czech	TXT 3.9G	50,632,492	692,110,883
Danish	TXT 2.8G	34,207,155	555,049,001
German	TXT 24G	278,310,907	4,269,352,871
Greek	TXT 2.4G	21,402,042	340,479,785
Spanish	TXT 24G	269,394,967	4,374,060,920
Estonian	TXT 709M	8,539,879	136,598,517
Finnish	TXT 2.5G	31,315,287	454,355,512
French	TXT 21G	216,646,826	3,761,995,609
Irish	TXT 284M	3,245,618	56,703,820
Croatian	TXT 421M	3,240,420	79,062,603
Hungarian	TXT 2.9G	36,432,544	509,256,374
Icelandic	TXT 212M	2,967,519	45,093,876
Italian	TXT 8.8G	96,975,991	1,682,841,134
Lithuanian	TXT 1.1G	13,191,973	185,525,058
Latvian	TXT 1.1G	13,063,804	197,361,114

MaCoCu uses the same bitext harvesting technology as ParaCrawl (just updated versions)

github.com/bitextor/bitextor

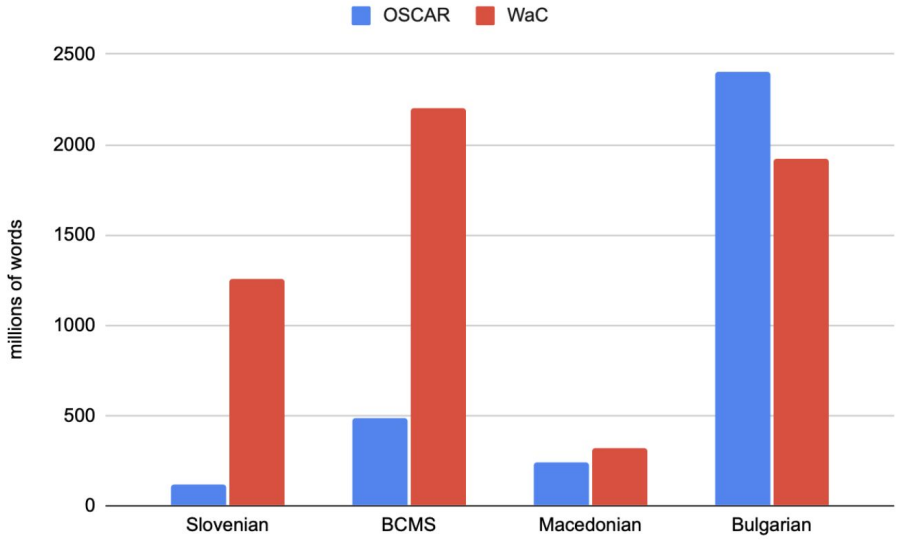
github.com/bitextor/bicleaner-ai

github.com/bitextor/bifixer

macocu

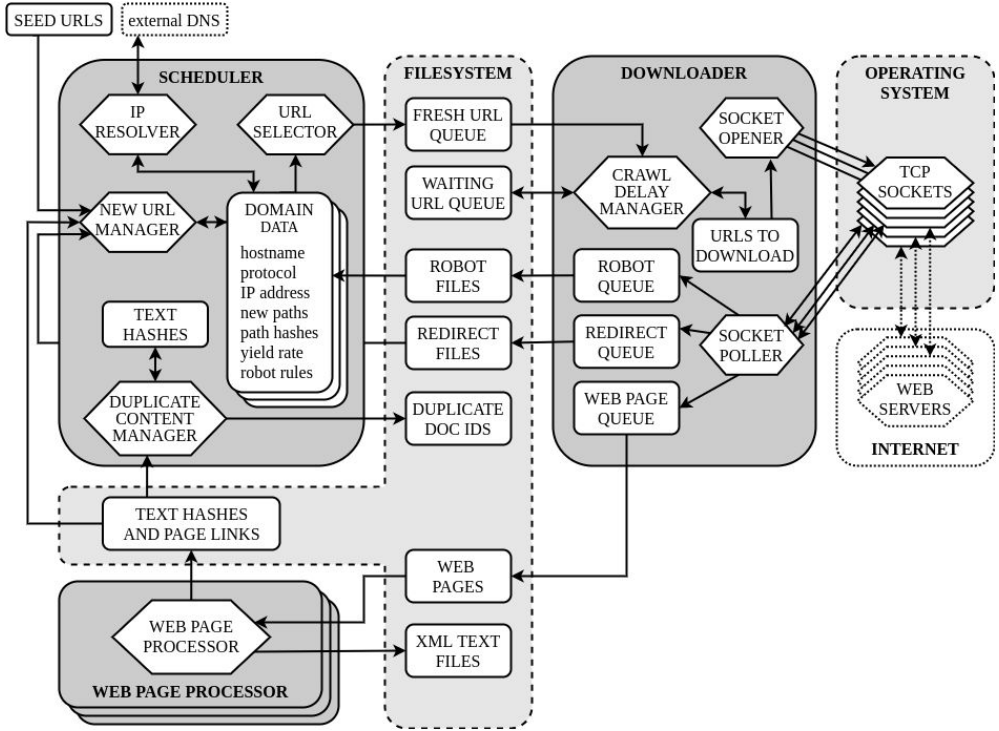
Why top-level-domain crawling?

