Journal of King Saud University – Computer and Information Sciences 29 (2017) 171–178

Contents lists available at ScienceDirect

Journal of King Saud University – Computer and Information Sciences

journal homepage: www.sciencedirect.com

Towards a standard Part of Speech tagset for the Arabic language

Imad Zeroual [a,*], Abdelhak Lakhouaja [a], Rachid Belahbib [b]

[a] Department of Mathematics and Computer, Science Faculty of Sciences, University Mohamed First, B-P 717, Oujda 60000, Morocco
[b] Doha Historical Dictionary of the Arabic Language, Doha, Qatar

ARTICLE INFO

Article history:
Received 13 April 2016
Revised 13 January 2017
Accepted 23 January 2017
Available online 2 February 2017

Keywords:
Natural Language Processing
Part of Speech
Tagging
Arabic tagset
TreeTagger

ABSTRACT

Part of Speech (PoS) tagging is still not very well investigated with respect to the Arabic language. Determining the PoS tags of a word in a particular context is difficult, primarily because there is no use of diacritics in most of contemporary texts. Consequently, the same word may be spelled in different ways. Further, detecting the difference between Arabic derivatives represents a very challenging issue for the majority of PoS taggers. Hence, the task of tagging the correct PoS tags requires advanced processing and the use of considerable resources. This study aims to design detailed hierarchical levels of the Arabic tagset categories and their relationships. These hierarchical levels allow easier expansion when required and produce more accurate and precise results. They are based on a comparative study and important references in Arabic grammar; they are also validated by experts in this field. In addition, the proposed tagset is implemented in a PoS tagger and tested via various experiments. We believe that our study makes a significant contribution to the literature because this work is an advancement in the direction of achieving a standard, rich, and comprehensive tagset for Arabic.

1. Introduction

Part of Speech (PoS) tagging is an important research area and the basis for a number of Natural Language Processing (NLP) tasks. Unfortunately, there is no standard PoS tagset used for Arabic Language Processing (ALP). In fact, only a small number of researchers are interested in the question of standards, especially in ALP. Consequently, it is difficult to benefit from existing PoS taggers or compare and evaluate different tagging approaches under the same conditions. Yet, several researchers have proposed tagsets that comply with their suitable objectives without considering Arabic grammatical features.

The majority of currently used tagsets are derived from English, which is a drawback for a morphologically complex language such as Arabic. The adaptation of such tagsets is problematic for Semitic languages as Zitouni (2014) claimed. "Approaches to PoS tagging were limited to English, resources for other languages tend to use 'tag sets', or inventories of categories that are minor modifications of the Standard English set". Moreover, the most widely used tagsets as standard guidelines, namely those recommended by the Expert Advisory Group on Language Engineering Standards (EAGLES), are designed for Indo-European languages. These guidelines are not entirely suitable for Arabic. Further, several of the current systems tend to target a PoS tagset that is not sufficiently suitable for different applications (Habash et al., 2009) (e.g., (Khoja, 2001; Darwish, 2002; Diab, 2007)).

The main challenge involved in constructing any NLP system for Arabic is amplified by the lack of language resources such as tagged corpora, which are fundamental for research and development in statistical computational linguistics (Farghaly and Shaalan, 2009). PoS tagging is one of the first processes that directly reflects the performance of other subsequent text processing (Albared et al., 2011). Habash and Sadat (2006) studied the effects of PoS tagging and demonstrated a positive influence on the quality of statistical machine translation.

Before addressing the PoS tagging process, the first requirement for the annotation of Arabic text is the compilation of a tagset that can accurately describe and address all the information regarding the language (Khoja et al., 2001). Further, an investigation of PoS tagging for Arabic indicates that using a complex tagset and then converting the resulting annotation to a smaller tagset provides a higher accuracy than tagging using the smaller tagset directly (Kübler and Mohamed, 2012).

* Corresponding author.
E-mail address: mr.imadine@gmail.com (I. Zeroual).
Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

http://dx.doi.org/10.1016/j.jksuci.2017.01.006
1319-1578/

The present paper aims to develop the finest possible PoS tagset for Arabic and to produce more accurate and precise results that can be used to maximize the performance of subsequent ALP tasks such as syntactic parsing.

The proposed tagset is tested using a probabilistic tagging method. This method estimates the transition probabilities using a decision tree, which differs from other probabilistic taggers. Based on this method, a language-independent PoS tagger called TreeTagger is then adapted to use this tagset.

This paper is organized as follows: in Section 2, we provide background information regarding PoS tagging and specific approaches that attempt to solve the problem of PoS ambiguities. Furthermore, some popular PoS taggers for Arabic are presented. We illustrate the relevant existing tagsets with their drawbacks in Section 3. In Section 4, we describe the proposed tagset based on standard design criteria and compare it to similar projects. In Section 5, we present the usability test of the tagset via various experiments and discuss the findings. We conclude this paper in Section 6.

2. Background information

After providing a definition of PoS tagging, various approaches that have been adopted for this process are presented. Further, examples of Arabic PoS taggers are cited with their performance results.

2.1. Definition of Part of Speech tagging

PoS tagging is the ability to computationally determine what PoS tag of a word is activated by its use in a particular context (Albared et al., 2011). It is the task that involves managing ambiguity in processed text.

2.2. Tagging approaches

The task of identifying all the possible PoS tags of a word is not difficult, thanks to the existence of efficient Morphosyntactic analysers for Arabic words such as "AlKhalil Morpho Sys" (Boudchiche et al., 2017), Madamira (Pasha et al., 2014), and (Buckwalter, 2004). However, it remains difficult to achieve disambiguation.

In earlier interesting works by other researchers (Farghaly and Shaalan, 2009; Maamouri and Bies, 2010), the reason why ambiguity exists on numerous levels in Arabic is presented. For example, analysing the Arabic word " ثمن "<tmn> using Buckwalter Arabic Morphological Analyzer (BAMA) produced 21 different analyses. Further, it was estimated that the average number of ambiguities for a token in the majority of languages is 2.3, whereas in Modern Standard Arabic, it is 19.2 (Farghaly and Shaalan, 2009). When the same process for the same word "ثمن "<tmn> using "AlKhalil Morpho Sys" is executed, the analyser determines 40 different analyses, considering all possible diacritical marks.

