Lexical and Morphological Statistics of an Arabic POS-Tagged Corpus

Hamdy Mubarak[*1], Kareem Shaban[*2], Forat Adel[*3]

[*]*Arabic NLP Researches, Sakhr Software,*

Cairo, Egypt

[1]hamdys@sakhr.com

[2]kshaban@sakhr.com

[3]forat@sakhr.com

Abstract: ***Part-Of-Speech (POS) tagging is a basic component necessary for many Natural Language Processing (NLP) applications. Building a manually tagged corpus helps in studying key statistics of a given language which form the basis for POS tagging systems. In this paper, we present both lexical and morphological statistics for Arabic that are derived from the Sakhr's POS manually tagged corpus. It covers text (7 M words) from a wide range of Arab countries in different domains over the years 2002-2004. The derived statistics are used as heuristics and preferential rules within a statistical Diacritizer which achieves a high accuracy in stem diacritization and POS disambiguation. Statistics includes information related to sentence and word lengths, punctuation marks, distribution of Arabic letters and diacritics, in addition to lexical and morphological information for POS distribution, stems, prefixes, suffixes, roots, morphological patterns, and morphosyntactic features like gender, number, person, and case ending. Modern Standard Arabic (MSA) is studied by analyzing the coverage of stems, roots, morphological patterns, prefixes, and suffixes. Comparisons with an arbitrary English corpus are shown in applicable cases.***

Key words: ***Corpus Statistics, Arabic NLP, POS Tagging, Diacritization, MSA.***

1 INTRODUCTION

Part-Of-Speech (POS) tagging is assigning a specific tag to each word of a sentence to indicate its function in the specific context [1]. POS tagging is considered as one of the basic components necessary for any robust Natural Language Processing (NLP) infrastructure [2], and it is needed in many tasks such as syntax and semantic analysis, text to speech (TTS), natural language parsing, information retrieval (IR), information extraction (IE), and machine translation (MT) [3]. A manually tagged corpus can be used for innumerable studies of word-frequency and POS. It also inspires the development of similar "tagged" corpora. Statistics derived by analyzing such corpus formed the basis of the latest POS tagging systems.

In this paper we will describe many lexical and morphological statistics that are derived from Sakhr's Arabic manually POS-Tagged corpus (POST) hand tagged by human annotators. These statistics include POS distribution, usage of stems, prefixes, suffixes, roots, morphological patterns, and also the usage of morphosyntactic features like gender, number, person, case ending, etc.
The benefits of these statistics were gained when they are considered as heuristics and preferential rules while building a Statistical Diacritizer which successfully disambiguates Arabic sentences by selecting the appropriate morphological analysis including POS, stem diacritics and morphosyntactic features. This Diacritizer also suggests the final case ending for each word which represents the syntactic function of words in context.

A comparison between Arabic and English corpora is conducted which considered some aspects like sentence length, word length, unique words, and punctuation marks. As a matter of fact, POST had a significant impact on training the statistical diacritizer's models whose stem diacritization and POS disambiguation accuracy reached 97%, and final case ending diacritization reached 92%.

This paper is organized as follows: Section 2 is a brief introduction to some aspects of Arabic language. Sections 3 and 4 describe Sakhr's morphological analyzer and POST. Sections 5 through 17 present detailed language statistics. Finally, section 18 gives some concluding remarks.

2 ARABIC LANGAUGE

Arabic is one of the six official languages of the United Nations and the mother tongue of more than 300 million people. It is the official language in 25 countries (also widely studied and used throughout the Islamic world), and the third most after English and French. Arabic is the largest living Semitic language whose main characteristic feature is that most

words are built up from roots by following certain fixed morphological patterns (which specify the vowels that can follow each consonant of root letters) and adding infixes, prefixes and suffixes. Arabic includes 28 letters and it is written cursively from right to left [4]. Arabic morphology is rather complex because of the morphological variation and the agglutination phenomenon. Letters change forms according to their position in the word (beginning, middle, end and separate) [5]. The modern form of Arabic is called Modern Standard Arabic (MSA), which is a simplified form of Classical Arabic, and it is the form used by all Arabic-speaking countries in publications, workplaces, government and media [6]. MSA is very often written without diacritics, which leads to a highly ambiguous text. Arabic readers could differentiate between words having the same writing form (homographs) by the context of the script [7].

3 MORPHOLOGICAL ANALYSIS

Sakhr's Morphological Analyzer is a morphological analyzer-synthesizer that provides basic analyses of a single Arabic word, covering the whole range of modern and classical Arabic. For each analysis, it provides its morphological data such as stem, root, morphological pattern, POS, prefixes, suffixes and also its morphosyntactic features like gender, number, person, case ending, etc. In addition to its high accuracy (99.8%), the Morphological Analyzer sorts the word analyses according to the usage frequency (using manual ordering of analyses for commonly-used words as appeared in an Arabic corpus of 4G words, or ordering according to stem frequency, otherwise). This morphological analyzer is integrated in most Sakhr products like TTS, MT, Search Engine and Text Mining.

