# OSIAN: Open Source International Arabic News Corpus - Preparation and Integration into the CLARIN-infrastructure

**Imad Zeroual[*], Dirk Goldhahn[†], Thomas Eckart[†], Abdelhak Lakhouaja[*]**
[*]Computer Sciences Laboratory, Mohamed First University, Morocco
{mr.imadine, abdel.lakh}@gmail.com
[†]Natural Language Processing Group, University of Leipzig, German
{dgoldhahn, teckart}@informatik.uni-leipzig.de

## Abstract

The World Wide Web has become a fundamental resource for building large text corpora. Broadcasting platforms such as news websites are rich sources of data regarding diverse topics and form a valuable foundation for research. The Arabic language is extensively utilized on the Web. Still, Arabic is relatively an under-resourced language in terms of availability of freely annotated corpora. This paper presents the first version of the Open Source International Arabic News (OSIAN) corpus. The corpus data was collected from international Arabic news websites, all being freely available on the Web. The corpus consists of about 3.5 million articles comprising more than 37 million sentences and roughly 1 billion tokens. It is encoded in XML; each article is annotated with metadata information. Moreover, each word is annotated with lemma and part-of-speech. The described corpus is processed, archived and published into the CLARIN infrastructure. This publication includes descriptive metadata via OAI-PMH, direct access to the plain text material (available under Creative Commons Attribution-Non-Commercial 4.0 International License - CC BY-NC 4.0), and integration into the WebLicht annotation platform and CLARIN's Federated Content Search FCS.

## 1 Introduction

The Arabic language is spoken by 422 million people, making it the fourth most used language on the Web[1]. Its presence on the Web had the highest growth of the ten most frequent online languages in the last 18 years. However, a few years ago, Arabic was considered relatively an under-resourced language that lacks the basic resources and corpora for computational linguistics, not a single modern standard Arabic tagged corpus was freely or publicly available. Since then, major progress has been made in building Arabic linguistic resources, primarily corpora (Zeroual and Lakhouaja, 2018a); still, building valuable annotated corpora with a considerable size is expensive, time-consuming, and requires appropriate tools. Therefore, many Arabic corpora builders produce their corpora in a raw format.

For building the Open Source International Arabic News (OSIAN) corpus, the typical procedures of the Leipzig Corpora Collection were utilized. Furthermore, a language-independent Part-of-Speech (PoS) tagger, Treetagger, is adapted to annotate the OSIAN corpus with lemma and part-of-speech tags.

The prime motivation for building OSIAN corpus is the lack of open-source Arabic corpora that can cope with the perspectives of Arabic Natural Language Processing (ANLP) and Arabic Information Retrieval (AIR), among other research areas. Hence, we expect that the OSIAN corpus can be used to answer relevant research questions in corpus linguistics, especially investigating variation and distinction between international and national news broadcasting platforms with a diachronic and geographical perspective.

After this introduction, the remainder of the paper is structured as follows: In section 2, we highlight the state-of-the-art of web-crawled

[1] http://www.internetworldstats.com/stats7.htm



*Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 175–182
Florence, Italy, August 1, 2019. 

corpora of the Arabic language. Further, the methodology and tools used to build the OSIAN corpus are presented in Section 3. In Section 4, the OSIAN corpus is described in more detail, yet, some data analyses are performed and discussed. Finally, Section 5 contains some concluding remarks and future work.

## 2 Literature review

The World Wide Web is an important source for researchers interested in the compilation of very large corpora. A recent survey (Zeroual and Lakhouaja, 2018b) reports that 51% of corpora are constructed based, totally or partially, on Web content. Web corpora continue to gain relevance within the computational and theoretical linguistics. Given their size and the variety of domains covered, using Web-derived corpora is another way to overcome typical problems faced by statistical corpus-based studies such as data-sparseness and the lack of variation.

The web corpora continue to gain relevance within the computational and theoretical linguistics. Given their size and the variety of domains covered, using web-derived corpora is another way to overcome typical problems faced by statistical corpus-based studies such as data-sparseness and the lack of variation. Besides, they can be used to evaluate different approaches for the classification of web documents and content by text genre and topic area (e.g., (Chouigui et al., 2017)). Furthermore, web corpora have become a prime and well-established source for lexicographers to create many large and various dictionaries using specialised tools such as the corpus query and corpus management tool Sketch-Engine (Kovář et al., 2016). Moreover, some completely new areas of research, for which they deal exclusively with web corpora, have emerged. Indeed, the aim was to build, investigate, and analyse corpora based on online social networks posts, short messages, and online forum discussions.