- Tradition in the computational / corpus linguistic circles - Web-as-Corpus initiative - crawl selected portions of the web
- Amount of data for languages of interest sometimes very low in the best monolingual collection OSCAR



macocu

The SpiderLing crawler



macocu

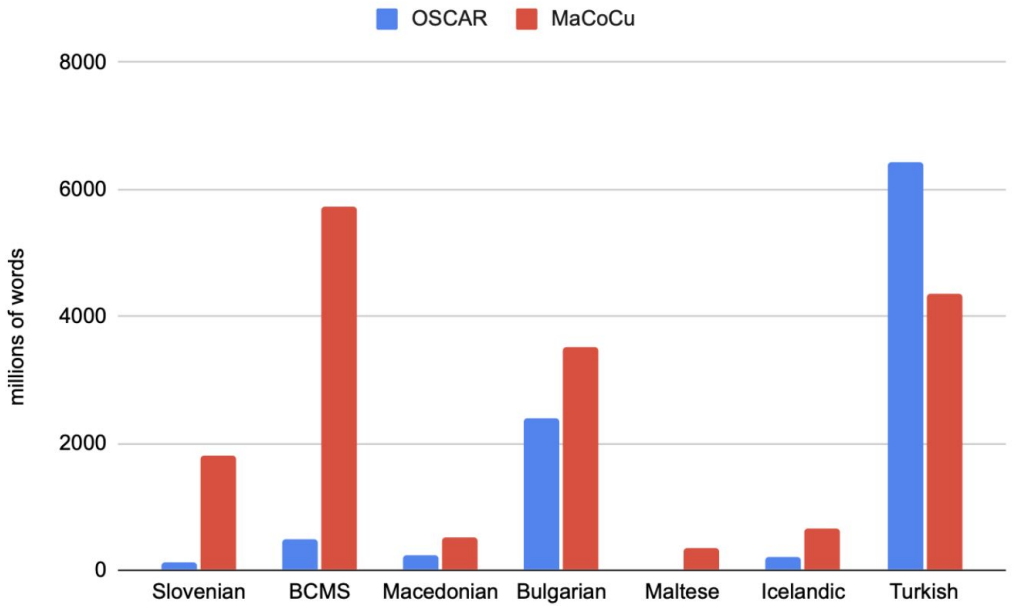
The SpiderLing crawler

Main features:

- Post-processing done on-the-fly - we know something about the quality of the crawled text already during crawling
  - Character encoding detection
  - Text extraction / boilerplate removal / quality prediction
  - Basic language identification
  - Text de-duplication
- Domain prioritization – the more high-quality and unique textual data a domain provides, the more it is crawled
  - Domains not yielding enough high-quality data stop being crawled
- Measures against low-quality content
  - Shorter URLs and shorter domains are crawled earlier and more often
  - Domains close to seed domains are prioritized

