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

Miquel Esplà-Gomis
mespla@dlsi.ua.es
Universitat d'Alacant

5th ELRC conference | March 10, 2021

1

Introduction

What is **MaCoCu**? Who is involved?

MaCoCu: What, when, who?

- What is MaCoCu?
- When will MaCoCu happen?
- Who is involved?

MaCoCu: What, when, who?

What is MaCoCu? It is a CEF action that focuses on collecting **monolingual** and **parallel data** from the Internet. It is specially focused on **under-resourced** languages and on **DSI-specific** data.

When will MaCoCu happen?

Who is involved?

MaCoCu: What, when, who?

What is MaCoCu?

When will MaCoCu happen? The action will start on June 2021 and will last for 2 years.

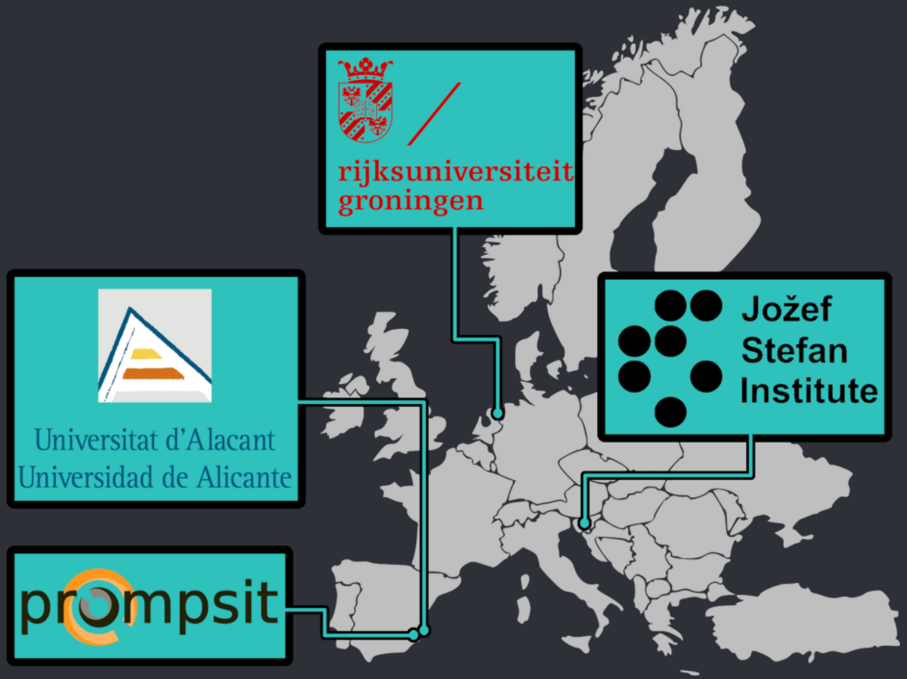
Who is involved?

MaCoCu: What, when, who?

What is MaCoCu?

When will MaCoCu happen?

Who is involved?



Consortium background

АБУМАТРАН ABUMATRAN

When: 2013 to 2016

Focus: Automating resources collection and MT building

Targeted languages: Croatian and other South-Slavic languages from Candidate States

ParaCrawl

When: 2017 to 2021

Focus: Building parallel corpora from documents available on the Internet: crawling + mining pre-existing resources (Common Crawl, Internet Archive)

Targeted languages: EU-official languages + Icelandic, Norwegian, Basque, Galician & Catalan

gourmet

When: 2019 to 2022

Focus: Building MT for under-resourced languages for the news domain

Targeted languages: Under-resource languages (many from Asia and Africa): Gujarati, Swahili, Tamil, Amharic, Kyrgyz, Hausa, Igbo, Tigrinya, Turkish, Bulgarian, Serbian, Macedonian, etc.

2

Objectives of the Action

What is **MaCoCu** aimed at? Which are the expected outcomes?

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

So... what do we mean by *relevant data*?

Relevant languages

- **Under-resourced languages of member states**: Maltese, Bulgarian, Slovenian, and Croatian + Icelandic
- **Under-resourced languages of candidate states**: Turkish, Albanian, Macedonian, Montenegrin, and Serbian

Relevant domain

- **Identifying data relevance for 10 Digital Single Infrastructures (DSIs):** e-Health, e-Justice, Online Dispute Resolution, Europeana, Open Data Portal, Business Registers Interconnection System, e-Procurement, Safer Internet, Cybersecurity, and Electronic Exchange of Social Security Information

Expected amount of data to be produced

Our objective is a minimum size of **5M** tokens per parallel and **10M** per monolingual corpus:

- focus on **quality**
- **enriched** with additional information
- **new**: TLD crawling, not re-using previously crawled data

Enriched parallel and monolingual data

MaCoCu data will be enriched with:

1. **Quality scores** and other indicators from **ELRC guidelines** for a cleaner corpus
2. **Language variety** identification
3. Information for **anonymisation**
4. For parallel data: **Source language** identified (*translationese*)

● Relevant data for DSIs

○ Classification of documents regarding their relevance for **10 targeted DSIs**

Initial evaluation on existing **Spanish** and **Dutch** data

Extension to the **10 under-resourced languages** targeted

Our final outcome

10 monolingual corpora & **10 parallel** corpora with which one can generate MT training data...