There are some approaches designed to achieve this disambiguation. The best known ones are statistical/probabilistic approaches, rule-based methods, and hybrid systems that using a combination of both statistical and rule-based methods:

- Statistical/probabilistic methods: Almost all of these are based on Markov models where training consists of learning both lexical and contextual probabilities. This approach is based on a large manually annotated corpus from which we extract probabilities.
- Rule-based methods: They function using rules that have been defined by linguists. A rule-based method is composed of three tasks:
  1. Morphological analysis: This consists of segmenting a sequence of input words into morphemes with respect to the language grammar. This process is accomplished by morphological analysers;
  2. Lexicon research: Lexicons include words that cannot be analysed in the morphological task, such as some stop words, proper nouns, Arabized nouns, and misclassified words;
  3. Sentence structure (El Hadj et al., 2009): This is based on the relationship between untagged words and their adjacent words. The Arabic language has relationships between adjacent words. For example, prepositions and interjections are usually followed by nouns. The word position in the sentence is an effective indicator to identify nouns. Some words always followed by nouns construct a linguistic rule to identify them in the text such as "إن وأخواتها", "كان وأخواتها ", and some of these words are mainly used when recognizing proper nouns such as "السيد " and "الجامعة " 'Mr. University'.
- Hybrid automatic system: This involves combining different methods such as rule-based methods with statistical/probabilistic methods. This system is used to assign the best tag for each of the words of the input text.

2.3. Arabic PoS taggers

A significant part of the work has been undertaken in the area of Arabic PoS tagging (Al-Sughaiyer and Al-Kharashi, 2004; Sawalha and Atwell, 2010); other projects have been developed by companies (Xerox, Sakhr, RDI) as commercial software. In this section, we summarize some of the most relevant works on PoS tagging.

The stochastic PoS taggers provide the appropriate tags based on the most likely tag sequence in tagged corpora; many developed algorithms are employed (Altabba et al., 2010), such as the Viterbi algorithm (Viterbi, 1967) using a Hidden Markov Model (HMM).

The most relevant PoS taggers based on this approach (Diab et al., 2004) are based on Support Vector Machine (SVM), a supervised learning algorithm that uses LDC's PoS tagset, consisting of 24 tags. Another SVM-based, Yamcha, which uses Viterbi decoding, was developed by Habash and Rambow (2005). The approach of Maamouri and Cieri (2002)) is based on the automatic annotation output produced by Tim Buckwalter's morphological analyser; it achieved an accuracy of 96%. Banko and Moore (2004) presented an HMM tagger that exploits context on both sides of a word to be tagged. It is evaluated in both the unsupervised and supervised cases and achieved an accuracy of approximately 96%. Another PoS tagger, similar to the one integrated into the Stanford PoS Tagger, adopted a maximum entropy approach by enriching the information sources used for tagging. Its end result accuracy on the Penn Treebank achieved 96.86% overall, and 86.91% on previously unseen words (Toutanova and Manning, 2000). Another probabilistic tagger was adapted for Arabic (Zeroual and Lakhouaja, 2016a); it differs from other probabilistic taggers in the manner the transition probabilities are estimated, namely with a decision tree. The authors report that the obtained accuracy rates were 99.4%, 92.6%, and 81.9% for the Quranic-vowelled corpus "Al-Mus'haf" (Zeroual and Lakhouaja, 2016b), unvowelled "Al-Mus'haf" corpus, and the NEMLAR corpus (Attia et al., 2005), respectively.

The Qutuf (Altabba et al., 2010) tagger is based on a system that consists of two tagging phases: premature and overdue (usual tagging). The premature tagging occurs before the morphological analysis phase, whereas the usual tagging happens after, and requires rules from a linguistic expert or manually annotated corpus to statistically generate the rules. The Qutuf tagset is based on the Sawalha tagset with refinement and expansion. Brill's "transformation-based" or "rule-based" PoS tagger for Arabic (Freeman, 2001) uses a machine-learning approach based on the Brown cor-

pus; a tagset of 146 tags was used. A similar work was developed by AlGahtani et al. (2009) using a transformation-based learning method, which is an error-driven approach to induce the retagging rules from a training corpus. The corpus used in this experiment was the Arabic Treebank and the morphological analyser BAMA. Based on two different algorithms during the tagging phase, the accuracy achieved 96.9%.

The APT tagger (Khoja, 2001) uses a hybrid technique of statistical and rule-based techniques and a tagset of 131, basically derived from the British National Corpus (BNC) (Leech, 1992) English tagset. An example of a hybrid method was made by a combination of a rule-based and memory-based learning method for tagging Arabic words (Tlili-Guiassa, 2006). This tagger uses a tagset extracted from Khoja's tagger with other new ones added. It was reported that the performance was 85%. Another system was presented by El Hadj et al. (2009) for Arabic PoS tagging that relies on the Arabic sentence structure and combines morphological analysis with HMM. Unlike the previous hybrid PoS tagging systems, a different system was established based on a probabilistic model and a morphological analyser to identify the correct tag in the context; it achieved an accuracy of 94.02% (Ababou and Mazroui, 2015). Finally, the Arabic Morphosyntactic Tagger (AMT) developed by Alqrainy (2008), uses a pattern-based, lexical, and contextual technique. Alqrainy built on traditional Arabic grammar books to design a new PoS tagset called ARBTAGS that followed the criteria proposed by Atwell (2008).

Virtually all Arabic PoS taggers use a tagset derived from English (e.g., (Diab et al., 2004)) or a summary of all Arabic features, which are based on theoretical than practical reasons. Except for the tagger developed by Al Shamsi and Guessoum (2006), other taggers do not generally consider the structure of the Arabic sentence during the tagging process.

3. Existing PoS tagsets

A tagset is a set of tags representing information regarding parts of speech and values of grammatical categories (e.g., case, gender) of word forms.

