# ParaCrawl: Web-scale parallel corpora for the languages of the EU

**M. Esplà-Gomis, M.L. Forcada**
Dept. Lleng. i Sist. Inform.
Universitat d'Alacant
E-03690 St. Vicent del Raspeig
Spain
{mlf,mespla}@dlsi.ua.es

**G. Ramírez-Sánchez**
Prompsit Language Engineering
Av. Universitat, s/n
E-03202 Elx
Spain
gema@prompsit.com

**H. Hoang**
School of Computing
University of Edinburgh
Edinburgh HE8 4AB
UK
Hieu.Hoang@ed.ac.uk

## Abstract

We describe two projects funded by the Connecting Europe Facility, *Provision of Web-Scale Parallel Corpora for Official European Languages* (2016-EU-IA-0114, completed) and *Broader Web-Scale Provision of Parallel Corpora for European Languages* (2017-EU-IA-0178, ongoing), which aim at harvesting parallel corpora from the Internet for languages used in the European Union. In addition to parallel corpora, the project releases successive versions of the free/open-source web crawling software used.

## 1 Introduction

Two projects are described in this abstract: *Provision of Web-Scale Parallel Corpora for Official European Languages* (Action 2016-EU-IA-0114, September 2017–March 2019, completed) or **Paracrawl**, and *Broader Web-Scale Provision of Parallel Corpora for European Languages* (Action 2017-EU-IA-0178, September 2018–September 2020, ongoing), or **Paracrawl-2**. Both are funded by the Connecting Europe Facility and have the same objective: to harvest parallel data from the Internet for languages used in the European Union. Namely, the first action focuses on parallel data between English and the other 23 official languages of the European Union, while the second one includes new pairs of languages, such as the pairs consisting of Spanish and the three regional languages recognized by Spain (Catalan, Basque, and Galician) or the two Norwegian languages (Bokml and Nynorsk). In addition to parallel corpora (see section 2, the project periodically releases versions of the free/open-source web-crawling software used, Bitextor (see section 3).

### 1.1 The consortium

Five partners are involved in these projects; two academic partners and three companies: The University of Edinburgh (coordinator), Edinburgh (UK); Universitat d'Alacant, Alacant (Spain); Prompsit Language Engineering S.L., Elx (Spain); TAUS B.V, Amsterdam (the Netherlands); Omniscien Technologies (trading) B.V., Zoetermeer (the Netherlands, only *Paracrawl2*).

## 2 Corpora built

Table 1 summarizes the most recent release, version 4, of parallel data between English and the remaining 23 languages of the European Union.

ParaCrawl corpora are publicly available under the Creative Commons CC0 license and can be found at the ParaCrawl website[1] and the ELRC-share repository. [2]

Random samples of 2,000 sentences for each language combination were validated by language experts for version 3 of the corpora allowing to tackle some of the most prominent issues before the release of version 4. Also, an extrinsic evaluation through MT was performed for some language pairs. It consistently confirms the positive impact of adding ParaCrawl corpora to baseline systems.[3]

## 3 Free/open-source crawling pipeline

One of the outputs of these projects is the free/open-source pipeline implemented to build the corpora in Table 1. The last version of this pipeline has been

[1] https://paracrawl.eu/releases.html

[2] https://elrc-share.eu/

[3] Adding Paracrawl 4 corpora to the WMT 2018 baseline improved the BLEU score in 11 out of 12 language pairs tested, https://paracrawl.eu/releases.html.