macocu

Final monolingual dataset sizes



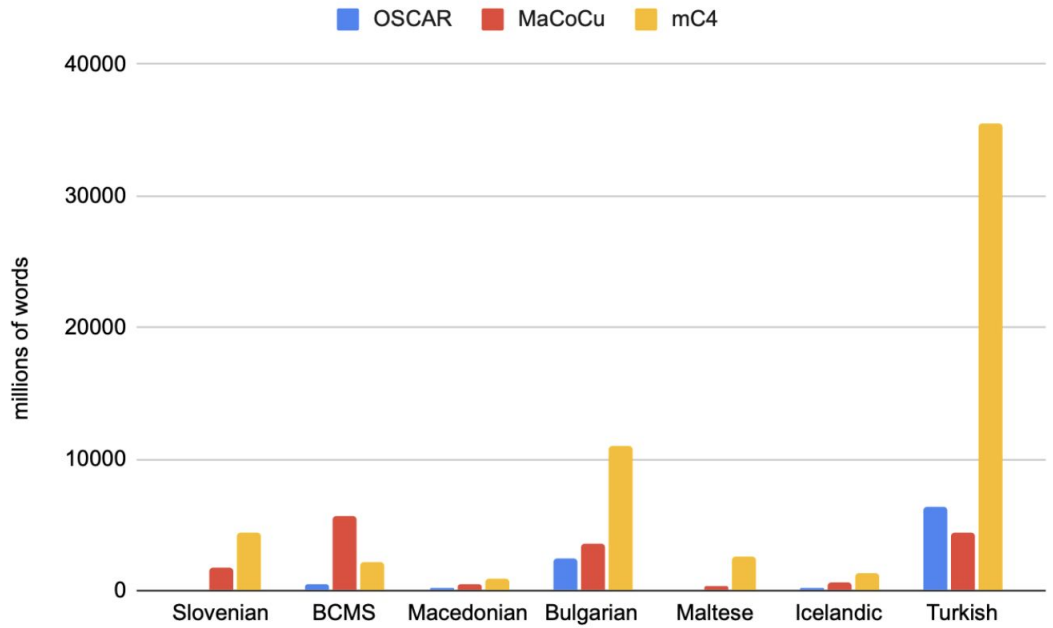
Still finalizing

- Albanian
- Catalan
- Greek
- Ukrainian

Similar trend in size

macocu

Let’s be fair and compare to mC4 as well



macocu

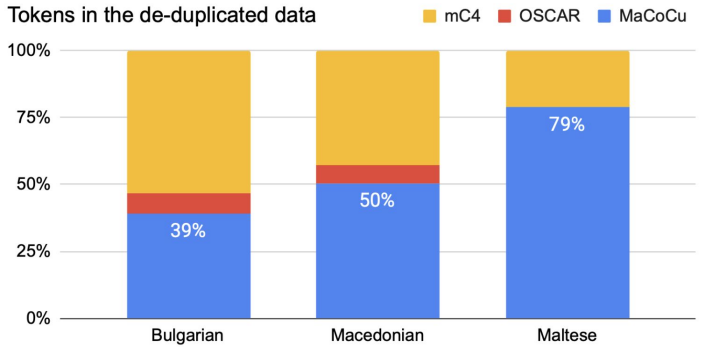
Overlap between monolingual corpora

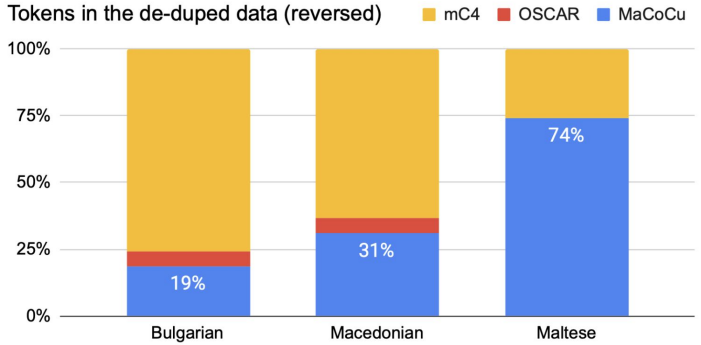
van Noord et al. (2021) near-deduplicate MaCoCu, OSCAR and mC4 data

- ~25% data due to stricter near-deduplication
- ~25% of data due to overlap to other datasets

