Building TALAA, a Free General and Categorized Arabic Corpus

Essma Selab[1] and Ahmed Guessoum[2]
[1]*Natural Language Processing and Machine Learning Research Group (TALAA), Université des Sciences et de la Technologie Houari Boumediene (USTHB), BP 32, El Alia 16111 Bab Ezzouar, Algiers, Algeria*
[2]*Laboratory for Research in Artificial Intelligence (LRIA), Université des Sciences et de la Technologie Houari Boumediene (USTHB), BP 32, El Alia 16111 Bab Ezzouar, Algiers, Algeria*

Keywords: Corpora, Arabic Natural Language Processing, Corpus Metrics.

Abstract: Arabic natural language processing (ANLP) has gained increasing interest over the last decade. However, the development of ANLP tools depends on the availability of large corpora. It turns out unfortunately that the scientific community has a deficit in large and varied Arabic corpora, especially ones that are freely accessible. With the Internet continuing its exponential growth, Arabic Internet content has also been following the trend, yielding large amounts of textual data available through different Arabic websites. This paper describes the TALAA corpus, a voluminous general Arabic corpus, built from daily Arabic newspaper websites. The corpus is a collection of more than 14 million words with 15,891,729 tokens contained in 57,827 different articles. A part of the TALAA corpus has been tagged to construct an annotated Arabic corpus of about 7000 tokens, the POS-tagger used containing a set of 58 detailed tags. The annotated corpus was manually checked by two human experts. The methodology used to construct TALAA is presented and various metrics are applied to it, showing the usefulness of the corpus. The corpus can be made available to the scientific community upon authorisation.

1 INTRODUCTION

Arabic is a Semitic language that has been in use since the 2nd millennium BC. It is today the language of about 350 million people and is used by one billion six hundred million Muslims. Classical Arabic is the language of the Qur'an, the holy book of Islam, and other religious literature, while Modern Standard Arabic (MSA) or "Fus'ha" is the formal Arabic used in the literature and media (Habash, 2010).

Arabic is the 4th language used on the Internet, with more than 135 million Arabic speaking users online. Recent statistics have registered the highest Internet use growth rate (5,296.6%) for users of Arab over the period 2000-2013 compared to 132.9% for Japanese, 1,910.3% for Chinese and 468.8% for English (Miniwatts Marketing Group, 2014).

Arabic natural language processing (ANLP) has gained increasing interest. Various approaches have been used over the last ten to fifteen years to develop several ANLP tools. Some are rule-based, while others are statistical or machine-learning-based. A number of tools are commercial, while others were implemented by researchers for the needs of the scientific community. But, unlike the English language, Arabic still lacks NLP tools that can cover the various applications with high quality, except for a few cases (Al-Taani et al., 2012; Shaalan et al., 1999).

Given the techniques used, the development and the quality of NLP tools are nowadays largely based on the availability of voluminous corpora. These can indeed be used for the analysis of the language sentences in large quantity and sufficient variations in order to attest the richness of the language (Véronis, 2001; Rastier, 2005), but also for the purposes of linguistic investigations of the language. Unfortunately, the scientific community has a deficit in large and varied Arabic corpora that are freely accessible (Marton et al., 2013; Othman et al., 2003).

Due to the availability of large amounts of Arabic data and unstructured information on the Internet, we have decided to use these electronic resources to build our Arabic corpus, TALAA. We present in this paper the methodology used to automatically collect and structure Arabic texts from daily Arabic newspaper websites. We also show the

Selab E. and Guessoum A..
Building TALAA, a Free General and Categorized Arabic Corpus.
DOI: 10.5220/0005352102840291
In *Proceedings of the International Conference on Agents and Artificial Intelligence* (PUaNLP-2015), pages 284-291
ISBN: 978-989-758-073-4

process used to annotate, structure and validate our corpus.

In Section 2, we present some of the research effort related to corpus building. The process of data collection, annotation, validation and structuring is presented in Section 3. Statistics about the corpus are given in Section 4 with an attempt to show its usefulness. The conclusion is given in Section 5.

2 RELATED WORK

Over the last decade, various corpora have been built, but most of them are used for commercial purposes or are not sufficiently large to represent the Arabic language.

