Alkhalil Corpus: An Open-Source Thematic and Lemmatized Corpus for Modern Standard Arabic

Samir BELAYACHI
Department of Computer Science
Faculty of Sciences
Mohammed First University
Oujda, Morocco
samirbelayachi@gmail.com

Azzeddine MAZROUI
Department of Computer Science
Faculty of Sciences
Mohammed First University
Oujda, Morocco
azze.mazroui@gmail.com

Abstract

The availability of large annotated corpora remains a major challenge for the development of natural language processing systems for under-resourced languages such as Arabic. In this paper, we present two annotated corpora dedicated to Modern Standard Arabic. These corpora are open-source and freely available on the Hugging Face platform. The first corpus, annotated by theme and designed to provide a balanced representation of contemporary Arabic usage, comprises approximately 76 million words collected from diverse sources covering multiple domains and geographical regions. The second corpus, containing approximately one million words, is a sub-corpus extracted from the first. It was annotated with lemma tags using a semi-automatic approach that combines automatic annotation with the Alkhalil lemmatizer and MADAMIRA, followed by manual validation.

1 Introduction

In the field of Natural Language Processing (NLP), corpora refer to structured collections of written texts or spoken transcripts, collected according to specific criteria in order to represent real-world language use. They are crucial in language teaching because they enable the identification of the most frequent and relevant words and structures, as well as the formulation of linguistic hypotheses. These data are generally organized and annotated according to the types of linguistic applications to be developed (morphological, syntactic, or semantic). Two main types of corpora are distinguished according to their linguistic coverage: monolingual corpora, which focus on a single language and multilingual corpora, which encompass several languages, such as parallel corpora consisting of texts in a given language aligned with their translations into one or more other languages. The latter are particularly useful for comparative studies and machine translation.

In NLP, corpora are indispensable for numerous applications (Manning, 1999). They play a central role in machine learning, where systems require corpora during the training phase. In supervised learning, the availability of rich and carefully annotated linguistic corpora strongly influences system performance. For Arabic, considered an under-resourced language, building such corpora faces several challenges, including the language's morphological richness, the near-systematic absence of diacritics (short vowels) and the coexistence of orthographic and linguistic variants, particularly dialectal forms.

Among the various possible annotations, topic identification is of particular importance, as it improves information retrieval by optimizing document organization. Lemmatization also plays a crucial role in many applications (Manning et al., 2008). It involves reducing each word to its canonical form, corresponding to dictionary entries. By reducing morphological variation, lemmatization improves lexical disambiguation and enhances the performance of many NLP tasks that incorporate it as a preprocessing step, such as sentiment analysis (Touahri and Mazroui, 2021) or text readability assessment (Nassiri et al., 2018).

To enrich existing resources for the Arabic language, we collected and annotated two corpora of Modern Standard Arabic (MSA). The first corpus comprises approximately 76 million words and its collection was guided by three dimensions of representativeness. The temporal dimension involved limiting the corpus to texts published after 1850 to ensure a faithful representation of MSA. The second dimension is geographical, encompassing the entire Arab world, both East and West. Finally, the thematic dimension concerns the topics of the texts, with ten different themes well represented in this corpus. The second corpus, containing approximately one million words, is a sub-corpus extracted from the first and annotated with lemma tags. In the

Proceedings of the 2nd Workshop on NLP for Languages Using Arabic Script (AbjadNLP 2026), pages 192–197
March 28, 2026. ©2026 Association for Computational Linguistics

annotation process, we adopted a semi-automatic approach that begins with analyzing the texts using the two lemmatizers Alkhalil[1](Boudchiche et al., 2017) and MADAMIRA (Maamouri et al., 2004). Words for which both analyzers provide the same lemma are considered correctly lemmatized, while the remaining words are manually lemmatized by linguistic experts. These corpora, available as open-source resources, can be used by NLP researchers in supervised learning tasks.

The remainder of this paper is organized as follows: the second section presents the state of the art on available corpora. The third section describes the methodology adopted for collecting the theme-annotated corpus, along with the corresponding statistics. The fourth section is devoted to the lemma-annotated corpus. Finally, the paper concludes with a summary and perspectives for future research.

2 Related Work

The creation of Arabic language corpora increasingly relies on web resources. According to (Alrayzah et al., 2024; Zeroual and Lakhouaja, 2018), the majority of existing Arabic corpora are built entirely or partially from online data. This trend is explained by the richness and diversity of available content, which enables coverage of a wide range of domains and lexical varieties. However, this approach also raises several challenges, particularly in terms of data cleaning, standardization and linguistic quality control.

