

Article

Associative Root–Pattern Data and Distribution in Arabic Morphology

Bassam Haddad ^{1,*} , Ahmad Awwad ¹, Mamoun Hattab ²  and Ammar Hattab ³

¹ Department of Computer Science, University of Petra, 11196 Amman, Jordan; awwad@uop.edu.jo

² Arabic Textware, 11181 Amman, Jordan; mamoun.h@texum-me.com

³ Brown University, Providence, RI 02912, USA; ammar_hattab@brown.edu

* Correspondence: haddad@uop.edu.jo; Tel.: +9626-579-9555 (ext. 1840)

Received: 28 January 2018; Accepted: 26 March 2018; Published: 29 March 2018



Abstract: This paper intends to present a large-scale dataset for Arabic morphology from a cognitive point of view considering the uniqueness of the root–pattern phenomenon. The center of attention is focused on studying this singularity in terms of estimating associative relationships between roots as a higher level of abstraction for words meaning, and all their potential occurrences with multiple morpho-phonetic patterns. A major advantage of this approach resides in providing a novel balanced large-scale language resource, which can be viewed as an instantiated global root–pattern network consisting of roots, patterns, stems, and particles, estimated statistically for studying the morpho-phonetic level of cognition of Arabic. In this context, this paper asserts that balanced root-distribution is an additional significant key criterion for evaluating topic coverage in an Arabic corpus. Furthermore, some additional novel probabilistic morpho-phonetic measures and their distribution have been estimated in the form of root and pattern entropies besides bi-directional conditional probabilities of bi-grams of stems, roots, and particles. Around 29.2 million webpages of ClueWeb were extracted, filtered from non-Arabic texts, and converted into a large textual dataset containing around 11.5 billion word forms and 9.3 million associative relationships. As this dataset is predominantly considering the root–pattern phenomenon in Semitic languages, the acquired data might be significant support for researchers interested in studying phenomena of Arabic such as visual word cognition, morpho-phonetic perception, morphological analysis, and cognitively motivated query expansion, spell-checking, and information retrieval. Furthermore, based on data distribution and frequencies, constructing balanced corpora will be easier.

Keywords: cognitive linguistics; Arabic morphology; corpus linguistics; associative root–pattern network; bi-gram analysis

1. Introduction and Motivation

Language modeling based on text corpora is of essential importance for natural language processing (NLP) and corpus linguistics. Since such corpora provide researchers with resources for performing different computational tasks in many fields of informatics and computational linguistics, such as machine learning and text mining, cognitive info communications [1], information retrieval, and others.

On the other hand, despite the importance of this fact for Arabic as a key language¹, it still lacks sufficient resources in this field, particularly in corpora devoted to cognitive aspects, compared to

¹ Arabic is an official language of 27 states, the third most prevalent after English and French, spoken by 420 million speakers, and it is one of six official languages of the United Nations [2].

other Indo-European languages. Reasons for this shortage might reside in the poor overall awareness of this field in the Arabic NLP community, besides the fact that building a corpus is costly and time-consuming [3].

However, in the last few years, there have been many reports on the release of new, free Arabic corpora. According to the survey conducted by the authors of [4], there are 66 freely available corpora with different categories, such as raw texts, annotated, speech, scanned and annotated documents, questions/answers, summaries corpora, and lexicons. The LDC (Linguistic Data Consortium) catalog shows the availability of 116 corpora, and the ELDA (Evaluations and Language Resources Distribution Agency) provides 80 corpora of various types. For comparison, there is even an earlier study [5], a report of 19 Arabic corpora ranging in size from 1–3 billion words. In spite of that, Arabic is still poor in resources in terms of quality and availability [6,7]. By contrast to the current available corpora, our approach emerges from a different perspective of building the APRoPAT² Corpus in the following senses:

- Statistical language modeling requires a large-scale corpus to statistically consider the implicit, inherent differences in a language.
- The cognitive aspects and abstraction of a language should be as accurate as possible. The present research concerned with cognitive word identification and recognition for Semitic languages is steadily providing evidence of being root–pattern-based during visual and textual word processing [8,9].

For considering the first aspect we have adopted a strategy to rely on a big data web collection gathered by ClueWeb09 within the Lemur Project at Carnegie Mellon University³. Moreover, the ClueWeb datasets are used by several tracks of the TREC conferences and were created to support research on information retrieval and related human language technologies.

The cognitive aspect has been considered by adopting the cognitive and modular aspects of the APRoPAT Language Model for Arabic [10,11] (see Figure 1). In this context, the focus has been on statistically studying the singularity of the root–pattern phenomenon. A key benefit of this approach resides in providing researchers with a novel large-scale language resource, which can be considered a global associative network considering roots, patterns, stems, and particles estimated statistically. We believe such resources are statistically still neglected in Arabic natural language processing (Arabic NLP) research for investigating the morpho-phonetic level of cognition for Arabic as an abstract level in terms of word cognition and perception⁴.

To the best of our knowledge, most Arabic corpora techniques rely on approaches dealing with linear or concatenative morphology, which is predominantly centered on term frequency, co-occurrence, and part of speech (POS) tagging on the word level of analysis. Our adopted approach is based on a *higher level of abstraction*, in the sense that *associative relationships* between *abstract representations* of words in the form of roots, patterns, stems, and particles are investigated rather than counting direct term co-occurrences. More specifically:

- The cognitive and statistical dimension of the non-linearity of the Arabic morphology has been considered in the form of handling a phonetic pattern as a statistical variable and not relying only on the co-occurrence of its instantiated word forms. For example, a root might be

² Stands for “Arabic Associative Probabilistic RooT PATtern Model.”

³ <http://www.lemurproject.org/lemur.php>

⁴ In cognitive science, the human language can be divided into different linguistic levels or strata, such as phonology, morphology, syntax, and semantics. For example, the Cognitive Grammar Theory relies merely on *two levels*, namely *phonology* and *semantics*. A grammar is, in this context, meaningful and does not represent an autonomous level, and lexicon and grammar form a continuum consisting in assemblies of such structures [12,13]. Construction-grammar-based theories recognize constructions mediating the two levels (phonetic and semantic), such as morphology and syntax [12]. Arabic and Semitic word cognition seems to handle the constructions in a unique way. Phonology and morphology of Arabic words are *non-concatenative* or, more precisely, are predominantly *non-linear* [8,9]

incorporated with a large number of patterns delivering multiple word syntactical forms with different meanings. This analysis is rather intended to provide researchers with basic and initial predictive values based on the association between a root and a pattern.

- A term is considered as an applicative function instantiating a pattern for some root [11] unless it is not derivable; i.e., if it cannot be reduced to a known root.
- Operating on root–pattern, root–root, root–stem, root–particle, and pattern–pattern associative probabilistic relationships represent a higher level of abstraction than computing co-occurrences of their instances. For example, root–root co-occurrence as an abstract associative relationship implies multiple term–term co-occurrences⁵.