Raw text corpora consist of a collection of texts with no added information such as tagging, parsing, etc. This kind of corpora is divided into 1) monolingual corpora, 2) parallel corpora, and 3) dialectal Corpora. The European Language Resources association (ELRA) (ELRA, 2008) provides more than 83 Arabic corpora in several categories (monolingual, multilingual, speech, annotated, etc.) such as An-Nahar Corpus (An-Nahar Corpus, 2014), an Arabic corpus collected from the Lebanese newspaper in the period between 1995 and 2000 and stored in HTML files. This corpus contains 45000 articles consisting of 24 million words for each year. The Al Hayat corpus (Al Hayat corpus, 2014) is another written corpus collected form Al-Hayat newspaper. It was developed at Essex University and covers articles from 1998. The Al Hayat corpus contains more than 18 M distinct tokens and 42,591 articles distributed into 7 domains (all punctuations and special characters having been removed). Unfortunately the corpora available on ELRA are not free.

Rafalovitch and Dale (2009) present a free parallel corpus available online (Parallel Corpus, 2014) that contains a collection of 2100 United Nations General Assembly Resolution documents with their parallel translations in the six UN official languages (Arabic, Chinese, English, French, Russian, and Spanish). The corpus contains about 3M tokens per language. Al-Sulaiti (2004), from the university of Leeds, developed a Contemporary Arabic free corpus (Contemporary corpus, 2014) in which the articles are categorized into different topics. The corpus contains written and spoken data of 1 million words. Graff and Walker (2003), from the University of Pennsylvania LDC, developed Arabic Gigaword, a written corpus built from texts taken from Agence France Press, Al Hayat Newspaper, Al Nahar Newspaper and Xinhua News Agency. The size of the corpus is approximately 1.1GB in compressed form and contains 391,619 tokens. Arabic Gigaword is available from the Linguistic Data Consortium, but it is not free. Alrabiah et al. (2013) built KSUCCA King Saud University Corpus of Classical Arabic, which contains over 50 Million words from classical Arabic. The corpus was developed as part of the PhD work on building a distributional lexical semantic model for classical Arabic, and investigating its applications to The Holy Quran. KSUCCA corpus can be used in several Arabic linguistic and computational linguistic researches. Almeman and Lee (2013) built an Arabic multi-dialect (Gulf, Levantine, Egyptian and North African) text corpus from web resources. The corpus contains 48M tokens.

Annotated corpora include POS-tagged corpora, parsed corpora, semantically annotated corpora, etc. LDC (LDC, 2014) and ELRA provide a set of Arabic annotated corpora and parallel annotated corpora, which are unfortunately not free. Khoja (Khoja, 2001), from Lancaster University, built an annotated corpus that contains manually-tagged Arabic newspaper texts. The first collection includes 50,000 tagged words using general tags (noun, verb, particle, number). The second contains 1,700 tagged words with more detailed tags (tense, gender, number, etc.). American and Qatari Modeling of Arabic (AQMAR) Wikipedia Dependency Corpus (AQMAR, 2014) is a hand-annotated corpus. The POS tagging and dependency parse information were collected from Arabic Wikipedia articles, consisting of 1262 sentences and more than 36,202 tokens. The corpus was developed as part of the AQMAR project.

3 DATA PREPARATION

The process of development of the TALAA corpus was divided into two main steps as presented below: 1) Data collection and 2) Data pre-processing.

3.1 Data Collection

The methodology used to build and structure the Arabic corpus consisted in developing an automatic system, a robot, to collect Arabic newspaper articles from different websites (see Table 1). Figure 1 presents the process used to extract and organize the data from the websites.

Table 1: Newspaper collection description.

Newspaper collection	Url	Country
Al Djazeera	www.aljazeera.net	Qatar
Al Ahram	www.ahram.org.eg	Egypt
El Khabar	www.elkhabar.com	Algeria
Al Sharq al Awsat	www.aawsat.com	U.K.
Al Bayan	www.albayan.ae	United Arab Emirates
Al Qabas	www.alqabas.com.kw	Kuwait
Al Arabiya	www.alarabiya.net	United Arab Emirates
Al Hayat	www.alhayat.com	Lebanon
An Nahar	www.annahar.com	Lebanon

