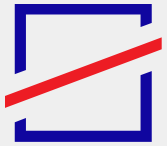
# Constructing Parallel Corpora from COVID-19 News using MediSys Metadata

Dimitris Roussis,
Vassilis Papavassiliou,
Sokratis Sofianopoulos,
Prokopis Prokopidis,
& Stelios Piperidis

**LREC 2022,**
**Marseille, France**

ATHENA Research & Innovation Information Technologies

# Outline of the Presentation

- Introduction
- MT for Crisis Response
- MediSys Infrastructure & Metadata
- Our Approach
  - Outline of our Methodology
  - Filtering Pipeline of the Parallel Corpora
  - Alternative Translations
- Statistics of Parallel Corpora
- Statistics of Monolingual Corpora
- Domain of the Corpus
- Conclusion

# Introduction

- The COVID-19 pandemic had a tremendous impact on the globe and increased our dependence on digital platforms. The abundance of related textual information led to many AI (Artificial Intelligence) applications and multidisciplinary initiatives (e.g. CORD-19).
- MT (Machine Translation) models focused on translating COVID-19 related information made use of newly created (TICO-19, TAUS Corona Crisis Corpus, etc.) and existing (e.g. OPUS EMEA) parallel corpora.
- We exploited ~57 million URLs gathered by the MediSys (Medical Information System) infrastructure to create comparable monolingual corpora and mine ~11.2 million sentence pairs related to COVID-19 for 26 EN-X language pairs.

# MT for Crisis Response

- MT can prove a useful tool in emergency situations. The COVID-19 pandemic unfolded as a global crisis with extended duration. The term "**infodemic**" was used to describe the scale of misinformation campaigns (e.g. "fake news", conspiracy theories) during the pandemic.
- There have been initiatives (4th LoResMT, COVID-19 MLIA-Eval, etc.) which focused on facilitating access to guidelines, news, announcements, high-quality articles, etc. This is vital for a wide range of stakeholders, such as healthcare professionals, researchers, immigrants, the elderly, etc.
- Some of the corpora we constructed have been used at the **COVID-19 MLIA-Eval**; the monolingual corpora for the Multilingual Semantic Search task and the parallel corpora for the MT task.

# MediSys Infrastructure and Metadata

- The EMM (Europe Media Monitor) / **MediSys** infrastructure processes news media to automatically identify potential public health threats.
- A dataset of metadata focused on COVID-19 related news articles was made publicly available in RSS/XML format.
- Originally, the MediSys dataset comprised of automatically extracted metadata (e.g. URL, title, language, identified named entities, etc.) categorized into groups according to the publication date of the news articles from which they originate.