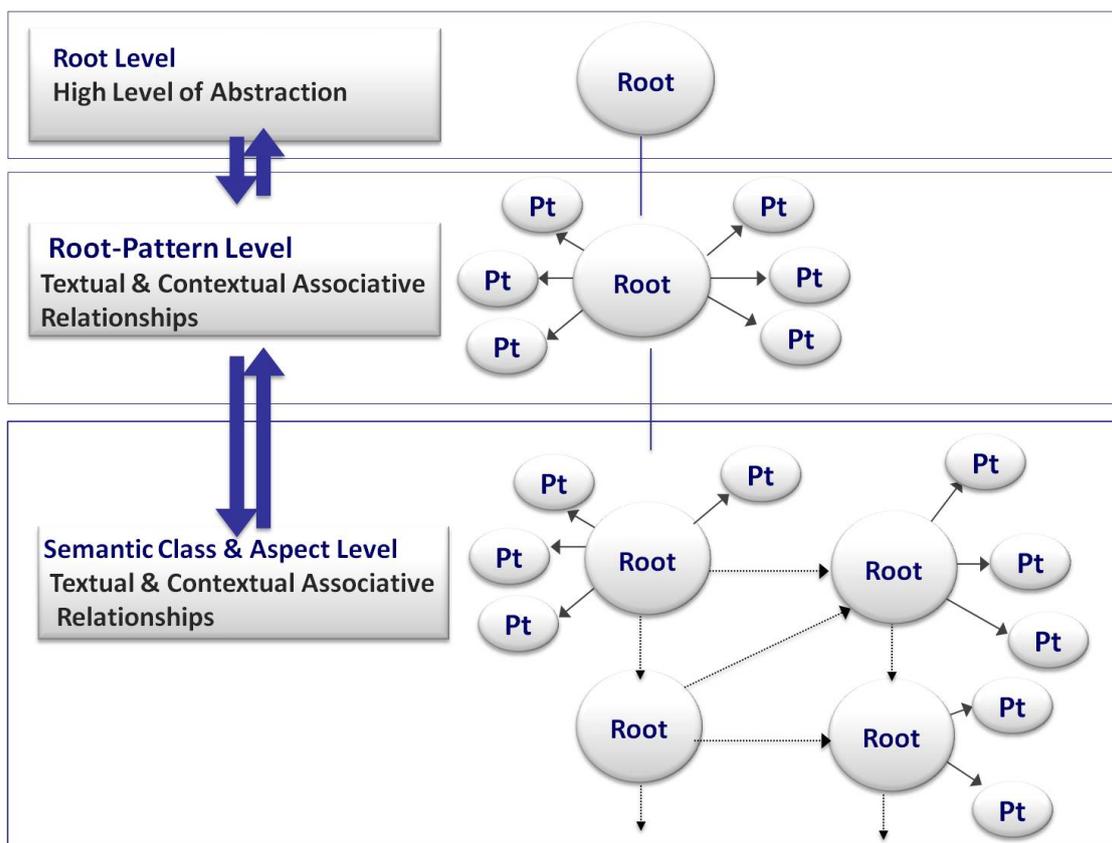


Figure 1. Levels of Abstraction in the APRoPAT Model in Context Bi-Directional Root-Patten Associative Relationships: Root and Root-Textual-Contextual Associative Relationships Levels [10].

1.1. A Preliminary Note on Non-Linear and Cognitive Aspects of the Arabic Morphology

The phonology and morphology of Arabic words are predominantly *non-linear*; i.e., *non-concatenative*. The recent research concerned with cognitive linguistics of Semitic languages shows that Arabic has a strong root–pattern relationships. This topic is still problematic in the Arabic computational community.

According to [10,15], a word form consists of the following aspects:

⁵ In the APRoPAT Model, such relations are defined as binary associative relationships on the morpho-phonetic level of cognition. The degree of association is associated with some COgnitive Degree of Association Strength Function; i.e., $CODAS(\mathcal{X}, \mathcal{Y})$, where \mathcal{X} and \mathcal{Y} can be a ROOT, PATTERN, STEM, or a PARTICLE [14].

- A root level, which mostly consists of three consonants symbolizing the highest preeminent semantic abstraction and is unmodifiable. An associative set of roots conveys an abstract semantic description without *clear phonetic information*.
- A templatic pattern in the form of consonant–vowel arrangement picturing the morpho-phonetic structure of a word form completing possible phonetic, syntactic, and semantic information⁶.
- An associative root–pattern relationship. A templatic pattern can be perceived by establishing the most plausible associative relationship between a pattern and a root relying on a bi-directional cognitive process. (see Figure 2)

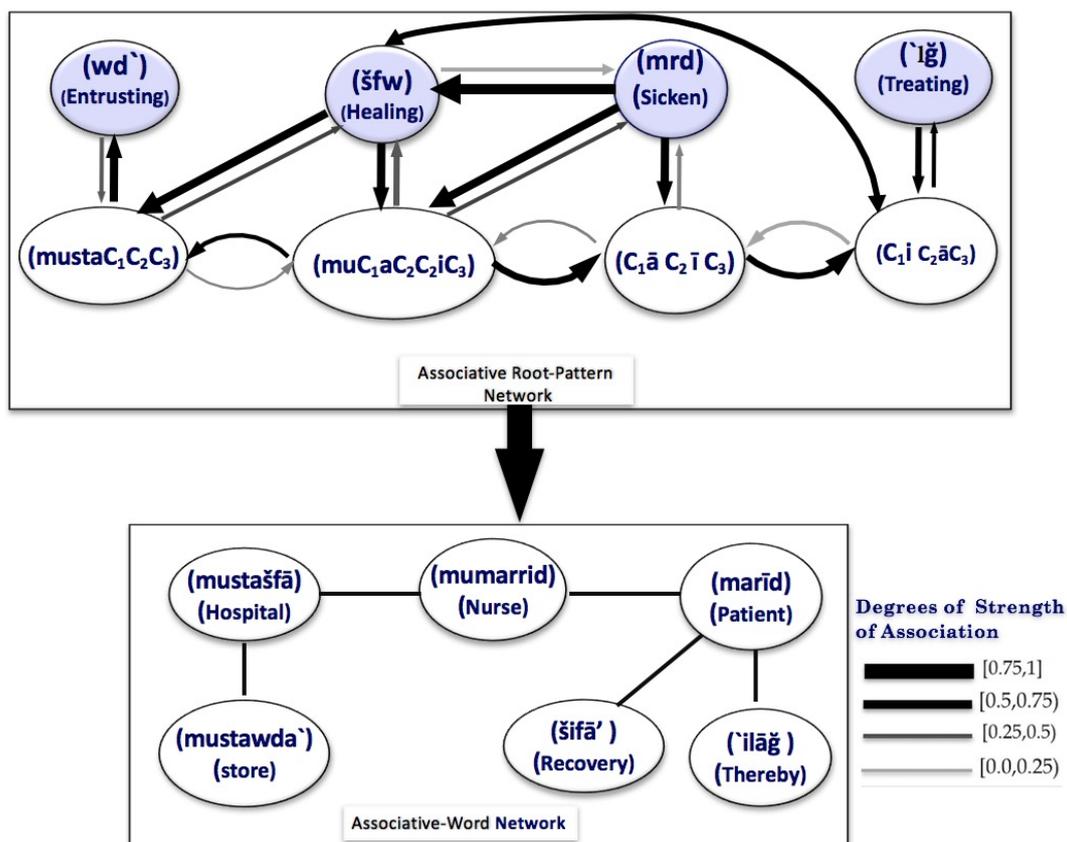


Figure 2. A sample of a bi-directional associative root–pattern network and a possible word network instantiation. The pattern *mustaC₁C₂C₃* might sound acoustically like the word (مُسْتَشْفَى, *mustašfā*, Hospital), but it is cognitively imperceptible without an instantiation with a root such as شَفَوْ , *šfw*, *Healing*. The strength of association is depicted according to a value within the unit interval, which can be estimated and locally computed. This presentation is concerned with the estimation of such prior values.

The *non-linearity* of the Arabic morphology is the principal aspect behind the difference between Indo-European and Semitic languages.

For example, the grade of association between the root *وضع, wd', Setting* and the pattern */maC₁C₂wC₃/* (see Appendix A, Adopted Transliteration) as a canonical template describing the morpho-phonetic structure of a word integrating *phonetic, syntactic, and semantic* information is estimated in the form of different values. Examples include as follows:

⁶ In the statistical analysis of each pattern, syntactical suffixes are also considered to form multiple phonetic patterns conveying additional syntactical information at the end of a pattern.

- The degree of association for the pattern / $maC_1C_2\bar{u}C_3$ / considering the root / $wḍ'$ /, *setting* is estimated by

$$P((/maC_1C_2\bar{u}C_3/) | (/وضع, wḍ' /, setting)) \approx 0.618375.$$

- The degree of association for the root / $وضع, wḍ'$ /, *setting* considering the pattern / $maC_1C_2\bar{u}C_3$ / is estimated by

$$P((/وضع, wḍ' /, setting) | (/maC_1C_2\bar{u}C_3/)) \approx 0.393711.$$

A bi-directional associative relationship between a root and a pattern can be interpreted as a **forward rule** $R \rightarrow P$ and a **backward rule** $R \leftarrow P$ with a *root-predictive value* and a *pattern-predictive value* as possible initial values.