... that is **anonymised**

... for specific **language variants**

... for a specific **DSI domain**

... with the optimal **compromise size vs. clean**

And... all the **code** developed available at **Github**!

Thank you very much for your attention!

Any questions?

You can also contact me: **mespla@dlsi.ua.es**

3

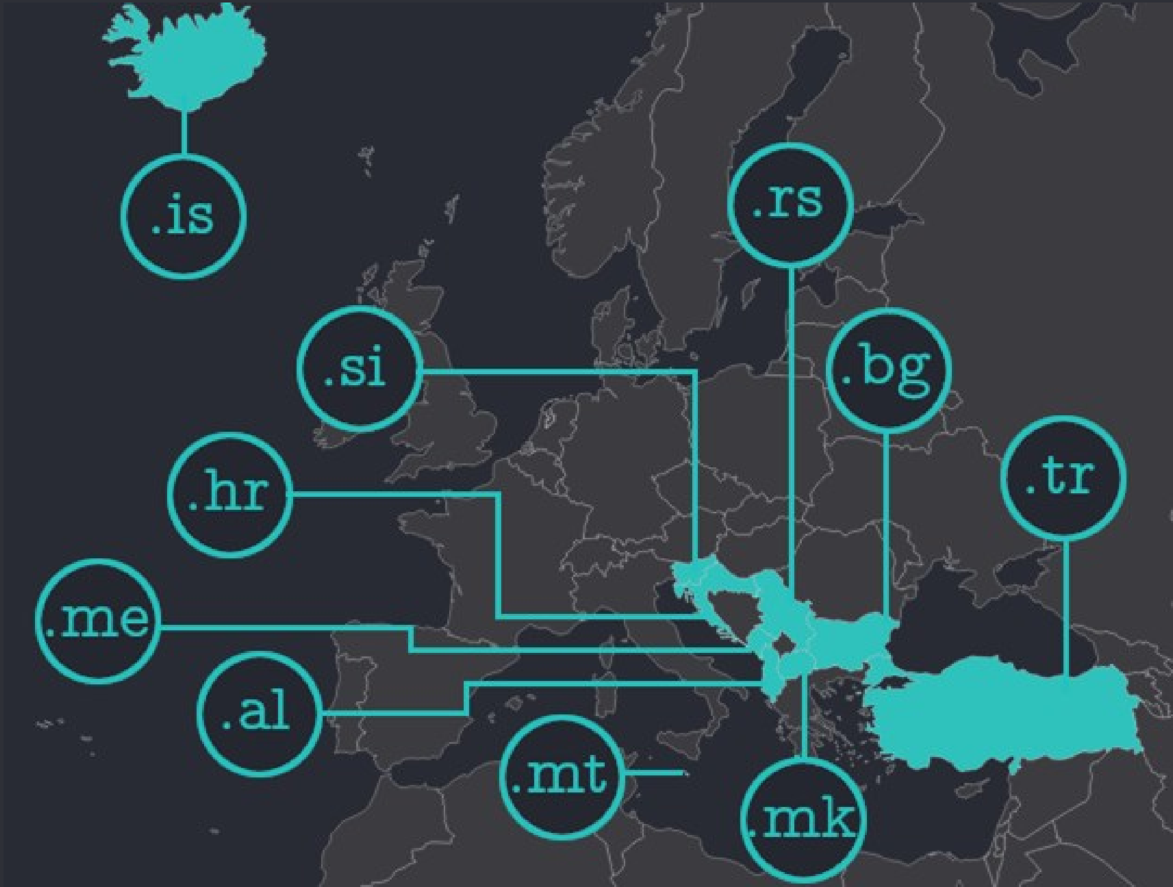
Backup slides

Support slides for discussion

Targeted languages: top-level domain crawling

Our strategy to acquire data is to crawl 10 TLDs looking for parallel and monolingual content

For parallel, mostly data aligned to English, but also other possible language pairs



Existing parallel corpora for targeted languages

Two main parallel data corpora:

- Paracrawl: **~15M** tokens for **Maltese**, **~30M** tokens for **Icelandic** (mined from the Internet Archive)
- MultiCCAligned: **~25M** tokens for languages such as **Albanian** or **Macedonian** (mined from Common Crawl)
  - ✓ ~46% of English and ~0.003% of Maltese: **slow** addition of under-resourced languages

Milestones of the Action 1/

June 2021 Starting date for the Action

Sep. 2021 Public website and social media: follow MaCoCu progress!

May 2022 First version of parallel and monolingual corpora for Bulgarian, Croatian, Slovenian, Icelandic, Maltese and Turkish. DSI-specific data for Spanish and Dutch

June 2022 Public evaluation report & initial software release

Milestones of the Action 2/

May 2022 Final version of parallel and monolingual corpora for Bulgarian, Croatian, Slovenian, Icelandic, Maltese and Turkish + Albanian, Macedonian, Montenegrin and Serbian. DSI-specific data for all these 10 languages

June 2022 Public evaluation report & final software release

and then... New projects? New languages? Larger datasets? Who knows!