Monolingual and Parallel Specialized Corpora from Wikipedia

Gemma Segués[1], Gonzalo López-Sánchez[1] and Antoni Oliver[1]

[1]Universitat Oberta de Catalunya, Rambla del Poblenou, 154-156, 08018 Barcelona, Spain

Abstract

This paper introduces a novel and user-friendly toolkit for the compilation of monolingual and parallel corpora from Wikipedia. The suite of programs leverages Wikipedia's category system to enable the automated generation of specialized, domain-specific corpora for any discipline. We demonstrate the practical application of this toolkit by creating a substantial corpus for training Neural Machine Translation (NMT) systems for the Catalan-English language pair. The results validate the quality of the generated data for high-performance model training. This methodology is particularly valuable for low-resource languages, addressing the persistent challenge of sourcing specialized parallel texts. By providing an accessible method for creating tailored datasets, our work significantly lowers the barrier to developing robust NLP tools for under-represented languages.

Keywords

monolingual corpora, parallel corpora, specialized corpora, Wikipedia

1. Introduction

This paper outlines a methodology and an accompanying software tool for compiling domain-specific monolingual and parallel corpora from Wikipedia. The primary objective is to process raw Wikipedia data dumps to construct comparable corpora in English and Catalan across a diverse range of subjects. A key feature of our approach is the generation of these resources at three distinct levels of granularity: document, paragraph, and sentence. Subsequently, the sentence-level comparable corpora serve as input for a bitext mining process. Following the work of Artetxe and Schwenk [1], we employ Translated Sentence Mining to identify and extract parallel sentence pairs, thereby converting the initial monolingual sources into a parallel corpus.

We plan to create Catalan-English parallel corpora for the subjects in the first level of the UNESCO nomenclature for fields of science and technology[1]: 11 Logic, 12 Mathematics, 21 Astronomy and astrophysics, 22 Physics, 23 Chemistry, 24 Life Sciences, 25 Earth and Space Sciences, 31 Agricultural Sciences, 32 Medical Sciences, 33 Technological Sciences, 51 Anthropology, 52 Demographics, 53 Economic Sciences, 54 Geography, 55 History, 56 Juridical Sciences and Law, 57 Linguistics, 58 Pedagogy, 59 Political Science, 61 Psychology, 62 Science of Arts and Letters, 63 Sociology, 71 Ethics and 72 Philosophy.

The research presented in this paper is part of a broader initiative, the project Eines d'intel·ligència artificial per al foment de la comunicació científica en català (Artificial Intelligence Tools for the Promotion of Scientific Communication in Catalan), funded by the Institut d'Estudis Catalans. This initiative seeks to address the challenges faced by scientists who wish to disseminate their research in Catalan. To this end, the project aims to develop and evaluate a suite of AI-driven translation tools (including text-to-text, speech-to-text, and speech-to-speech) between Catalan and English. A critical prerequisite for building such specialized systems is the availability of large, domain-specific parallel datasets. Therefore, the methodology and corpora described represent a core contribution to the project's foundational goals.

SEPLN 2026: 42nd International Conference of the Spanish Society for Natural Language Processing, León, Spain, 22-25 September 2026.

gseguesm@uoc.edu (G. Segués); glopezsanch@uoc.edu (G. López-Sánchez); aoliverg@uoc.edu (A. Oliver)

0009-0004-7287-6615 (G. Segués); 0009-0006-1486-838X (G. López-Sánchez); 0000-0001-8399-3770 (A. Oliver)

[1]https://skos.um.es/unesco6/00/html

2. Wikipedia for corpora creation

Wikipedia has long been recognized as a fundamental resource for building large-scale, multilingual corpora due to its vast size, broad domain coverage, and permissive licensing. The state of the art in this area can be framed around three key themes directly relevant to our objectives: the creation of large-scale annotated corpora, the mining of parallel sentences from comparable sources, and the critical challenges of data preprocessing and quality assurance.