Ljubešić and Lauc (2019) have a similar observation while training BERTić

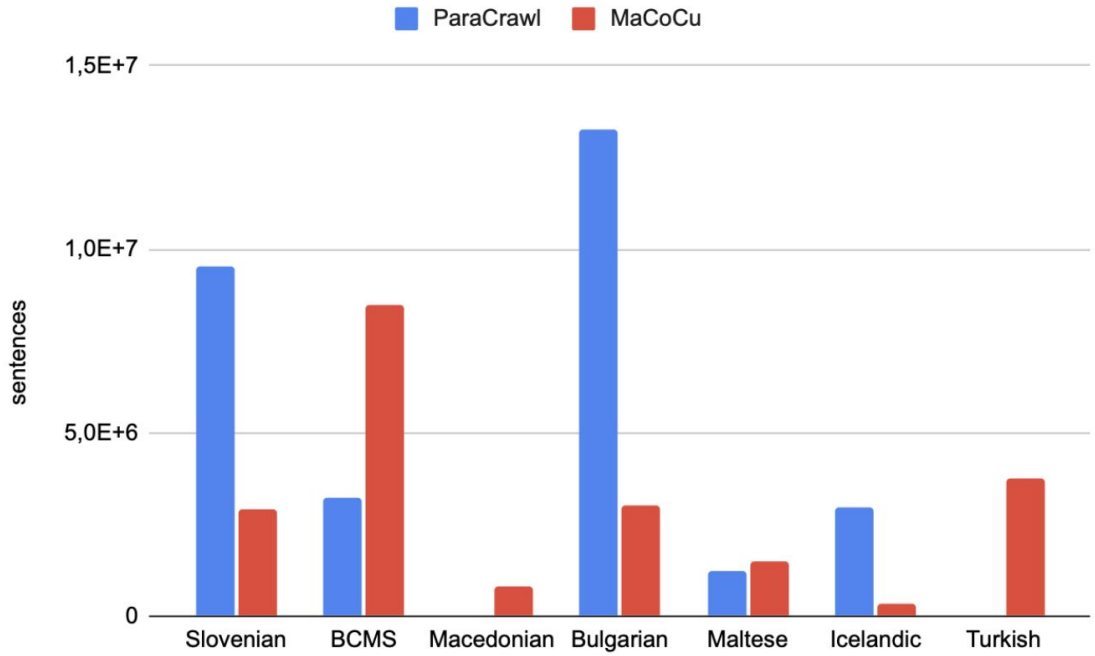
- cc100 had only 15% overlap with TLD crawls from 2011, 2014 and 2019
- the three TLD crawls had <10% overlap between each other





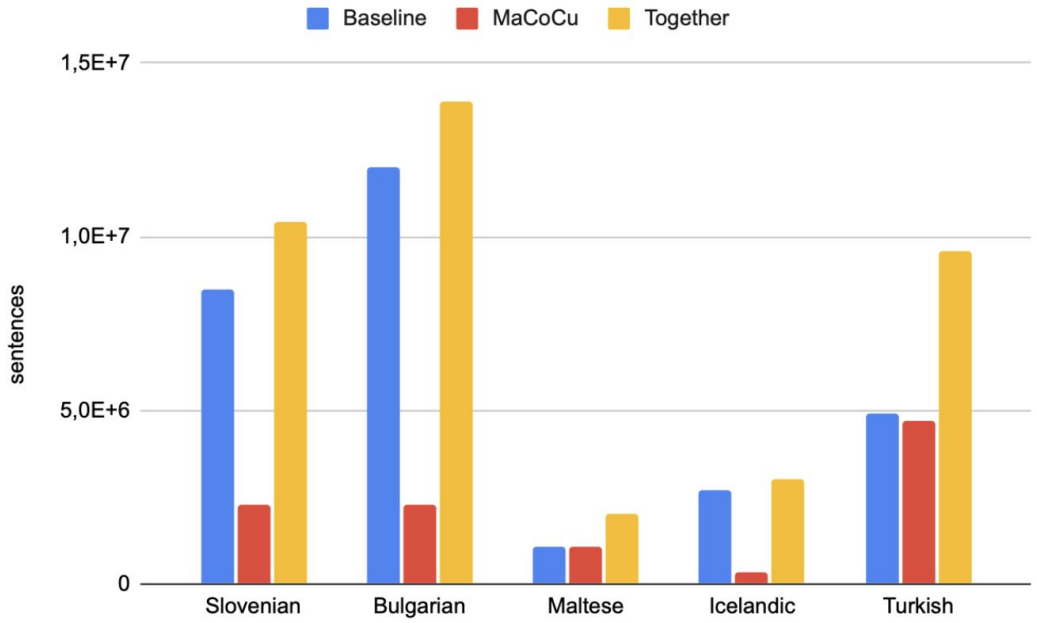
macocu

Final parallel dataset sizes



macocu

Overlap between parallel corpora



macocu

macocu

Data enrichment

Co-financed by the Connecting Europe Facility of the European Union

Language identification

FastSpell - extension of fasttext, using HunSpell dictionaries to double-check fasttext close calls between similar languages

https://github.com/mbanon/fastspell

Internal MaCoCu evaluation	
CLD2	95.3
CLD3	95.0
langid	94.2
fasttext	95.8
FastSpell	96.4
HeLI-OTS	97.1

Jauhiainen et al. 2022 (HeLI-OTS) report even more striking differences on benchmarks focused on smaller languages, observing that fasttext favours large languages

