Targum

A Multilingual New Testament Translation Corpus

Maciej Rapacz · **Aleksander Smywiński-Pohl**
AGH University of Kraków
{mrapacz, apohllo} @ agh.edu.pl

LREC 2026 · Palma de Mallorca

Why another Bible corpus?

Polish Bible Translations Online

A popular Polish Bible aggregator:
bibliepolskie.pl

- **45 translations** with full NT text indexed
- ~20 readable online
- How many are available in existing corpora?

Stary i Nowy Testament [24]

Biblia Leopolity	1561	Biblia warszawska [brytyjka]	1975	Śląskie Towarzystwo Biblijne	2012
Biblia Brzeska	1563	Biblia Poznańska	1975	Stare i Nowe Przymierze [EIB] - liter.	2016
Biblia Szymona Budnego	1572	Biblia Lubelska	1991	Biblia pierwszego Kościoła	2016
Biblia Wujka	1599	Interlinia - Vocatio	1993	Przekład toruński	2017
Biblia Gdańska	1632	Biblia warszawsko-praska	1997	Biblia Impulsy	2017
Przekład Mariawitów	1925	Przekład Nowego Świata	1997	Biblia Ekumeniczna	2018
Komentarze KUL	1959	Nowy Komentarz Biblijny	2005	Stare i Nowe Przymierze [EIB] - dosł.	2018
Biblia Tysiąclecia	1965	Biblia Paulistów	2008	Nowy Przekład Dynamiczny [NPD]	2021
Biblia Królowej Zofii	1455	Przekłady Izaaka Cylkowa	1883	Biblia Hebrajska	2021

Tylko Nowy Testament [21]

NT Królewiecki	1553	NT z Wulgaty - E.Dąbrowski	1947	Ekumeniczny Przekład Przyjaciół	2012
NT Krakowski	1556	NT - Seweryn Kowalski	1957	Przekład Odzyskiwania	2018
NT Marcina Czechowica	1577	NT z greckiego - E.Dąbrowski	1961	Przekład Filologiczny	2019
NT Rakowski	1606	Słowo Życia - Parafraza NT	1989	NT - Andrzej Mazurkiewicz	2019
NT Rzewuskiego	1868	Współczesny przekład	1991	Oblubienica.eu - Przekład dosłowny	2020
NT TBS - Karol Węgierski	1876	Komentarz żydowski NT	2004	Przekład prawosławny	2022
Biblia Gdańska - Rewizja warsz.	1881	Oblubienica.eu - Przekład interlinearny	2011	Biblia Króla Jakuba	2024

Source: https://bibliepolskie.pl/przeklady.php

New Testament Translations in Existing Corpora

Corpus	Polish
Christodouloupoulos & Steedman (2015)	1
eBible (Akerman, 2023)	2
Mayer & Cysouw (2014)	6

New Testament Translations in Existing Corpora

Corpus	French	Polish
Christodouloupoulos & Steedman (2015)	1	1
eBible (Akerman, 2023)	4	2
Mayer & Cysouw (2014)	17	6