Particularly relevant to our work is Wikicorpus [2], a trilingual corpus of over 750 million words that includes both Catalan and English. A key contribution of this work was the automatic enrichment of the corpus with advanced linguistic annotations, such as part-of-speech (PoS) tags and word sense disambiguation, thereby providing a high-quality resource for lexical semantics research. This and other projects like WikiWoods [3] established the feasibility of creating valuable, structured corpora from raw Wikipedia dumps.

More recent research has focused on extracting parallel sentences from comparable corpora, a technique central to our methodology. While initial approaches relied on metadata or document-level alignment, the state of the art has shifted decisively towards methods based on multilingual sentence embeddings. This line of work, pioneered by Schwenk [4] and Artetxe and Schwenk [5], demonstrated that sentences from different languages could be mapped into a single vector space, allowing for efficient similarity searches.

Our approach specifically follows the Translated Sentence Mining (TSM) strategy proposed by Artetxe and Schwenk [1], which uses this embedding space to identify parallel sentences with high precision. The power of this technique when applied to Wikipedia at a massive scale was demonstrated by WikiMatrix [6], which employed multilingual sentence representations to mine parallel data across all possible language pairs. This method yielded an unprecedented 135 million parallel sentences for 16,720 language pairs, significantly expanding resources beyond English-centric alignments. The quality of this extracted bitext was validated by training Neural Machine Translation (NMT) systems that achieved strong BLEU scores, demonstrating its high value, particularly for distant and low-resource language pairs. Our work applies this proven methodology to generate domain-specific parallel data for Catalan-English, a contribution of significant value for specialized applications.

A persistent challenge in leveraging Wikipedia, and a primary motivation for our software tool, is the preprocessing of its raw content. The source material's format, WikiText, is a complex markup that mixes its own syntax with HTML and advanced templates. Improperly handling this structure introduces significant noise, which can be "utterly detrimental to any downstream use of our corpus" [7]. Indeed, research confirms that corpus quality and accurate sentence segmentation —which must be aware of markup— are often more critical for model performance than raw data size [8].

Consequently, achieving our target granularity of document, paragraph, and sentence requires a sophisticated understanding of WikiText's structural elements. While prior work has addressed noise filtering and segmentation, our methodology uniquely integrates these steps. We do so to create the aligned, multi-granular, and domain-specific comparable corpora that are essential for building the specialized translation tools of the project Eines d'intel·ligència artificial per al foment de la comunicació científica en català.

3. Methodology and algorithms

3.1. UNESCO nomenclature and Wikipedia categories

Wikipedia articles are organized using a system of user-assigned categories. While the platform allows for the unrestricted creation of new categories, users predominantly adhere to an established set. Wikipedia provides its own high-level classification structured around academic disciplines.[2] Additionally, DBpedia offers a more granular taxonomy of these articles,[3] enabling the recursive retrieval of subcategories from a given parent category down to a specified depth.

3.2. The CCWikipedia database

To speed up the process of finding suitable Wikipedia pages we use a SQLite database created from several files of the English Wikipedia dumps[4]: the `pages-articles.xml.bz2` and the `langlinks.sql.gz`. We also need the `skos_categories_en.ttl.bz2` file.[5] The schema is entity-centric, with a central numeric identifier (`ident`) used across multiple tables to link an entity to its associated metadata. This design efficiently organizes information related to titles, categorization, inter-language links, qualitative attributes, and quantitative metrics, facilitating complex queries and data retrieval.

