MaCoCu: Massive collection and curation of monolingual and bilingual data: focus on under-resourced languages

Marta Bañón[†], 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[†]

[‡]Jožef Stefan Institute, [†]Prompsit, [♦]Rijksuniversiteit Groningen, [★]Universitat d'Alacant

[‡]{taja.kuzman,nikola.ljubesic,peter.rupnik}@ijs.si, [‡]vit.suchomel@sketchengine.eu, [†]{mbanon,gramirez,jzaragoza}@prompsit.com,
[♦]{r.i.k.van.noord,a.toral.ruiz,t.n.van.der.werff}@rug.nl, [★]{mespla,mlf,cgarcia,lpla}@dlsi.ua.es

1.MaCoCu at a glance

MaCoCu's objective is **collecting and curating relevant data to improve MT** for the European Commission (and also for everyone else!).

A 2-year action started in June 2021. Aimed at collecting **monolingual and parallel data** from the Internet for **under-resourced languages**.

Consortium:

Universitat d'Alacant
Universidad de Alicante

university of groningen

Jožef Stefan Institute

prompsit

2. Targeted languages

Ten languages covered by MaCoCu, including:

- especially low-resourced;
- South-Slavic languages;
- official languages of EU candidate states.

Covered during the **first** year:

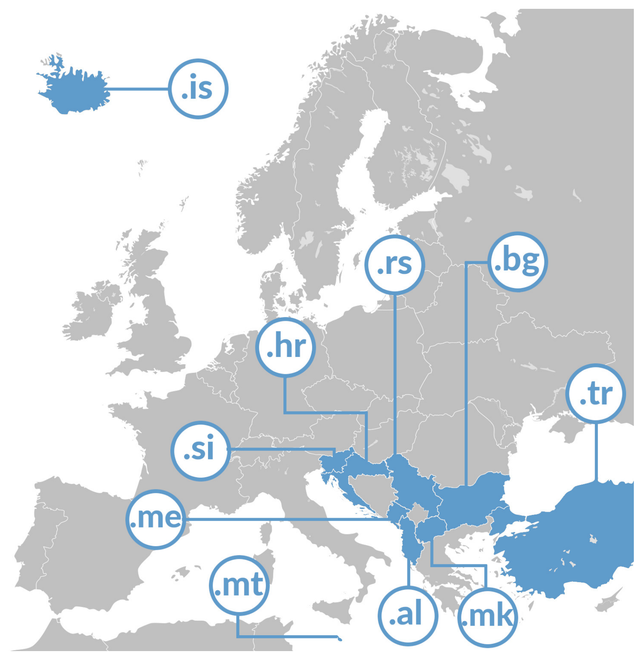
- Maltese (mt)
- Bulgarian (bg)
- Slovenian (sl)
- Croatian (hr)
- Turkish (tr)
- Macedonian (mk)
- Icelandic (is)

Covered during the **second** year:

- Montenegrin (cnr)
- Albanian (sq)
- Serbian (sr)

3. Data acquisition

Textual data in MaCoCu is obtained by **crawling the national top-level domains** of the countries where targeted languages are official.



The tools and models used in MaCoCu are publicly available: `https://github.com/MaCoCu/documentation`.

4. Monolingual data

Monolingual corpora are made of documents split by paragraphs. Sample from the Maltese corpus:

```
<corpus id="MaCoCu-mt">
  <doc id="macocu.mt.1" title="GWIDA.mt - Magħkom bl-aħħar aħbarijiet mid-dinja tal-ispettaklu"
crawl_date="2021-07-12" lang_distr="[('mt', 1.0)]" url="https://www.gwida.mt/" domain="gwida.mt" file_type="html"
lm_score="0.506">
    <p id="macocu.mt.1.1" quality="neargood" lm_score="0.499" lang="mt" heading="yes">
    Bil-Filmat: Diversi Żgħażagħ Barranin Qaqoċċa Ħdejn Xulxin Jiżfnu Mal-Mużika Fil-Headphones
    </p>
    <p id="macocu.mt.1.2" quality="good" lm_score="0.460" lang="mt">
    Issa Bil-Filmat: Diversi Żgħażagħ Barranin Qaqoċċa Ħdejn Xulxin Jiżfnu Mal-Mużika Fil-Headphones https://-
www.gwida.mt/mt/bil-filmat-diversi-g-a-ag-aqo-a-dejn-xulxin-ji-fnu-mal-mu-ika-fil-headphones Shanice Ahar 8 ta'
Lulju
    </p>
    <p id="macocu.mt.1.3" quality="neargood" lm_score="0.397" lang="mt" heading="yes">
    Bil-Filmat: JD Taqbiżlu Ma' Tina Waqt Xandira Diretta! Ara X'Għamel...
    </p>
    <p id="macocu.mt.1.4" quality="neargood" lm_score="0.592" lang="mt" heading="yes">
    Bil-Filmat: Ara Kif Irreaġixxew Dawn Il-Barranin Għall-Punti Ta' Malta Fil-Eurovision Song Contest!
    </p>
...
```

Metadata at document level:

- URL of the original document;
- the crawling date;
- content type;
- score from language model;
- paragraph language distribution in the document.

Metadata at paragraph level:

- identifier of the position of the paragraph within the document;
- paragraph estimated quality;
- score from language model;
- paragraph-level language.

Sizes in millions of documents and words for the 7 languages already released are:

	tr	bg	hr	sl	mk	is	mt
# docs.	16.0	10.5	7.3	5.8	2.0	1.7	0.5
# words	4346.3	3508.9	2318.3	1779.1	524.1	644.5	347.9

5. Parallel data

Parallel corpora are made of segment pairs. Sample from the Maltese-English corpus:

```
<?xml version="1.0"?>
<tmx version="1.4">
...
  <tu tuid="5" datatype="Text">
    <prop type="score-bicleaner-ai">0.534</prop>
    <prop type="biroamer-entities">No</prop>
    <prop type="translation-direction">mt-orig</prop>
    <prop type="type">1:1</prop>
    <tuv xml:lang="en">
     <prop type="source-document">https://www.etwinning.net/en/b-trendsissuespolicies-an.htm</prop>
     <prop type="checksum-seg">51f1cea479c4ed40</prop>
     <prop type="paragraph-id">p44s10</prop>
    <prop type="english-variant-document">B</prop>
    <prop type="english-variant-domain">B</prop>
     <seg>Where should mobile phones be kept during class?</seg>
    </tuv>
    <tuv xml:lang="mt">
     <prop type="source-document">https://www.etwinning.net/mt/b-trendsissuespolicies-an.htm</prop>
     <prop type="checksum-seg">9ca0cef421f41e10</prop>
     <prop type="paragraph-id">p44s9</prop>
     <seg>Fejn għandhom jinżammu l-mobajls waqt l-iskola?</seg>
    </tuv>
   </tu>
...
```

Metadata at sentence pair level:

- Quality score (between 0.0 and 1.0) provided by Bicleaner;
- Personal data detected in the segment pair with BiROAMer;
- Translation direction automatically detected;
- English variant (British or American) automatically identified.

Sizes in millions of segment pairs and words for the 7 pairs (aligned with English) already released are:

	tr	bg	hr	sl	mk	is	mt
# segs.	10.3	3.9	3.1	3.2	0.5	0.4	1.2
# words	513.5	158.7	134.9	137.0	23.9	14.4	69.6

6. Acknowledgements

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. The contents of this publication are the sole responsibility of its authors and do not necessarily reflect the opinion of the European Union.