# Multilingual Evaluation of Main Content Extractors for Web Pages

Aurélien Bournonville
aurelien.bournonville@unicaen.fr
University of Caen Normandy
Caen, Normandy, France
Babbar.tech
Petit-Quevilly, Normandy, France

Gaël Dias
gael.dias@unicaen.fr
University of Caen Normandy
Caen, Normandy, France

Thomas Largillier
thomas.largillier@babbar.tech
Babbar.tech
Petit-Quevilly, Normandy, France

Emmanuel Marchand
emmanuel.marchand@babbar.tech
Babbar.tech
Petit-Quevilly, Normandy, France

Fabrice Maurel
fabrice.maurel@unicaen.fr
University of Caen Normandy
Caen, Normandy, France

Guillaume Pitel
guillaume.pitel@babbar.tech
Babbar.tech
Petit-Quevilly, Normandy, France

François Rioult
francois.rioult@unicaen.fr
University of Caen Normandy
Caen, Normandy, France

## Abstract

Tools designed to extract main content from web pages require thorough evaluation, yet existing benchmarks disproportionately focus on English-language datasets. Consequently, previous studies have shown that while these extractors are well-optimized for English, their effectiveness partially or entirely declines in other languages. This study reproduces and extends recent benchmarks by incorporating multilingual datasets as a key factor. We analyze extractor performance across five languages—Greek, English, Polish, Russian, and Chinese—highlighting the need to adapt extraction models to linguistic variation. Our results show that while some extractors maintain stable performance, others suffer significant drops in precision and recall on non-English or structurally irregular pages.

## CCS Concepts

• **Information systems** → **Content analysis and feature selection**; **Test collections**; **Document filtering**.

## Keywords

Boilerplate removal, Multilingual settings, Web content extraction

**ACM Reference Format:**
Aurélien Bournonville, Gaël Dias, Thomas Largillier, Emmanuel Marchand, Fabrice Maurel, Guillaume Pitel, and François Rioult. 2025. Multilingual Evaluation of Main Content Extractors for Web Pages. In *Proceedings of Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '25).* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3726302.3730234


*SIGIR '25, Padua, Italy.*

ACM ISBN 979-8-4007-1592-1/25/07
https://doi.org/10.1145/3726302.3730234

## 1 Introduction

Isolating the *main content* of web pages enables the extraction of meaningful information, making the web a valuable source of high-quality data. However, web pages also contain *boilerplate* elements such as navigation menus, footers, and advertisements. Bevendorff et al. [3] define main content as the central article or any non-redundant material, excluding both global page elements and user comments. Alternatively, main content can be viewed as the information users seek when visiting a page, with all else considered boilerplate. While comments are usually secondary, this is not universal—on some sites, user discussions are central. In this study, we adopt the conventional extractor perspective, which targets authored, textual content as main content.

Extractors typically follow either heuristic or machine learning approaches. Heuristic methods, used in tools like Readability and Trafilatura [1], rely on rules based on HTML structure and content density, offering efficiency but limited adaptability. Machine learning methods, such as those in Boilerpipe [4] and BoilerNet [7], use classifiers trained on labeled data to capture structural and textual patterns. While more flexible, they require annotated datasets and are computationally intensive.

Previous studies show that extractor performance varies significantly with language and HTML structure [2], with most systems optimized for English due to the availability of evaluation datasets, leading to reduced effectiveness on non-English content.

To explore the multilingual dimension of main content extraction, we replicated the Web Content Extraction Benchmark (WCEB) [3] under multilingual conditions. We first identified biases in existing datasets, notably the inclusion of comments in gold standards. We then introduced the DAnIEL dataset[1] [6], comprising English,

[1] https://github.com/rundimeco/waddle/tree/master/corpora/Corpus_daniel_v2.1

Chinese, Greek, Polish, and Russian pages, to assess language impact, including both dominant and underrepresented languages.

Our results emphasize the importance of adapting extraction models to linguistic diversity. By incorporating four additional languages alongside English, we observed structural differences—such as sentence length—that significantly affected performance. While some languages (e.g., Greek) occasionally improved results, others led to notable degradation. Lastly, we found that web page complexity does not reliably predict extraction quality.

## 2 Related Work

### 2.1 Heuristic and Machine Learning Extractors

Main content extractors use two kinds of approaches, based on *heuristics* or *machine learning*. Heuristic-based extractors rely on predefined assumptions about the structure of web pages. In order to identify the main content, these approaches often analyze HTML structure, such as tag density, content density, and link presence. They are computationally efficient and require no training data but depend heavily on human expertise to craft these rules. Examples include Readability[2][3], which uses handcrafted rules to optimize extraction for article-like pages, Trafilatura [1], which combines XPath queries with fallbacks and tools such as jusText [10].