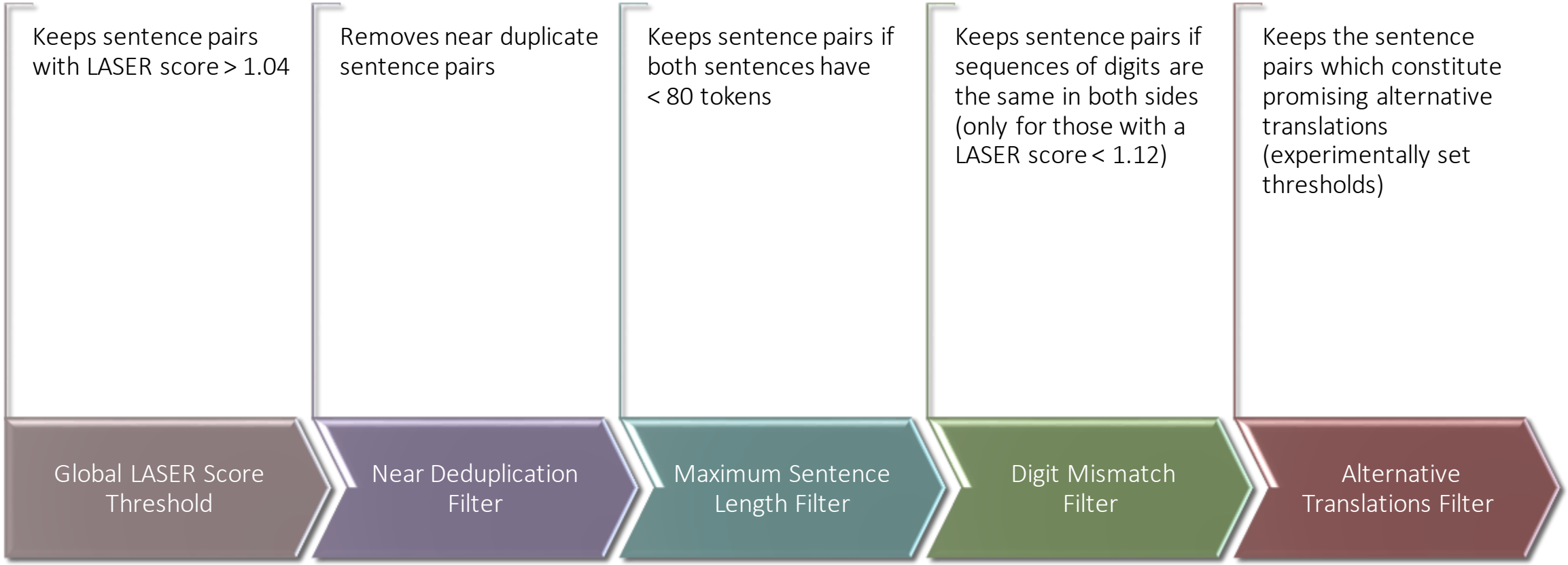
# Our Approach

- We selected batches spanning across 10 months (December 2019 to September 2020), parsed the metadata and downloaded ~57 million extracted URLs for several selected languages.
- We considered that two monolingual corpora are comparable since:
  1. They were published in a specific period (e.g. Greek news of March & English news of March)
  2. They originated from COVID-19 news articles (many were republished on several portals and languages)
- Therefore, we used them to mine sentence pairs.

# Outline of our Methodology

- **Parse** — Parse MediSys dataset (language and URL of each entry)
- **Fetch** — Get ~57 million URLs (12% failed to download)
- **Clean** — Remove boilerplate with ILSP-FC Cleaner Module
- **Merge** — Create comparable monolingual corpora (language-period)
- **Process** — Sentence Splitting, Language Identification, Deduplication
- **Mine** — Use LASER to mine parallel sentences (assigns alignment score)
- **Filter** — Apply a filtering pipeline to increase MT-readiness of data

# Filtering Pipeline of the Parallel Corpora



# Alternative Translations

- We investigated cases in which a source sentence has multiple translations in a target language (or vice versa).
- In **Table 1**, we can see an example from the EN-NL (English-Dutch) bilingual corpus in which the Dutch sentence has two alternative translations in English.
- As we can see, both are valid.
  The alternative translations filter was used to preserve such cases as they could prove useful in the construction of paraphrase datasets (e.g. ParaBank).

| Source Sentence | Target Sentence |
|---|---|
| T cells might be able to recognize virally infected cells and destroy them, limiting the virus's spread in the body. | T-cellen herkennen met virussen geïnfecteerde cellen en vertellen deze cellen dat ze zichzelf moeten vernietigen, zodat het virus zich niet door het hele lichaam kan verspreiden. |
| T-cells recognize virus-infected cells and tell those cells to self-destruct, preventing the virus from spreading throughout the body. | |

Table 1: Examples of alternative translations

# Statistics of Parallel Corpora

| EN- | **AR** Arabic | **BG** Bulgarian | **CS** Czech | **DA** Danish | **DE** German | **EL** Greek | **ES** Spanish | **ET** Estonian | **FI** Finnish |
|---|---|---|---|---|---|---|---|---|---|
| | 355,536 | 662,595 | 244,503 | 171,727 | 1,076,666 | 529,518 | 1,488,765 | 70,803 | 111,589 |
| **EN-** | **FR** French | **HR** Croatian | **HU** Hungarian | **IS** Icelandic | **IT** Italian | **LT** Lithuanian | **LV** Latvian | **MK** Macedonian | **NL** Dutch |
| | 1,134,809 | 230,857 | 111,992 | 4,265 | 738,917 | 149,897 | 123,893 | 182,438 | 462,361 |
| **EN-** | **NO** Norwegian | **PL** Polish | **PT** Portuguese | **RO** Romanian | **SK** Slovak | **SL** Slovenian | **SQ** Albanian | **SV** Swedish | **Total** |
| | 111,642 | 565,915 | 1,062,473 | 693,614 | 289,454 | 122,186 | 344,204 | 283,122 | **11,248,573** |