Publicly available Arabic web corpora are quite limited, which greatly impacts research and development of Arabic NLP and IR. However, some research groups (Zaghouani, 2017) have shown potentials in building web-derived corpora in recent years. Among them are:

- Open Source Arabic Corpora[2] (OSAC) (Saad and Ashour, 2010): It is a collection of large and free accessible raw corpora. The OSAC corpus consists of web documents extracted from over 25 Arabic websites using the open source offline explorer, *HTTrack*. The compilation procedure involves converting HTML/XML files into UTF-8 encoding using "Text Encoding Converter" as well as removing the HTML/XML tags. The final version of the corpus comprises roughly 113 million tokens. Besides, it covers several topics namely Economy, History, Education, Religion, Sport, Health, Astronomy, Law, Stories, and Cooking Recipes.
- arTenTen (Arts et al., 2014): It is a member of the TenTen Corpus Family (Jakubíček et al., 2013). The arTenTen is a web-derived corpus of Arabic crawled using Spiderling (Suchomel et al., 2012) in 2012. The arTenTen corpus is partially tagged. i.e., one sample of the corpus, comprises roughly 30 million, is tagged using the Stanford Arabic part-of-speech tagger. While, another sample, contains over 115 million words, is tokenised, lemmatised, and part-of-speech tagged using MADA system. All in all, the arTenTen comprises 5.8 billion words but it can only be explored by paying a fee via the Sketch Engine website[3].
- ArabicWeb16: Since 2009, the ClueWeb09 web crawl (Callan et al., 2009), that includes 29.2 million of Arabic pages, was considered the only and largest Arabic web crawl available. However, in 2016, a new and larger crawl of today's Arabic web is publicly available. This web crawl is called ArabicWeb16 (Swuaileh et al., 2016) and comprises over 150M web pages crawled over the month of January 2016. In addition to addressing the limitation of the ClueWeb09, ArabicWeb16 covers both dialectal and Modern Standard Arabic. Finally, the total size of the compressed dataset of ArabicWeb16 is about 2TB and it is available for download after filling a request form[4].
- The GDELT Project[5] is a free open platform for research and analysis of the global

[2] https://sites.google.com/site/motazsite/corpora/osac
[3] https://www.sketchengine.co.uk/
[4] https://sites.google.com/view/arabicweb16
[5] https://www.gdeltproject.org/

database. All the datasets released are free, open, and available for unlimited and unrestricted use for any academic, commercial, or governmental use. Also, it is possible to download the raw datafiles, visualize it, or analyse it at limitless scale. Recently, the GDELT Project is starting to create linguistic resources. In fact, 9.5 billion words of worldwide Arabic news has been monitored over 14 months (February 2015 to June 2016) to make a trigram dataset for the Arabic language. Consequently, an Arabic trigram table of the 6,444,208 trigrams that appeared more than 75 times is produced[6].

It is worth mentioning that larger corpora in the region of billions of words are usually created by downloading texts from the web unselectively with respect to their text type or content. Therefore, the content of such corpora cannot be determined before their construction, thus, it is necessary to filter, clean, and evaluate it afterwards.

## 3 Methodology and tools

In this section, we describe the crawling, processing and annotation tasks alongside with the tools used.

### 3.1 Data acquisition

In a first step the data needs to be crawled from the World Wide Web. Since the crawled data are often duplicated or in other ways problematic, they need to be cleaned and filtered. Therefore, the following processing steps were executed.

#### 3.1.1 Leipzig Corpora Collection

The Leipzig Corpora Collection (LCC) (Goldhahn et al., 2012; Quasthoff et al., 2014) started as "Projekt Deutscher Wortschatz[7]" in the Nineties as a resource provider for digital texts in the German language mostly based on newspaper articles and royalty-free text material.

Today, the LCC offers corpus-based monolingual full form dictionaries in more than 200 languages mainly based on online accessible text material, divided under several aspects like the year of acquisition, text genre, country of origin and more. Since June 2006, LCC can be accessed at http://corpora.uni-leipzig.de. In addition to direct access via a Web interface, LCC data is also offered for free download.