In contrast, machine learning-based extractors use classification models to locate the main content or the boilerplate (binary classification). These models can leverage features such as text density, tag patterns, or word frequencies. For example, Boilerpipe [4] uses decision trees on shallow text and structural features, while BoilerNet [7] employs sequence-labeling models based on LSTMs. However, these models are computationally expensive and require labeled datasets for training, which are scarce.

### 2.2 Web Content Extraction Benchmarks

In Bevendorff et al. [3], 8 datasets were evaluated through 14 main content extractors and 5 HTML-to-text extractors. The datasets were: CETD, CleanEval, CleanPortalEval, Dragnet[4], Google-Trends-2017, L3S-GN1, Readability[5] and Scrapinghub. The main content extractors did their best to ignore the boilerplate. The HTML-to-text conversion tools were used as baselines, their role were to extract all the content from the page, including boilerplate.

The metrics used for evaluation were ROUGE-L [8] and the Levenshtein distance [5]. ROUGE-L evaluates textual similarity by capturing the longest common subsequence between two texts, while respecting word order. The Levenshtein distance measures the minimum number of operations (insertions, deletions and substitutions) required to transform one character string into another, regardless of linguistic structures.

Following their study, a page is considered *complex* if it contains a high proportion of boilerplate content. As the ground truth locates the main content, the complexity $c$ of a web page is related to the expert annotation and it is defined in Equation 1 where $T$ is a multiset of DOM text tokens and $truth(t)$ returns 1 if the token $t$ belongs to the ground truth, otherwise 0.

$$c = 1 - \frac{|\{t \in T : truth(t) = 1\}|}{|T|} \quad (1)$$

Bevendorff et al. show that while no single method outperforms all others for main content extraction, heuristic approaches generally prove more robust and efficient than machine learning models, especially on complex pages. Combining several tools improves overall performance by reducing the variance of results. But the authors mention that the field remains limited by a lack of recent large datasets.

Although the study is effective in terms of evaluation, we will add a few limitations. Firstly, there is indeed a lack of datasets specifically tuned for this task, and there is an over-representation of datasets targeting blogs and press articles genre, which obviously guides the creation and benchmarking of extractors. Similarly, the most widely used datasets are concentrated on the English language.

In our experiments, we propose to select three main content extractors: Trafilatura and Readability, due to their strong heuristic performance, and Boilerpipe as the top machine learning extractor. We also included html_text[6] for its high recall as an HTML-to-text tool, using it as the sole baseline.

### 2.3 Multilingual Evaluation of Extraction Tools

The multilingual study of the main content extractors by Barbaresi and Lejeune [2] compared their performance on the DAnIEL dataset, which comprises web pages in Greek, Chinese, Polish, Russian, and English.

This study concluded that the performance of web extraction tools varies considerably according to the languages and HTML structures of the web pages. Tools designed primarily for English showed superior results in that language but inferior performance on multilingual pages or pages written entirely in another language. Although this is highly language-dependent, heuristic approaches such as jusText and Readability were the most stable.

Although the findings of Barbaresi and Lejeune are instructive, they lack the same strong comparison with the widely used existing benchmarks as proposed by Bevendorff et al. [3]. As a consequence, in this paper, we propose to combine the best aspects of both studies to perform a robust evaluation of web content extractors in a multilingual setting, so that conclusive findings can be drawn.

## 3 Benchmark Datasets

In this section, we detail the modifications we have undertaken to improve experimental conditions, i.e. corpus cleaning by comments removal and introduction of a multilingual dataset.

### 3.1 Comments Removal

Upon reviewing web pages with poor performance, we observed that comments in blogs and news articles were still present in the ground truth of certain datasets. By default, extractors are generally configured not to extract comments, which can introduce a bias in the actual performance of the extractors. Since the ground truth in the Dragnet dataset was labeled, we were able to quickly remove the comments and retest the extractors. All extractors performed better on the cleaned dataset, with Readability benefiting the

[2]https://github.com/masukomi/arc90-readability
[3]https://github.com/mozilla/readability
[4]https://github.com/dragnet-org/dragnet
[5]https://github.com/mozilla/readability
[6]https://github.com/TeamHG-Memex/html-text

most. Finally, we compared the size of Dragnet before and after cleaning, based on the number of characters. It turned out that 48% of Dragnet content were comments. The CETD and CleanEval datasets suffered from a similar issue. As the comments were not marked in the ground truth, we preferred to exclude them from the study. Table 1 shows extraction results with ROUGE-L Precision, Recall and $F_1$ scores evidencing the positive impact of comment removal. The $F_1$ score clearly improves when the cleaned version of Dragnet is used, exclusively obtained by the increase in Recall. Indeed, while Precision is steadily higher with the original Dragnet, Recall is boosted when comments are removed, thus showing the importance of curated benchmarks.