Several research teams have therefore created Arabic corpora using web data. For example, the general-purpose ArTenTen corpus (Belinkov et al., 2013) is built from massive web resources. Several versions of this corpus have been developed (arTenTen12, arTenTen18, arTenTen24): the first version contains approximately 7 billion words, while more recent versions are enriched with morphosyntactic annotations, including part-of-speech (POS) tags and lemmas (Arts et al., 2014). Similarly, the Nemlar corpus[2] (Boudchiche and Mazroui, 2015) is a valuable resource for the study of the Arabic language. It contains nearly 500,000 words distributed across 489 files and covers 13 diverse thematic areas. The words in the corpus are annotated with multiple linguistic features, such as vowel form, lemma, stem, clitics attached to the stem, grammatical category and morphological pattern.

The OSIAN corpus *OSIAN* (Zeroual et al., 2019) is an open-source resource collected from 32 popular Arabic newspapers. It consists of approximately 3.5 million articles containing over one billion words and each article is annotated with descriptive metadata.

The Tashkeela corpus *Tashkeela* (Zerrouki and Balla, 2017) is a large collection of approximately 80 million fully diacritized Arabic words. It consists primarily of classical and contemporary texts and was developed to support research on automatic diacritization.

Despite recent efforts to develop annotated resources for the Arabic language, such resources remain limited, as do standardized annotation tools and guidelines (Alayba, 2025). Consequently, Arabic continues to be under-resourced compared to many other languages. Therefore, the creation of new, high-quality resources remains essential for the development of effective NLP applications (Senator et al., 2025).

In this work, we first created a large MSA corpus of approximately 76 million words, annotated with theme labels. Next, a second corpus of approximately one million words was extracted from the first and annotated with lemma tags. Both corpora are open-source and can be used by NLP researchers for supervised model training and evaluation.

3 Methodology for Constructing the First Corpus

Our primary objective was to create a corpus representative of Modern Standard Arabic. During its construction, we therefore considered the following three criteria:

- **Temporal dimension**: all collected texts were produced after 1850 in order to ensure a faithful representation of contemporary MSA.
- **Geographical coverage**: to maximize lexical diversity, we ensured the inclusion of texts from different regions of the Arab world, notably the Maghreb, the Middle East and the Gulf region.
- **Thematic diversity**: particular attention was paid to thematic variety, with ten major themes well represented in the corpus.

[1] https://alkhalil.oujda-nlp-team.net/AlKhalil-Lemmatizer.php

[2] https://alkhalil.oujda-nlp-team.net/Nemlar.php

Table 1 presents all the sources used, including both news websites and works from digital libraries.

News websites	Digital libraries
Al Yaoum24[3]	Hindawi[4]
Al Jazeera[5]	Ashamela[6]
BBC Arabic[7]	
CNN Arabic[8]	
Hespress [9]	
Al Bayan[10]	
Arsco[11]	
Nature Research Journal[12]	

Table 1: Corpus sources

3.1 Data Cleaning

Data collected from the web generally contain elements that are irrelevant to linguistic analysis, such as HTML tags, JavaScript code, metadata, or text fragments in other languages. To obtain a homogeneous and usable corpus, we carried out a preprocessing phase, which is an essential step in building a linguistic corpus suitable for NLP. The objective of this phase is to improve the quality and consistency of the textual data by eliminating noise and ensuring that the selected documents accurately reflect authentic language use.

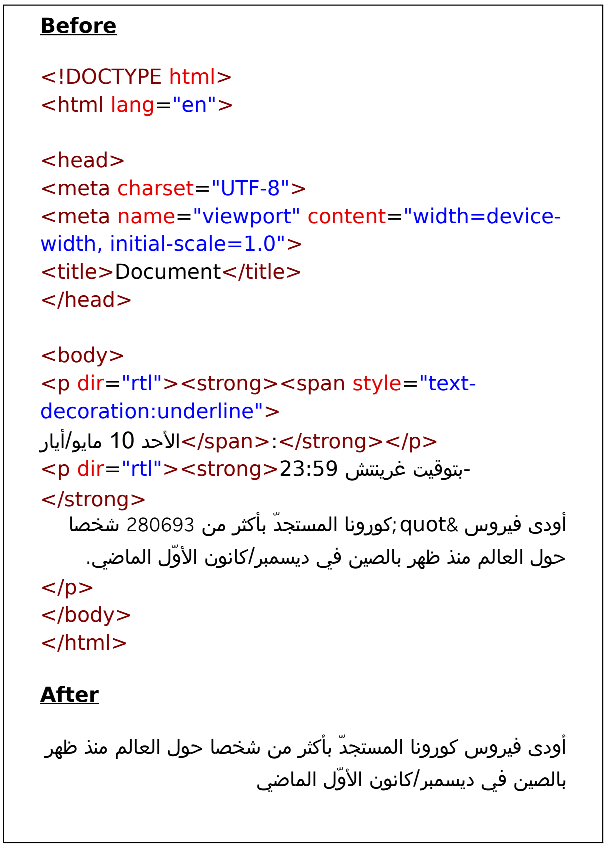
The preprocessing process we adopted is structured around the following axes:

- **Removal of non-textual elements**: all HTML tags, scripts, advertisements, hyperlinks and multimedia content were removed in order to retain only plain text.