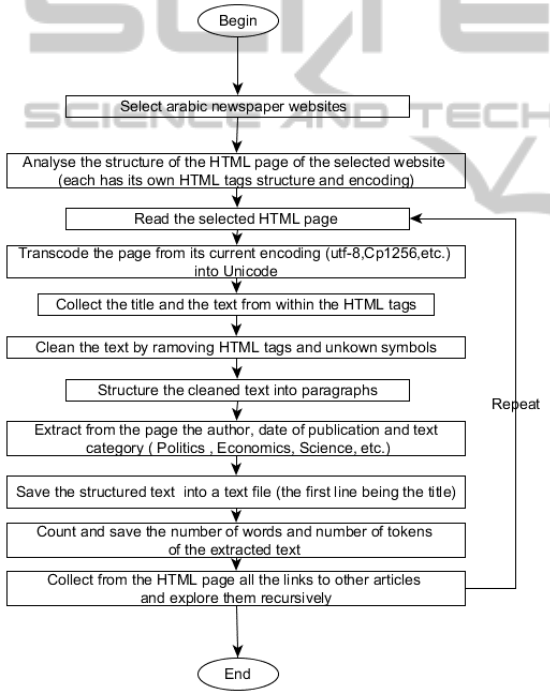

Figure 1: The process of TALAA corpus building.

First we select a daily Arabic newspaper website from which we want to extract our data. We then analyse the structure of the HTML pages of the selected newspaper website since each e-newspaper has its own HTML tag structure and its own encoding (utf-8, Cp1256, etc.).

Next, the robot reads the HTML pages of the selected newspaper, transcodes them from their current encoding to UNICODE and extracts the Arabic text and the article properties (author, publication date, type, etc.), all of which are contained within specific pre-identified HTML tags.

For each page, the extracted text is cleaned of unknown symbols and tags, structured into paragraphs and finally saved into a text file the first line of which is the newspaper article title.

The following general Syntax is used to name any created file:

Name of the collection_category article serial number _publishing date

Example from the Algerian daily newspaper (El-Khabar, 2010):

KH_PO1_2014208: Collection El-Khabar, article number 1 of type politics published on August 24th, 2014.

The files are saved into different directories according to the articles categories (Politics, Economics, Science, etc.). On the one hand, this helps broaden the coverage of the corpus and, on the other hand, it makes the built corpus useful for various purposes, such as text categorization and other ANLP applications. The number of tokens and words is also calculated and saved.

The process is repeated by having the robot follow the various links found on any page to extract the pages pointed to.

The robot has been programmed to run continuously so as to collect Arabic newspaper archives that are as large as possible.

3.2 Data Pre-processing

In order to refine and structure the data, the following data pre-processing steps have been performed:

a. Segmentation: each collected article was segmented into sentences. The sentence length varies from 2 to 25 words.

b. Pos-tagging of the sentences: To extract different features from our sentences, we have used the POS-tagger used in the SAIE "Statistical Arabic Information Extraction" system, a System for Arabic Named Entity extraction using statistical language models in the form of Hidden Markov Models. The SAIE architecture consists in a NLP module: a tokenizer, the Buckwalter stemmer (Buckwalter, 2002), and an HMM-based POS tagger of Arabic text, along with an NE extraction module. LDC's Arabic Treebank (Maamouri et al., 2005) was used in the training step of SAIE. The latter uses a POS-tag set of 58 tags and was reported to have a 97% F-measure (Al Shamsi and Guessoum, 2006).

c. Data Validation: Despite the fact that (Al Shamsi and Guessoum, 2006) reported a 97% F-

measure for the SAIE POS-Tagger, this measure also implies that the output of the tagging needs to be manually checked and corrected. The TALAA corpus having been POS-tagged by using SAIE, the TALAA corpus was manually checked by two human experts in the Arabic language. They had to carefully check the semantics of every sentence and validate/correct the POS tags.

Table 2: POS-Tag set of the SAIE POS-Tagger.