The above example stresses that there is actually a conceptual difference between using an instance and operating with terms on their abstraction level as variables. An instantiated form of the pattern $\langle maC_1C_2\bar{u}C_3, maf'ul \rangle$ in the context of the root $\langle وضع, wḍ', setting \rangle$ such as $\langle مَوْضُوع, mawḍ'ūr, Subject \rangle$ might have different co-occurrence values when operating in their abstract form. The concept of interpreting bi-directional grades of association between root and pattern as predictive and root-predictive values has been successfully employed in the detection and correction of non-words and spelling errors—in particular, to detect cognitive spelling errors [16]. For example, the estimated values in Table 1 support the view that humans would not construct a word based on the root $\langle سلم, slm, being safe \rangle$ and the pattern $\langle مَفْعُول, maC_1C_2\bar{u}C_3 \rangle$. However, it is conceivable to construct a word using another root such as $\langle علم, 'lm, knowing \rangle$. Analogously, we can construct word vectors based on this morpho-phonetic characteristic of Arabic and Semitic languages.

Table 1. Sample of estimated bi-directional grades for associative root–pattern relationships. Root occurrences and their predictive values as a degree of association are represented.

| ROOT | Root-Occ | $\langle \text{مُسْتَفْعَلَةٌ, } mustaC_1C_2aC_3h - uN \rangle$ | | $\langle \text{فَاعِلٍ, } C_1\bar{a}C_2iC_3, \rangle$ | | $\langle \text{مَفْعُولٌ, } maC_1C_2\bar{u}C_3u \rangle$ | |
|--|----------|---|-------------|---|--------------|--|-------------|
| | | R → P | R ← P | R → P | R ← P | R → P | R ← P |
| $\langle \text{عمل, 'ml, Working} \rangle$ | 39485264 | 0.00228212 | 0.2650689 | 0.0000116728 | 0.0038521 | 0.000578 | 0.00008432 |
| $\langle \text{علم, 'lm, Knowing} \rangle$ | 85620930 | 0.00000691 | 0.00067412 | 0.0192878 | 0.0694434 | 0.0012404 | 0.01846788 |
| $\langle \text{درس, drs, Learning} \rangle$ | 18789183 | 0.000000001 | 0.000000001 | 0.00962323 | 0.0003241 | 0.00149652 | 0.00027534 |
| $\langle \text{قيل, qyl, Working} \rangle$ | 60361866 | 0.0000018767 | 0.02654881 | 0.00049551695 | 0.0050278676 | 0.00005864 | 0.000008413 |
| $\langle \text{جمع, ḡm', Gathering} \rangle$ | 61034613 | 0.0000002227 | 0.000399877 | 0.00000097145 | 0.0004955169 | 0.0027966 | 0.00241713 |
| $\langle \text{كتب, ktb, Working} \rangle$ | 49405697 | 0.00000003237 | 0.000004704 | 0.0000024709 | 0.00102028 | 0.030609488 | 0.02141586 |
| $\langle \text{سلم, slm, Working} \rangle$ | 67039191 | 0.0000078228 | 0.015421753 | 0.000000180 | 0.000101108 | 0.000000001 | 0.000000001 |

1.2. Scope of the Presentation

This paper is primarily concerned with the process of creating the APRoPAT Corpus and presenting some quantitative features of a large-scale web dataset and their importance in the context of providing researchers with basic statistical data concerned with studying the root–pattern phenomena in Arabic and semitic languages. Non-derivative word forms analysis will not therefore be a major part of this presentation. Furthermore, this presentation is not intended to present linguistic or lexical definitions or the steps involved in the morphological analysis used by employing the morphological analyzer Petra-Morph or other stemmers. Details of the potential application and technical complexity of the involved resources and the APRoPAT Language Model as a cognitive model are beyond the scope of this paper. Details can be found in [10,14,17].

The remaining part of this paper is organized as follows. An overview of the processes involved in constructing the corpus will be discussed. In particular, web dataset pre-processing, initial bi-gram indexing, arabic text filtering, and root–pattern statistics. Finally an overview of quantitative features of the corpus and distribution will be introduced with a concluding outlook and discussion.

2. Related Work

Corpus-based language modeling opens new potential approaches to studying different phenomena of a language. It is not sufficient to calculate the frequency of some linguistic features. More complicated purpose-driven analysis are expected to be performed (e.g., the investigation of word usage such as corpus pattern analysis, co-occurrence analysis, ranking aspects, and others). In this paper, associative relationships between linguistic constituents, such as roots and patterns in Semitic languages and consequently the cognitive aspects or priming effects, are investigated [18]. In general, such deep search is focused on discovering implicitly inherited *linguistic* or *data patterns* by investigating *reduplication* and *repetition* at different levels in light of the *functional descriptions* of a language [19,20].

In this context, our approach attempts therefore to investigate certain qualitative and quantitative features of Arabic on the morpho-phonetic level of word recognition based on large-scale web text datasets such as ClueWeb.

In spite of the fact that processes of analyzing and building a corpus are similar in many aspects [19,21], our initial goal at this *stage of research* is directed toward *exploring* and *concretizing* a corpus-based cognitive language model for Arabic, i.e., the APRoPAT-Model, and possibly for other Semitic languages by estimating the involved associative relationships by conditional probabilities in the form of bi-gram analysis on an abstract level of the non-linearity of Arabic morphology. Another aim is to refine the obtained values toward future data mining and information retrieval tasks such as query expansion based on predicting phonetic query vectors [17,18].

Furthermore, as the concept of modeling of Arabic morphology is based on abstract associative entities of the mental lexicon in the form of an associative bidirectional probabilistic network between roots, patterns and particles, to the best of our knowledge, there have been no reports directly connected to this research except for [10,17] and other related work.

Furthermore, in light of the current status of corpus-based computational linguistic research and the availability of multipurpose comprehensive Arabic corpora, we decided to create and analyze our own corpus to facilitate and promote the potential applications related to the APRoPAT Language Model.

On the other hand, we believe that the resulting APRoPAT Corpus is useful in different ways, as it considers, besides the traditional corpora features, (e.g., a cleaned, large text collection, annotation, and n-gram frequencies), novel associative relationships at the morpho-phonetic level of cognition [17]. These relationships are estimated by bi-directional prior conditional probabilities for roots, patterns, root–pattern relationships, stems, and root–particle relationships, forming an associative probabilistic network acting as an abstract model for word recognition.

In addition, proceeding from a large-scale dataset such as ClueWeb09⁷, a standard textual collection for our research, would support the assumption that resulted features are comprehensive and representative. Datasets containing a huge collection of raw webpages provide researchers with a representative alternative to creating big data corpora rather than relying merely on *n-gram web corpora* [22]. Following the Zipf law⁸ for word distribution in human languages, we find that few words are used frequently and that many or most words are used rarely. As is the case for all *language*

⁷ Clueweb 12 was released during the production of this manuscript. Unfortunately it contains a very small collection of Arabic webpages: <http://www.lemurproject.org/clueweb12.php/>.

⁸ http://en.wikipedia.org/wiki/George_Kingsley_Zipf.

constituents and phenomena, the *more text* available in a corpus, the *more language phenomena* can be determined.

In the context of scale, we found reports on building a 70-billion-word corpus for English from ClueWeb [23] as well as a similarly large corpus for Chinese [22] and a smaller one for Arabic [24,25].

However, in the sense of the used techniques, our approach consists in employing similar basic languages resources (a morphological analyzer, a POS tagger, and other special tools and algorithms). The NEMLAR project involved the BLARK (Basic Language Resource Kit [26,27] Arabic Treebank, the KSUCCA Corpus [28], and others. As mentioned earlier, more details about these corpora can be found in [4,5]. Furthermore, some additional probabilistic morpho-phonetic measures and their distribution have been estimated in the form of *root* and *pattern entropies*, besides bi-directional conditional probabilities, such as the conditional entropy of roots given a specific pattern for illustrating the entropy of the conditional distribution in the dataset. These values reflect the disorder or the impurity for classifying roots under observation different patterns and vice versa. We believe these widely negated probabilistic values are useful in resolving different types of ambiguities occurring in Arabic NLP.

3. Building APRoPAT Corpus

As stated before, we decided to use the Arabic subset of ClueWeb09 as the source for the corpus data to simplify the process of collecting webpages as a departure resource for creating the corpus. Furthermore, the existence of free and robust crawlers, particularly for crawling very large-scale webpages, is actually rare [24].