Several works have been undertaken for developing PoS tagsets and have been implemented by taggers as mentioned previously. Examples of these tagsets are: the tagset used by Khoja et al. (2001), that used by El Hadj et al. (2009), and the Penn Arabic Treebank (PATB) (Maamouri and Bies, 2004) tagset. Moreover, other projects have resulted from the PATB PoS tagset such as Sawalha (2009) and Diab (2007). All these tagsets have been developed for different purposes. In general, however, they were for the enrichment of text corpora with linguistic analyses to maximize their use in a wide range of NLP applications.

In contrast to the majority of existing tagsets, only a small number of works have suggested a tagset for standard use. To the best of our knowledge, these relevant works were developed by Khoja (2001), Alqrainy (2008), and Sawalha (2009), in addition to the suggested universal tagsets (Petrov et al., 2011) and (Rambow et al., 2006) that are meant to be used for multiple languages including Arabic. The number of basic tags used in these tagsets fluctuates from 12 to 114; the number of possible combined tags is composed of over 2000 tag types in the PATB tagset.

Typically, it is difficult to compare tagsets, primarily because every PoS tagger aims at attaining its own objective. However, after a comparative investigation into the mentioned tagsets, we conclude that the PATB tagset is the most appropriate. The PATB tagset addresses important grammatical information; however, it has limitations and requires refinement. Moreover, this tagset is for morphological features rather than PoS tagging and some of its basic tags are more related to semantic categories than morphosyntactic categories.

Based on the PATB tagset, Sawalha (2009) proposed a new fine-grained tagset. Subsequently, additional refinement and expansion were performed on the Sawalha tagset. However, Aliwy (2013) claims that this latter tagset continues to include tags that have more theoretical than practical features.

3.1. Drawbacks of existing tagsets

The previously mentioned tagsets suffer from several problems that we identify below:

- Almost all Arabic PoS taggers use tagsets derived from English, which is inconvenient for Arabic (Diab et al., 2004; Zitouni, 2014);
- The compilation of a tagset does not accurately describe and address all the information regarding Arabic grammar. For example, Hadni (Hadni et al., 2013) used only three tags (Noun, Verb, and Particle) as a tagset;
- Some PoS tagsets are a summary of all Arabic features, which is more theoretical than practical. For example, Atwell (Atwell, 2008) proposed tags that are difficult to determine except if we already know the morphological feature of the verb root "صَحِيح " 'Sound verb' and "مُضَعَّف " 'Doubled verb' or semantically know the sentence context such as "نُونُ الْوِقَايَةُ " 'nūn of protection' and "الْعَاقِل " 'Rational, which express Humanness'.
- The tagsets used for several PoS taggers are not comparable with each other, which does not permit a valid comparison of the accuracy;
- Tagsets lack suitable documentation that illustrates the decision made for each design of its dimensions;
- The most widely used tagsets are based on the proposed recommendations of EAGLES, which are designed for Indo-European languages and are based on Latin as a common ancestor. As Atwell (2008) stated, "Corpus linguists have not attempted to apply EAGLES standards to Arabic, a non-Indo-European language. If they did, the tag set arrived at might well seem alien to Arabic linguists and grammarians". We, as other researchers (Khoja et al., 2001; Ahmad et al., 2006; Gharaibeh and Gharaibeh, 2012) agree with Atwell's claim. Therefore, we suggest that Arabic should have its own tagset, and that the tagset should not be based on the EAGLES guidelines only, but also on specific Arabic grammar features;
- The behaviour of certain categories in Arabic substantially differs from Indo-European languages or certain categories may simply not exist. For example:
  - The Gerund tag in the EAGLES recommendations is considered as a form attribute of the verb; whereas in Arabic, it is a noun subcategory and has, in turn, six subcategories (verbal noun, gerund with initial mïm, gerund of instance, gerund of state, gerund of emphasis, and gerund of profession).
  - For number features, EAGLES uses only Singular and Plural. In addition to these two tags, Arabic uses Dual tag. Moreover, the Plural has six subcategories (sound plural, broken plural, plural of paucity, plural of multitude, ultimate plural, and plural of plural).
- The impossibility of combining diverse taggers for improving accuracy.

4. Suggested PoS tagset

In this study, we were able to design hierarchical levels of an Arabic PoS tagset based on important references in Arabic grammar, such as the Lexicon of Arabic Language Grammar in tables and tablets "معجم قواعد اللغة العربية في جداول ولوحات " by Al-Dahdah (1989), Tatbiq Al-Nahwi "التطبيق النحوي " by Rajhi (2000), and

Mu'jam al-I'rāb wa al-imlā' "معجم الإعراب والإملاء" by Ya'qūb (1983). We also collaborated with experts in this field.

4.1. Criteria for a standard Arabic tagset

We propose recommendations and design criteria for morphosyntactic categories for the Arabic language considering both formal and functional aspects. These recommendations are as follows:

- Traditional Arabic grammar rules: the tagset should follow the Arabic grammatical system rather than those derived from other languages;
- Identifying categories/subcategories: ability to distinguish different levels of word categories for the morphosyntactic tagset. For that reason, we use a hierarchical taxonomy, because traditional Arabic grammarians recognize only three main parts-of-speech that map approximately to Noun, Verb, and Particle. Hence, all PoS tags including EAGLES categories are considered as subcategories of these three main categories. For example, pronoun (ضَمِير)<Damiyr>, adjective (صِفَةٌ مُشَبَّهَة)<SifatN muxab~ahat>, and adverb (ظَرْف)<zarof> are subcategories of noun.
- Target users and/or applications: the PoS tagset should be sufficiently general for different applications;
- Unambiguity: the tagset should be clearly defined;
- Reusability: the tagset should be amenable to be used again by other researchers;
- Extensibility: the tagset should be easily expandable to include more Arabic features, whenever required;
- Processability: it should be possible to use a reduced version of the original tagset based on practical than theoretical reasons;
- Comparability: it should make room for an improved comparison evaluation between different PoS taggers;
- Interchangeability: it should allow forward/backward conversion between the main categories and subcategories.