- `titles` This table serves as the primary lookup for entity names. It maps a unique numeric identifier (`ident`) to its corresponding canonical or default title (`title`), thereby establishing the core identity of each record in the database.
- `categories` This table manages the classification of entities. Each entry links an entity via its `ident` to a specific `category` name. This structure facilitates a many-to-many relationship, allowing a single entity to be assigned to multiple distinct categories.
- `categoryrelations` Designed to build a hierarchical category structure, this table defines parent-child or other relationships between categories. Each row links a `category` to a related one (`categoryREL`), which is essential for navigating the category tree and understanding its topology.
- `langlinks` This table provides crucial internationalization support by linking an entity's `ident` to its titles in various languages. It stores the foreign `title` alongside a corresponding language code (`lang`).
- `qualities` This table is employed to assign specific attributes or metadata tags—referred to as a `quality`—to entities. It allows for the classification of records based on predefined characteristics, such as an article's assessment status (e.g., "Featured Article" or "Stub").
- `sizes` A straightforward table for storing quantitative data about each entity. It maps an `ident` to a numerical `size`, which could represent various metrics such as the article's byte length, word count, or another relevant measure of scale.

While users have the option to construct the database themselves via a Python script or a dedicated graphical user interface (GUI) bundled with the tool, this procedure can be lengthy. To facilitate access, we periodically publish an updated version of the database, which can be downloaded and utilized directly, bypassing the need for manual generation.

While using this database expedites corpora generation, its origin from the English Wikipedia introduces an English-centric bias. Specifically, searches are performed using English categories, and page titles in other languages are derived from the English Wikipedia's langlinks. This approach can inadvertently omit non-English pages that lack an English equivalent, though this is generally not a significant issue due to the comprehensive nature of the English Wikipedia. To mitigate this bias, users have the option to generate a new database from any other language edition of Wikipedia and subsequently perform all steps using it.

3.3. Creation of the corpora

The process of creating specialized monolingual and parallel corpora from Wikipedia is divided into the following steps, all of which can be performed with a single Python script or the provided GUI program. If we want to create a monolingual corpora we will perform the following steps:

1. **Category Selection and Scoping:** The initial step involves selecting the categories to explore and defining the desired taxonomic depth for a specific language. The tool allows for experimentation with various configurations, providing real-time figures on the total number of resulting categories, subcategories, and pages. Once these metrics are deemed satisfactory, the process can proceed, concluding with the generation of a file that lists the titles of all selected Wikipedia pages.
2. **Text Extraction:** Using the title file generated in the previous step, this stage involves extracting the full text of the selected pages from a complete Wikipedia dump file (specifically, the

[2] https://en.wikipedia.org/wiki/Outline_of_academic_disciplines
[3] https://downloads.dbpedia.org/3.9/en/skos_categories_en.ttl.bz2
[4] https://dumps.wikimedia.org/enwiki/
[5] Available at https://downloads.dbpedia.org/

`pages-articles.xml.bz2` archive) for each language. The process reads the entire dump and retrieves the content for each matching page title.

3. **Text Segmentation:** The extracted page content is then segmented into smaller linguistic units. This operation is performed on the text files using a set of rules defined in a Segmentation Rules eXchange (SRX) file. The output is a collection of segmented files ready for further processing.

4. **Paragraph Deduplication:** This step processes all the generated files to select and deduplicate their content at the paragraph level. The goal is to create a non-redundant corpus of paragraphs from the selected Wikipedia pages.

5. **Segment Deduplication:** Finally, a similar filtering and deduplication process is applied at the segment level. This ensures that the final corpus consists of unique sentences or segments.

The successful execution of this procedure results in a set of deliverables: (1) a directory of files, each containing the full text of a selected page; (2) another directory of files, each containing the corresponding page's content segmented on a per-line basis; (3) a comprehensive file with all deduplicated paragraphs; and (4) a final file comprising all unique, deduplicated segments.

To create a parallel corpus, the previously described steps must be performed twice—once for each language of interest. It is crucial to remember that equivalent Wikipedia pages in two different languages are independent texts and are not necessarily translations of each other, although some content may overlap. Consequently, creating a parallel corpus from this source is not a classical text alignment task. Instead, it is a process of bitext mining, which involves searching for parallel segments within comparable, non-parallel documents.