ClueWeb09⁹ is a collection of ca. 1 billion webpages in 10 languages collected by Carnegie Mellon University within the Lemur Project. The dataset is used by several tracks of the TREC conferences and was created to support research on information retrieval and related human language technologies. For building the APRoPAT Corpus, we used the Arabic subset, which contains ca. 29.2 million webpages.

The process for constructing the corpus included different steps. However, the focus was on the part concerned with statistical analysis for estimating bi-directional conditional probabilities involved in the APRoPAT Language Model. At this step *roots*, *patterns*, *stems*, and *particles* are involved. Special and deep morphological pre-processing is required before considering any n-gram analysis. However, for generality purposes and *computational issues*, *uni- and bi-gram analysis* were considered as well.

The procedure of creating the corpus includes the following processes in a simplified form (see also Figure 3):

- Extraction. HTML file extraction from compressed WARC archives, provided by ClubWeb.
- HTML File-Parsing. Converting HTML files into plain text files.
- Bi-Gram Analysis and Initial Indexing: Establishing a vector space of all distinct terms of the parsed text with frequencies of each term. In addition, this step includes generating a two-dimensional matrix from the vector space with frequencies of every two consecutive terms as a basic feature.
- Filtering. Removing all non-Arabic terms that might still be remaining (e.g., Persian words). The Arabic subset of ClueWeb is actually not clean.
- Morphological Analysis and POS Tagging. Analyzing all corpus terms considering all possible roots, stems, patterns, and parts of speech tagging. At this step, Petra-Morph was predominantly utilized.
- Core APRoPAT Statistical Analysis, which includes the following:
 - Computing Basic Quantitative Features.

⁹ <http://lemurproject.org/clueweb09.php/>.

- Bi-Directional Root–Pattern Analysis.
- Bi-Directional Root–Root Analysis.
- Bi-Directional Root–Stem Analysis.
- Pattern–Pattern Analysis.
- Particle Tagging. Specifying the involved particles dataset.
- Bi-Directional Root–Particle Analysis. Computing root–particle probabilities.
- Computing Bi-Directional Specific Entropies.

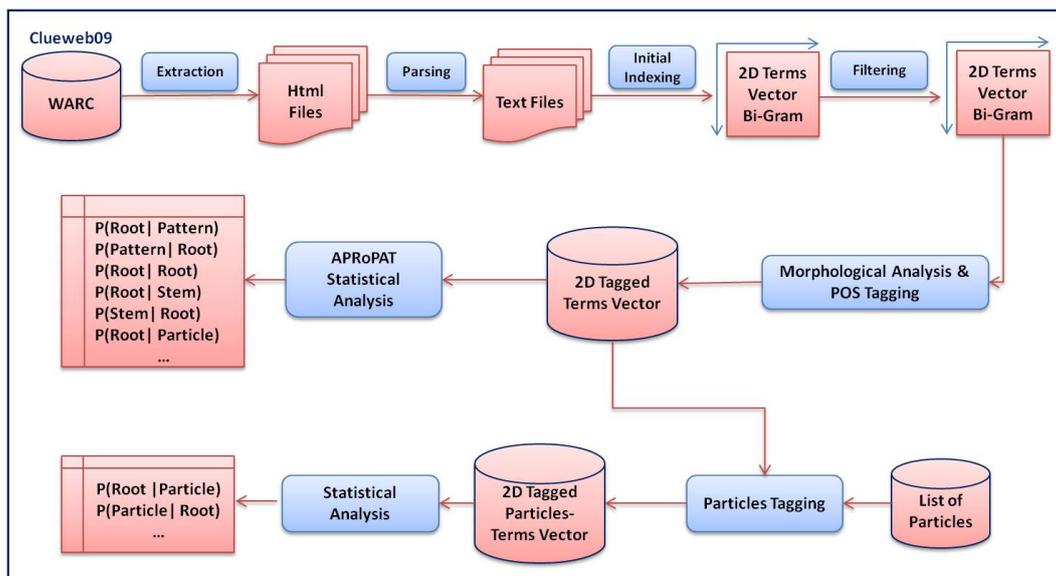


Figure 3. Processes involved in constructing APRoPAT corpus.

3.1. Web Dataset Pre-Processing

The Arabic ClueWeb dataset is organized in web archives forming around 29.2 million webpages. It contains 975 WARC files, 240 GB of which are compressed and 1 TB of which are uncompressed. Each WARC file contains on average about 30,000 webpages. To extract the contained *text files* from such a large-scale compressed web collection, an efficient algorithm was required to process them in a reasonable time.

On the other hand, we had to cope with this problem using the available computer system as single server, (see Appendix B).

To extract the HTML files from the WARC archive files, a *modified* and *optimized* version of the Java code provided by ClueWeb¹⁰ was utilized¹¹. Furthermore, the extracted HTML files needed to be converted to plain text files. This step was performed by HTML parsing¹². At this phase, the HTML DOM (document object model) was traversed and, for each DOM's element, the plain text was extracted while all tags and HTML formatting were removed. At this step, the 975 WARC files were converted into 975 text files and stored, comprising 180 GB.

¹⁰ <http://boston.lti.cs.cmu.edu/clueweb09/wiki/tiki-index.php?page=Working+with+WARC+Files>.

¹¹ As reading and writing to a hard disk is a slow process, the code was changed so that only the HTML file in memory was extracted and then parsed, and the extracted text result was then saved to disk

¹² Jsoup HTML Parser, written in Java, was used, as it allowed us to process the HTML file while in memory without the need to call another program or to save it to hard disk

3.2. Initial N-Gram Indexing and Non-Arabic Text Filtering

The main goal of this process was to create a *unique 2D vector space* of all words in the extracted text files from the *ClueWeb Dataset*. The obtained results are important to reduce the amount of involved data for further statistical analysis, besides its value in creating a big dataset of conditional probabilities for each two terms for future work.

Before starting any n-gram analysis, we needed to go through all the Arabic words in the dataset to be indexed. The main concern at this step was the *performance*, so writing an efficient algorithm was required to finish looping through the text files within a reasonable time.

Basically, the implemented code uses *two large arrays* that are *static* and stored in memory to increase the performance; one is used to store a *one-word distinct list*, while the second is for a *two-word distinct list*. Both arrays use the dictionary class in C#. The uni-gram dictionary uses the words themselves as the dictionary key and uses the frequency as the dictionary value.

These pre-processing steps were important to deal with the large amount of data under the used computing system.

Non-Arabic words, in particular *Persian*, were still present in the resulting 2D vector space, which was resolved using the free available list of Persian words collected from Wikipedia and some Persian dictionaries.

An overview of word distribution is summarized in Figure 4, which complies with the aforementioned Zipf law. Figure 5 shows a balanced root distribution. This fact supports the initial assumption that Arabic roots represent a higher level of abstraction, capturing basic word meanings. Moreover, the corpus size seems to be large enough to cover most derivable word and related topics (for comparison, see Section 5 and Appendix C).

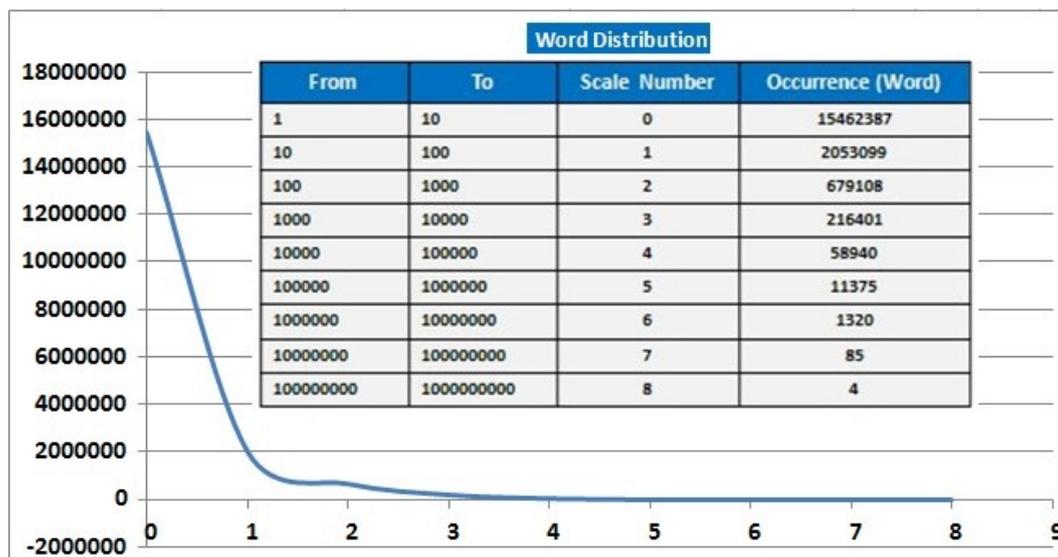


Figure 4. Distribution of word frequencies in log-scale. Following the Zipf law for word distribution in human languages, we find that few words are used frequently and that most words are used rarely. The table in the figure shows 9 log-scales: Scale 0 shows that there are 15,462,387 words in the corpus of 1–10 frequencies. On the other hand, Scale 8 shows that there are only 4 words with very high frequencies—from 100,000,000 to 1,000,000,000.

