

ParaCrawl: Provision of Web-Scale Parallel Corpora for Official European Languages

paracrawl.eu

Kenneth Heafield, University of Edinburgh

neural.mt



Funding



Co-financed by the European Union
Connecting Europe Facility



Large, Broad Corpora from the Web

	ParaCrawl	ILSP
Translated words	4 million–1 billion	Manually evaluate 3%
Coverage	Broad	Focused
Web Domains	510,482	1/crawl

Data in Release 1

Language	Websites	Filtered Sentences	Filtered Words (en)
German	49,656	36,351,593	476,398,001
French	35,512	27,622,881	546,401,428
Spanish	27,194	16,001,341	325,745,201
Italian	21,940	8,318,493	155,973,063
Portuguese	14,786	2,809,381	57,392,721
Dutch	10,212	2,560,472	45,149,412
Polish	10,212	1,275,162	22,092,316
Czech	8,429	10,020,250	78,743,955
Romanian	7,104	2,459,752	32,800,110
Finnish	5,990	624,058	9,485,039
Estonian	1,784	1,298,103	13,134,231
Latvian	1,725	242,227	4,250,040

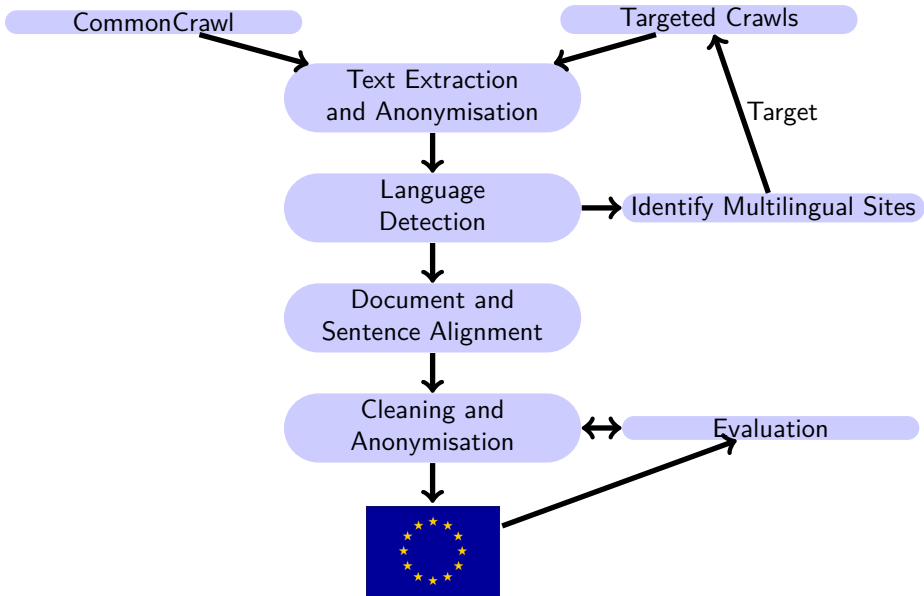
June 2018: 18 EU languages

March 2019: 24 EU languages

Quality Impact

Pair	Baseline	+ParaCrawl	Gain
Czech→English	25.7	26.3	+0.6
Finnish→English	21.7	24.2	+2.5
German→English	29.7	31.4	+1.7
Latvian→English	15.6	16.4	+0.8
Romanian→English	29.2	32.4	+3.1

BLEU scores on 2017 Conference on Machine Translation test



Improved Filtering

Release 1.1, 14 April 2018

- Supervised classifier trained on 50k good, 50k bad sentences
- Test set ensures consistent cutoff across languages
- Pattern-based filtering

Sent to ELRC-SHARE, posting in process.

Get it now: tinyurl.com/ybtda3dk

Corpus filtering shared task with ParaCrawl data:


statmt.org/wmt18/parallel-corpus-filtering.html

Legal Issues


Personal Data

Copyright

ELRA sells webcrawl parallel corpora:

[Browse Resources](#) [Information](#) Cart total [View cart](#) [Register](#) [Login](#)

Linguatools Webcrawl Parallel Corpus German-English 2015

12  1





▸ View resource name in all available languages





ISLRN: 800-190-274-236-9

ID: [ELRA-W0091](#)

The corpus consists of 10 million German-English parallel sentences that were crawled from the internet between 10/2013 and 04/2015. The sentences were gathered from over 112,000 different hosts. An elaborate multi-step quality filtering was applied, including language identification filter, machine translation filter, grammaticality filter, etc. to get as clean data as possible. There are no duplicate sentence pairs, and there is no overlap with existing publicly available corpora like europarl, DGT-TM, etc. Web pages have been automatically categorized for subject area. The corpus is available in TMX and Moses format (encoding UTF-8). [Read Less](#)

▸ View resource description in French

MEMBER	academic	commercial
Licence: Non Commercial Use - ELRA END USER	1000.00 € 	4800.00 € 
Licence: Commercial Use - ELRA VAR	4800.00 € 	4800.00 € 

NON MEMBER	academic	commercial
Licence: Non Commercial Use - ELRA END USER	1200.00 € 	5000.00 € 
Licence: Commercial Use - ELRA VAR	5000.00 € 	5000.00 € 

`tinyurl.com/y7tttpuya` (ELRA URLs are long)

Did ELRA anonymise or ask permission?