- **Encoding normalization**: all documents were converted to UTF-8, ensuring compatibility with annotation and analysis tools.

- **Linguistic filtering**: segments containing primarily Latin characters, numerals, or other non-Arabic alphabets were removed. This step is crucial to avoid introducing linguistic noise, especially in news articles that may contain foreign terms or phonetic transcriptions.

An illustrative example of the data cleaning process is presented in Figures 1, showing the text *before* and *after* preprocessing.



Before

```
<!DOCTYPE html>
<html lang="en">

<head>
<meta charset="UTF-8">
<meta name="viewport" content="width=device-
width, initial-scale=1.0">
<title>Document</title>
</head>

<body>
<p dir="rtl"><strong><span style="text-
decoration:underline">
الأحد 10 مايو/أيار</span>:</strong></p>
<p dir="rtl"><strong>23:59 بتوقيت غرينتش-
</strong>
  أودى فيروس &quot;كورونا المستجدّ بأكثر من 280693 شخصا
  حول العالم منذ ظهر بالصين في ديسمبر/كانون الأوّل الماضي.
</p>
</body>
</html>
```

After

أودى فيروس كورونا المستجدّ بأكثر من شخصا حول العالم منذ ظهر
بالصين في ديسمبر/كانون الأوّل الماضي

Figure 1: Example of raw text before and after the preprocessing steps.

3.2 Similarity-Based Document Filtering

Web data frequently contain repeated content or slightly modified variants (for example, the same news article published on several websites). Such redundancies can skew lexical statistics and reduce the effective diversity of the corpus vocabulary.

To remove duplicate and similar documents, we used a method combining TF-IDF (*Term Frequency – Inverse Document Frequency*) weighting with the cosine similarity measure. This deduplication process was structured into three main steps:

1. **Document vectorization**: each document is represented as a TF-IDF-weighted vector, where each dimension corresponds to a distinct term in the corpus.

2. **Similarity calculation**: for each pair of documents, we compute the cosine similarity be-

[3] https://alyaoum24.com/
[4] https://www.hindawi.org/
[5] https://www.aljazeera.net/
[6] https://shamela.ws/
[7] https://www.bbc.com/arabic
[8] https://arabic.cnn.com/
[9] https://www.hespress.com/
[10] https://albayane.press.ma/
[11] https://arsco.org/
[12] https://www.natureasia.com/ar/nmiddleeast/

Topic	Document	Word	Vocabulary
Art	32	1125830	126446
Culture	3532	1313605	125990
Policy	29786	10662091	258740
Sport	45167	11066958	216636
Science	25653	11215501	263702
Society	16109	4631100	176369
Sociology	111	5838694	274323
Economics	27898	9321691	192013
Literature	337	13928747	595120
Health	21293	7154819	182215
Total	169918	76259036	

Table 2: Descriptive statistics of the corpus

tween their respective vectors. This measure ranges from 0 (completely different documents) to 1 (identical documents).

3. **Filtering**: when the similarity exceeds an empirical threshold set at 0.75, the two documents are considered similar and one of them is removed.

This method significantly reduced the raw size of the corpus while preserving a high level of lexical and thematic diversity, thereby avoiding the overrepresentation of certain terms or expressions due to content repetition.

We thus obtained a cleaned corpus C, whose statistics are presented in Table 2.

It is important to note that the term vocabulary refers to the set of distinct words present in the corpus.

To facilitate its use by the scientific community, the corpus has been made publicly available on our Hugging Face account[13], thus providing a practical and reusable resource for researchers and developers interested in the automatic processing of the Arabic language.

4 Annotation Methodology for the Second Corpus

Our second objective was to construct an Arabic corpus annotated with lemma tags. To achieve this, we adopted a hybrid annotation approach that combines automatic analysis with expert manual validation, in order to ensure high annotation quality while optimizing human effort.

In the first step, corpus C was analyzed using the lemmatizers of two Arabic morphological platforms, the Alkhalil Platform for Arabic Language Processing[14] (Boudchiche et al., 2017) and MADAMIRA (Pasha et al., 2014), both widely recognized for their reliability and accuracy. Each tool generated a lemma for every word in the analyzed corpus.