For each word the dictionaries contain:

- Word frequency information.
- Sample sentences.
- Statistically significant word co-occurrences (based on left or right neighbours or whole sentences).
- A semantic map visualizing the strongest word co-occurrences.
- Part of speech information (partially).
- Similar words and other semantic information (partially).

#### 3.1.2 Crawling and processing of data

For corpus creation, an adapted version of the CURL-portal (Crawling Under-Resourced Languages[8]) (Goldhahn et al., 2016) of the LCC was utilized. CURL allows creating Web-accessible and downloadable corpora by simply entering URLs into the portal. In order to build a balanced corpus of international Arabic news, the data have been drawn from a wide range of reliable sources around the world. Six million webpages were downloaded, three and a half million pages which contain Arabic text were extracted and sub-corpora for several Arabic speaking countries were created.

The crawling was conducted in March 2018 using Heritrix, the crawler of the Internet Archive. Further processing was carried out according to the language independent processing chain described in (Goldhahn et al., 2012) and involved steps as extracting raw text from the Web ARChive file format, sentence separation and removal of non-sentences using regular expressions. Finally, texts were extracted based on Web domain and assigned to the respective country. Furthermore, since the crawler writes the data in one large file, we developed a tool for extracting the texts based on the Web domain. For each Web domain, the tool extracts and saves each article/page in a single file. Finally, these articles are assigned to the respective country. A list of the crawled Web domains, the number of articles extracted, and the countries covered are provided in the Appendix "A".

[6] https://goo.gl/MZZkDJ
[7] http://wortschatz.uni-leipzig.de
[8] http://curl.corpora.uni-leipzig.de/

The number of articles extracted from the crawled data is varying from one website to another. Some domains were only restricted by the short duration of the crawling, whereas others ran out of crawlable URLs early due to a low amount of crawlable resources, robots.txt-restrictions or external links to other domains which were not followed.

### 3.2 Corpus annotation

Among the widely used and relevant types of corpus annotations are e.g. lemma and part of speech. Lemmatization is a basic morphological analysis to deal with derivation paradigms, whereas part-of-speech tagging is part of a further syntactic analyses (i.e., parsing) to determine the sentence's syntactic structure. Both annotation forms affect the performance of subsequent text analysis in NLP and IR.

For both part of speech tagging and lemmatization tasks, we used a previously adapted and well-established version of Treetagger for the Arabic language (Imad and Abdelhak, 2016). Further, we improved this model and retrained it using new linguistic resources namely the Frequency Dictionary of Arabic (Buckwalter and Parkinson 2014). This frequency dictionary contains the top 5,000 words that were derived from a collection of representative corpora that include 30 million words of both written texts and transcribed speech.

A sample of 10,000 words of the corpus has been manually checked to evaluate the performance of Treetagger and the achieved accuracy rate is 95.02%.

## 4 The OSIAN Corpus

Instead of using unselected data from the Web, the aim of the OSIAN corpus is to build a balanced corpus in which the data must be drawn from a wide range of reliable and open sources. Therefore, this corpus is compiled based on 31 different international Arabic news broadcasting platforms, all being freely available on the Web.

We extracted six million webpages. After cleaning and filtering, we were left with about three and half million articles comprising more than 37 million sentences and roughly 1 billion tokens.

### 4.1 Word length distribution

The average length of words varies from 7 to 12 letters in many languages[9]. According to Mustafa (2012), the average length of Arabic words in a normal text is five letters. When analyzing the OSIAN corpus the length of 36% of the words is above six letters, this percentage is increased to 75% if duplicate words are considered. This makes the corpus a good soil to evaluate techniques that aim to reduce a word to its base form.

It is worth mentioning that tokens with length superior to 10 letters are not considered since news articles contain phrases written without space characters between words as well as non-derived and concatenated words, such as "الأورومتوسطي"/Euro-Mediterranean, "الكهرومغناطيسية"/Electromagnetism, etc. This explains why we found more than two million unique tokens that consist of over 11 letters which is an irrational result for the Arabic language.