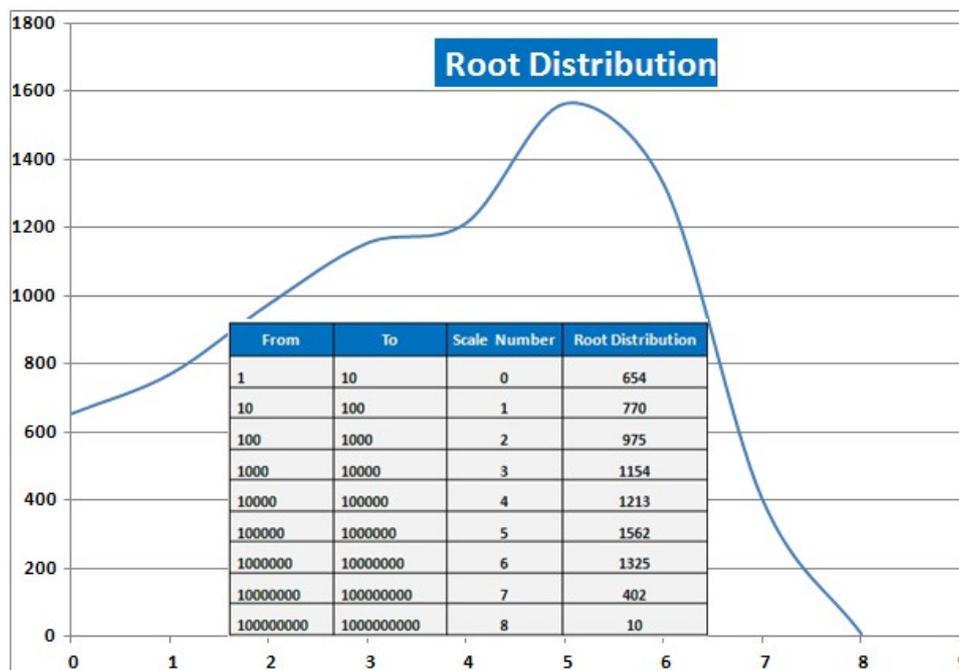


Figure 5. Root distribution in log-scale. Most frequent Arabic roots occur between Scales 3 and 7; e.g., Scale 5 shows that there are 1562 roots having 100,000 to 1,000,000 frequencies. Around 40% are frequent roots. The distribution of roots in the corpus is relatively balanced, reflecting that most derivative words of Arabic are present in the dataset, as around 8065 roots, i.e., most roots in modern standard Arabic were considered.

4. The APRoPAT Morpho-Phonetic Dataset

As mentioned earlier, the corpus is preliminarily designed to reflect specific aspects of the APRoPAT Model as a cognitive model for Arabic word recognition and identification. At this stage of research, we are interested in estimating certain morphological units on the cognitive morph-phonetic level of word cognition. POS Tagging was partially considered in building the corpus. Roots, patterns, stems, and particles were considered as basic morpho-phonetic elements at this level. We primarily utilized the *Petra-Morph Analyzer*¹³. However, to verify the *Petra-Morph* root–pattern performance, optimized versions of the *Al-Khalil Analyzer*¹⁴ and a *Kholja light-stemmer*¹⁵ were also employed for comparison¹⁶.

¹³ *Petra-Morph* is a morphological analyzer and POS tagger that is based on the work of Arabic Textware 'morphological analyzer. Arabic Textware (ATW) developed its morphological analyzer for the purpose of indexing Arabic text in 2001. It was used in the Addaall search engine [29], the company's enterprise search product. ATW's morphology analyzer uses a finite state approach, utilizing a database of more than 6000 Arabic roots and more than 600 Arabic patterns. It includes lists of prefixes, affixes, and suffixes and a list of special words. The speed of the morphological analyzer suited the purpose of fast indexing, analyzing thousands of words per second. When applied to Arabic ClueWeb, its coverage reached 82%. This percentage of coverage required a University of Petra research team to make modifications to the system by adding two layers to enhance its coverage and quality of results. The first layer was to normalize the input words, and the second to modify its stemmer, and to redirect the analyzer to process the new stem. The new product we named *Petra-Morph* reached a better percentage of coverage reaching 87%. The last modification under the umbrella of *Petra-Morph* was to reduce the number of roots to a list related to Modern Standard Arabic, excluding the abandoned traditional roots. This process enhanced the quality of the results by reducing the number of possibilities the analyzer generated for each word.

¹⁴ <http://sourceforge.net/projects/alkhalildotnet/>.

¹⁵ a modified version of <http://zeus.cs.pacificu.edu/shereen/ArabicStemmerCode.zip>.

¹⁶ While testing the *Al-khalil* morphological analyzer, we found that it was relatively slow, especially for big data, so we accelerated the *Al-Khalil* morphological analyzer in .Net in several steps (using parallel processing and pre-loading). For *Kholja* stemmer, we also had to rewrite it in C# to enhance its performance (done by preloading and caching all of the required files instead of reading them from the disk each time). To compare the coverage of a file of 2.16 GB that contains stems, and possibly some roots and patterns with an indicator to the source of the analysis for each word (*Petra-Morph*,

At this phase, it was expected that many words in the ClueWeb Dataset cannot be morphologically analyzed. There were multiple reasons for this:

- Simple orthographic errors, e.g., مشكو و ور, *instead of* مشكور, cognitive errors, e.g., يعطيج, *instead of* يعطيك, and run-on errors; e.g., سلام عادل, *instead of* سلام عادل. More details of such error types and how to deal with them can be found in [16].
- Arabized proper names, e.g., جنكزخان, *Genkiskhan*.
- Arabized technical terms, e.g., دوت نت, *dot net*.
- Fictive words (non-words), e.g., مكريش, *هههه*.
- Technical and political abbreviations, e.g., اتش د, *HD*.
- Dialects, e.g., على الطائر, *Quickly*, ما ودي, *Idon't want*

These types of word occurrences were light-stemmed to be subject to future statistical analysis. However this aspect lies outside the scope of the present paper.

Tables 2 and 3 summarizes some *quantitative features* of the root–pattern relationship, stem, and particle analysis of ClueWeb.

Table 2. Considered root–pattern features in the dataset.

| Morpho-Constituent | Size | Comment |
|--|------------|--|
| Roots | 8065 | Predominantly based on Petra-Morph and the Alkalil Database |
| Patterns | 5737 | Patterns and all of their syntactical variations based on Petra-Morph. |
| Particles | 281 | |
| Word forms without Root–Pattern Relation | 11,322,272 | These Data represent the non-derivative part of the corpus; i.e., words without a clear root–pattern relationships such as certain country's names, proper names, and others. Frequency distribution shows that this type has a much lower frequency in the corpus compared to the words that have root–pattern relationships. |

Root–Pattern Statistics

Based on basic frequencies of the APRoPAT Morph dataset; i.e., roots, stems, patterns, particles, and root–pattern relationships extracted from ClueWeb, initial statistical analysis was performed to estimate conditional probabilities of the proposed associative root–pattern relationships in the APRoPAT Model.

Basically, bi-directional binary relations were established between roots, patterns, stems, and particles, as explained in Section 3.