New Testament Translations in Existing Corpora

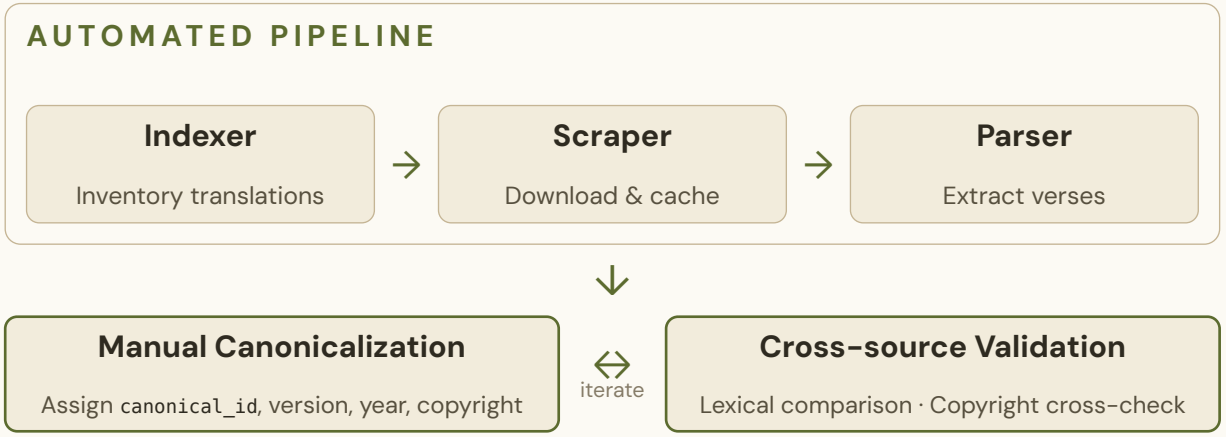
Corpus	English	French	Italian	Polish	Spanish	Total
Christodouloupoulos & Steedman (2015)	2	1	1	1	1	6
eBible (Akerman, 2023)	24	4	2	2	5	37
Mayer & Cysouw (2014)	39	17	7	6	22	91

New Testament Translations in Existing Corpora

Corpus	English	French	Italian	Polish	Spanish	Total
Christodouloupoulos & Steedman (2015)	2	1	1	1	1	6
eBible (Akerman, 2023)	24	4	2	2	5	37
Mayer & Cysouw (2014)	39	17	7	6	22	91
Targum	**390**	**78**	**33**	**48**	**102**	**651**
Targum (no duplicates)	**194**	**41**	**17**	**29**	**53**	**334**
vs. prior biggest	**5.0×**	**2.4×**	**2.4×**	**4.8×**	**2.4×**	**3.7×**

Corpus Construction Pipeline

- First **scrape & cache**, then **parse** — fetch each site once
- Manually label each text: **which translation**, **which edition**
- Validation loop: low similarity between "same" texts → fix parser or relabel



Manual Canonicalization — Example

`canonical_id`
Translation family, e.g. King James Version, Biblia Tysiąclecia

`canonical_version`
Specific edition, e.g. 1613, british-1769, catholic-edition

`canonical_year`
Integer year of the edition; placement on the time axis

`copyright_status`
License tier: public-domain · open-license · all-rights-reserved

```
canonical_id: king-james-version
canonical_version: british-1769
canonical_year: 1769
copyright_status: public-domain
```

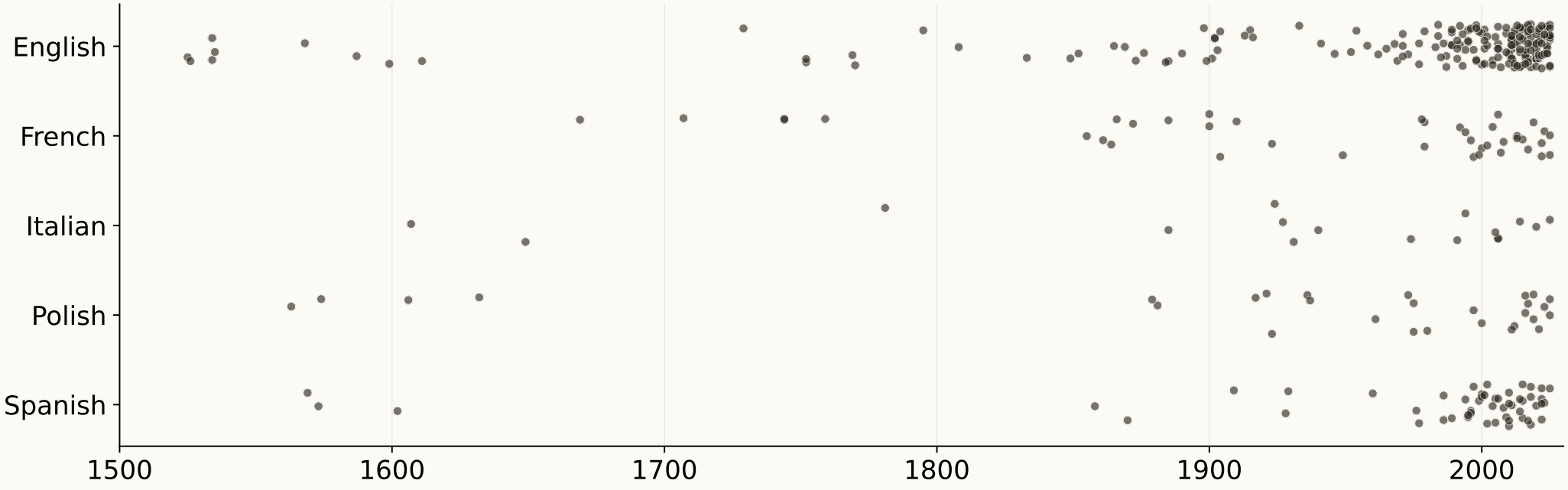
Four fields assigned manually to all 651 translations.