| Word length | Occurrence (Unique) | Percentage | Occurrence (Duplicated) | Percentage |
|---|---|---|---|---|
| 2 | 4,180 | 0,03% | 113,129,168 | 12,22% |
| 3 | 45,723 | 0,28% | 148,295,530 | 16,03% |
| 4 | 412,528 | 2,52% | 154,159,209 | 16,66% |
| 5 | 1,550,485 | 9,48% | 175,925,523 | 19,01% |
| 6 | 2,877,426 | 17,59% | 133,290,941 | 14,40% |
| 7 | 3,353,777 | 20,50% | 107,877,916 | 11,66% |
| 8 | 2,864,584 | 17,51% | 54,007,298 | 5,84% |
| 9 | 1,919,115 | 11,73% | 20,526,042 | 2,22% |
| 10 | 1,196,370 | 7,31% | 9,072,780 | 0,98% |
| >10 | 2,137,492 | 13,06% | 9,050,623 | 0,98% |
| Total | 16,361,680 | 100% | 925,335,030 | 100% |

Table 1: Word length statistics

[9] http://www.ravi.io/language-word-lengths

Table 1 displays the percentage of words covered in the OSIAN corpus with respect to their lengths, including unique and duplicate words.

### 4.2 Word frequency list

Calculating word frequencies enables us to indicate the distribution of words across the text categories. Besides, it is feasible to produce word frequency lists using the tokens' PoS tags instead of their orthographic status.

Obviously, function words will be at the top of the frequency wordlist. Nevertheless, the words thematically organized in Table 2 are also among the most frequent words.

In the context of IR and corpus linguistics, many of the top frequently words have no value or effect on further analyses since they are typical in news articles; examples include "العالم" (World: F=1,182,181; R=37), "الحكومة" (Government: F=667,862; R=73), and "مفاوضات" (Negotiations: F=524,035; R=101). However, the words listed in Table 2 are a result of the circumstances of the Middle East in recent years, FIFA World Cup, and the Brexit, which make these words occur frequently in various world news. Using LancsBox to analyze the corpus data, it was possible to calculate frequencies of words that are obvious collocates such as "كأس العالم" (World Cup), "الاتحاد الأوربي" (European Union), and "البيت الأبيض" (White House). Moreover, it is also possible to calculate statistical information about the association, the strength of collocation, and the comparative frequencies of word forms in the overall data of the OSIAN corpus or in country-separated data.

### 4.3 Corpus format

The XML-format is used to facilitate the use of the corpus. This is the first version of the OSIAN corpus which consists of separate directories for each country. Furthermore, each directory includes the articles in XML format, where the sentences are lemmatized and PoS tagged. Moreover, the XML files contain metadata to provide information about domain names, webpage location, and the date of extraction. For more illustration, Figure 1 presents a sample of the XML files.

Note that some Web domains include in their URLs the topic of the published articles like the sample provided in Figure 1 where the word "Science and tech" appeared in the article's URL. This is another feature that can be used to classify the articles based on their topics, one among other techniques, to prepare them for classification and topic detection. Unfortunately, not all the URLs include such information; therefore, the topic label remains "unknown" till a solution is found (using topic detection and tracking methods).

### 4.4 CLARIN Integration

CLARIN[10] (**C**ommon **L**anguage **R**esources and **T**echnology **I**nfrastructure) is a European Research Infrastructure established in 2012 and took up the mission to create an online environment to provide access to language resources (in written, spoken, or multimodal form)

| Theme | Word | Frequency (F) | Rank (R) |
|---|---|---|---|
| Persons | ترامب (Trump, President of USA) | 608,176 | 81 |
| | سلمان (Salman, King of Saudi) | 380,086 | 164 |
| | السيسي (El-Sisi, President of Egypt) | 114,586 | 687 |
| Countries | سوريا (Syria) | 960,732 | 51 |
| | بريطانيا (United Kingdom) | 862,156 | 57 |
| | قطر (Qatar) | 704,457 | 70 |
| Topics | الانتخابات (Election) | 482,688 | 117 |
| | بريكست (Brexit) | 434,376 | 134 |
| | كأس العالم (World Cup) | 349,873 | 188 |
| Organizations | الناتو (NATO) | 387,174 | 161 |
| | الاتحاد الأوربي (European Union) | 177,383 | 448 |
| | البيت الأبيض (White House) | 124,762 | 648 |