**Table 1: ROUGE-L average comparison between cleaned Dragnet and the original Dragnet.**

| Model | Original Dragnet | | | Cleaned Dragnet | | |
|---|---|---|---|---|---|---|
| | Precision | Recall | $F_1$ | Precision | Recall | $F_1$ |
| HTML_TEXT | **0.474** | 0.995 | **0.604** | 0.368 | **0.997** | 0.501 |
| BOILERPIPE | **0.888** | 0.750 | 0.773 | 0.852 | **0.858** | **0.838** |
| READABILITY | **0.929** | 0.771 | 0.806 | 0.916 | **0.899** | **0.896** |
| TRAFILATURA | **0.916** | 0.834 | 0.839 | 0.858 | **0.925** | **0.861** |

## 3.2 Adding DAnIEL dataset

To improve the experimental conditions, we added the DAnIEL dataset –Diverse And Novel Information Extraction from Languages, introduced by Lejeune et al.[6]. DAnIEL was designed for news extraction and monitoring in a multilingual context. The data originates from multilingual news feeds collected from various online sources, and covers five languages. In the version of the dataset we accessed[7], the distribution was as follows: 476 pages in English, 400 in Chinese, 274 in Polish, 273 in Greek, and 266 in Russian. The corpus was filtered using a language detector[8], followed by a secondary manual review to ensure correct language attribution.

Figure 1 shows the complexity levels of pages from the DAnIEL dataset and the 8 other datasets included in the web Content Extraction Benchmark (WCEB). Interestingly, every sub-corpus of DAnIEL has a significantly higher level of complexity than all the other datasets included in WCEB proposed by Bevendorff et al. [3].

# 4 Results

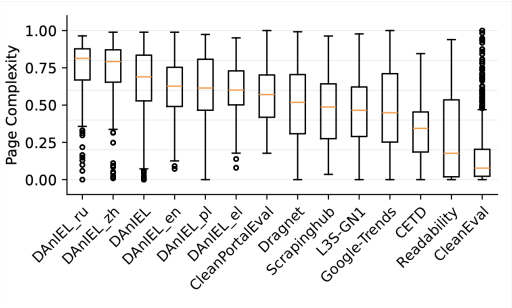
The evaluation of main content extractors was conducted using the ROUGE-L metric, with its Precision, Recall and $F_1$ scores. This metric allows for a robust comparison of performance across different models and datasets.

## 4.1 Language-independent Performance

The language-independent performance of the content extraction models, evaluated on both macro- and micro-average, is summarized in Table 2. Macro-average stands for the mean of each dataset performance, ignoring size, whereas micro-average weights each dataset performance by its size. To ensure a balanced evaluation

[7]https://github.com/rundimeco/waddle/tree/master/corpora/Corpus_daniel_v2.1
[8]https://pypi.org/project/langdetect/

**Figure 1: Web page complexity boxplots per dataset.**




across linguistic diversity, macro-averaging is employed by grouping classes equally based on their respective languages.

Models such as READABILITY and BOILERPIPE consistently demonstrated high performance across Precision, Recall, and $F_1$ scores, with READABILITY achieving a macro$F_1$ score of 0.838 and BOILERPIPE achieving 0.805. TRAFILATURA performs much worse than expected with a macro$F1_1$ score achieving only 0.77, which means it is less stable on non-English languages. The baseline extractor, HTML_TEXT, had of course very good Recall but poor Precision.Nevertheless, note that micro-averaging less penalizes TRAFILATURA, although globally same tendencies are observed between macro- and micro-averaging.