Generally, each item was treated as a *random variable*, where *add-one smoothing* was applied to estimate all the above-mentioned *binary relations*. This step is comparable to computing all involved entities in an associative sub-network [17].

Definition 1 shows the proposed procedure heuristic for estimation of the involved associative probabilistic relations:

Al-Khalil, or Khoja), the Petra-Morph covered about 25% of the total number of words, while Al-Khalil covered 0.05% more words, and the Khoja stemmer covered the remaining 70% of the words. We should notice that these percentages do not reflect necessarily the actual representation of these words in the Clueweb09 database, so these words might only have stems and non-words and might have much lower frequencies in the Clueweb database compared to the words that have roots and patterns.

Definition 1. Let \mathfrak{M} be the set of all variables involved in the APRoPAT morpho-phonetic dataset involved types: $\mathfrak{M} = \{\text{ROOT}, \text{PATTERN}, \text{STEM}, \text{PARTICLE}\}$ and $\mathcal{X}, \mathcal{Y} \in \mathfrak{M}$, then the values of a probabilistic associative relationship are estimated by

$$P(\mathcal{X} | \mathcal{Y}) \approx \frac{|P(\mathcal{X} \cap \mathcal{Y})| + 1}{|P(\mathcal{Y})| + \mathcal{Y}_V} \tag{1}$$

where \mathcal{Y}_V is the vocabulary of the variable $\mathcal{Y} \in \mathfrak{M}$, $\forall x_i$ and $y_j \in \mathcal{X}$ and \mathcal{Y} , respectively. and

$$P(\mathcal{Y} | \mathcal{X}) \approx \frac{|P(\mathcal{X} \cap \mathcal{Y})| + 1}{|P(\mathcal{X})| + \mathcal{X}_V} \tag{2}$$

where \mathcal{X}_V is the vocabulary of the variable $\mathcal{X} \in \mathfrak{M}$, $\forall x_i$ and $y_j \in \mathcal{X}$ and \mathcal{Y} , respectively.

The resulting relations are summarized in Table 4.

Table 3. The overall considered size.

| Quantitative Feature | Size | Comment |
|----------------------------|----------------|---|
| Document (webpages) | 29,192,662 | 975 text files (180 GB) for each, with 30,000 webpages. ClueWeb 975 WARC in 240 GB compressed files and in 1 TB uncompressed |
| Word-Forms | 11,437,025,140 | |
| Vocabulary | 18,482,719 | Distinct words types stored in one 398 MB file. |
| Persian Bigrams Word Forms | 122,516,474 | Total frequency of Persian bigrams. ClueWeb is not clean; it contains multiple Persian and Urdu word forms. |
| Persian Bigram Types | 10,739,392 | These data were excluded from the Arabic vocabulary. It has a relatively small ratio; approx. 2% of the unique Arabic bi-grams and around 8% of the total count of the bi-gram analysis. However, it constitutes a large portion of the vocabulary |
| Roots | 8065 | The large number of roots represents how many abstract concepts there are in the corpus. Furthermore, the concepts seems to be nearly balanced and normally distributed (see also Figure 5). Around 40% of the Arabic roots were very frequent. However, the occurrence of stems and non-derivative word forms have not been considered in this analysis, as the focus of the presentation was set on computing associative root–pattern relationships. |

Table 4. Associative relationships in the APRoPAT morpho-phonetic dataset.

| Morpho-Associative Feature | Size |
|--------------------------------|-----------|
| Bi-Directional Pattern–Pattern | 1,152,310 |
| Bi-Directional Root–Pattern | 6,992,022 |
| Bi-Directional Root–Root | 149,870 |
| Bi-Directional Root–Particle | 513,538 |
| Bi-Directional Root–Stem | 503,506 |

Furthermore, the term-specific ROOT and PATTERN entropy has also been considered, as a novel measure to capture the entropy of a specific root or pattern occurring bi-directionally with multiple patterns or roots, respectively.

Definition 2. Let ROOT and PATTERN be the sets of all involved roots and patterns in the APRoPAT morpho-phonetic dataset¹⁷, then the specific ROOT Entropy for some pattern Pt_j is estimated by

$$H(\text{ROOT} \mid \text{PATTERN} = Pt_j) \triangleq - \sum_{i=1} P(r_i \mid \text{PATTERN} = pt_j) \log P(r_i \mid \text{PATTERN} = Pt_j) \quad (3)$$

$\forall r_i \in \text{ROOT for some } Pt_j \in \text{PATTERN}$

and the specific PATTERN entropy for the root r_j is estimated by

$$H(\text{PATTERN} \mid \text{ROOT} = r_j) \triangleq - \sum_{i=1} P(Pt_i \mid \text{ROOT} = r_j) \log P(Pt_i \mid \text{ROOT} = r_j) \quad (4)$$

$\forall Pt_i \in \text{PATTERN for some } r_j \in \text{ROOT}$

Similarly, the other possible specific entropies have been computed.

5. Corpus Size and Distribution

To attain an overview of the main characteristics of the APRoPAT Corpus, Tables 2 and 3 summarize some quantitative features. It shows that the resulted corpus is of large-scale containing ca. 11.5 billion word forms with ca. 18.5 million distinct ClueWeb types.

An overview of word distribution is summarized in Appendix C. Figure 4 shows word distribution, which complies with the aforementioned Zipf law. On the other hand, Figure 5 shows a balanced root distribution supporting the initial assumption that the corpus covers balanced semantic aspects and the basic concept of APRoPAT as a cognitive model [10,17].

6. Outlook and Conclusions

In this paper, the process of constructing a large-scale corpus was described. Around 29.2 million webpages from ClueWeb were extracted, filtered, and converted into a large-text corpus containing around 11.5 billion word forms and estimating around 9.3 million instances of morpho-associative relationships in the form of prior conditional priorities.

Different language resources and tools were also created, modified, and employed to deal with large-scale computing. We believe that such resources provide researchers with balanced large-scale text corpora containing implicitly different functional descriptions of Arabic over the web as a dynamic platform for collecting qualitative features of a living language such as Arabic. This work also proposes considering root distribution as an key evaluation criteria for measuring topic coverage of a corpus in the context of Arabic.

Furthermore, as a considerable part of the dataset can be represented as a semantic network capturing bi-directional associative relationships on the morpho-phonetic level of word cognition, the dataset is ideally suited to study the cognitive aspects related to root–pattern phenomena of Arabic. Creating corpora for Arabic on a large scale is nearly missing in the current state of research, particularly those considering the cognitive aspect of word organization in the mental lexicon.

On the other hand, the availability of such resources paves the way for multiple potential applications for studying Semitic languages from different points of view. This aspect has already been successfully employed in developing a cognitively motivated non-word detection and correction system and in improving precision and ranking of morphological analysis. Currently, it is intended to consider the results of developing a query expansion model based on an associative root–pattern network. Furthermore, the resulting balanced associative semantic network provides a crucial resource for developing a probabilistic word network.

¹⁷ In terms of the APRoPAT model, these represent cognitive variables on morpho-phonetic level of cognition.

Acknowledgments: The authors would like to thank University of Petra, Amman, Jordan for the financial support, which made this research possible, and Arabic Textware Company, Amman, Jordan for providing the researchers with resources.

Author Contributions: B.H. conceived, designed the Model, and wrote the paper; A.A. checked Hardware environment and the speeding-up process; M.H. checked and analysed the morphological context and it's appropriateness for the rquirements; A.H., M.H., and B.H. implemented the required tools.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A Adopted Transliteration

In this paper, transcription of Arabic letters is based on DIN and [30]. Long vowels are represented through the letters (\bar{a}), (\bar{i}), and \bar{u} , while short vowels are as follows: *fathā*, *a*, *kasrah*, *i*, and *ḍammah*, *u*.

- A root $r_i \in \text{ROOT}$, the set of all roots is depicted as three arguments (Arabic, Latin Transliteration, and Abstract Meaning), e.g., the root instance (كتب, *ktb*, *Writing*).
- A pattern $pt_j \in \text{PATTERN}$, the set of all patterns, is also depicted as two arguments: (Latin transliteration and root radical non-linear template positions), whereas C_1 , C_2 , and C_3 represent root radical variables such as in $/maf'ul/$, $maC_1C_2\bar{u}C_3$; i.e., $f = C_1$, $' = C_2$, and $l = C_3$.