ADJ	CONJ	EXCEPT	PRON_2S
PRON_2MP	DPRON_FS	FUNC_WORD	SUFF_SUBJ_MP
CVERB	DEF	FUTURE	IV2
DPRON_MD	DPRON_MP	INTERROGATE	SUFF_SUBJ_2FP
SUFF_SUBJ_ALL	SHORT_FORM	DPRON_FP	SUFF_SUBJ_2MP
SUFF_SUBJ_2D	PPRON_2FP	DPRON_FD	SUFF_SUBJ_2S
IV3	MOOD_SJ	PRON_2	SUFF_F_D
IVERB	PVERB	PRON_2D	SUFF_M_P
Num	PREP	PRON_2FP	NEGATION
NOUN	PRON	IV1P	PRON_3MP
MOOD_I	PRON_1P	PRON_3D	DPRON_F
PRON_3FP	PRON_1S	PNOUN	PRON_3FS
SUFF_SUBJ_1P	SUFF_SUBJ_3FD	PRON_3MS	SUFF_S_INDEF
SUFF_SUBJ_FP	DPRON_MS	SUFF_M_D	PPRON_3FP
SUFF_F_S	SUFF_F_P		

d. Corpus structuring using XML: XML (eXtensible Markup Language) is a meta-markup language used to represent and structure data in a textual document (Cunningham, 2005). Today XML is considered the suitable data exchange format. Its representation is used by Microsoft Office (Office Open XML), OpenOffice.org and LibreOffice (OpenDocument), and Apple's iWork. Figure 2 shows how our data is structured in XML format.

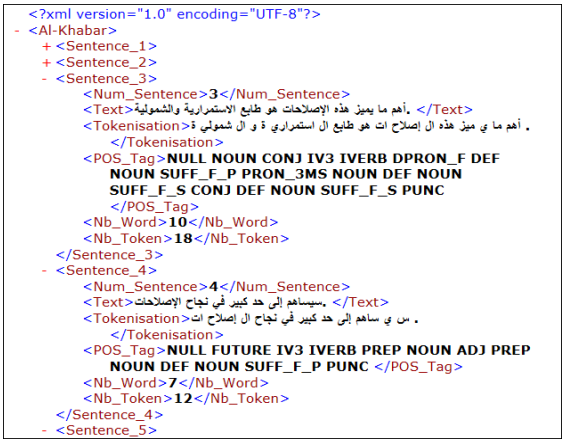

Figure 2: Structure of an XML file in TALAA.

Every top parent node of the XML file represents a collection from the database, the elements of the collection being the sentences extracted from this collection and having the following attributes:

- Num_sentence: sentence number in the database.
- Text : input sentence.
- Tokenisation: the sentence after the tokenisation step.
- POS_Tag: Pos-tagging of the sentence using the SAIE Tagger.
- Nb_words: number of words in the sentence.
- NB_Tokens: number of tokens in the sentence.

4 DESCRIPTION OF THE TALAA CORPUS

In this section, we present some corpus statistics to assess corpus quality as proposed by (Biemann et al., 2013). These are: size of the data; empirical law; distribution of word, sentence and document length; and distribution of characters, words, n-grams, etc.

4.1 Size of the Corpus

The methodology and the process followed to develop the TALAA corpus have enabled us to build a large and varied Arabic corpus. It is a collection of 57,827 articles published in newspaper websites during the 5-year period 2010 to 2014. The articles were taken from eight different categories as shown in Table 4. Figure 3 presents the number of articles present in each collection category. The TALAA corpus contains (so far) 14,068,407 words and 582,531 types (Table 3).

Researchers can use the TALAA corpus as an entire raw collection, as a set of categorised raw collections (politics, economics, sports, etc.), or as a collection of Arabic sentences, Pos-tagged and structured into XML format (7000 tokens).

Table 3: Description of the TALAA corpus.

Features	Corpora
Number of articles	57,827
Number of categories	8
Number of words	14,068,407
Number of types	582,531
Number of tokens	15,891,729
Tagged and validated tokens	7000

Table 4: Corpus categories.

Category	Number of articles
Culture	5322
Economics	8768
Politics	9620
Religion	4526
Society	9744
Sports	9103
World	6344
Other	4400

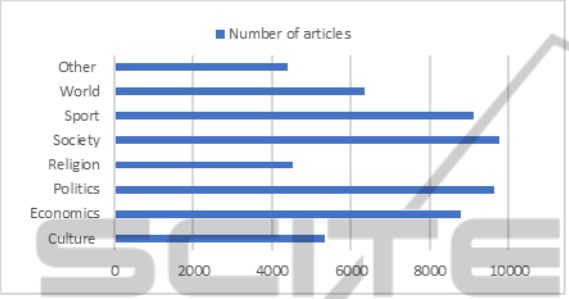

Figure 3: Number of articles per category.