**Table 2: ROUGE-L metrics, Precision, Recall and $F_1$ over all 10 datasets by macro- and micro-averaging per language.**

| Macro-averaging per language | | | | | | |
|---|---|---|---|---|---|---|
| Model | Mean | | | Median | | |
| | Precision | Recall | $F_1$ | Precision | Recall | $F_1$ |
| HTML_TEXT | 0.300 | **0.970** | 0.414 | 0.280 | **1.000** | 0.412 |
| BOILERPIPE | 0.826 | 0.830 | 0.805 | 0.909 | 0.928 | 0.898 |
| READABILITY | **0.862** | 0.849 | **0.838** | **0.935** | 0.927 | **0.914** |
| TRAFILATURA | 0.749 | 0.880 | 0.770 | 0.828 | 0.955 | 0.852 |
| **Micro-averaging per language** | | | | | | |
| Model | Mean | | | Median | | |
| | Precision | Recall | $F_1$ | Precision | Recall | $F_1$ |
| HTML_TEXT | 0.407 | **0.983** | 0.531 | 0.393 | **1.000** | 0.562 |
| BOILERPIPE | 0.869 | 0.827 | 0.820 | 0.965 | 0.973 | 0.945 |
| READABILITY | **0.894** | 0.839 | **0.841** | **0.988** | 0.954 | **0.953** |
| TRAFILATURA | 0.835 | 0.883 | 0.823 | 0.968 | 0.964 | 0.934 |

## 4.2 Language-dependent Performance

As detailed in Table 3, the performance of three different extractors varies significantly across languages. In particular, we focused on READABILITY, BOILERPIPE and TRAFILATURA leaving out HTML_TEXT due to lack of space. READABILITY outperforms most models for the vast majority of languages, achieving an $F_1$ score of 0.962 (Greek), 0.862 (Polish), 0.840 (Russian) and 0.672 (Chinese). The only exception is evidenced by TRAFILATURA for English, with a maximum $F_1$ score value of 0.883. All models show similar behavior with respect

to different languages, showing that Greek and English are the best performing settings, while Chinese poses the greatest challenge.

**Table 3: ROUGE-L, Precision, Recall and $F_1$ over DAnIEL subcorpora for Readability, Boilerpipe and Trafilatura.**

| Readability | | | | | | |
|---|---|---|---|---|---|---|
| Model | Mean | | | Median | | |
| | Precision | Recall | $F_1$ | Precision | Recall | $F_1$ |
| Greek | 0.969 | 0.964 | **0.962** | 0.992 | 0.986 | **0.983** |
| English | 0.912 | 0.865 | 0.862 | 0.987 | 0.974 | **0.972** |
| Polish | 0.875 | 0.888 | **0.862** | 0.970 | 0.981 | 0.937 |
| Russian | 0.855 | 0.849 | **0.840** | 0.960 | 0.945 | **0.937** |
| Chinese | 0.688 | 0.702 | **0.672** | 0.759 | 0.762 | **0.750** |
| **Boilerpipe** | | | | | | |
| Model | Mean | | | Median | | |
| | Precision | Recall | $F_1$ | Precision | Recall | $F_1$ |
| Greek | 0.954 | 0.975 | 0.961 | 0.975 | 1.000 | **0.983** |
| English | 0.893 | 0.854 | 0.847 | 0.968 | 0.987 | 0.959 |
| Polish | 0.870 | 0.891 | 0.861 | 0.963 | 0.991 | **0.966** |
| Russian | 0.740 | 0.825 | 0.750 | 0.926 | 0.983 | 0.925 |
| Chinese | 0.664 | 0.622 | 0.611 | 0.711 | 0.688 | 0.667 |
| **Trafilatura** | | | | | | |
| Model | Mean | | | Median | | |
| | Precision | Recall | $F_1$ | Precision | Recall | $F_1$ |
| Greek | 0.833 | 0.933 | 0.868 | 0.934 | 0.984 | 0.938 |
| English | 0.899 | 0.911 | **0.883** | 0.984 | 0.968 | 0.956 |
| Polish | 0.784 | 0.886 | 0.800 | 0.899 | 0.989 | 0.922 |
| Russian | 0.729 | 0.856 | 0.759 | 0.841 | 0.951 | 0.857 |
| Chinese | 0.501 | 0.836 | 0.555 | 0.479 | 0.889 | 0.588 |

### 4.3 Low ROUGE-L F1 Web Pages

To highlight cross-language performance differences, we measured the proportion of problematic web pages in the DAnIEL dataset—those for which one or more extractors yield a ROUGE-L $F_1$ score below a given threshold. We tested thresholds from 0.1 to 0.5 (step 0.1), excluding html_text, used solely as a baseline. Table 4 shows the proportions of problematic pages per language and threshold, revealing a substantial impact of language on extraction. At the lowest threshold, almost all problematic pages are in Chinese, while Greek remains consistent with its overall performance.

## 5 Discussion

To strengthen our evaluation, we analyze the correlation between web page complexity and performance scores, and draw conclusions on feature language agnosticity.