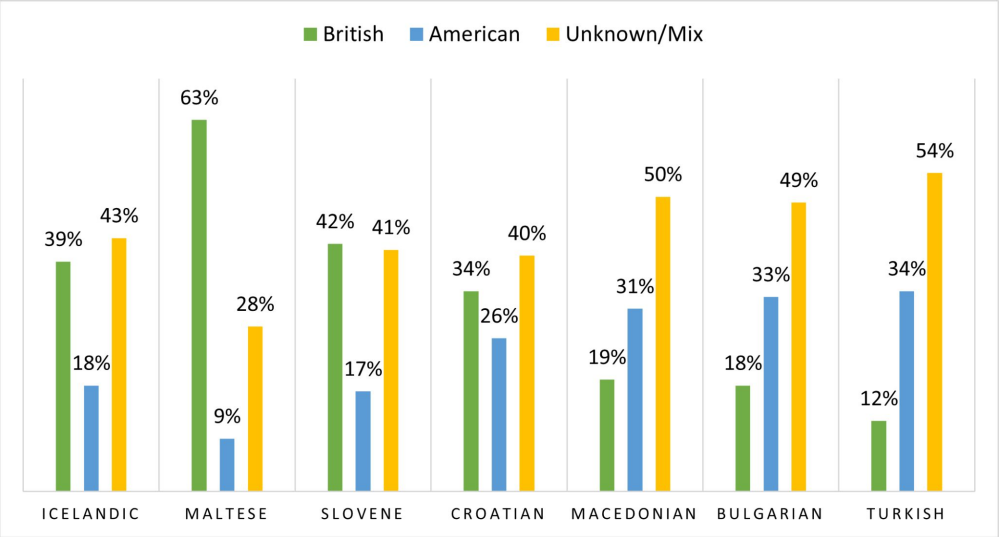
For MaCoCu corpora we use four language identifiers and robust heuristics

macocu

Language variety identification

Distinguishing American vs. British English in the English side of our parallel data via a lexicon-based method

https://github.com/macocu/American-British-variety-classifier



macocu

Language variety identification

BCMS to be classified into the four responding varieties

Off-the-shelf solutions cannot handle the problem

Training classifiers even on parallel data (news) does not transfer well to other domains (Twitter, web)

Heavy surface-feature-selection on web-scale data results in stable cross-domain results (semi-supervised transformer approaches, such as Caswell et al. (2020) still to be investigated)

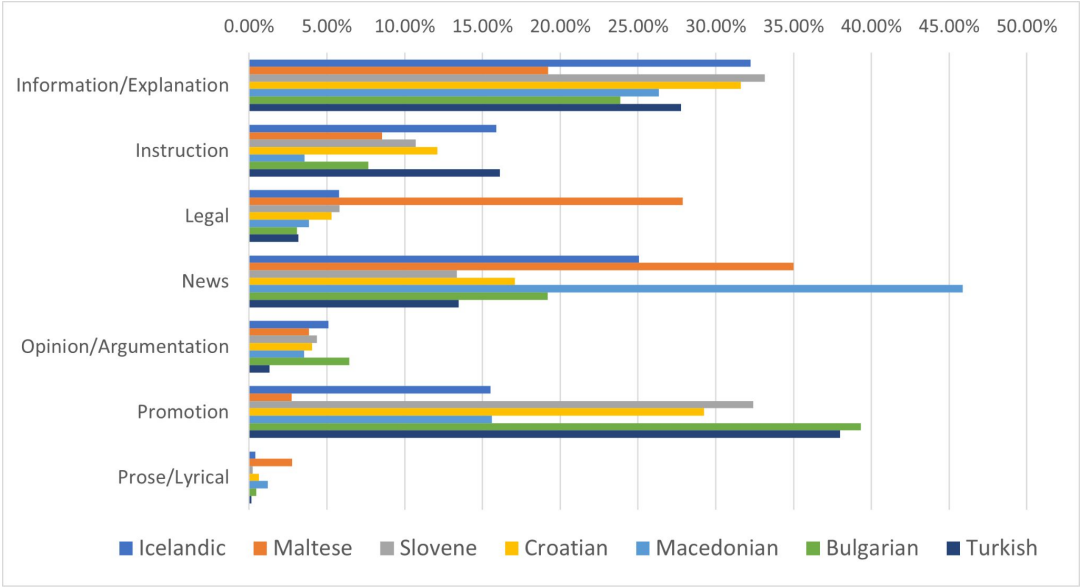
VarDial is celebrating its 10th birthday at EACL in Dubrovnik!