4.2 Variety of the TALAA Corpus

In order to study the impact of the number of documents of the TALAA corpus on the size and the diversity of the corpus, we have calculated the number of words and the number of types added by each document. From Table 5 and Figure 4, we can see that the number of documents contributes to the variety of our corpus since the number of distinct words increases in relation to the number of documents, which is as expected. However, Figure 4 gives a more precise dependence co-relation between the number of distinct words (diversity) as a function of the number of documents. It shows that the word diversity significantly increases beyond 2000 documents.

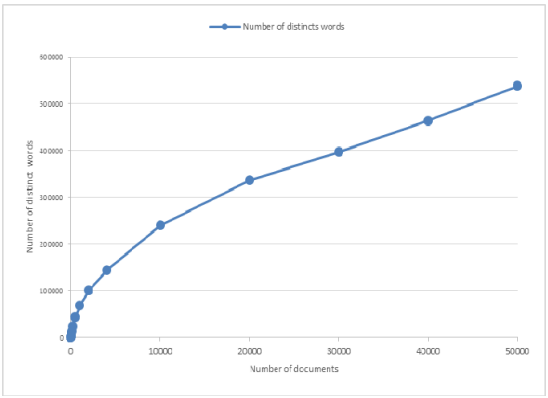

Figure 4: Representation of the number of distinct words in relation to the number of documents.

Table 5: Impact of the number of document in the corpus.

Number of documents	Number of words	Number of distinct words
1	973	584
2	2236	1210
3	3395	1735
4	4717	2222
5	5890	2662
10	7321	3247
25	10788	4830
50	17909	7881
100	36847	14068
200	71433	23755
500	174676	44108
1000	342666	68943
2000	612791	101192
4000	1119603	143666
10000	2745939	240491
20000	5778160	336914
30000	8787350	397609
40000	10222218	465093
50000	12138321	538418

4.3 Modeling the Distribution of Terms using Zipf's Law

Zipf's law (Zipf, 1949) describes the relationship between the frequency distribution of a word and its rank in a corpus that represents a language. Zipf's law is an empirical law based on an observation which states that the frequency distribution of any word in a corpus is inversely proportional to its rank. Zipf's law does not care about the words, but about their rank and frequency only. It is given in (1)

$$r * Prob(r) = A \quad (1)$$

$$Prob(r) = Freq(r) / N \quad (2)$$

(1) and (2) imply (3)

$$r * Freq(r) = C \quad (3)$$

where r is the rank of words in descending order with respect to their frequency (the most frequent word having rank 1); Freq(r) is the frequency of the word at rank r; Prob(r) is the probability of a word at rank r; N is the number of words in the corpus; A and C are constants.

(3) states that, for any word in the corpus, computing *r * Freq(r)* gives a constant. Zipf's law is not an exact law and can be formulated as (4)

$$r^{\alpha} * Freq(r) = C \ / \ \alpha, \ C \text{ are constants and } \alpha \text{ is a close to 1} \quad (4)$$

Table 6 below gives the top 50 most frequent words in the corpus. We find that the most frequent words in TALAA are conjunction particles and prepositions as well as a few specific words (see the words in bold in Table 6).

In order to verify that the quality of the TALAA corpus follows zipf's law, we plot in Figure 5 the frequency of terms as a function of the rank, in a log-log graph. For comparison purposes, Figure 5 also contains the ideal Zipf's law of TALAA, which is a straight line with slope -1. We thus conclude that the TALAA corpus follows Zipf's law.

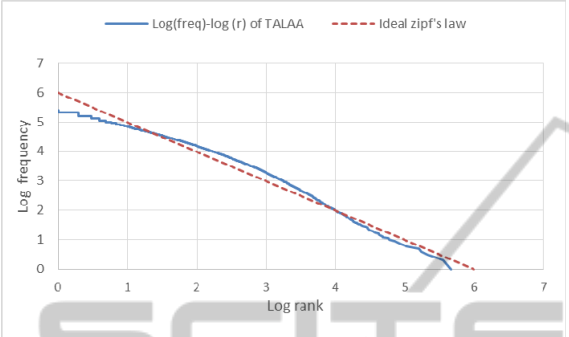

Figure 5: The log-log graph of the frequency and ideal zipf's law of the TALAA corpus.