6. **Bitext Mining:** This step involves performing bitext mining on the deduplicated segment files for each language. The process is compatible with systems both with and without GPU acceleration; however, utilizing a GPU will dramatically speed up the computation. The primary output of this stage is a set of aligned, parallel segments.

7. **Language Verification and Rescoring:** The initial alignment from the previous step may contain errors. Common issues include segments not in the expected language (e.g., an English sentence incorrectly aligned with itself, as non-English Wikipedia articles often contain English titles of books or films) and other misalignments. To address these problems and refine the output, we apply the rescoring algorithm described in Oliver and Álvarez [9] to clean up the resulting alignment. This process provides you with a text file with all the segments with the following scores attached: (1) detected source language and confidence of the language detection; (2) detected target language and confidence of the language detection and (3) cosine similarity of the source and target segments using SBERT.

8. **Filtering and Selection of Final Segments:** Based on the output of the rescoring algorithm, we establish minimum threshold values for each score. Only segments that meet or exceed these thresholds are selected for the final corpus. In our experiments, we set a uniform minimum score of 0.75, a value determined from previous empirical work.

A significant challenge arises when creating parallel corpora between a language with an extensive Wikipedia (e.g., English) and a less-resourced one. If the article lists are compiled separately for each language, the resulting corpora will likely have a substantial size imbalance. To mitigate this issue, the scripts and programs are equipped with a feature that limits the articles extracted from the larger Wikipedia to only those corresponding to articles present in the smaller edition.

3.4. Command-Line Interface

The tool to create parallel corpora is implemented as a command-line interface (CLI) organized into six subcommands each corresponding to a distinct operation: `create`, `segment`, `align`, `rescore`, `select`, and `pipeline`. Figure 1 shows the help message of the CLI. Its design around subcommands rather than a set of flags affords significant readability and operational flexibility: each phase can be run independently—for instance, running only `create` and then `segment`—permitting the inspection of intermediate outputs before proceeding to the subsequent stages of alignment and quality filtering. Alternatively, the whole process can be executed monolithically in a single run by using the `pipeline` subcommand. The tool is also structured as a Python package and can be installed and executed with the command `cfw` or `corpus-from-wikipedia`.

Subcommands

These activate the specific stages of the pipeline, allowing for either stepwise or monolithic execution.

`create` Initiate the pipeline by extracting pages from the dumps based on the specified categories.

`segment` Activate the creation of a segmented version of the extracted text.

`align` Perform the bitext mining (alignment) process between the L1 and L2 segmented corpora.

`rescore` Trigger the rescoring phase, where alignments are evaluated using more computationally expensive models (Language Detection and SBERT).

`select` Filter the rescored segments based on the defined quality thresholds.

`pipeline` Execute all the previous commands in one go.

Each subcommand defines a set of required positional arguments, and optional ones that modify the default behavior of the tool, as detailed in the following sections. Positional arguments, i.e., necessary parameters, need to be input in the correct order for the tool to work, while optional ones are specified with a short (`-`) or long flag (`--`) and can be in any order. When executing the pipeline command, many parameters (such as input and output) are handled internally. The following section highlights the most useful parameters for each subcommand, showcasing the full capabilities of the tool.

```
$ cfw -h
usage: cfw [-h] {create,segment,align,rescore,select,pipeline} ...

Tool that allows the creation, segmentation, alignment, rescoring and segment selection of
parallel corpora from Wikipedia. It supports step-by-step execution and full pipeline execution.
That is, commands may be run individually, e.g., in order to inspect results between steps,
or execute the entire workflow at once using the 'pipeline' command.

positional arguments:
  {create,segment,align,rescore,select,pipeline}
    create              Create a corpus from Wikipedia.
    segment             Segment the extracted corpus.
    align               Perform bitext mining and alignment between two corpora.
    rescore             Rescore the alignment using more computationally expensive models.
    select              Filter the rescored parallel segments
    pipeline            Execute the whole pipeline: create > segment > align > rescore > select

options:
  -h, --help            show this help message and exit
```

Figure 1: Script help output from `corpusFromWikipedia.py`.