Table 3 presents selected statistics from this analysis. The two analyzers agree on the same lemma in 76.86% of cases and provide different lemmas in 18.97% of the words. In the remaining cases (4.17%), at least one of the analyzers fails to analyze the word. The coverage rates (percentage of words analyzed) of the Alkhalil and MADAMIRA lemmatizers are very high, reaching 97.18% and 97.05%, respectively.

Metric	Percentage
Concordance rate (between the two lemmatizers)	76.86%
Non-concordance rate	18.97%
Alkhalil coverage	97.18%
Madamira coverage	97.05%

Table 3: Analysis statistics

We then adopted a convergence-based validation strategy, inspired by standard manual annotation practices in linguistics [Carletta, 1996]. A word is considered correctly analyzed when both tools produce the same lemma. This assumption is grounded in the idea that agreement between two independent systems, relying on different models and lexical resources, constitutes a reliable indicator of morphological correctness. To assess the validity of this assumption, we randomly selected a set of sentences in which both analyzers assigned the same lemma to 2,218 words. A linguistic expert subsequently performed a manual verification of these joint decisions, revealing an error rate of approximately 0.81% (i.e., an accuracy of 99.19%) in cases where the two systems agreed on the lemma. These results indicate that agreement between the two tools provides a strong and reliable indicator of annotation correctness.

The analysis of sentences for which both analyzers produce exactly the same lemmas for all words shows that the concordance rate between Alkhalil and MADAMIRA is particularly high for short sentences, especially those composed of five

[13] https://huggingface.co/datasets/oujda-nlp-team/TopicClassifiedTexts

[14] https://alkhalil.oujda-nlp-team.net/

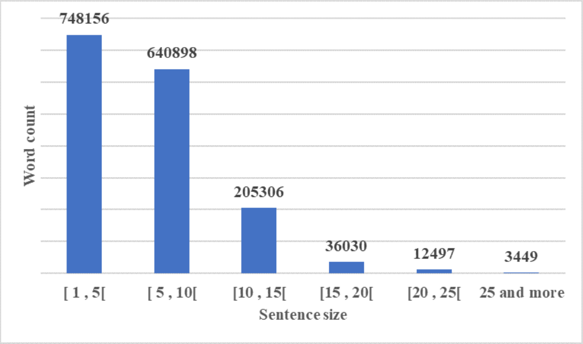


Figure 2: Word count by sentence length intervals for sentences with complete agreement between the two analyzers.

words or fewer. Conversely, a progressive decrease in concordance is observed as sentence length increases, due to greater contextual complexity and more frequent morphological ambiguities. Figure 2 presents the distribution of sentences showing complete agreement between the two lemmatizers according to sentence length.

These results directly guided the selection of sentences submitted for manual annotation. In order to minimize the workload of linguistic annotators while maximizing annotation quality, we prioritized all sentences for which agreement between the two lemmatizers was complete and whose length exceeded four words. This automatic validation covered 898,180 words, which were considered correctly annotated based on tool agreement.

Furthermore, among longer sentences, for which automatic validation proved less reliable, we selected a subset of 5 449 sentences, representing approximately 100 000 words. Among these, around 15 000 words exhibited lemma discrepancies between the two lemmatizers and therefore required manual intervention. These sentences were assigned to two expert linguistic annotators. The annotation was performed using an online collaborative platform, which facilitated coordination between annotators, ensured traceability of decisions and reduced the cognitive load associated with the task.

It should be noted that during this manual annotation phase, only words for which the two lemmatizers proposed different lemmas were examined, in order to resolve inconsistencies and improve the overall consistency of the corpus.

At the end of this iterative process, combining automatic validation guided by statistical analysis and targeted human expertise, we compiled a high-quality annotated corpus comprising approximately one million lemmatized words. This resource provides a reliable basis for training and evaluating Arabic natural language processing systems.

5 Conclusion et travaux futurs

To enrich the linguistic resources available for the Arabic language, we constructed two annotated corpora in this work.

The first corpus, annotated by theme, comprises approximately 76 million words and is representative of Modern Standard Arabic. The texts were collected from diverse sources, covering multiple geographical regions and thematic domains, thereby reflecting the richness and diversity of contemporary Arabic usage.

The second corpus is a sub-corpus annotated with lemma tags. It was constructed using a semi-automatic approach that combines automatic lemmatization with manual validation. This corpus consists of approximately one million words.

Both corpora are open-source and freely available on the Hugging Face platform[15]. They can be used for training and evaluation in a wide range of natural language processing applications for Arabic.