4.4 Punctuation

Figure 6 shows the frequency distribution of the punctuation: comma, colon, dot, exclamation and question marks in the TALAA corpus. The frequent use of punctuation is another indicator of the corpus quality (Felice, 2012).

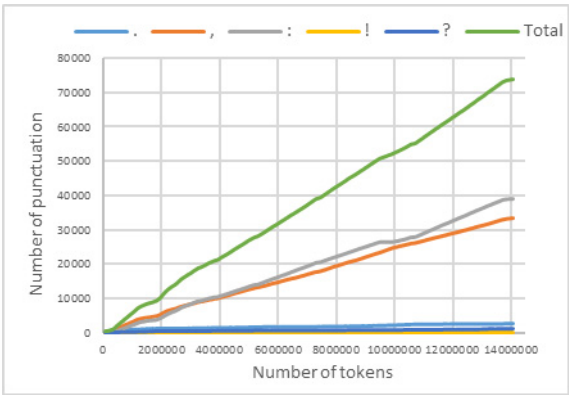

Figure 6: Distribution of punctuation in terms of the number of tokens in the TALAA corpus.

4.5 Length of Sentences

Figure 7 represents the distribution of the sentences according to their length: number of words per sentence and number of tokens per sentence.

The mode of the sentence length for the TALAA corpus is 26 words and the average sentence length is 49 words. The mode of sentence length in letters for TALAA corpus is 286 letters per sentence and the average is 477.6 letters per sentence.

Table 6: The top 50 most frequent words in the TALAA corpus.

Rank	Word	Freq(r)	Prob (%)	r * Prob
1	في	466045	3,313	0,0331
2	من	355626	2,528	0,0506
3	على	215023	1,528	0,0680
4	أن	171898	1,222	0,0854
5	إلى	158860	1,129	0,1028
6	التي	107959	0,767	0,1203
7	عن	88549	0,629	0,1377
8	الذي	79681	0,566	0,1551
9	ما	70826	0,503	0,1726
10	مع	59369	0,422	0,1900
11	لا	57945	0,412	0,2074
12	بعد	48337	0,344	0,2249
13	**الجزائر**	48168	0,342	0,2423
14	هذا	44095	0,313	0,2597
15	هذه	42464	0,302	0,2771
16	حيث	42122	0,299	0,2946
17	بين	38107	0,271	0,3120
18	لم	35934	0,255	0,3294
19	**أمس**	35093	0,249	0,3469
20	**الخبر**	33114	0,235	0,3643
21	كان	30773	0,219	0,3817
22	خلال	30513	0,217	0,3991
23	أو	29585	0,210	0,4166
24	كل	28937	0,206	0,4340
25	كما	28647	0,204	0,4514
26	قبل	28430	0,202	0,4689
27	**الوطني**	27253	0,194	0,4863
28	**رئيس**	27072	0,192	0,5037
29	إن	24639	0,175	0,5212
30	أنه	24304	0,173	0,5386
31	بأن	23518	0,167	0,5560
32	غير	22546	0,160	0,5734
33	ذلك	21550	0,153	0,5909
34	**الله**	21279	0,151	0,6083
35	و	20696	0,147	0,6257
36	أمام	20533	0,146	0,6432
37	منذ	19362	0,138	0,6606
38	عبد	19229	0,137	0,6780
39	كانت	18954	0,135	0,6955
40	**سنة**	18318	0,130	0,7129
41	أي	18225	0,130	0,7303
42	**العام**	18095	0,129	0,7477
43	الذين	18004	0,128	0,7652
44	هو	17802	0,127	0,7826
45	عليه	17727	0,126	0,8000
46	قد	17622	0,125	0,8175
47	**الجزائرية**	17014	0,121	0,8349
48	**الرئيس**	16863	0,120	0,8523
49	**خاصة**	16548	0,118	0,8697
50	**قال**	16296	0,116	0,8872