`Create` **parameters**

`lang1` Name or two letter ISO code of the source language.

`lang2` Name or two letter ISO code of the target language. Keep empty to create a monolingual corpus.

`-c, --categories` Wikipedia categories to search. Must be in between quotation marks ("") and separated by a comma (,).

`-d, --depth` Recursion depth for traversing the category tree.

`--restrict` Restrict L2 pages to equivalent L1 pages.

`Segment` **parameters**

`--force-srx-lang` Override the default SRX language configuration if your language is not available in the file.

`--force-segmenter` Use a Stanza sentence segmenter instead of SRX. Note that it is slower than the SRX implementation, but useful if your language is supported in this library and not in the SRX file.

`--chunk` Compile the whole corpus in chunks before segmenting. If the corpus is made up of millions of files, the time saved is massive. Note that individual segmented articles will not be available.

`Align` **parameters**

`-dev, --device` Device used to align the segments, i.e., the GPU or the CPU. Default is GPU.

`--mode` Preset performance mode for computation of embeddings. 'safe' is meant for consumer hardware (8GB VRAM), 'balanced' is meant for workstations (24GB VRAM), 'fast' is meant for high-performance computing servers (80GB+ VRAM). Default: safe.

`Rescore` **parameters**

`--SEmodel` Define the SentenceTransformer model used to calculate cosine similarity. Default is `LaBSE`.

`--LDmodel` Specify the path to the FastText language detection model. Defaults to `lid.176.bin`.

`Select` **parameters**

`--sldc/--tldc` The minimum confidence threshold (default 0.75) for the source language detection and target language detection.

`--minSBERT` The minimum SBERT cosine similarity score required to select a segment pair (default 0.75).

`--min-chars` Minimum character length of selected segments. By default, segments of any length are selected. Short segments can be ignored using this flag. Default: 0.

`Pipeline` **parameters**

In addition to the previous commands, `pipeline` enables end-to-end execution of the whole process, making it much more convenient and simple since the user only needs to specify a few positional parameters: the source and target language, the Wikipedia categories to search for in the dumps and the depth. Optional parameters of each previously mentioned subcommand that modify the default behavior (such as `--restrict`, `--device` or `--sldc`) are also available. It might be worth mentioning that the output data of each process the pipeline goes through is also saved separately (i.e., the initial corpora, the segments, etc.), not only the end result. Here follows an example of the pipeline command:

```
cfw pipeline ca en -c "Biochemistry,
Neuroscience" --depth 3
```

3.5. GUI program

To enhance usability, a GUI-based wrapper for the script was developed using Tkinter, mirroring the command-line functionality.

UNESCO	Wikipedia	Cat.	P. ca	P. en
11 Logic	Logic	151	859	4,433
12 Mathematic	Mathematics	178	1,984	10,148
21 Astronomy and astrophysics	Astronomy, Astrophysics	288	1,877	7,579
22 Physics	Physics	407	4,333	16,744
23 Chemistry	Chemistry	504	4,561	21,457
24 Life Sciences	Biology	180	1,476	8,998
25 Earth and Space	Earth sciences, Astronomy, Astronautics	550	3,397	15,373
31 Agricultural sciences	Agriculture	542	2,193	13,930
32 Medical Sciences	Medicine, Health	1,109	4,476	30,111
33 Technological Sciences	Engineering	853	4,935	29,967
51 Anthropology	Antrhopology	321	2,285	12,072
52 Demographics	Demographics	278	451	3,586
53 Economic Sciences	Economics, Business	1,110	3,675	35,172
54 Geography	Geography	302	1,496	7,034
55 Histroy	History	1,065	3,464	21,610
56 Juridical sciences and Law	Law	841	1,779	15,944
57 Linguistics	Linguistics	426	2,456	15,226
58 Pedagogy	Education	554	1,458	13,685
59 Political Science	Political science	150	1,783	6,851
61 Psychology	Psychology	174	900	6,414
62 Sciences of Arts and Letters	Architecture, Performing arts, Visual arts, Literature	3,407	10,077	70,666
63 Sociology	Sociology	417	2,608	15,249
71 Ethics	Ethics	129	1,135	5,855
72 Philosophy	Philosophy	230	1,526	6,763

