SciPar: A Collection of Parallel Corpora from Scientific Abstracts

Dimitris Roussis,
Vassilis Papavassiliou,
Prokopis Prokopidis,
Stelios Piperidis,
& Vassilis Katsouros

LREC 2022, Marseille, France

ATHENA Research & Innovation Information Technologies

Outline of the Presentation

- Introduction
- The Importance of SciPar
- The Sources of SciPar
- Data Acquisition Methodology & Tools
- Statistics of Parallel Corpora
- Statistics of Monolingual Corpora
- Qualitative Characteristics
- Conclusion

Introduction

- Large-scale parallel corpora are vital for MT (Machine Translation) systems. However, most language pairs and domains are underrepresented.
- Scientific parallel corpora generally **focus on the biomedical domain**. CAPES, SciELO and ASPEC are some notable exceptions, although they concern high-resource languages (English, French, Spanish, Japanese, etc.).
- Institutional/National repositories contain titles and abstracts (translated in at least 2 languages) from bachelor/master theses & doctoral dissertations.
- We exploited these resources to construct high-quality bilingual corpora in the scientific research domain.

The Importance of SciPar

- Recent empirical evidence suggests that the research performance of a university declines *as the linguistic distance of its local language from English increases*. Improving the quality of translations related to scientific research could accelerate research in a more equitable manner.
- **SciPar** is constituted by a wide variety of scientific subjects and disciplines and addresses language pairs which are relatively under-resourced (e.g. EN-EL, EN-FI, EN-HR, EN-LV, EN-SL, EN-SV)
- As a multilingual corpus related to scientific research, it has many applications: Training NMT systems, domain adaptation of existing systems, cross-lingual plagiarism detection, transfer learning

The Sources of SciPar

We used the metadata of **86 institutional repositories**. Most have been built with DSpace or EPrints software.

We experimented with more repositories which did not yield any data due to *limited data accessibility.*