Table 2: Corpus statistics for all EN-X language pairs – 1.04 filter

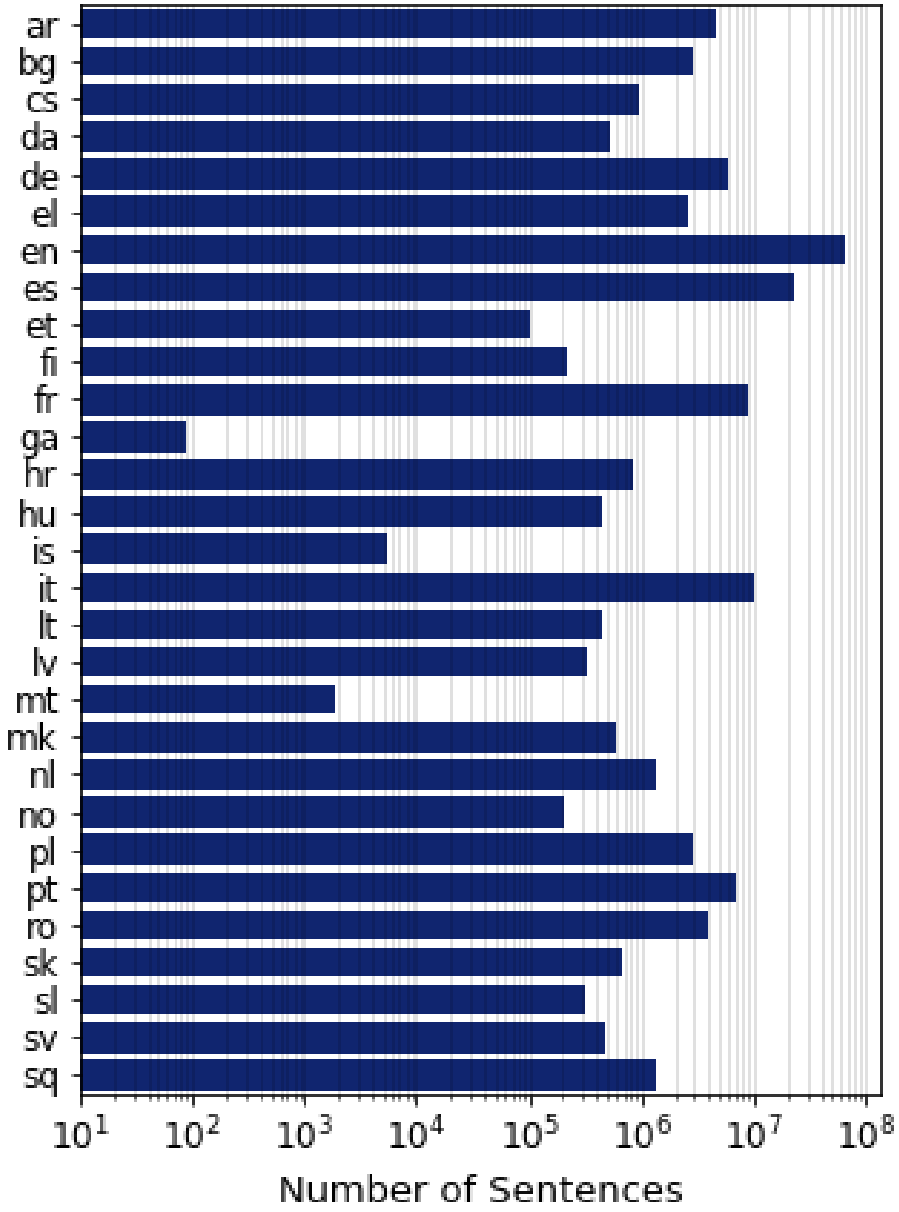
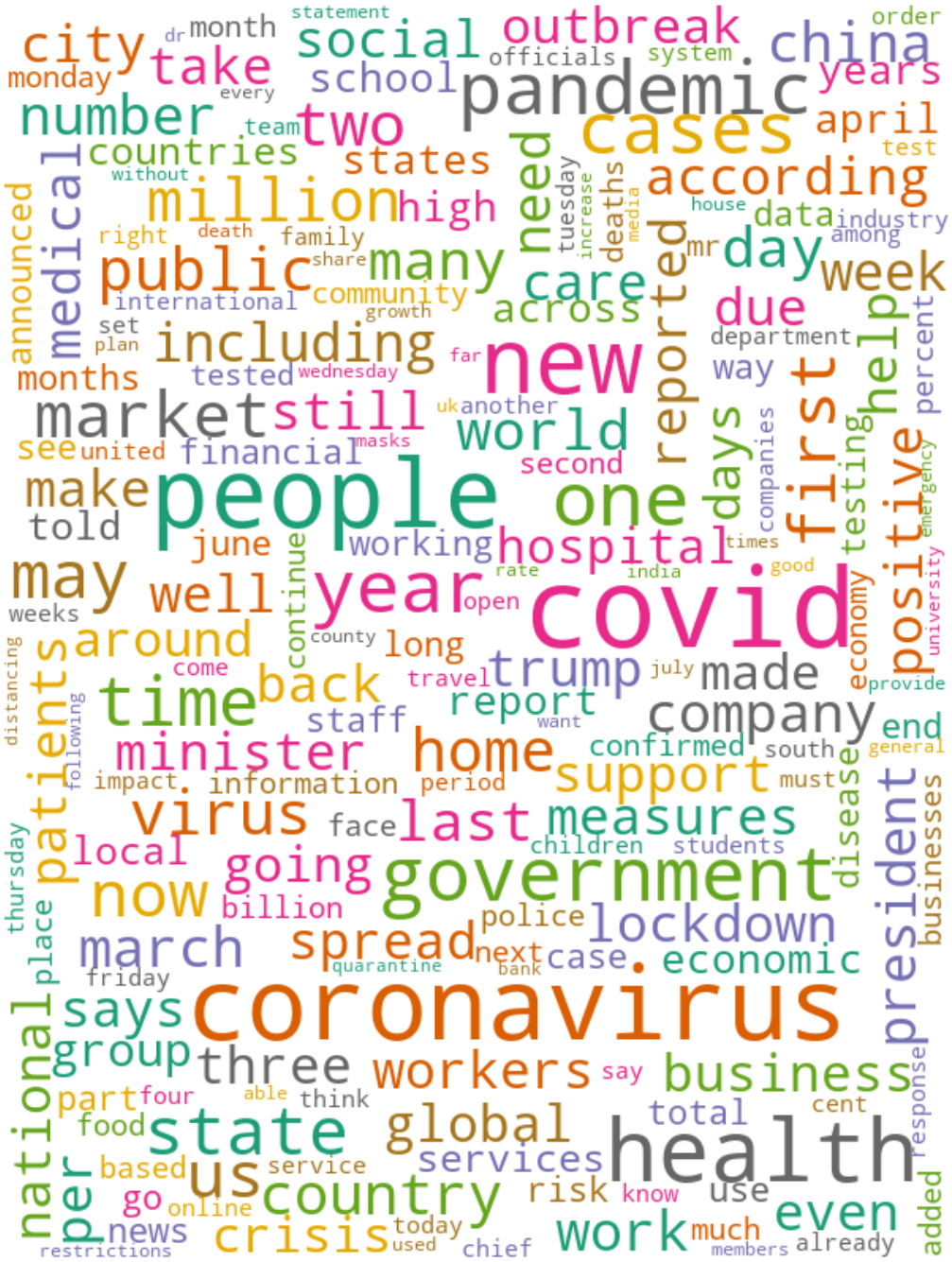



Figure 2: Total number of sentences per language

# Statistics of Monolingual Corpora

In **Figure 2**, we can see the number of total number of sentences for each language in the corpus, after deduplication and language identification (horizontal axis is in log-scale).




# Domain of the Corpus

Word cloud of the English monolingual corpus (~60M sentences) which is indicative of the **domain** of the parallel resource:

News articles related to COVID-19 and its effects on society, governance and the global economy.

# Conclusion

- We have presented a new COVID-19 related parallel resource with 11.2 million sentence pairs generated for 26 language pairs.
- It is based on the openly available metadata which have been created by the EMM / MediSys processing chain of news articles.
- We believe that the corpus can prove useful in training or adapting MT systems for COVID-19 related information and it has already been used in shared tasks.
- The methodology for constructing and filtering the corpus can be seen as an application of simulating a rapid response of the MT community to the COVID-19 crisis or future similar crises.
- We plan to further augment the resource with newly published content (e.g. batches of data for November and December 2021) and other language pairs.

Thank you for your time!

Please do not hesitate to share any questions, ideas and remarks!