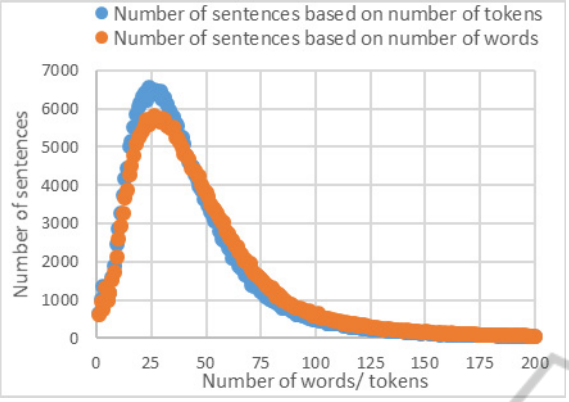

Figure 7: Distribution of sentences according to their length (number of words/ tokens) in the TALAA corpus.

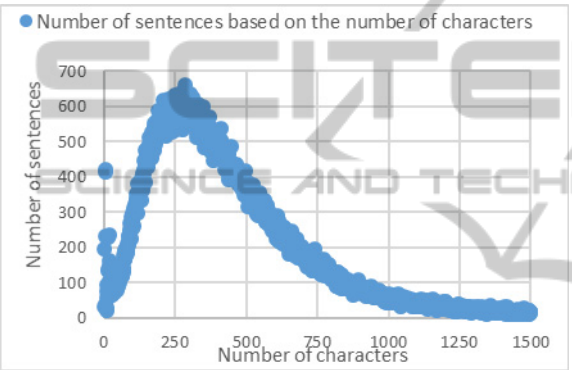

Figure 8: Distribution of sentences according to the number of letters.

4.6 Arabic Words

Since Arabic is an agglutinative language, a word can be formed by joining affixes morphemes together. Figure 9 shows the distribution of the Arabic words according to their length in TALAA corpus.

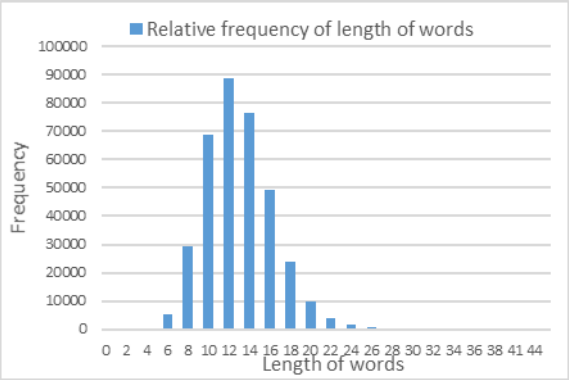

Figure 9: Distribution of words according to their length in TALAA corpus.

The mode of the word length in the corpus is 12 letters and the average length of words is 13.02 letters. The words length distribution in the corpus is as follows: 5.75% of the words have between 0 and 8 letters, 60.33% between 9 and 15 letters and 33.88 % between 16 and 23 letters. 0.04 % of the Arabic words in TALAA contain more than 23 letters; they have turned out to be concatenations of several words where the space was omitted. These words were checked and corrected.

5 CONCLUSIONS

In this paper we have presented the methodology we have followed to exploit electronic resources to build the TALAA corpus, a large and varied general Arabic corpus.

The robot that we have implemented has helped us construct a collection of more than 14 million words (582,531 types). The corpus contains 57,827 articles and 15,891,729 tokens. An XML file was structured to contain 7000 tagged tokens which have been manually checked and corrected by two human experts.

As future work, we intend to use the TALAA corpus in the development of a grammatical induction module for Arabic. The corpus being rich, it will also be used to improve the current accuracy of the Arabic parsers.

REFERENCES

Al Hayat corpus, Catalogue of Language Resources, http://catalog.elra.info/product_info.php?products_id=632&language=fr (Last visited September 2014).

Almeman, K., and Lee, M., 2013. Automatic building of Arabic multi dialect text corpora by bootstrapping dialect words. In *Communications, Signal Processing, and their Applications (ICCSPA)*. Sharjah.

Alrabiah, M., Alsalman, A. M., and Atwell, E., 2013. KSUCCA a cornerstone to study the semantics of the Quranic words in the light of distributional lexical semantics, *NOORIC'1435-2013*, Almadinah Almonawwrah, Saudi Arabia.