Translations by Source

Source	English	French	Italian	Polish	Spanish	Total
bible.audio	3	15	–	–	1	**19**
bible.com	75	19	7	8	28	**137**
bible.is	18	8	1	2	15	**44**
biblegateway.com	64	4	5	3	19	**95**
biblehub.com	46	–	–	–	–	**46**
bibles.org	72	6	3	3	14	**98**
biblestudytools.com	37	–	–	–	–	**37**
bibliepolskie.pl	–	–	–	20	–	**20**
crossbible.com	12	1	–	–	1	**14**
ebible.org	24	4	2	2	5	**37**
jw.org	1	1	1	1	1	**5**
laparola.net	9	3	11	–	2	**25**
obohu.cz	29	17	3	9	16	**74**
Total	**390**	**78**	**33**	**48**	**102**	**651**

No single site is enough — coverage needs aggregators + national sites.

Diachronic Distribution of Translations



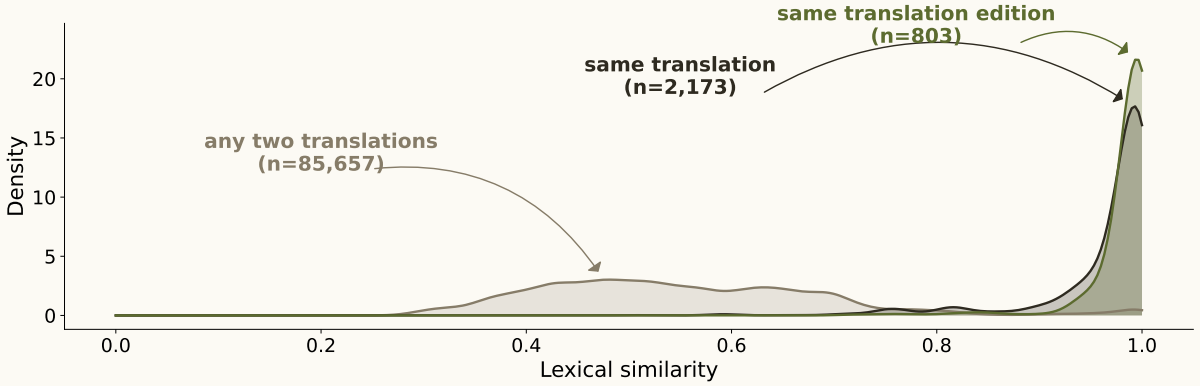
Each dot = one unique translation (canonical_id × version), positioned by edition year.