Appendix B Performance and Running Time

As stated before, the main concern of creating such a large-scale corpus was the performance and the running time. The data was tested and created on a single server: a Quad-core CPU, 3.3 GHZ, 16 GB Memory, 2500 Hard Drive under Linux Ubuntu 12.0, 64-bit. We have adopted the following strategies for enhancing the performance:

- Reducing reading and/writing from the hard disk and relying on memory as much as possible.
- Avoiding excessive processing on the word level as much as possible, and postponing it until the unique 2D vector space for each associative relation was created during bi-gram analysis as these 2D vectors contain much less data to process.
- Taking advantage of the ability of the CPU to execute instructions in parallel as much as possible.
- Implementing efficient data structures such as hash tables to hold the words and their frequencies.

The first phases of the projects (dataset pre-processing, in particular HTML file extraction and parsing, initial bi-gram texting, and non-Arabic text filtering) were very complex and contained substantial amount of data, so they required long running times to be performed. However, the remaining steps were also huge and complex enough, so that some resources were modified and even rewritten to make them capable of large-scale computing, such as parallel processing and pre-loading. Technical details are beyond the scope of this paper. Appendix Table A1 shows running times of some of the phases with comments on the amount of the involved data and their storage size.

Table A1. Process running time and results.

| Process | Running Time | Input Data | Output Data |
|-----------------------------------|--------------|--|--|
| HTML files Extraction and Parsing | ~ 4 days | 975 WARC files (240 GB compressed, 1 TB uncompressed) | 975 text files (180 GB) |
| Initial Indexing | ~ 2 days | 975 text files (180 GB) | 39 index files (26.5 GB) |
| Morphological Analysis | ~ 6 h | 1 unigram list file (408 MB), and 15 Petra-Morph Database Files (7.98 GB) | 1 Morph Database File (2.16 GB) |
| Uni-gram Frequencies | ~ 3 h | 1 unigram list file (408 MB), and 1 Morph Database File (2.16 GB), and 1 Particles List File | 5 Frequencies Files (364 MB) |
| Statistical Probabilities | ~ 45 min | 11 Frequencies Files 401.8 MB | 8 probabilities files 420 MB |
| APRoPAT Text Dataset Indexing | 5 days | 975 WARC files (240 GB compressed, 1 TB uncompressed) | 44 index files (89.5 GB), and 1 main index file (438 MB) |

Appendix C Dataset Distribution

The following figures show different associative relations and their distribution in the dataset depicted in log-scale:

- Root–Pattern Distribution in Dataset; Figure A1
- Pattern Distribution in Dataset; Figure A2
- Pattern–Pattern in Dataset; Figure A3

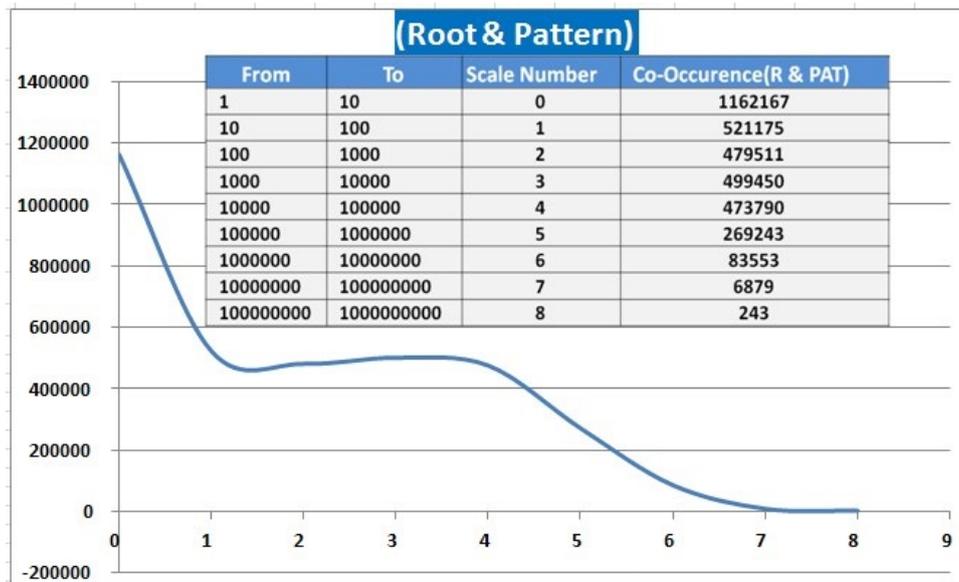


Figure A1. Root–pattern distribution in log-scale reflecting that the majority of Arabic word forms are derivative.

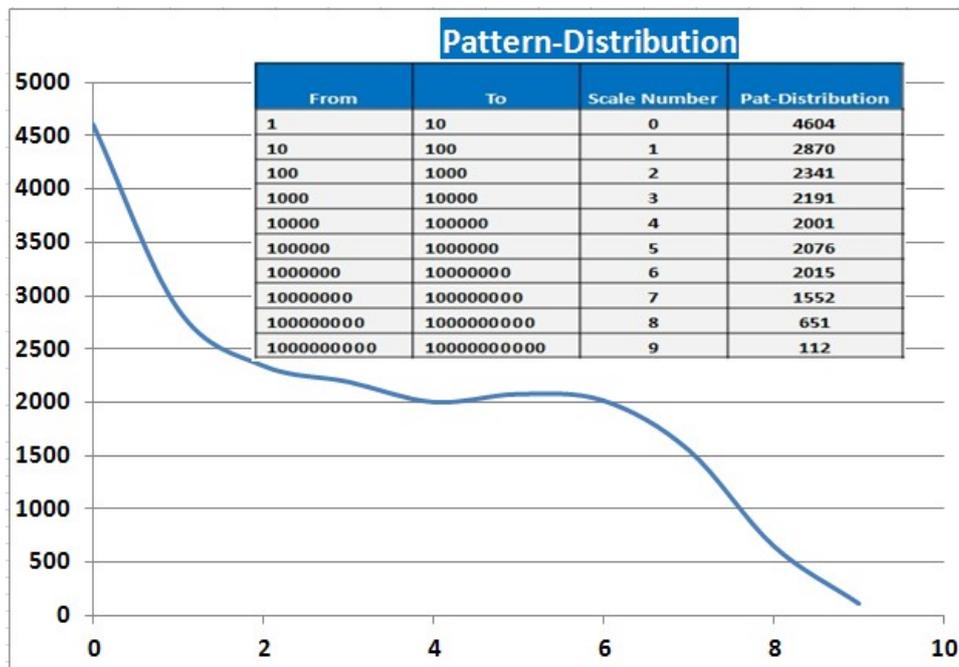


Figure A2. Pattern distribution in log-scale in the corpus reflecting their importance in Arabic.

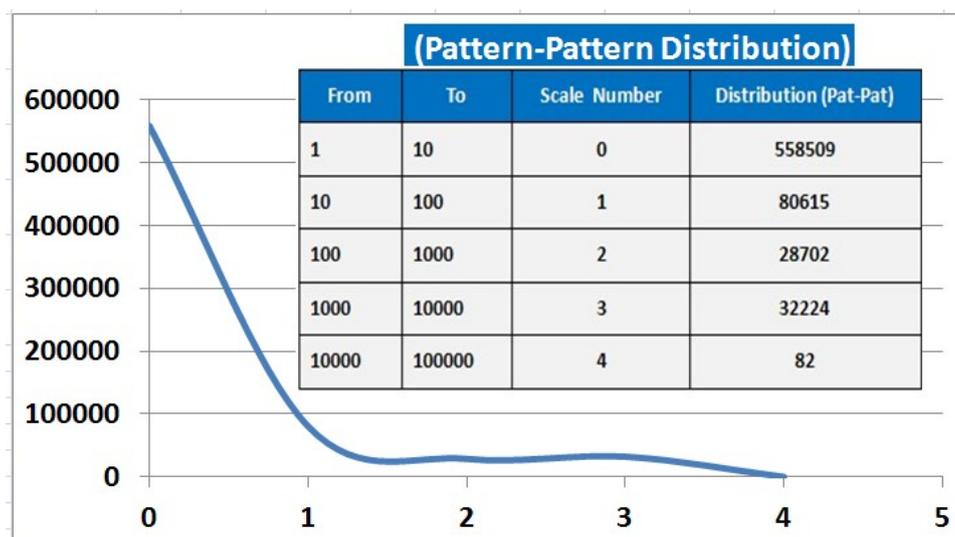


Figure A3. Pattern–pattern distribution in log-scale reflecting the co-occurrence of word forms on variable levels.