4.2. Suggested PoS tagset

The Arabic PoS inventory consists of three main categories (Al-Dahdah, 1989; Ghalayini, 2013): Noun (إسْم)<Aisom>, Verb (فِعْل)<fiEol>, and Particle (حَرْف)<Harof>. Each one of these categories has many subcategories.

Badawi et al. (2013) claimed that nouns are all those elements with nominal inflection or function (including indeclinable forms). The noun category also includes adjectives and adverbs (which are formally nouns in particular functions), demonstratives, relatives, and pronouns of all types (which are nouns in status yet not in form). Verbs are all those elements with verbal inflection, including fossilized items. Because they incorporate an agent pronoun, they may stand alone as complete sentences. Finally, particles are morphologically indeterminate and can only be defined by their function. They are frequently bound and comprise all the bound morphemes not included in the other two categories, such as prepositions, and conjunctions. Particles are uninflected and devoid of number, gender, and definiteness.

4.2.1. Noun Part-of-Speech

A noun conveys lexical meaning (as opposed to grammatical meaning) but gives no indication of time. Nouns are classified according to five main subcategories: definiteness (definite or indefinite), number (Singular, Dual, Plural), gender (Masculine, Feminine), inflection (Derived or Primitive) and declension (Declined or Invariable). Each class contains several tags. Fig. 1 presents the hierarchical levels of the noun categories and their tags.

4.2.2. Verb Part-of-Speech

A verb conveys lexical meaning (as opposed to grammatical meaning), and gives indication of time. Verbs are classified according to two main subcategories: the perfect verb (which consists of another six subcategories: declension, conjugation, inflection, transitivity, voice, and vocalic) and the imperfect verb (which consists of another four subcategories: declension, conjugation, inflection, and vocalic). Further, each category consists of several tags. Fig. 2 displays the hierarchical levels of the verb categories and their tags.

4.2.3. Particle Part-of-Speech

A particle conveys no lexical meaning, which is conveyed through the parts of speech it relates to. Particles are classified according to two main subcategories: common and specific. The specific class consists of two others subcategories (specific to noun and specific to verb). Moreover, each class consists of several tags. Fig. 3 displays the hierarchical levels of the particle categories and their tags.

4.3. Discussion

Many surveys and comparative studies have been completed for relevant approaches for tagging and PoS tagsets. It is not easy to compare and determine the best tagset, primarily because every PoS tagger addresses its specific objectives. This methodology of work creates two major problems. The first is the absence of an efficient standard PoS tagset for Arabic that can unify the efforts made in this field. The second is the difficulty of benefitting from different PoS taggers simultaneously.

Some Arabic PoS tagging systems use tagsets derived from English, which make them not very well suited to address the long-established Arabic language features. Other PoS taggers used tagsets based only on a summary of Arabic grammatical rules, including tags that have more theoretical than practical features. To overcome these weaknesses and other mentioned in Drawbacks Section 3.1, the proposed tagset has 110 basic tags classified into four different levels (see Table 1), which accurately describe and address Arabic language features considering both formal and functional aspects.

In contrast with other proposed Arabic tagsets, we suggest an appropriate PoS tagset for Arabic based on the standardization criteria we mentioned above. These criteria require designing detailed hierarchical levels. Furthermore, our work considers the morphological complexity of Arabic, allowing the proposed tagset to address both Modern Standard Arabic and Classical Arabic. The proposed hierarchical levels allow the tagset to be easily expandable to refine and include additional Arabic features, whenever required.

5. Usability test of the tagset

In this section, we highlight the use of the proposed tagset via various experiments on text from both Modern Standard Arabic and Classical Arabic. In this regard, a language independent PoS tagger (called TreeTagger) was implemented.

5.1. TreeTagger

TreeTagger is a decision tree based tagger for annotating text with PoS tags and lemma information. It was developed by Schmid (1995)). This tagger has been officially used to tag more than 20 different languages other than Arabic (TreeTagger, n.d.); however, it is adaptable to other languages if a lexicon and a tagged training corpus are available. Fortunately, a language model