- Spans the **16th century** to today
- Most translations from **recent decades** — and still growing

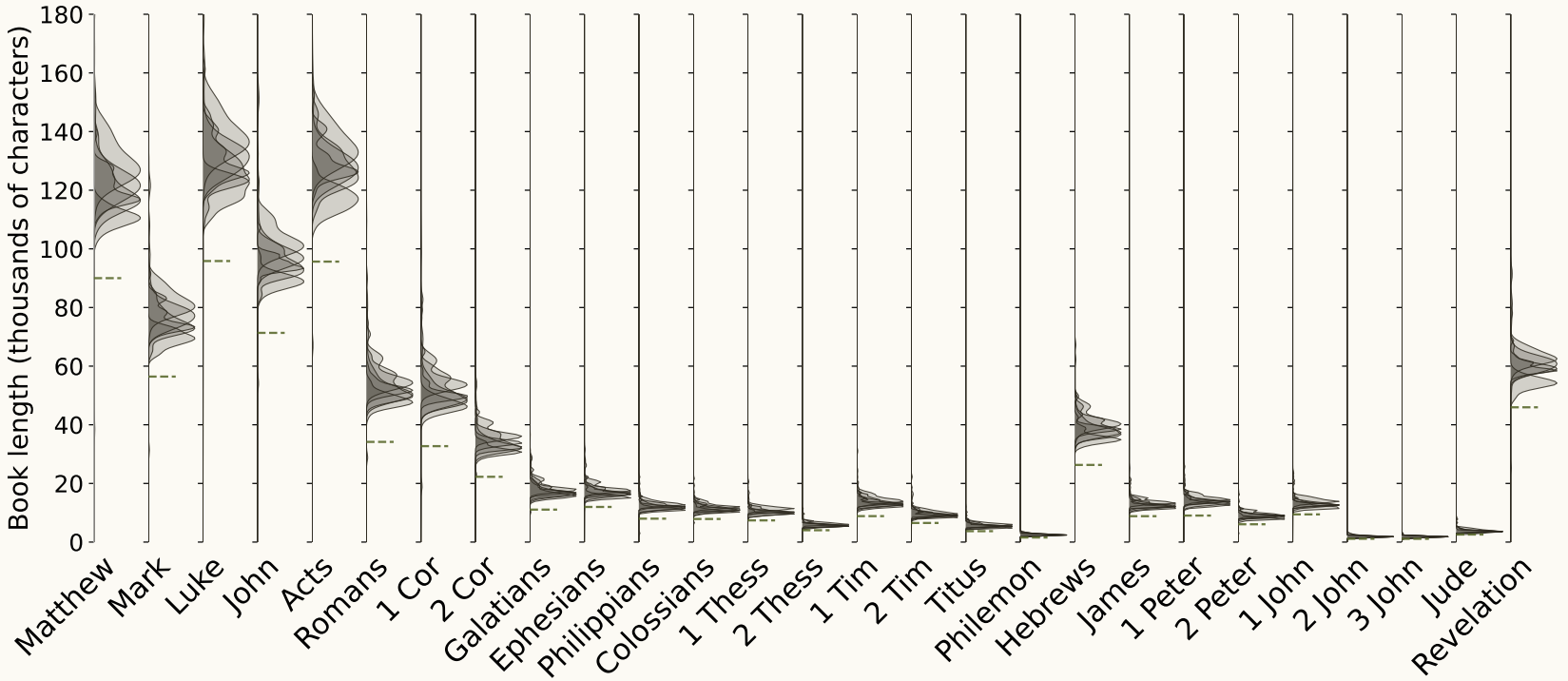
Pairwise Variation: Validating Canonicalization

Three tiers of translation relatedness, pooled across all five languages:

- **Same edition**, different sources → near-identical (~1.0): canonicalization works
- **Same translation**, different editions → wide spread; some editions are genuinely different texts
- **Any two** → the full intra-lingual diversity