“Or check the video done by Steve Huff:”

“Oder auch der Beitrag von Steve Huff:”

Did ELRA anonymise or ask permission?

“Or check the video done by Steve Huff:”

“Oder auch der Beitrag von Steve Huff:”

“All information published on this website is copyright protected and may not be used without written permission from Schaper & Brümmer.”

“Alle auf dieser Internetseite veröffentlichten Informationen sind urheberrechtlich geschützt und dürfen ohne schriftliche Freigabe des Unternehmens Schaper & Brümmer nicht genutzt werden.”

National Library of the Netherlands

The Netherlands has no legal deposit law.

National library archives the web anyway.

Report on legal issues: tinyurl.com/ycn6daap

Conclusion: “adopted a pragmatic way to handle the copyright issues: the opt-out approach. This approach assumes implicit permission for web archiving.”

ELRC Report's Body has Viable Options

Every EU case cited by the report was won by the crawler.

“not impossible to organize the crawling process in such a way as to comply”

“the creation of statistical language models would also qualify as lawful use”

ELRC Report's Body has Viable Options

Every EU case cited by the report was won by the crawler.

“not impossible to organize the crawling process in such a way as to comply”

“the creation of statistical language models would also qualify as lawful use”

ELRC Summary Ignores Options

Jumps from: “It seems that the most viable way”

To: “Only the sources that pass this validation procedure”

ELRC Report's Body has Viable Options

Every EU case cited by the report was won by the crawler.

“not impossible to organize the crawling process in such a way as to comply”

“the creation of statistical language models would also qualify as lawful use”

ELRC Summary Ignores Options

Jumps from: “It seems that the most viable way”

To: “Only the sources that pass this validation procedure”

Please do not read just the summaries.

ELRC-endorsed Temporary Exemption

Legal Basis

InfoSoc directive: mandatory exemption for temporary acts of reproduction
“it is not impossible to organize the crawling process in such a way as to comply with the conditions of the exception” –ELRC report

Option

We provide URLs, sentence positions, and checksums (Forcada et al 2006)
eTranslation runs our script to download pages again.
Script creates model, automatically deletes web pages.
Ok to keep model “lawful use”

ELRC-endorsed Temporary Exemption

Legal Basis

InfoSoc directive: mandatory exemption for temporary acts of reproduction
“it is not impossible to organize the crawling process in such a way as to comply with the conditions of the exception” –ELRC report

Option

We provide URLs, sentence positions, and checksums (Forcada et al 2006)
eTranslation runs our script to download pages again.
Script creates model, automatically deletes web pages.
Ok to keep model “lawful use”

This doesn't appear in the ELRC report summaries!

Saner Approach: ROAM

Randomise Shuffle the sentences.

Omit Remove data. In Germany, “up to 15% of a work can be reproduced and communicated to the public” –ELRC

Anonymise Replace phone numbers, e-mail addresses, etc. with a constant.

Mix Jumble sentences from different sources.

Case law from ELRC report: Thumbnailing

Google Image search provides thumbnails of images from the web.

- ① Uploading on the open internet (without any measures to prevent indexing) yields an **implied consent** (Vorschaubilder I)

Case law from ELRC report: Thumbnailing

Google Image search provides thumbnails of images from the web.

- 1 Uploading on the open internet (without any measures to prevent indexing) yields an **implied consent** (Vorschaubilder I)
- 2 Even if uploaded without the rightholder's permission (Vorschaubilder II)

Case law from ELRC report: Thumbnailing

Google Image search provides thumbnails of images from the web.

- 1 Uploading on the open internet (without any measures to prevent indexing) yields an **implied consent** (Vorschaubilder I)
- 2 Even if uploaded without the rightholder's permission (Vorschaubilder II)
- 3 Don't link if you "knew or ought to have known" it infringes. Commercial sites ought to know. (CJEU in GS Media)

Case law from ELRC report: Thumbnailing

Google Image search provides thumbnails of images from the web.

- 1 Uploading on the open internet (without any measures to prevent indexing) yields an **implied consent** (Vorschaubilder I)
- 2 Even if uploaded without the rightholder's permission (Vorschaubilder II)
- 3 Don't link if you "knew or ought to have known" it infringes. Commercial sites ought to know. (CJEU in GS Media)
- 4 Google can't be expected to know, despite being commercial (Vorschaubilder III)

Case law from ELRC report: Thumbnailing

Google Image search provides thumbnails of images from the web.

- 1 Uploading on the open internet (without any measures to prevent indexing) yields an **implied consent** (Vorschaubilder I)
- 2 Even if uploaded without the rightholder's permission (Vorschaubilder II)
- 3 Don't link if you "knew or ought to have known" it infringes. Commercial sites ought to know. (CJEU in GS Media)
- 4 Google can't be expected to know, despite being commercial (Vorschaubilder III)

GS Media and Vorschaubilder III

GS Media Commercial \implies ought to know about infringement.
Vorschaubilder III Google cannot be expected to know.

ELRC: Court based reasoning on GS Media \implies abandoned implied consent.

Alternative: case based on GS Media. Doesn't mean implied consent is dead.

Conclusion

ParaCrawl provides:

- Broad coverage
- Very large corpora
- Demonstrated quality gains.

Copyright:

Temporary copy exemption is viable, ROAM is saner.
ELRC should separate legal opinions from summaries.