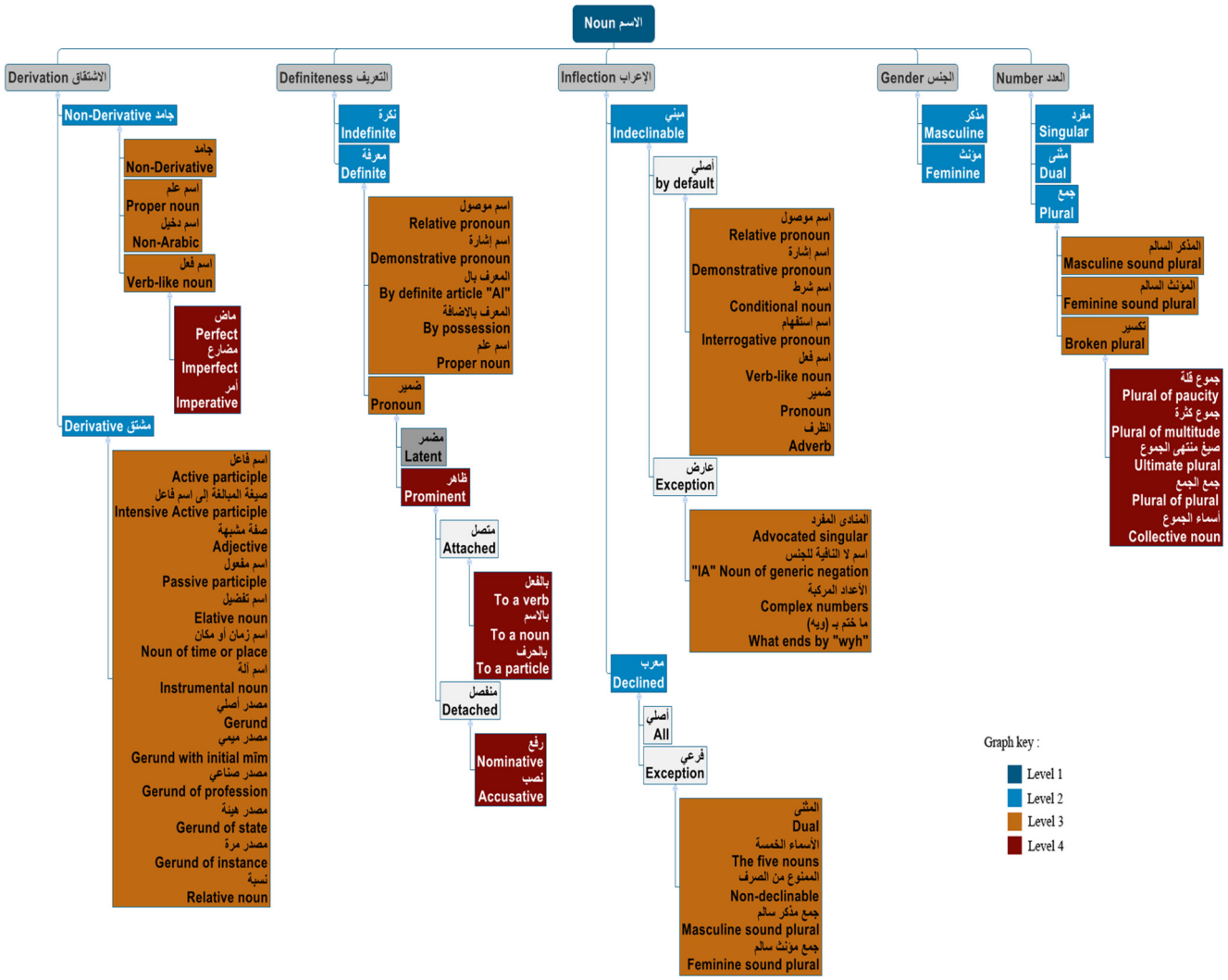

Fig. 1. Hierarchical levels of noun categories.

was created recently to adapt TreeTagger for Arabic (Zeroual and Lakhouaja, 2016a). Nevertheless, a manual system was developed to convert the tagset used in TreeTagger to the proposed hierarchical levels and to convert the tagset from one level to another.

5.2. Experiments and discussion

The performance of TreeTagger was tested on data from the NEMLAR (Attia et al., 2005) and Al-Mus'haf (Zeroual and Lakhouaja, 2016b) corpora. Ninety percent of the words were used for the training phase and the remaining 10% of the words were used for testing. In this section, we focus on various important results that we achieved during our experiments. Notice that, in addition to the main three tags (Noun, Verb, and Particle) on Level 1, there are two others for punctuation and non-Arabic words. Table 2 provides the different results of tagging accuracy for each level of our suggested tagset.

The tagging process achieved satisfactory results for different forms of Arabic text and for each level of the suggested tagset. Moreover, tagging with a complex tagset did not cause a sharp degradation of the accuracy. In fact, accuracy increased when more complex tagsets were used as it is the case for Al-Mus'haf corpus (from Level 3 to Level 4) and for NEMLAR corpus (from previous levels to Level 4). Consequently, we concluded that the ambiguity in the PoS tagging process begins to increase after Level 1 and decreases on complex levels such as Level 4. Furthermore, the probability of a given trigram is determined by following the corresponding path through the tree until a leaf is reached. This means that if we attempt to obtain the probability of a particular tag, we must first answer the test at the root node. For this reason, a change of the tagset has a significant impact on the training process of TreeTagger. For example, the probability of a tag preceded by a Verb (VERB) and a Particle (PRT) changes from one level to another. Table 3 presents an example of a probability change for the word "فهم" <fhm>.

Table 3 presents the probabilities of a tag preceded by a verb and a particle. Even for the same word, such as "فهم" <fhm>, this can easily change based on the level adopted. Consequently, the change of adopted tagsets significantly affects the probabilities estimated by TreeTagger during the training process, which is reflected directly in the accuracy results.

Based on an investigation of PoS tagging for Arabic, Kübler and Mohamed (2012) believed that using a complex tagset and then converting the resulting annotation to a smaller tagset provides a higher accuracy than tagging using the smaller tagset directly. Fortunately, the suggested hierarchical levels also allow a similar investigation. Thus, an experiment was performed to test Kübler's investigation results. Table 4 describes in detail a comparison