Book Length Distribution

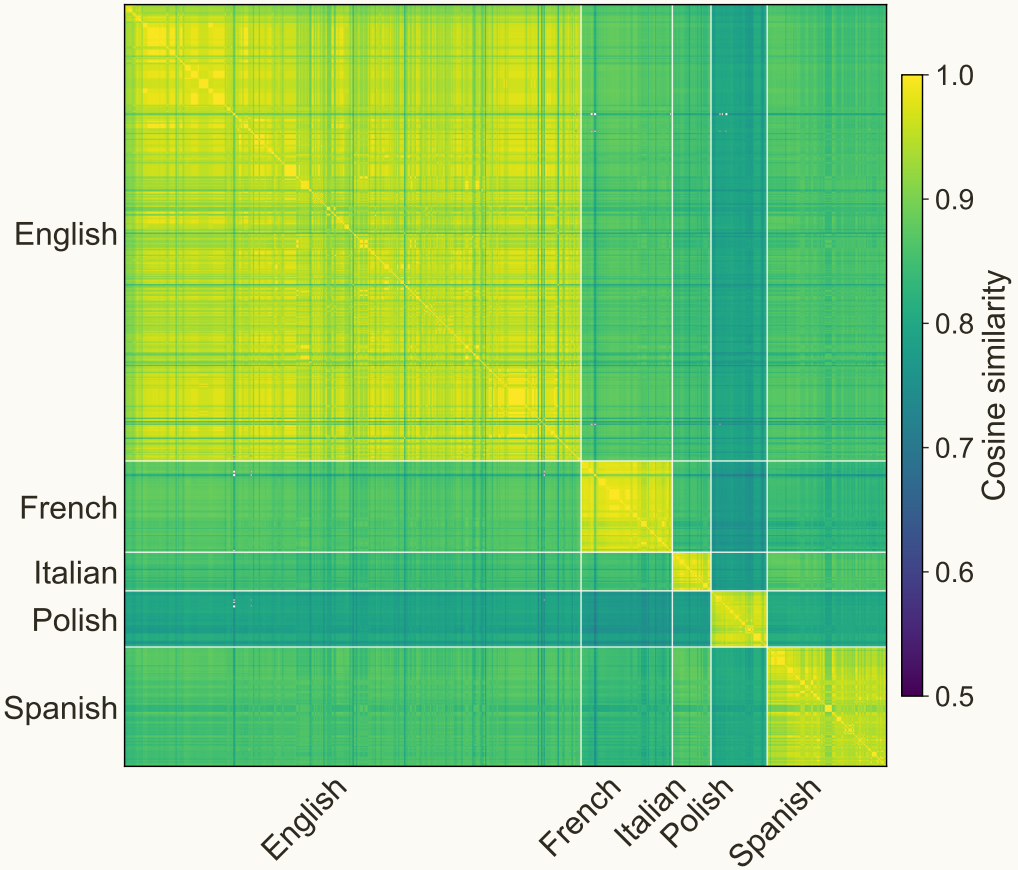


Distribution of book lengths across unique translations. Dashed line = Greek source (SBLGNT).

- Every language sits **above the Greek source**
- **French** consistently longest, **Polish** shortest

Cross-Language Semantic Similarity

- Translations cluster by language first
- Romance languages cluster together
- Polish most distant; English between



Pairwise similarity for all translations, sorted by language and year.

Next Steps

- **Richer metadata** — source text (critical edition; Latin or Greek?), confessional tradition (Catholic, Protestant, ecumenical)
- **More translations**
- **More languages?**

Resources

Openly released

- Public-domain & open-license **texts** — GitHub, HuggingFace
- **Embeddings** & **pairwise similarity** for all translations — including copyrighted ones

On request

- Full **copyrighted texts** — for non-commercial research

Thank you

mrapacz.com/lrec2026

This research was supported by the National Science Centre, Poland, under project number 2025/57/N/HS2/04961. We gratefully acknowledge the Polish high-performance computing infrastructure PLGrid (HPC Center: ACK Cyfronet AGH) for providing computational facilities and support under grant no. PLG/2026/019145. This work was supported by the "Excellence Initiative – Research University" program and the Faculty of Computer Science at AGH University of Kraków.