macocu

Genre identification

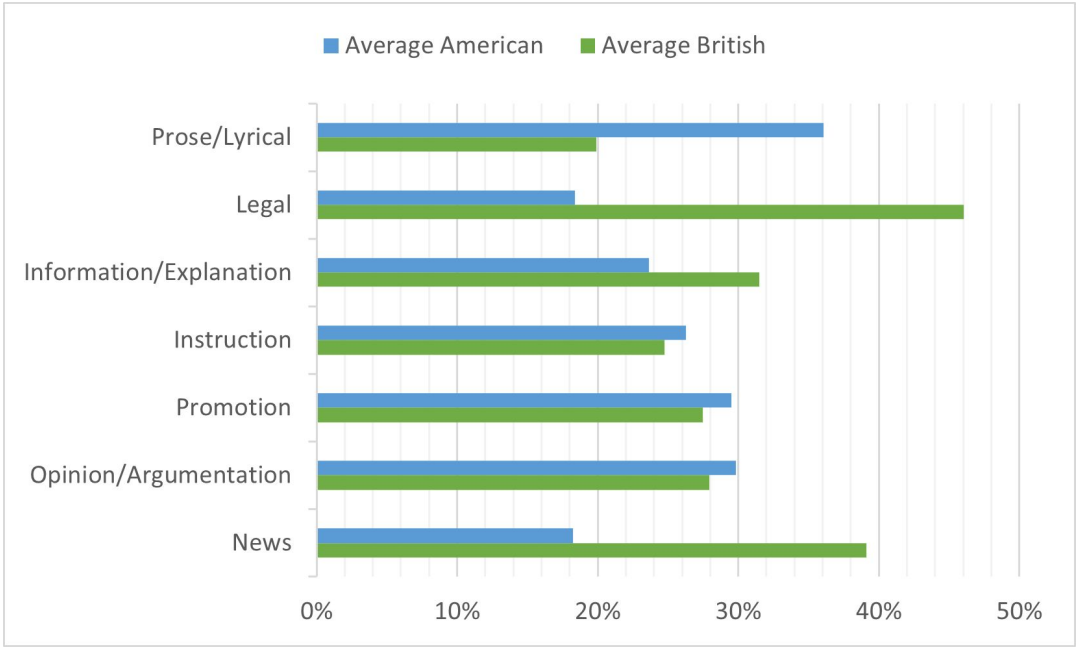
Performance on this task significantly improved with transformers (macro F1 0.36 to 0.62), combine syntactic and lexical features (Kuzman et al. 2021)

Very good results in cross-lingual settings (Laipalla et al. 2022)



macocu

Variety vs. genre



macocu

Translation direction identification

- Fine-tuning XLM-R on Europarl or our own annotated data

Lang	Train data	Eval data	Accuracy
bg	Europarl (20) + MaCoCu	Europarl (bg)	81.5%
hr	MaCoCu	MaCoCu	87.2%
is	Europarl (20) + MaCoCu	WMT	70.0%
mt	Europarl (20) + MaCoCu	News	54.7%
sl	MaCoCu	MaCoCu	80.0%
tr	Europarl (20) + MaCoCu	WMT	79.1%

Human vs. machine translation

Method:

- Given that we know something is a translation, can we automatically determine whether a human or a machine produced it?
- Use human translations from WMT test sets, translate with Google or DeepL to create balanced data sets
- Experiments on DE→EN, pre-trained transformers

Findings:

- 65-70% accuracy on the sentence-level, even higher on document-level (DeBERTa-v3 77%, Longformer 82%)
- Clear drop in performance (3-10%) when evaluating cross-MT system

Automatic Discrimination of Human and Neural MT: A Study with Multiple Pre-Trained Models and Longer Context. EAMT 2022. Tobias van der Werff, Rik van Noord and Antonio Toral

macocu

Human vs. machine translation

Ongoing:

- Experiments on additional language pairs
- Classifier with multilingual LM
  - Use also source text (additional clues)
  - Train on multiple languages pairs together
- Robustness
  - Train on multiple MT systems together
- Test on web-crawled data

Bitext classification by DSI domain

DSI - Digital Service Infrastructure

e-Justice, Cybersecurity, Safer Internet, e-Health, Open Data Portal, e-Procurement, Europeana, Online Dispute Resolution, Electronic Exchange of Social Security Information

Crawl our own data as none on ELRC are of useful quality

Best results obtained with DeBERTa-v3 (macro F1 0.68), encode also class probability while processing all MaCoCu bitext

Rik van Noord, Cristian García-Romero, Miquel Esplà-Gomis, Leopoldo Pla Sempere, and Antonio Toral. 2022. Building Domain-specific Corpora from the Web: the Case of European Digital Service Infrastructures. In Proceedings of the BUCC Workshop within LREC 2022.

macocu

Modelling
(Large language models)

Co-financed by the Connecting Europe Facility of the European Union

Language models and data quality

How sensitive are large language models to data quality?