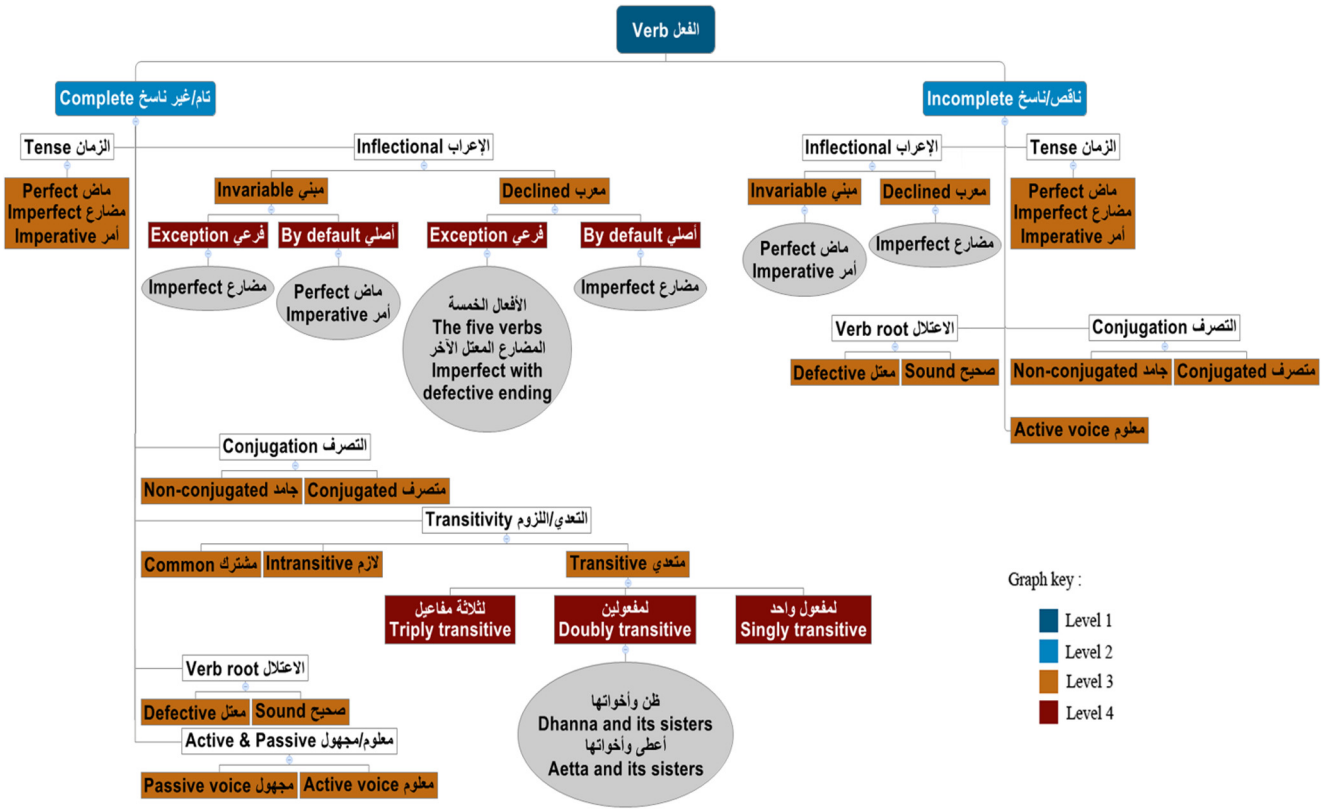

Fig. 2. Hierarchical levels of verb categories.

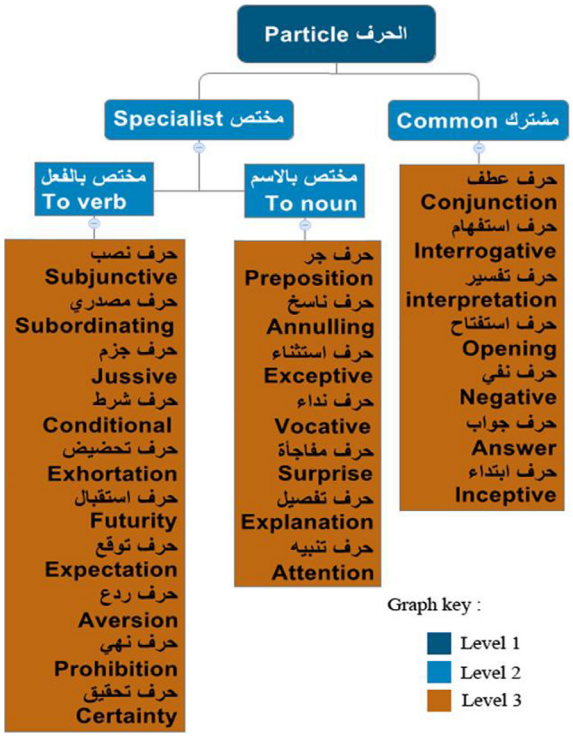

Fig. 3. Hierarchical levels of particle categories.

between the results achieved using the smaller tagset directly and the results achieved by converting the resulting annotation with the complex tagset to the smaller tagset.

All the conversion processes confirmed a marginal improvement in the accuracy, supporting Kübler's investigation. More precisely, this conversion improved the performance of PoS tagging from 0.01% (2 → 1 in Nemlar corpus) to 1.55% (4 → 1 in Al-Mus'haf corpus). This demonstrates that using rich morphosyntactic information in the PoS tagset is not necessarily an obstacle for PoS tagging. On the contrary, they may improve tagging accuracy. The reason for this may be that complex tagset precisely describes the distributional features of words. For example (see Table 3), the full tag (NOUN_Pronoun.3rd-person) describes the word's characteristics better than the simple tag (NOUN).

In addition to the previous experiments, we added an analysis regarding the level of PoS tagging ambiguity. Ambiguity can exist between the main categories (Noun, Verb, and Particle) and between subcategories of the same main category. We believe that the error rate is more acceptable during the tagging process if it is between the subcategories than if it is between the main categories. Table 5 exhibits the rate of ambiguity that is not solved during the tagging process.

Table 5 emphasizes that ambiguity exists with a high degree between the subcategories of the hierarchical levels (Levels 2, 3 and 4) in comparison with the main categories (Level 1).

5.3. Discussion

We have addressed several examples of Arabic PoS taggers (six as statistical taggers, two as rule-based taggers and five as hybrid taggers). However, this number of PoS taggers is very low when compared to other languages (e.g., English, French, and Spanish). Yet, only three works (Khoja et al., 2001; Maamouri and Bies, 2010; Sawalha and Atwell, 2010) have suggested a tagset for standard use.

Table 1
Basic tags of proposed tagset.

Levels	Number of basic tags			
	Noun	Verb	Particle	
Level 1	1	1	1	
Level 2	11	4	3	
Level 3	33	14	25	
Level 4	10	7	0	
Total	55	26	29	110

Table 2
Tagging accuracy analysis.

Corpora	Total words	Levels	Number of tags	Accuracy (%)
Al-Mus'haf	78,121	1	4	97.18
		2	26	94.02
		3	79	91.35
		4	95	91.65
NEMLAR	500,000	1	5	97.15
		2	12	93.86
		3	63	95.74
		4	107	97.55

Table 3
Example of probability change through levels.