Table 2: Relevant words from the frequency wordlist

[10] https://www.clarin.eu/

```
<?xml version="1.0" encoding="UTF-8"?>
<Article num="1">
<Source name="BCC">
  <Date>2018-03-19</date>
  <Location>http://www.bbc.com/arabic/scienceandtech/2014/08/140829_smart_watches_samsung_lg
  </Location>
  <Topic> Science and Tech</Topic>
  <Language>ara</Language>
</Source>
<Text>
أعلنت شركتا سامسونغ وإلى جي الكوريتين الجنوبيتين طرح المزيد من الساعات الذكية…
</Text>
<Annotation>
  <Sentence id="1">
    <Word Surfaceform="أعلنت" PoS="VERB" Lemma="أَعْلَنَ" />
    <Word Surfaceform="شركتا" PoS="NOUN" Lemma="شَرِكَة" />
    <Word Surfaceform="سامسونغ" PoS="PN" Lemma="سَامْسُونْغ" />
    <Word Surfaceform="وإلى" PoS="PRT" Lemma="إِلى" />
    <Word Surfaceform="جي" PoS="ABR" Lemma="جى" />
    <Word Surfaceform="الكوريتين" PoS="ADJ" Lemma="كُورِيّ" />
    <Word Surfaceform="الجنوبيتين" PoS="ADJ" Lemma="جَنُوبِيّ" />
    <Word Surfaceform="طرح" PoS="NOUN" Lemma="طَرْح" />
    <Word Surfaceform="المزيد" PoS="NOUN" Lemma="مَزِيد" />
    <Word Surfaceform="من" PoS="PRT" Lemma="مِنْ" />
    <Word Surfaceform="الساعات" PoS="NOUN" Lemma="سَاعَة" />
    <Word Surfaceform="الذكية" PoS="ADJ" Lemma="ذَكِيّ" />
    …
  </Sentence>
…
</Annotation>
</Article>
```

Figure 1: A sample of OSIAN corpus encoded in XML format

for the support of scholars in the humanities and social sciences, and beyond (de Jong et al., 2018). Currently, CLARIN also offers advanced tools to discover, explore, exploit, annotate, analyse, and combine such data sets wherever they are located.

Unsurprisingly, a strong focus of CLARIN has been laid so far on resources for European languages. The integration of more data for non-European languages will broaden and extend possible research questions that users of the infrastructure can approach. Among others, the CLARIN centre at the University of Leipzig is working on expanding available resources for a variety of languages with a dedicated focus on lesser-resourced ones.

Based on standard procedures and workflows that have been proven effective for "in-house" resources, the OSIAN corpus is processed, archived and published into the CLARIN infrastructure. This publication includes descriptive metadata via OAI-PMH[11], direct access to the plain text material (available under Creative Commons Attribution-NonCommercial 4.0 International License - CC BY-NC 4.0), and integration into the WebLicht annotation platform and CLARIN's Federated Content Search FCS. In the future, the corpus will be made available via the KonText advanced corpus query interface for the Manatee-open corpus search engine (as used in the NoSketchEngine). This will enable compatibility with the FCS-QL specification v2.0 and will allow querying text and annotation layers such as part of speech and lemmas.

## 5 Conclusion and future work

In this paper we presented a new open source corpus based on well-known and reliable international broadcasting platforms. After cleaning and filtering processes, the datasets are automatically annotated with lemma and PoS tags.

[11] See for example http://hdl.handle.net/11022/0000-0007-C65C-3

At the moment, this corpus comprises roughly 1 billion tokens that have been stored in a uniform XML format. The XML format of the OSIAN corpus will be publicly available for download and use in research. In addition, the current version and any updates of the OSIAN corpus can be found through the CLARIN research infrastructure, connecting them to central services such as VLO and FCS for metadata and content search.

In the future, we will extend the OSIAN corpus to cover more international Arabic news with a diachronic and geographical perspective to make the corpus an ideal choice to explore language change and variation. Additionally, we will aim to improve the accuracy of the used tools as well as to adopt new and meaningful forms of annotation. Regarding CLARIN-integration, FCS 2.0 and the querying of annotation layers is planned to be supported. Furthermore, we will explore the usage of the OSIAN corpus in corpus linguistics, ANLP, and AIR.

## References


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.

Tim Buckwalter and Dilworth Parkinson. 2014. A frequency dictionary of Arabic: Core vocabulary for learners. *Routledge*.

Jamie Callan, Mark Hoy, Changkuk Yoo, and Le Zhao. 2009. *Clueweb09 data set*. Available at http://lemurproject.org/clueweb09/.