Table 1
Categories, and pages for each subject area in Wikipedia for depth 2

UNESCO	cat segments	eng segments	aligned-segments
11 Logic	32,425	81,444	4,786
12 Mathematic	71,305	169,838	12,835
21 Astronomy and astrophysics	44,784	91,285	8,176
22 Physics	142,155	390,979	33,378
23 Chemistry	142,879	376,931	25,313
24 Life Sciences	58,717	148,976	8,377
25 Earth and Space	109,532	236,327	16,930
31 Agricultural sciences	81,547	176,683	7,391
32 Medical Sciences	151,070	383,682	19,916
33 Technological Sciences	67,188	170,733	10,742
51 Anthropology	85,411	188,722	93,60
52 Demographics	14,303	33,676	1,308
53 Economic Sciences	131,124	328,324	17,431
54 Geography	35,574	71,074	3,666
55 History	142,366	290,218	17,158
56 Juridical sciences and Law	55,536	154,772	5,692
57 Linguistics	84,774	197,118	7,648
58 Pedagogy	53,300	120,886	6,058
59 Political Science	46,113	113,028	3,440
61 Psychology	35,985	91,369	4,605
62 Sciences of Arts and Letters	347,661	768,447	42,221
63 Sociology	93,985	243,023	10,177
71 Ethics	45,093	106,630	5,037
72 Philosophy	67,756	131,217	5,797

Table 2
Number of Catalan, English and aligned segments for each UNESCO category.

4. Experimental part

4.1. Setting

Our experiments utilized the English and Catalan Wikipedia dumps dated September 1, 2025. Table 1 details the UNESCO categories and the corresponding Wikipedia categories employed to identify the associated pages. As is evident from the table, the number of English pages significantly surpasses that of Catalan pages across all UNESCO categories. Therefore, for the construction of the corpora, we restricted the English collection to only those pages possessing a Catalan equivalent.

4.2. Results and evaluation

Table 2 details the number of unique segments extracted for Catalan and English, along with the count of parallel segments identified after a cleaning process. This cleaning utilized a 0.75 confidence score for language detection and SBERT cosine similarity. As is evident, despite restricting the English pages to those with a Catalan equivalent, the number of English unique segments is considerably larger than that of Catalan segments, indicating that the English pages are, on average, larger.

Table 2 also reveals a significant disparity: the number of aligned segments is substantially lower than the total

Domain	Correct	Incorrect	Correct With Variations
Life Sciences	94	6	8
Philosophy	91	9	5
Juridical sciences and Law	87	13	9

Table 3
Manual Evaluation of Alignment Accuracy for 100-Segment Samples per Domain.

number of available segments in both Catalan and English. This finding is expected, as equivalent Wikipedia pages are not necessarily direct translations of one another, even if some fragments happen to be. Nonetheless, since pages in both languages discuss the same subject matter, it remains likely that translation equivalents can be found.

To assess the quality of the generated corpora, a manual evaluation was conducted on the fields of Philosophy (72), Life Sciences (24), and Juridical sciences and Law (56). From each domain-specific corpus, a random sample of 100 sentence pairs was extracted and reviewed by a professional translator, who provided a binary judgment for each pair: correct or incorrect.