Levels	Sentences in training data	Tags for the word "فهم "	Probability (%)
1	إنْ جَاؤُوا فَهُمْ . . .	NOUN	60
	إنْ كَانَ فَهُمْ . . .	NOUN	
	إنْ كَانَ فَهِمْ . . .	NOUN	
	إنْ اسْتَوْعَبَ فَهِمَ . . .	VERB	40
	إنْ شَاءَ فَهِمَ . . .	VERB	
3	إنْ جَاؤُوا فَهُمْ . . .	NOUN_PRON.3P*	20
	إنْ كَانَ فَهُمْ . . .	NOUN_V.GERN**	20
	إنْ كَانَ فَهِمْ . . .	NOUN_ADJ	20
	إنْ اسْتَوْعَبَ فَهِمَ . . .	VERB_PER.AC***	40
	إنْ شَاءَ فَهِمَ . . .	VERB_ PER.AC***	

* Noun_pronoun.3rd-person.
** Noun_Verbal-noun.Gerund.
*** Verb_Perfect.Active-voice.

Our experiments confirm that the implementation phase of the proposed tagset in the PoS tagging process using TreeTagger is satisfactory. Furthermore, the designed hierarchical levels allow various tests without significant degradation of the accuracy for both Classical and Modern Standard Arabic text. Moreover, these hierarchical levels improve the accuracy because of the interchangeability between the levels, which facilitates the conversion process.

Regarding ambiguity, these hierarchical levels, which identify the main categories and their subcategories through different levels, accurately determine where the ambiguity intensifies. Therefore, this tagset can be easily implemented in PoS tagging; it also improves the accuracy owing to the interchangeability and extensibility of its designed hierarchical levels, which are based on carefully chosen standard design criteria.

Table 5
Ambiguity between main categories and subcategories.

Corpora	Ambiguity between main categories (%)	Ambiguity between subcategories (%)
Al-Mus'haf	1.27	7.08
NEMLAR	1.97	3.30

6. Conclusion

In this paper, we highlighted the importance of PoS tagging for NLP. We presented approaches used for PoS tagging of Arabic text and the most relevant PoS taggers. Furthermore, we discussed the well-known tagsets used in this field and their drawbacks. Then, we suggested a range of criteria for a tagset to be generalized and standardized, which can be used in the process of PoS tagging different text types such as Classical Arabic and Modern Standard Arabic.

Table 4
Tagging accuracy with converting process.

Corpora	Levels	Accuracy in direct tagging	Converting process	New accuracy (%)
Al-Mus'haf	1	97.18%	2 → 1	98.20
			3 → 1	98.31
			4 → 1	98.73
	2	94.02%	3 → 2	94.12
			4 → 2	94.45
	3	91.35%	4 → 3	91.69
NEMLAR	1	97.15%	2 → 1	97.16
			3 → 1	97.47
			4 → 1	98.03
	2	93.86%	3 → 2	94.18
			4 → 2	94.86
	3	95.74%	4 → 3	96.49

Our methodology is based on the fact that Arabic has many morphological and grammatical features. Thus, the purpose of this study was to target the finest possible PoS tagset for Arabic and to increase the accuracy of PoS taggers. Hence, we designed detailed hierarchical levels of the Arabic tagset categories that capture long-established Arabic grammar distinctions and facilitate expansion to include more Arabic tags when required.

The usability of the proposed tagset was verified using TreeTagger. Although the results demonstrated a satisfactory accuracy, the conversion processes improved the performance of the PoS tagging which lends support to Kübler's investigation.

This work is an advancement in the direction of achieving a standard, rich, and comprehensive tagset for Arabic. We evaluated this tagset using new data to refine it as required.

References

Ababou, N., Mazroui, A., 2015. A hybrid Arabic POS tagging for simple and compound morphosyntactic tags. Int. J. Speech Technol., 1–14

Ahmad, K., Cheng, D., Almas, Y., 2006. Multi-lingual sentiment analysis of financial news streams. In: Proc. of the 1st Intl. Conf. on Grid in Finance.

Al Shamsi, F., 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.

Albared, M., Omar, N., Ab Aziz, M.J., 2011. Developing a competitive HMM arabic POS tagger using small training corpora. Intelligent Information and Database Systems. Springer, pp. 288–296.

Al-Dahdah, A., 1989. The grammar of the Arabic language in tables and lists. Beirut Maktabat Lebnan Arab.

AlGahtani, S., Black, W., McNaught, J., 2009. Arabic part-of-speech tagging using transformation-based learning. In: Proceedings of the Second International Conference on Arabic Language Resources and Tools, Cairo, Egypt.

Aliwy, A.H., 2013. Arabic Morphosyntactic Raw Text Part of Speech Tagging System. Repozytorium Uniwersytetu Warszawskiego.

Alqrainy, S., 2008. A morphological-syntactical analysis approach for Arabic textual tagging. De Montfort University.

Al-Sughaiyer, I.A., Al-Kharashi, I.A., 2004. Arabic morphological analysis techniques: a comprehensive survey. J. Am. Soc. Inf. Sci. Technol. 55, 189–213.

Altabba, M., Al-Zaraee, A., Shukairy, M.A., 2010. An Arabic morphological analyzer and part-of-speech tagger. Fac. Inform. Eng. Arab Int. Univ Damascus Syr. Thesis Present.

Attia, M., Yaseen, M., Choukri, K., 2005. Specifications of the Arabic Written Corpus produced within the NEMLAR project.

Atwell, E.S., 2008. Development of tag sets for part-of-speech tagging.

Badawi, E.S., Carter, M., Gully, A., 2013. Modern written Arabic: a comprehensive grammar. Routledge.

Banko, M., Moore, R.C., 2004. Part of speech tagging in context. In: Proceedings of the 20th International Conference on Computational Linguistics. Association for Computational Linguistics, p. 556.

Boudchiche, M., Mazroui, A., Ould Abdallahi Ould Bebah, M., Lakhouaja, A., Boudlal, A., 2017. AlKhalil Morpho Sys 2: A robust Arabic morpho-syntactic analyzer. J. King Saud Univ. - Comput. Inf. Sci. 29, 141–146.

Buckwalter, T., 2004. Buckwalter Arabic Morphological Analyzer Version 2.0. Linguistic Data Consortium, University of Pennsylvania, 2002. LDC Cat alog No.: LDC2004L02. ISBN 1-58563-324-0.