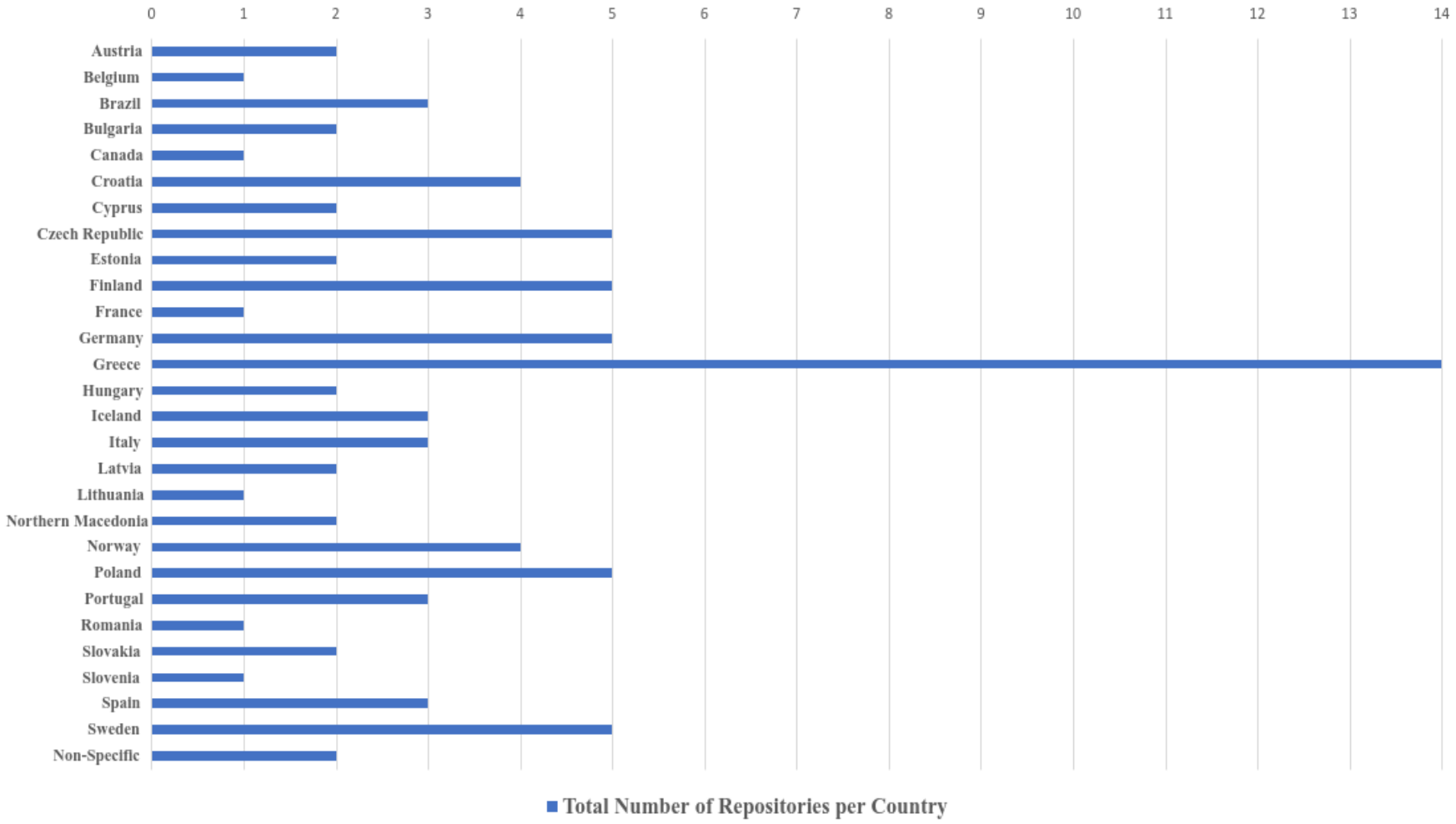

Figure 1: Total number of repositories per country of origin

Data Acquisition Methodology & Tools (1/2)

Our data acquisition methodology was modified for each repository.
An outline of the initial basic steps of our methodology and the tools we used:

- **Locating** multilingual repositories and **identifying** the pattern of their URL structure, i.e. the way in which their records are stored.
- **Fetching records** (metadata of theses, dissertations, etc.) as HTMLs using the GNU Wget package.
- **Parsing HTMLs** with the Beautiful Soup Python package and custom scripts to extract parallel titles/abstracts as document pairs.
- Using **language identification** software (fastText) to verify the language of each document.

Data Acquisition Methodology & Tools (2/2)

After acquiring records from the repositories, extracting the text from the abstracts, and creating document pairs, the next steps of our methodology were the following:

- **Splitting documents into sentences** using the NLTK library.
- **Mining parallel sentences** from document pairs using LASER.
- **Concatenating** all sentence pairs into a single file for each language pair.
- **Filtering** sentence pairs of limited use to MT from each bilingual corpus (e.g. duplicates, sentences containing only digits, etc.).

Statistics of Parallel Corpora

- The resource consists of **9,172,462 sentence pairs** in **31 language pairs** covering 25 languages.
- **Table 1:** Sentence pairs of the 24 bilingual EN-X corpora.
- **Table 2:** Sentence pairs of the 7 bilingual corpora in other language pairs.

EN-	BG	CS	DE	EL	ES	ET	FI	FR
	1,790	989,912	822,364	681,666	337,915	74,495	377,718	1,076,057
EN-	HR	HU	IS	IT	LT	LV	MK	NB
	802,961	15,007	95,782	29,588	158,622	335,316	4,065	51,528
EN-	PL	PT	RU	SK	SL	SQ	SV	NN
	798,094	943,157	2,818	59,690	289,454	7,653	621,892	2,172