Artetxe et al. (2022)

- Experiments on Basque
- Compare cc100 data (416M) and newly crawled, higher quality data (423M)
- The resulting language models perform very similarly on downstream evaluation tasks

van Noord et al. (2022)

- Experiments on Maltese (MaCoCu+OSCAR+mC4)
- Modelling results on a clean (146M) and mixed-quality (including MT-ed content, 439M) dataset
- Comparable results on downstream evaluation

macocu

Current best MaCoCu language models

XLM-R additionally pre-trained on MaCoCu data, Icelandic

	UPOS		XPOS		NER		COPA
Model	**Dev**	**Test**	**Dev**	**Test**	**Dev**	**Test**	**Test**
XLM-R-base (baseline)	96.8	96.5	94.6	94.3	85.3	89.7	55.2
XLM-R-large (baseline)	97.0	96.7	94.9	94.7	88.5	91.7	54.3
IceBERT (baseline)	96.4	96.0	94.0	93.7	83.8	89.7	54.6
XLM-R + MaCoCu (75k)	97.3	97.0	95.4	95.1	90.8	93.2	59.6

Current best MaCoCu language models

XLM-R additionally pre-trained on MaCoCu data, Turkish

	UPOS		XPOS		NER		COPA	
Model	**Dev**	**Test**	**Dev**	**Test**	**Dev**	**Test**	**Test (MT)**	**Test (HT)**
XLM-R-base (baseline)	89.0	89.0	90.4	90.6	92.8	92.6	56.0	53.2
XLM-R-large (baseline)	89.4	89.3	90.8	90.7	94.1	94.1	52.1	50.5
BERTurk (baseline)	88.2	88.4	89.7	89.6	92.6	92.6	57.0	56.4
XLM-R + MaCoCu (70k)	89.1	89.4	90.7	90.5	94.4	94.4	60.7	58.5

Overall trend - harder to improve on linguistic tasks, easier on tasks requiring extralinguistic knowledge (NER) or reasoning (COPA)

macocu

macocu

Modelling
(Machine translation)

Co-financed by the Connecting Europe Facility of the European Union

Parallel datasets

Language	Baseline corpora	Baseline size	MaCoCu size	Total size
Bulgarian	ParaCrawl, CommonCrawl	14,424,709	2,326,861	16,245,592
Croatian	ParaCrawl, CommonCrawl, Tilde	5,106,096	2,114,409	7,029,981
Icelandic	ParaCrawl, CommonCrawl, Tilde	3,281,592	349,246	3,537,840
Maltese	ParaCrawl, Tilde	2,291,030	1,056,735	3,175,333
Slovene	ParaCrawl, CommonCrawl, Tilde	10,870,935	2,337,395	12,741,147
Turkish	CommonCrawl	4,927,668	4,709,457	9,575,968

macocu

Baseline + MaCoCu

Is there anything to be gained from MaCoCu?

	Bulgarian		Croatian		Icelandic		Maltese		Slovene		Turkish	
	CO	BL	CO	BL	CO	BL	BS	BL	CO	BL	CO	BL
Baseline	79.2	41.0	80.6	29.6	47.0	23.5	80.7	37.6	77.0	29.8	81.3	28.2
Baseline + MaCoCu	**+0.9**	**+0.3**	**+0.5**	**+0.2**	**+2.4**	**+0.4**	**+1.0**	**+2.1**	**+0.3**	**+0.1**	**-1.2**	**+0.2**

macocu

Baseline + MaCoCu

Is there anything to be gained from MaCoCu?