Darwish, K., 2002. Building a Shallow Morphological Analyzer in One Day. In: Proc. ACL Workshop Comput. Approaches Semit. Lang.

Diab, M., 2007. Towards an optimal POS tag set for Modern Standard Arabic processing. In: Proceedings of Recent Advances in Natural Language Processing (RANLP). pp. 91–96.

Diab, M.T., 2007. Improved Arabic base phrase chunking with a new enriched POS tag set. In: Proceedings of the 2007 Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources. Association for Computational Linguistics, pp. 89–96.

Diab, M., Hacioglu, K., Jurafsky, D., 2004. Automatic tagging of Arabic text: From raw text to base phrase chunks. In: Proceedings of HLT-NAACL 2004: Short Papers. Association for Computational Linguistics, pp. 149–152.

El Hadj, Y., Al-Sughayeir, I., Al-Ansari, A., 2009. Arabic part-of-speech tagging using the sentence structure. In: Proceedings of the Second International Conference on Arabic Language Resources and Tools, Cairo, Egypt.

Farghaly, A., Shaalan, K., 2009. Arabic natural language processing: Challenges and solutions. ACM Trans. Asian Lang. Inf. Process. TALIP 8, 14.

Freeman, A., 2001. Brill's POS tagger and a Morphology parser for {Arabic}.

Ghalayini, M.I.M.S., 2013. Jami'al-durus al-'arabiyah. Turath For Solutions.

Gharaibeh, I.K., Gharaibeh, N.K., 2012. Towards Arabic Noun Phrase Extractor (ANPE) using information retrieval techniques. Int. J. Softw. Eng. 2, 36–42. http://dx.doi.org/10.5923/j.se.20120202.04.

Habash, N., Rambow, O., 2005. Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, pp. 573–580.

Habash, N., Sadat, F., 2006. Arabic Preprocessing Schemes for Statistical Machine Translation. In: Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers, NAACL-Short '06. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 49–52.

Habash, N., Rambow, O., Roth, R., 2009. MADA+ TOKAN: A toolkit for Arabic tokenization, diacritization, morphological disambiguation, POS tagging, stemming and lemmatization. In: Proceedings of the 2nd International Conference on Arabic Language Resources and Tools (MEDAR), Cairo, Egypt, pp. 102–109.

Hadni, M., Ouatik, S.A., Lachkar, A., Meknassi, M., 2013. Hybrid Part-Of-Speech Tagger for Non-Vocalized Arabic Text. Int. J. Nat. Lang. Comput. 2, 1–15.

Khoja, S., 2001. APT: Arabic part-of-speech tagger. In: Proceedings of the Student Workshop at NAACL. pp. 20–25.

Khoja, S., Garside, R., Knowles, G., 2001. A tagset for the morphosyntactic tagging of Arabic. Comput. Dep. Lanc. Univ.

Kübler, S., Mohamed, E., 2012. Part of speech tagging for Arabic. Nat. Lang. Eng. 18, 521–548.

Leech, G., 1992. 100 million words of English: the British National Corpus (BNC). Lang. Res. 28, 1–13.

Maamouri, M., Bies, A., 2004. Developing an Arabic treebank: Methods, guidelines, procedures, and tools. In: Proceedings of the Workshop on Computational Approaches to Arabic Script-Based Languages. Association for Computational Linguistics, pp. 2–9.

Maamouri, M., Bies, A., 2010. The Penn Arabic Treebank.

Maamouri, M., Cieri, C., 2002. Resources for Arabic Natural Language Processing. In: International Symposium on Processing Arabic.

Pasha, A., Al-Badrashiny, M., Diab, M., El Kholy, A., Eskander, R., Habash, N., Pooleery, M., Rambow, O., Roth, R.M., 2014. Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of arabic. In: Proceedings of the Language Resources and Evaluation Conference (LREC), Reykjavik, Iceland.

Petrov, S., Das, D., McDonald, R., 2011. A universal part-of-speech tagset. ArXiv Prepr. ArXiv11042086.

Rajhi, A., 2000. Tatbiq Al-Nahwi (التطبيق النحوي). Dar Annahda Al Arabia for printing, publishing and distribution.

Rambow, O., Dorr, B., Farwell, D., Green, R., Habash, N., Helmreich, S., Hovy, E., Levin, L., Miller, K.J., Mitamura, T., others, 2006. Parallel syntactic annotation of multiple languages. In: Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC2006). Genoa, Italy.

Sawalha, M., 2009. Arabic Morphological Features Tag set.

Sawalha, M., Atwell, E.S., 2010. Fine-grain morphological analyzer and part-of-speech tagger for Arabic text. In: Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10). European Language Resources Association (ELRA), pp. 1258–1265.

Schmid, H., 1995. Treetagger| a language independent part-of-speech tagger. Inst. Für Maschinelle Sprachverarbeitung Univ. Stuttg. 43, 28.

Tlili-Guiassa, Y., 2006. Hybrid method for tagging Arabic text. J. Comput. Sci. 2, 245–248.

Toutanova, K., Manning, C.D., 2000. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. Proceedings of the 2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora: Held in Conjunction with the 38th Annual Meeting of the Association for Computational Linguistics, Vol. 13. Association for Computational Linguistics, pp. 63–70.

TreeTagger [WWW Document], n.d. URL http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/ (accessed 12.8.15).

Viterbi, A.J., 1967. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. Inf. Theory IEEE Trans. 13, 260–269.

Ya'qūb, I., 1983. Mu'jam al-I'rāb wa al-imlā' (معجم الإعراب والإملاء). Dar El Ilm Lilmalayin.

Zeroual, I., Lakhouaja, A., 2016a. Adapting a decision Tree based Tagger for Arabic. In: Presented at the 2nd International Conference on Information Technology for Organisation Development, IEEE, Fez.

Zeroual, I., Lakhouaja, A., 2016b. A new Quranic Corpus rich in morphosyntactical information. Int. J. Speech Technol. 19, 339–346.

Zitouni, I. (Ed.), 2014. Natural language processing of semitic languages, theory and applications of natural language processing. Springer, Berlin, Heidelberg.