Amina Chouigui, Oussama Ben Khiroun, and Bilel Elayeb. 2017. ANT Corpus: An Arabic News Text Collection for Textual Classification. In *proceedings of the 14th ACS/IEEE International Conference on Computer Systems and Applications (AICCSA 2017)*, pages 135–142, Hammamet, Tunisia.

Franciska de Jong, Bente Maegaard, Koenraad De Smedt, Darja Fišer, and Dieter Van Uytvanck. 2018. CLARIN: Towards FAIR and Responsible Data Science Using Language Resources. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 3259–3264.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. In *LREC*, pages 759–765.

Dirk Goldhahn, Maciej Sumalvico, and Uwe Quasthoff. 2016. Corpus Collection for Under-Resourced Languages with More than One Million Speakers. In *Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 67–73, Portorož.

Zeroual Imad and Lakhouaja Abdelhak. 2016. Adapting a decision tree based tagger for Arabic. In *proceedings of the International Conference on Information Technology for Organizations Development (IT4OD)*, pages 1–6. IEEE.

Miloš Jakubíček, Adam Kilgarriff, Vojtěch Kovář, Pavel Rychlý, and Vít Suchomel. 2013. The tenten corpus family. In *proceedings of the 7th International Corpus Linguistics Conference CL*, pages 125–127.

Vojtěch Kovář, Vít Baisa, and Miloš Jakubíček. 2016. Sketch Engine for bilingual lexicography. *International Journal of Lexicography*, 29(3):339–352.

Suleiman H. Mustafa. 2012. Word stemming for Arabic information retrieval: The case for simple light stemming. *Abhath Al-Yarmouk: Science & Engineering Series*, 21(1):2012.

Uwe Quasthoff, Dirk Goldhahn, and Thomas Eckart. 2014. Building large resources for text mining: The Leipzig Corpora Collection. In *Text Mining*, pages 3–24. Springer.

Motaz K. Saad and Wesam Ashour. 2010. Osac: Open source arabic corpora. In *proceeding of the 6th International Conference on Electrical and Computer Systems (EECS'10)*, volume 10.

Vít Suchomel and Jan Pomikálek. 2012. Efficient web crawling for large text corpora. In *proceedings of the seventh Web as Corpus Workshop (WAC7)*, pages 39–43.

Reem Swuaileh, Mucahid Kutlu, Nihal Fathima, Tamer Elsayed, and Matthew Lease. 2016. ArabicWeb16: A New Crawl for Today's Arabic Web. In *proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 673–676. ACM.

Wajdi Zaghouani. 2017. Critical survey of the freely available Arabic corpora. arXiv preprint arXiv:1702.07835.

Imad Zeroual and Abdelhak Lakhouaja. 2018a. Arabic Corpus Linguistics: Major Progress, but Still a Long Way to Go. In *Intelligent Natural Language Processing: Trends and Applications, Studies in Computational Intelligence*, pages 613–636. Springer, Cham.

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

# A Appendices

| **Region or country** | **Web-domain** | **Nb. of articles** |
|---|---|---|
| International | news.un.org<br>arabic.euronews.com<br>ara.reuters.com<br>namnewsnetwork.org<br>arabic.sputniknews.com | 693,629 |
| Middle-east | aljazeera.net<br>alarabiya.net | 366,211 |
| Algeria | djazairess.com | 588,514 |
| Australia | eltelegraph.com | 4,614 |
| Canada | arabnews24.ca<br>halacanada.ca | 30,135 |
| China | arabic.cctv.com | 1,365 |
| Egypt | alwatanalarabi.com | 85,351 |
| France | france24.com | 74,718 |
| Iran | alalam.ir | 344,011 |
| Iraq | iraqakhbar.com | 28,248 |
| Germany | dw.com | 117,261 |
| Jordon | sarayanews.com | 49,461 |
| Morocco | www.marocpress.com | 188,045 |
| Palestine | al-ayyam.ps | 81,495 |
| Qatar | raya.com | 8,986 |
| Russia | arabic.rt.com | 57,238 |
| Saudi Arabia | alwatan.com.sa | 1,512 |
| Sweden | alkompis.se | 33,790 |
| Syria | syria.news | 36542 |
| Tunisia | www.turess.com | 495,674 |
| Turkey | turkey-post.net<br>aa.com.tr | 76,638 |
| UAE | emaratalyoum.com | 25,081 |
| UK | bbc.com | 10,686 |
| USA | arabic.cnn.com | 113,557 |

Table 1: List of crawled web-domains