A segment pair was classified as correct if the target sentence was a complete and accurate translation of the source. This definition included close paraphrases and allowed for minor discrepancies, such as the addition of one to three tokens, provided that they were not crucial to the overall meaning. For example, the pair below was deemed correct as the additional tokens in the English segment do not alter the core message:

> Ajudats per aquest coneixement i juxtaposició de creences, els abbàssides consideraven valuós mirar l'islam amb ulls grecs, i mirar els grecs amb ulls islàmics.
>
> **Brickman 84-85** Aided by this knowledge and juxtaposition of beliefs, the Abbasids considered it valuable to look at Islam with Greek eyes, and to look at the Greeks with Islamic eyes.

Conversely, pairs were marked incorrect when additional tokens changed the sentence's meaning, as in the following case where "És la segona part" doesn't translate to "All this inspired":

> **És la segona part del** lema de la revolució francesa: llibertat, igualtat, fraternitat.
>
> **All this inspired the** motto of the French Revolution: Freedom, Equality, Fraternity.

The manual evaluation revealed a high alignment accuracy across all three domains. The Life Sciences corpus achieved the highest accuracy at 94% (94 correct, 6 incorrect pairs), followed by the Philosophy corpus with 91% accuracy (91 correct, 9 incorrect). The Juridical sciences and Law corpus also showed strong results with an 87% accuracy rate (87 correct, 13 incorrect).

Furthermore, the analysis quantified the subset of correct pairs that contained minor, non-essential token variations. This occurred in 8 pairs for Life Sciences, 9 for Juridical sciences and Law, and 5 for Philosophy, demonstrating the robustness of the alignment to slight paraphrasing.

4.3. Discussion

The manual evaluation results strongly validate the core methodology of this paper. The high accuracy scores, ranging from 87% to 94% across diverse domains, confirm that the combination of bitext mining followed by a rigorous rescoring and filtering process is highly effective for extracting quality parallel sentences from comparable Wikipedia corpora. The chosen 0.75 threshold for both language detection and SBERT similarity appears to strike an effective balance, filtering out noise while retaining a high volume of correct alignments.

An important finding is the nature of the "correct" alignments. Our evaluation criteria permissively included close paraphrases and minor variations. The fact that our SBERT-based filtering successfully identified these semantically equivalent (but not always literal) translations, as quantified in Table 3, is a significant strength. This suggests the resulting corpora are well-suited for training robust NMT systems, which would benefit from exposure to this type of linguistic diversity.

The results in Table 2 also shed light on the nature of the Wikipedia data itself. The observation that the number of unique English segments is substantially larger than Catalan segments, even after restricting the English pages to only those with Catalan equivalents, confirms that English articles are, on average, more comprehensive. This highlights the necessity of the article restriction feature to create a computationally feasible and relevant search space for the bitext mining.

Furthermore, the significant gap between the total number of unique segments and the final count of aligned segments is not a failure of the methodology but rather an expected outcome. It reinforces the premise that Wikipedia editions are comparable, not parallel, resources. Our work demonstrates that it is possible to "mine" these valuable, albeit sparse, parallel fragments from the much larger, non-parallel comparable texts.

The variance in accuracy between domains, with Juridical sciences and Law at 87% and Life Sciences at 94%, is also noteworthy. This could be attributable to the inherent nature of the subjects; legal language, for example, may contain more jurisdiction-specific concepts that are less likely to be directly translated between Catalan and English Wikipedia pages, resulting in fewer high-confidence alignments.

Ultimately, this methodology provides a practical solution to the core problem identified in the introduction: the scarcity of specialized, in-domain parallel data for low-resource languages. While the absolute number of segments for some fields (e.g., Demographics) is modest, this data is highly specialized. Such in-domain corpora are invaluable for the fine-tuning stage of NMT development, directly contributing to the goals of our project.

4.4. Limitations

The experimental evaluation presented in this paper has focused primarily on quantifying alignment accuracy; that is, the percentage of extracted parallel segments that constitute correct translations of each other. The high-precision results, ranging from 87% to 94%, serve to validate the efficacy

of the bitext mining and subsequent filtering pipeline.