### 5.1 Extractors Stability

Low performance is not necessarily related to high page complexity. We verify this by examining Pearson and Spearman correlations between complexity and ROUGE-L, with all p-values $< 0.001$ (Table 5). The Pearson coefficient [9] quantifies linear relationships, whereas Spearman [11] captures monotonic ones, making it more robust to non-linear dependencies.

**Table 4: Proportion (in percentage) of *problematic* web pages for one or several extractors by language.**

| $F_1 \leq$ n | n = 0.1 | n = 0.2 | n = 0.3 | n = 0.4 | n = 0.5 |
|---|---|---|---|---|---|
| Greek | 4 | 9 | 16 | 27 | 44 |
| English | 8 | 13 | 16 | 28 | 40 |
| Polish | 8 | 18 | 27 | 35 | 49 |
| Russian | 15 | 30 | 51 | 66 | 78 |
| Chinese | 67 | 87 | 95 | 98 | 99 |

The correlation between complexity $c$ and ROUGE-L $F_1$ is weak-to-moderate for the extractors, but strong for html_text. This analysis indicates that Readability is the most stable extractor across all complexity levels, while Boilerpipe and Trafilatura show similar stability. Unsurprisingly, html_text deteriorates as complexity increases.

**Table 5: Correlation between Complexity and ROUGE-L $F_1$ over 4233 pages. All p-values were below 0.001.**

| Extractor | Pearson $r$ | Spearman $\rho$ |
|---|---|---|
| Boilerpipe | -0.183 | -0.290 |
| Readability | **-0.084** | **-0.263** |
| Trafilatura | -0.294 | -0.345 |
| html_text | -0.924 | -0.949 |

### 5.2 Language Agnosticity of Features

Main content extractors score web page blocks to decide if they form part of the main content. For example, Readability employs character count, comma count, and link density. Pronounced differences in these cues, such as between English and Chinese, show that such features are not optimised for multilingual extraction. They heavily influence performance in languages that diverge from the ones targeted during design, while handcrafted rules may happen to suit others, like Greek.

## 6 Conclusion

In this article, we provided a robust analysis of the multilingual behavior of main web content extractors. Our study underscores the importance of adapting extraction models to linguistic nuances. By adding four languages, i.e. Greek, Polish, Russian and Chinese, to the dominant English datasets, structural differences appeared. The multilingual settings significantly impacted the extractors, occasionally enhancing performance (e.g., in Greek) but more often reducing it. In addition, we showed that the complexity of a web page does not determine the performance of the extraction task.

## Acknowledgments

This research is being carried out as part of the InDyX project, funded by the French government as part of France 2030.

## References


[1] Adrien Barbaresi. 2021. Trafilatura: A web scraping library and command-line tool for text discovery and extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*. 122–131.

[2] Adrien Barbaresi and Gaël Lejeune. 2020. Out-of-the-box and into the ditch? Multilingual evaluation of generic text extraction tools. In *Language Resources and Evaluation Conference (LREC 2020)*. 5–13.

[3] Janek Bevendorff, Sanket Gupta, Johannes Kiesel, and Benno Stein. 2023. An Empirical Comparison of Web Content Extraction Algorithms. In *Proceedings of the 46th International ACM SIGIR* (Taipei, Taiwan). ACM, 2594–2603.

[4] Christian Kohlschütter, Peter Fankhauser, and Wolfgang Nejdl. 2010. Boilerplate Detection Using Shallow Text Features. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining* (New York, USA). ACM, 441–450.

[5] V. I. Lcvenshtcin. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet Physics-Doklady*, Vol. 10. Issue 8.

[6] Gaël Lejeune, Romain Brixtel, Antoine Doucet, and Nadine Lucas. 2012. DAnIEL: Language Independent Character-Based News Surveillance. In *Advances in Natural Language Processing*. Vol. 7614. Springer Berlin Heidelberg, 64–75. Lecture Notes in Computer Science.

[7] Jurek Leonhardt, Avishek Anand, and Megha Khosla. 2020. Boilerplate Removal using a Neural Sequence Labeling Model. https://arxiv.org/pdf/2004.14294

[8] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*. 74–81.

[9] Karl Pearson. 1895. VII. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London* 58, 347 (1895), 240–242.

[10] Jan Pomikálek. 2011. Removing boilerplate and duplicate content from web corpora. *Disertacnı práce, Masarykova univerzita, Fakulta informatiky* (2011).

[11] Charles Spearman. 2010. The proof and measurement of association between two things. *International Journal of Epidemiology* 39, 5 (2010), 1137–1150.