Al Shamsi, F., and Guessoum, A., 2006. A Hidden Markov Model - Based POS Tagger for Arabic. In: *Proceeding of the 8th International Conference on the Statistical Analysis of Textual Data*, France, pp.31-42.

Al-Sulaiti, L., 2004. Designing and Developing a Corpus of Contemporary Arabic. University of Leeds, UK. *MSc Thesis*.

An-Nahar Corpus, Catalogue of Language Resources, http://catalog.elra.info/search_result.php?keywords=W0027&language=en (Last visited September 2014).

AQMAR, http://www.ark.cs.cmu.edu/ArabicDeps/, (Last visited November 2014).

Biemann, C., Bildhauer, F., Evert, S., Goldhahn, D., Quasthoff, U., Schäfer, R and Swiezinski, L., 2013. Scalable Construction of High-Quality Web Corpora. *Journal for Language Technology and Computational Linguistics*, vol. 28(2),pp. 23–59.

Buckwalter, T., 2002. "Buckwalter Arabic morphological analyzer version 1.0". LDC Catalog No: LDC2002L49. *Linguistic Data Consortium*, University of Pennsylvania.

Contemporary corpus, http://www.comp.leeds.ac.uk/eric/latifa/research.htm (Last visited September 2014).

Cunningham, L. A., 2005. Language, Deals and Standards: The Future of XML Contracts. Washington University Law Review, Boston College Law School *Research Paper* No. 93.

Diab, M., Hacioglu, K. and Jurafsky, D., 2004. Automatic Tagging of Arabic Text: From Raw Text to Base Phrase Chunks. In: *Proceedings of Human Language Technology conference / North American chapter of the Association for Computational Linguistics annual meeting*, USA, pp. 149–152.

El-Khabar, 2010. El-khabar newspaper online, Available at: www.elkhabar.com/ar/ (Last visited November 2014).

European Language Resources Association (ELRA), 2008, Catalogue of Language Resources [En ligne] // http://catalog.elra.info/index.php. (Last visited October 2014).

Felice, M., 2012. Linguistic Indicators for Quality Estimation of Machine Translations. *MSc Thesis in Natural Language Processing and Human Language Technology*. University of Barcelona.

Graff, D., 2003. Arabic Gigaword LDC2003T12. Web Download. Philadelphia: *Linguistic Data Consortium*.

Habash, N. Y., 2010. Introduction to Arabic Natural Language Processing. Columbia University: *A Publication* in the Morgan and Claypool Publishers series.

Khoja, S., 2001. APT: Arabic Part-of-speech Tagger. *Proceedings of the Student Workshop at the Second Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL2001)*, Carnegie Mellon University, Pittsburgh, Pennsylvania.

LDC, Linguistic Data Consortium-University of Pennsylvania, http://www.ldc.upenn.edu/. (Last visited October 2014).

Maamouri, M., Bies, A., Buckwalter, T., Jin, H. and Mekki, W., 2005. Arabic Treebank: Part 3 (full corpus) v 2.0 (MPG + Syntactic Analysis).

Manning, C. D, and Schütze, H., 1999. Foundations of Statistical Natural Language Processing, *MIT Press*, ISBN 978-0-262-13360-9, pp. 24.

Marton, Y., Habash, N. and Rambow, O., 2013. Dependency Parsing Of Modern Standard Arabic With Lexical And Inflectional Features. Computational Linguistics, 39(1).

Miniwatts Marketing Group, 2014. Internet World Stats. Available at: http://www.internetworldstats.com/stats7.htm, updated on Sept 18, 2014. (Last visited October 2014).

Parallel Corpus, United Nations General Assembly Resolutions: A Six-Language Parallel Corpus. http://www.uncorpora.org/.(Last visited October 2014).

Rafalovitch, A., and Dale, H., 2009. United Nations General Assembly Resolutions: A Six-Language Parallel Corpus. In: *Proceedings of the MT Summit XII*. Ottawa, Canada. pp. 292-299.

Rastier, F., 2005. Enjeux épistémologiques de la linguistique de corpus. *Williams*, pp. 31-46.

Véronis, J., 2001. Sense tagging: does it make sense? , Corpus Linguistics'2001 Conference, Lancaster, U.K.

Zipf, G. K., 1949. Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology. Cambridge, *MA: Addison-Wesley*.