However, a known limitation of this methodology is that the thematic purity of the generated corpora has not been formally evaluated. Our approach for compiling specialized corpora relies entirely on Wikipedia's user-assigned category system, supplemented by the more granular taxonomy provided by DBpedia.

While this structure is foundational to the toolkit, we have not performed a manual verification to ascertain whether all articles retrieved for a given domain were, in fact, correctly categorized. This reliance on the existing category tags means it is possible that the resulting domain-specific corpora contain a degree of thematic noise, potentially including articles that are not strictly relevant to the intended discipline.

5. Conclusions and future work

This study presents a user-oriented, modular software toolkit designed to construct domain-specific monolingual and parallel corpora from Wikipedia. The core methodology integrates Wikipedia's category hierarchy with modern bitext mining techniques grounded in multilingual sentence embeddings [1]. Using this approach, the authors generated specialized Catalan–English corpora for the fields of science and technology defined by the UNESCO nomenclature. An expert manual evaluation confirmed the high quality of these corpora, yielding an alignment accuracy of 87 to 94% in three distinct domains: Philosophy (72), Life Sciences (24), and Juridical Studies and Law (56).

The resulting framework provides a transparent and reproducible pipeline that reduces the barrier to corpus creation, thereby fostering linguistic diversity and adhering to open data principles. While acknowledging limitations such as potential thematic noise derived from out-of-domain content in Wikipedia pages, the authors highlight the need to train domain-specialized machine translation systems in order to assess whether the corpora developed in this work provide measurable advantages over general-purpose models.

In conclusion, the study establishes a scalable and adaptable framework to enrich linguistic resources for other under-represented languages.

This toolkit holds a free license (GNU-GPL v3) and can be downloaded from Github.[6]

Acknowledgements

This research received a grant (IEC-PR2025-4-REC) from the Institute for Catalan Studies (IEC) within the framework of the call for research grants 2025, and it is also partially supported by the OpenEU Alliance, with funding from the European Union (EUROPEAN EDUCATION AND CULTURE EXECUTIVE AGENCY - EACEA) under grant agreement No 101177241.

References

[1] M. Artetxe, H. Schwenk, Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond, Transactions of the Association for Computational Linguistics 7 (2019) 597–610. URL: http://dx.doi.org/10.1162/tacl_a_00288. doi:10.1162/tacl_a_00288.

[2] S. Reese, G. Boleda, M. Cuadros, L. Padró, G. Rigau, Wikicorpus: A word-sense disambiguated multilingual wikipedia corpus, in: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), 2010.

[3] D. Flickinger, S. Oepen, G. Ytrestøl, Wikiwoods: Syntacto-semantic annotation for english wikipedia., in: LREC, 2010.

[4] H. Schwenk, Filtering and mining parallel data in a joint multilingual space, arXiv preprint arXiv:1805.09822 (2018).

[5] M. Artetxe, H. Schwenk, Margin-based parallel corpus mining with multilingual sentence embeddings, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2019, pp. 3197–3203. URL: http://dx.doi.org/10.18653/v1/P19-1309. doi:10.18653/v1/p19-1309.

[6] H. Schwenk, V. Chaudhary, S. Sun, H. Gong, F. Guzmán, Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia, arXiv preprint arXiv:1907.05791 (2019).

[7] L. J. Solberg, A corpus builder for Wikipedia, Master's thesis, University of Oslo, 2012.

[8] T. P. Adewumi, F. Liwicki, M. Liwicki, Corpora compared: The case of the swedish gigaword & wikipedia corpora, arXiv preprint arXiv:2011.03281 (2020).

[9] A. Oliver, S. Álvarez, Filtering and rescoring the ccmatrix corpus for neural machine translation training, in: Proceedings of the 24th Annual Conference of the European Association for Machine Translation, 2023, pp. 39–45.

[6] https://github.com/mtuoc/MTUOC-corpusFromWikipedia