	Bulgarian		Croatian		Icelandic		Maltese		Slovene		Turkish	
	CO	BL	CO	BL	CO	BL	BS	BL	CO	BL	CO	BL
Baseline	79.2	41.0	80.6	29.6	47.0	23.5	80.7	37.6	77.0	29.8	81.3	28.2
Baseline + MaCoCu	**+0.9**	**+0.3**	**+0.5**	**+0.2**	**+2.4**	**+0.4**	**+1.0**	**+2.1**	**+0.3**	**+0.1**	**-1.2**	**+0.2**

What do humans think?

	bg	hr	is	mt	sl	tr
Without MaCoCu preferred (%)	9.2	25.7	22.9	26.1	37.4	36.2
Same quality (%)	79.8	43.9	49.5	36.2	22.9	22.3
With MaCoCu preferred (%)	11.0	30.5	27.6	37.7	39.6	41.5

macocu

ParaCrawl vs. MaCoCu

Are MaCoCu data better than ParaCrawl (we control for size)

	Bulgarian		Croatian		Maltese		Slovene	
	CO	BL	CO	BL	BS	BL	CO	BL
ParaCrawl	68.4	35.5	72.9	26.8	80.4	35.4	67.2	26.4
MaCoCu	-2.9	-1.4	-1.4	-0.5	+0.3	0.4	+2.9	+0.2

macocu

The importance of bitext deduplication

COMET scores on Flores:

	bg	hr	mt	sl	tr	avg
Lenient deduplication	64.0	71.5	51.7	65.3	51.6	60.8
Strict deduplication	**67.8**	**74.3**	**52.1**	**66.8**	**66.0**	**65.4**

Data set sizes:

	bg	hr	mt	sl	tr
Lenient deduplication	3.89M	3.09M	1.23M	3.16M	10.35M
Strict deduplication	2.16M	1.92M	0.97M	2.15M	3.80M

macocu

Main takeaways

In resource-building projects **human inspection of data** is crucial!

Next thing to do - enrich your data with automatic methods to understand what your data consist of, also enables downstream users the select subsets

Current technology favours large languages - downward spiral for small languages

The “solved” task of language identification seems to be one of the most burning issues in collecting and using large collections of web data

The CC and TLD collections seem to contain rather different data - complementarity to be exploited?

Snapshots / crawls a few years apart contain rather little duplicates

Both monolingual and bilingual MaCoCu data have added value for downstream models

MaCoCu datasets and models

Datasets are available from the CLARIN.SI repository (https://www.clarin.si/repository/xmlui/)

Parallel datasets will also be included in OPUS (https://opus.nlpl.eu)

Models are available through HuggingFace (https://huggingface.co/MaCoCu)

We still have 6 months ahead of us, so many additional datasets and models coming!

macocu

MaCoCu datasets -> CLASSLA corpora

We will publish the monolingual MaCoCu datasets of South Slavic languages as linguistically annotated corpora in the CLARIN.SI concordancers (CLASSLA is the CLARIN knowledge centre for South Slavic languages)

“WaC” are currently go-to corpora for most South-Slavic languages

Question for discussion: why do we not produce linguistic corpora from most web-based datasets? Linguists are very interested! What are the main hurdles?

Special Interest Group on Web as Corpus

ACL SIGWAC https://www.sigwac.org.uk

Set-up in 2005 by Marco Baroni, Stephanie Evert, Adam Kilgarriff et al.

Strong corpus-linguistic focus on exploiting web data

Since very recently SIGWAC is led by Veronika Laippala, Benoît Sagot, Pedro Ortiz Suarez and me, our plan is to widen the focus of the SIG on researchers using large-scale, (primarily) textual web data collections for different research aims and directions.

Pedro and I invite you to join us by signing up to the e-mail list (equals membership) here: http://devel.sslmit.unibo.it/mailman/listinfo/sigwac

macocu

Next steps

Iterative TLD crawling - infrastructure already set-up inside CLARIN.SI

Further data enrichment + research in the consequences of various types of data bias

Challenges of machine-generated data

Other modalities - speech, image, video

Further exploitation of the structural richness of the web

Focus on African and Asian languages

The MaCoCu crowd

Marta Bañón

Mălina Chichirău

Miquel Esplà-Gomis

Mikel L. Forcada

Cristian García-Romero

Taja Kuzman

Nikola Ljubešić

Rik van Noord

Leopoldo Pla Sempere

Gema Ramírez-Sánchez

Peter Rupnik

Vít Suchomel

Antonio Toral

Tobias van der Werff

Jaume Zaragoza

macocu

Thank you!

Co-financed by the Connecting Europe Facility of the European Union

This action has received funding from the European Union's Connecting Europe Facility 2014-2020 - CEF Telecom, under Grant Agreement No. INEA/CEF/ICT/A2020/2278341.
This communication reflects only the author's view. The Agency is not responsible for any use that may be made of the information it contains.

macocu