Table 1: Parallel sentences for all EN-X language pairs

DE-ES	DE-FR	DE-RU	ES-FR	ES-RU	FR-RU	MK-SQ
261	278	185	4,787	679	1,235	3,661

Table 2: Parallel sentences for all other language pairs

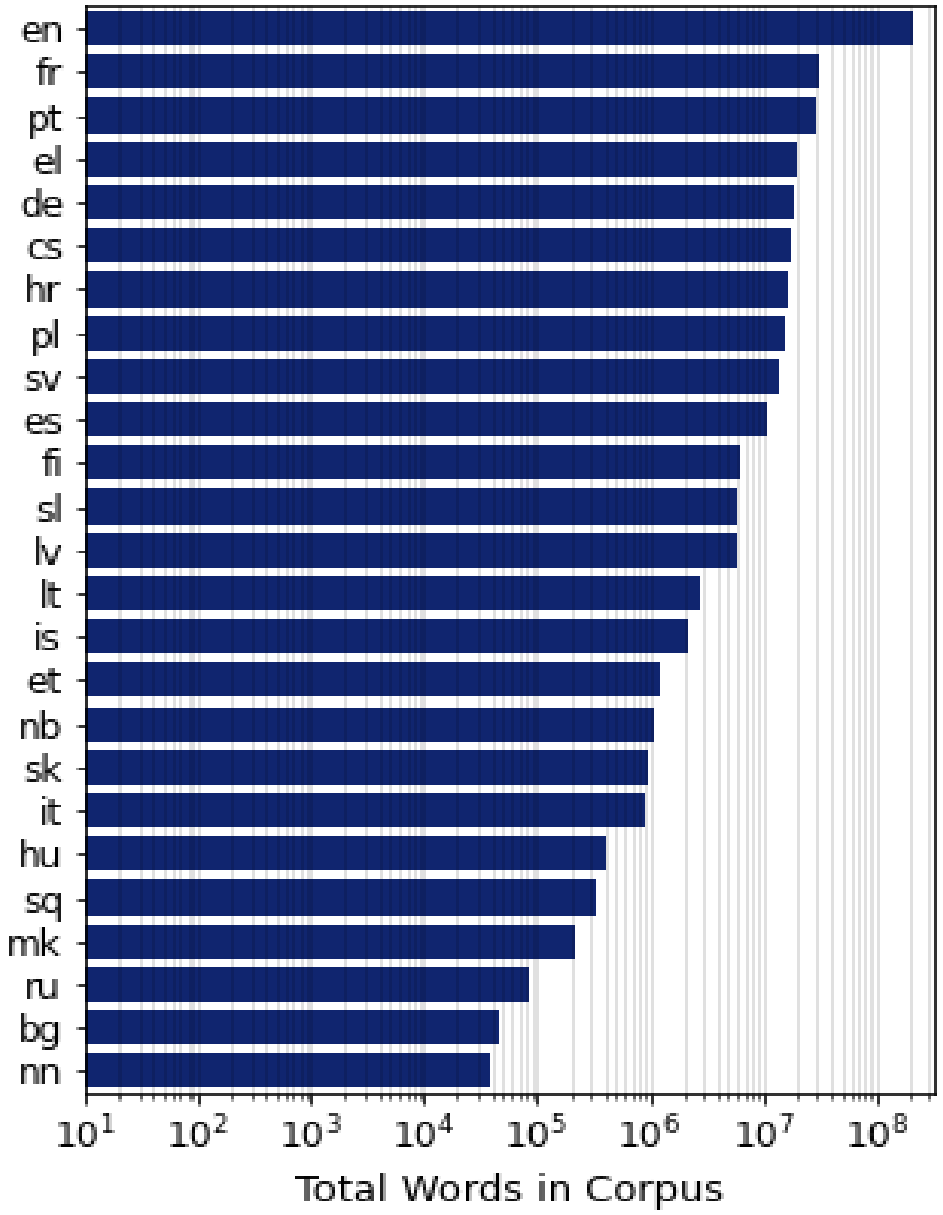

Figure 2: Total words in corpus by language

Statistics of Monolingual Corpora

In **Figure 2**, we can see the number of words for each language in the corpus (horizontal axis is in log-scale).

Monolingual sentences have many uses in NLP: back-translation, training language models or domain classifiers etc.

Qualitative Characteristics

- **Domain:**
  Scientific research
- **Scientific areas covered include:**
  Economics, social sciences, medicine, informatics, mathematics, engineering, history, gender studies, environmental studies, psychology, etc.
- **Concepts and terms have different meanings across different sub-disciplines (English-Slovak):**
  1st: beam -> lúča (laser beam; physics)
  2nd: beams -> nosníky (structural part)

Source Sentence (EN)	Target Sentence (SK)
Optoelectronic sensors are detectors based on the principle of scanning the **beam** in the range or they capture image.	Optoelektronické snímače sú detektory založené na princípe snímania svetelného **lúča** v príslušnom spektre či snímanie obrazu.
The main **beams** are made of welded structural upright profile.	Hlavné **nosníky** sú tvorené zo zváranej konštrukcie skriňového profilu.

Table 3: Examples of different word meanings in different scientific disciplines

Conclusion

- **SciPar** is a new parallel resource with **9.17 million sentence pairs** generated for **31 language pairs**. It is available via the **ELRC-SHARE** repository in TMX format.
- It is based on the openly available metadata on institutional repositories, digital libraries of universities and national archives.
- The corpus is related to the **broad domain of scientific research** as its data originate from titles and abstracts of various fields and sub-disciplines.
- We believe that the corpus can prove useful in training or adapting MT systems for scientific texts, especially regarding relatively under-resourced language pairs.
- We aim to harvest more repositories so as to further augment the resource with newly published content and other language pairs.

Thank you for your time!

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