Looking ahead, we plan to further enrich these resources by adding additional levels of linguistic annotation, such as stem, root, diacritized form and morphosyntactic tags (POS). This will broaden the scope of these corpora and enhance their value for the scientific community. Through this initiative, we aim to contribute to the development and dissemination of high-quality open-source Arabic linguistic resources.

References

Abdulaziz M Alayba. 2025. Arabic natural language processing (nlp): A comprehensive review of challenges, techniques, and emerging trends. *Computers*, 14(11):497.

Asmaa Alrayzah, Fawaz Alsolami, and Mostafa Saleh. 2024. Arafast: Developing and evaluating a comprehensive modern standard arabic corpus for enhanced natural language processing. *Applied Sciences*, 14(12):5294.

Tressy Arts, Yonatan Belinkov, Nizar Habash, Adam Kilgarriff, and Vit Suchomel. 2014. artenten: Arabic corpus and word sketches. *Journal of King Saud University-Computer and Information Sciences*, 26(4):357–371.

[15] https://huggingface.co/datasets/oujda-nlp-team/

Yonatan Belinkov, Nizar Habash, Adam Kilgarriff, Noam Ordan, Ryan Roth, Vıt Suchomel, and 1 others. 2013. artenten: a new, vast corpus for arabic. *Proceedings of WACL*, 20.

M. Boudchiche and A. Mazroui. 2015. Enrichment of the nemlar corpus with the lemma label. In *Study Day "Arabic Language Resources for NLP: Construction, Standardization, Management and Operation"*, Rabat, Morocco. November 26, 2015.

Mohamed Boudchiche, Azzeddine Mazroui, Mohamed Ould Abdallahi Ould Bebah, Abdelhak Lakhouaja, and Abderrahim Boudlal. 2017. Alkhalil morpho sys 2: A robust arabic morpho-syntactic analyzer. *Journal of King Saud University-Computer and Information Sciences*, 29(2):141–146.

Mohamed Maamouri, Ann Bies, Tim Buckwalter, and Wigdan Mekki. 2004. The penn arabic treebank: Building a large-scale annotated arabic corpus. In *NEMLAR conference on Arabic language resources and tools*, volume 27, pages 466–467. Cairo.

Christopher Manning. 1999. *Foundations of statistical natural language processing*. The MIT Press.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, UK.

Naoual Nassiri, Abdelhak Lakhouaja, and Violetta Cavalli-Sforza. 2018. Modern standard arabic readability prediction. In *ARABIC LANGUAGE PROCESSING: FROM THEORY TO PRACTICE*, volume 782 of *Communications in Computer and Information Science*, pages 120–133. Sidi Mohammed Ben Abdellah Univ, Natl Sch Appl Sci; Arabic Language Engn Soc Morocco; Ctr Natl Rech Sci Tech; Acad Hassan II Sci Tech; Ecole Natl Sci Appliquees Fes; LISA; Fac Sci Tech; Fac Sci Scharia. 6th International Conference on Arabic Language Processing (ICALP), Fez, MOROCCO, OCT 11-12, 2017.

Arfath Pasha, Mohamed Al-Badrashiny, Mona Diab, Ahmed El Kholy, Ramy Eskander, Nizar Habash, Manoj Pooleery, Owen Rambow, and Ryan Roth. 2014. MADAMIRA: A fast, comprehensive tool for morphological analysis and disambiguation of Arabic. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1094–1101, Reykjavik, Iceland. European Language Resources Association (ELRA).

Ferial Senator, Abdelaziz Lakhfif, Imene Zenbout, Hanane Boutouta, and Chahrazed Mediani. 2025. Leveraging chatgpt for enhancing arabic nlp: Application for semantic role labeling and cross-lingual annotation projection. *IEEE Access*.

Ibtissam Touahri and Azzeddine Mazroui. 2021. Studying the effect of characteristic vector alteration on arabic sentiment classification. *Journal of King Saud University-Computer and Information Sciences*, 33(7):890–898.

Imad Zeroual, Dirk Goldhahn, Thomas Eckart, and Abdelhak Lakhouaja. 2019. Osian: Open source international arabic news corpus-preparation and integration into the clarin-infrastructure. In *Proceedings of the fourth arabic natural language processing workshop*, pages 175–182.

Imad Zeroual and Abdelhak Lakhouaja. 2018. Data science in light of natural language processing: An overview. *Procedia Computer Science*, 127:82–91.

Taha Zerrouki and Amar Balla. 2017. Tashkeela: Novel corpus of arabic vocalized texts, data for autodiacritization systems. *Data in brief*, 11:147.