References

1. Baranyi, P.; Csapo, A.; Sallai, G. *Cognitive Infocommunications (CogInfoCom)*; Springer International Publishing: Basel, Switzerland, 2015.
2. Arabic. Available online: http://en.wikipedia.org/wiki/Arabic_language (accessed on 28 March 2018).
3. Al-Thubaity, A.; Khan, M.; Al-Mazura, M.; Al-Mousa, M. New Language Resources for Arabic: Corpus Containing More Than Two Million Words and A Corpus Processing Tool. In Proceedings of the International Conference on Asian Language Processing (IALP), Urumqi, China, 17–19 August 2013.
4. Zaghouan, W. Critical Survey of the Freely Available Arabic Corpora. In Proceedings of the Workshop on Free/Open-Source Arabic Corpora and Corpora Processing Tools, Reykjavik, Iceland, 27 May 2014.
5. Al-Sulaiti, L.; Atwell, E.S. The design of a corpus of contemporary Arabic. *Int. J. Corpus Linguist.* **2006**, *11*, 135–171.
6. El-Haj, M.; Kruschwitz, U.; Fox, C. Creating language resources for under-resourced languages: Methodologies, and experiments with Arabic. In *Language Resources and Evaluation*; Springer: Dordrecht, The Netherlands, 2014.
7. Haddad, B. Semantic Representation of Arabic: A logical Approach towards Compositionality and Generalized Arabic Quantifiers. *Int. J. Comput. Process. Orient. Lang.* **2007**, *20*, doi:10.1142/S0219427907001585.
8. Bentin, S.; Forst, R. Morphological Factors in word Identification in Hebrew. In *Morphological Aspects of Language Processing*; Feldman, L., Ed.; Erlbaum: Hillsdale, NJ, USA, 1994; pp. 271–292.
9. Boudelaa, S. Is the Arabic Mental Lexicon Morpheme-Based or Stem-Based? Implications for Spoken and Written Word Recognition. In *Handbook of Arabic Literacy, Literacy Studies 9*; Springer: Dordrecht, The Netherlands, 2014.
10. Haddad, B. Cognitive Aspects of a Statistical Language Model for Arabic based on Associative Probabilistic Root-PATtern Relations: A-APRoPAT. Available online: http://www.infocommunications.hu/documents/169298/393366/2013_4_2_Haddad.pdf (accessed on 28 March 2018).
11. Haddad, B. Probabilistic Bi-Directional Root-Pattern Relationships as Cognitive Model for Semantic Processing of Arabic. In Proceedings of the 3rd IEEE International Conference on Cognitive Infocommunication 2012, Kosice, Slovakia, 2–5 December 2012.
12. Croft, W.; Cruse, D.A. *Cognitive Linguistics*; Cambridge University Press: Cambridge, UK, 2004.
13. Langacker, R.W. An Introduction to Cognitive Grammar. *Cogn. Sci.* **1986**, *10*, 1–40.
14. Haddad, B. Cognitively-Motivated Query Abstraction Model based on Associative Root-Pattern Networks. To be published, draft is available upon request, 2018.

15. Haddad, B. Representation of Arabic Words: An Approach towards Probabilistic Root-Pattern Relationship. In Proceedings of the International Conference on Knowledge Engineering and Ontology Development, Madeira, Portugal, 6–8 October 2009.
16. Haddad, B.; Yaseen, M. Detection and Correction of Non-Words in Arabic: A Hybrid Approach. *Int. J. Comput. Process. Orient. Lang. IJCPOL* **2007**, *20*, 237, doi:10.1142/S0219427907001706.
17. Haddad, B.; El-Khalili, N.; Hattab, M. A Cognitive Query Model for Arabic based on Probabilistic Associative Morpho-Phonetic Sub-Networks. In Proceedings of the 5th IEEE Conference on Cognitive Infocommunications-CogInfoCom, Vietri sul Mare, Italy, 5–7 November 2014.
18. El-Khalili, N.; Haddad, B.; El-Ghalayini, H. Language Engineering for Creating Relevance Corpus. *Int. J. Softw. Eng. Appl.* **2015**, *9*, 107–116.
19. Meyer, C.F. *English Corpus Linguistics An Introduction*; Cambridge University Press: Cambridge, UK, 2002.
20. Wang, S.-p. Corpus-based approaches and discourse analysis in relation to reduplication and repetition. *J. Pragmat.* **2005**, *37*, 505–540.
21. Alansary, S.; Nagi, M.; Adly, N. Towards Analyzing the International Corpus of Arabic (ICA): Progress of Morphological Stage. Available online: https://www.researchgate.net/profile/Sameh_Alansary/publication/263541571_Towards_Analyzing_the_International_Corpus_of_Arabic_ICA_Progress_of_Morphological_Stage/links/0a85e53b2e2622211d000000/Towards-Analyzing-the-International-Corpus-of-Arabic-ICA-Progress-of-Morphological-Stage.pdf (accessed on 28 March 2018).
22. Yu, C.-H.; Chen, H.-H. Chinese Web Scale Linguistic Datasets and Toolki. Available online: <http://www.aclweb.org/anthology/C12-3063> (accessed on 28 March 2018).
23. Pomikalek, J.; Jakubicek, M.; Rychly, P. Building a 70 Billion Word Corpus of English from ClueWeb. Available online: http://www.lrec-conf.org/proceedings/lrec2012/pdf/1047_Paper.pdf (accessed on 28 March 2018).
24. Belinkov, Y.; Habash, N.; Kilgarriff, A.; Ordan, N.; Roth, R.; Suchomel, V. arTenTen: A new, vast corpus for Arabic. Available online: https://www.sketchengine.co.uk/wp-content/uploads/arTenTen_corpus_for_Arabic_2013.pdf (accessed on 28 March 2018).
25. Eckart, T.; Alshargi, F.; Quasthoff, U.; Goldhahn, D. Large Arabic Web Corpora of High Quality: The Dimensions Time and Origin. Available online: <http://www.lrec-conf.org/proceedings/lrec2014/workshops/LREC2014Workshop-OSACT%20Proceedings.pdf#page=35> (accessed on 28 March 2018).
26. Maamouri, M.; Bies, A.; Buckwalter, T.; Mekki, W. The Penn Arabic Treebank: Building a Large-Scale Annotated Arabic Corpus. Available online: https://www.researchgate.net/profile/Mohamed_Maamouri/publication/228693973_The_penn_arabic_treebank_Building_a_large-scale_annotated_arabic_corpus/links/0046351802c78190c5000000.pdf (accessed on 28 March 2018).
27. Yaseen, M.; Attia, M.; Maegaard, B.; Choukri, K.; Paulsson, N.; Haamid, S.; Krauwer, S.; Bendahman, C.; Fersoe, H.; Rashwan, M.; et al. Building Annotated Written and Spoken Arabic LR's in NEMLAR Project. Available online: <https://pdfs.semanticscholar.org/95d7/1fc0a2de2228d62372026ff0913cf2a83959.pdf> (accessed on 28 March 2018).
28. Alrabiah, M.; Al-Salman, A.; Atwell, E. The design and construction of the 50 million words KSUCCA. In Proceedings of the Second Workshop on Arabic Corpus Linguistics (WACL-2), Lancashire, UK, 22 July 2013.
29. Hattab, M.; Haddad, B.; Yaseen, M.; Duraidi, A.; Shmias, A.A. Addaall Arabic Search Engine: Improving Search based on Combination of Morphological Analysis and Generation Considering Semantic Patterns. Available online: http://fafs.uop.edu.jo/download/research/members/202_778_Mamo.pdf (accessed on 28 March 2018).
30. Fischer, W. *Grammatik des Klassischen Arabisch*; Harrassowitz Verlag: Wiesbaden, Germany, 1972.