| language paired with English | number of segment pairs | number of English tokens |
|---|---:|---:|
| Bulgarian | 1,039,885 | 21,109,546 |
| Croatian | 1,002,053 | 19,904,218 |
| Czech | 2,981,949 | 48,918,151 |
| Danish | 2,414,895 | 48,240,290 |
| Dutch | 5,659,268 | 108,197,376 |
| Estonian | 853,422 | 16,537,397 |
| French | 31,374,161 | 664,924,148 |
| Finnish | 2,156,069 | 41,564,859 |
| German | 16,264,450 | 307,786,150 |
| Greek | 1,985,233 | 38,322,532 |
| Hungarian | 1,901,342 | 30,835,267 |
| Irish | 357,399 | 8,241,515 |
| Italian | 12,162,239 | 260,361,435 |
| Latvian | 553,060 | 10,996,032 |
| Lithuanian | 844,643 | 15,087,805 |
| Maltese | 195,510 | 4,100,912 |
| Polish | 3,503,276 | 65,618,419 |
| Portuguese | 8,141,940 | 156,125,200 |
| Romanian | 1,952,043 | 39,882,223 |
| Slovak | 1,591,831 | 26,711,854 |
| Slovenian | 660,161 | 14,489,659 |
| Spanish | 21,987,267 | 476,409,854 |
| Swedish | 3,476,729 | 70,088,534 |

**Table 1:** Statistics for the Paracrawl corpus, version 4

released as version 7 of the parallel-data-crawling tool Bitextor.[4] This pipeline covers all the stages from crawling data from websites on the Internet to delivering a clean parallel corpus. Namely, the stages included in this process are: (1) downloading HTML documents from the Internet; (2) preprocessing, normalizing and augmenting information from these documents; (3) aligning documents that are parallel; (4) aligning the segments in each of the document pairs identified; (5) filtering noisy data, deduplicating and formatting the output.

From the beginning of Paracrawl actions, several tools and modules have been contributed by the partners of the consortium and integrated in Bitextor. After a partial re-implementation of the pipeline control module (workflow manager), the tool is now highly configurable, allowing to run the pipeline using alternative components for the different stages of processing. Bitextor 7 was designed for high scalability in order to tackle the challenges of dealing with large amounts of data coming from thousands, or hundreds of thousands, of websites crawled from the Internet. It is built to work with distributed clusters such as SLURM and PBS Pro and cloud computing, with specific support for Azure.

It is also worth to mention Bicleaner,[5] a tool to filter parallel data that has been integrated in the Bitextor pipeline. This tool, which ranked among the best systems in the shared task on parallel corpora cleaning at WMT 2018 (Sánchez-Cartagena et al., 2018), is especially useful when dealing with noisy corpora such as those obtained through massive crawling from the Internet.

All the modules released as part of Bitextor 7 can be used within the pipeline or independently, to create new pipelines for specific purposes.

## 4 Future work

Paracrawl2 action is ongoing; some of the most relevant objectives for the next months of project are: adding Icelandic, Norwegian (Bokmål and Nynorsk), Basque, Catalan/Valencian, and Galician to the languages already covered languages; covering formats different to HTML (PDF, DOCX, ODT, etc.);including domain identification to support the extraction of relevant data from ParaCrawl corpora; processing the Internet Archive[6] to gather new parallel data; improving Bitextor by improving document and segment alignment and corpus cleaning; improving data exploitation by repairing rather than simply discarding and by segmenting too long sentences into clauses; delivering randomised, anonymised, partially omitted and mixed data sets to extend their usage; formalising human and automatic evaluation for quality testing.

In addition to the work covered in the ongoing action, four of the partners of the Paracrawl2 consortium have recently been awarded a new Connecting Europe Facility action, *Continued Web-Scale Provision of Parallel Corpora for European Languages* (Action 2018-EU-IA-0063, *Paracrawl-3*), which will extend the results of the previous projects.

## References

Sánchez-Cartagena, Víctor M., Marta Bañón, Sergio Ortiz-Rojas, and Gema Ramírez-Sánchez. 2018. Prompsit's submission to the WMT 2018 parallel corpus filtering shared task. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 968–975, Brussels, Belgium, October. Association for Computational Linguistics.

[4] `https://github.com/bitextor/bitextor`

[5] `https://github.com/bitextor/bicleaner`

[6] `https://archive.org/`