4 ARABIC POS-TAGGED CORPUS

POST includes texts (from newspapers, news services, and magazines) from different Arabic-speaking countries in different domains (Politics, Economy, Sport, Religion, Science, Medicine, etc) over the years 2002-2004. The corpus size is about 7M words (~330K sentences).

In our study of Arabic spelling mistakes in newspapers, we found out that Common Arabic Mistakes (CAM) occur in initial Hamza, final Taa Marbuta, and final dotted Yaa with a percentage varying from 1% to 12%, with an average of 5% of words. So, preprocessing of Arabic text is necessary, before tagging process takes place, in order to correct and normalize Arabic text by removing diacritics and irrelevant characters.

For each word in a sentence and based on its surrounding context, human annotators select the appropriate morphological analysis from all analyses generated by the Morphological Analyzer for this word, and also determine the final case ending based on this context. Out-Of-Vocabulary (OOV) words and wrong analyses are also flagged during the tagging process and this gave a great feedback to the lexicon, proper nouns, and corrector databases.

For a comparison with an English corpus, we selected texts with same size (7M words) from famous news agencies.

Figure 1 shows the sentence length distribution in both Arabic and English corpora. The average length of sentence is 21 words in Arabic and 19 in English. In 95% of the cases, sentence length is in the range 2-37 words in Arabic and 2-42 in English.

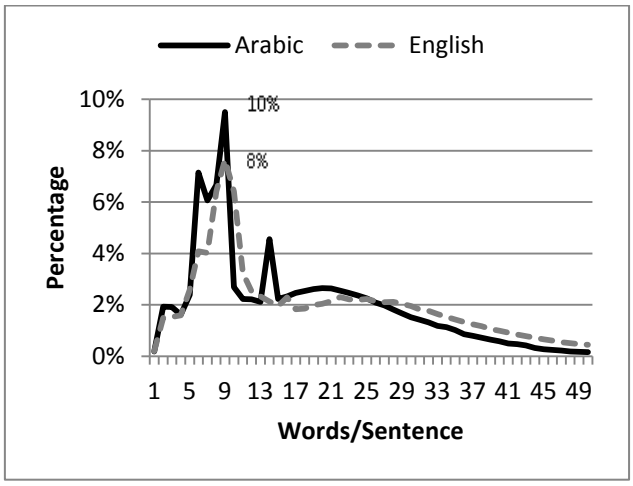

Figure 1: Sentence Length Distribution in Arabic and English

5 SENTENCE AND WORD LENGTHS

On the other hand, Figure 2 shows the word length distribution (in characters) in Arabic and English. The average length of word is 5 Characters in Arabic and 3 in English.

In 95% of the cases, word length is in the range 2-9 characters in Arabic and 2-11 in English.

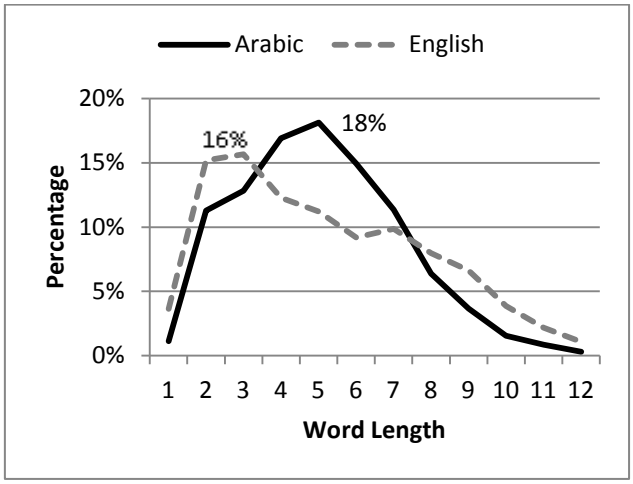

Figure 2: Word Length Distribution in Arabic and English

6 ARABIC LETTERS

Note: Buckwalter Arabic transliteration scheme (http://www.qamus.org/transliteration.htm) is used in all applicable cases.

Figure 3 shows the distribution of Arabic letters. It is notable that, in any Arabic document, only 2 letters ("ا *A*" and "ل *l*") represent 26% of the existing letters, and 6, represent 50%. These 6 letters are ("ا *A*", "ل *l*", "ي *y*", "م *m*", "ن *n*" and "و *w*") and they are used in the definite article ("ال *Al*"), long vowels ("ا *A*", "و *w*" and "ي *y*"), and the letters ("م *m*" and "ن *n*") that are frequently used in some function words and commonly in others.

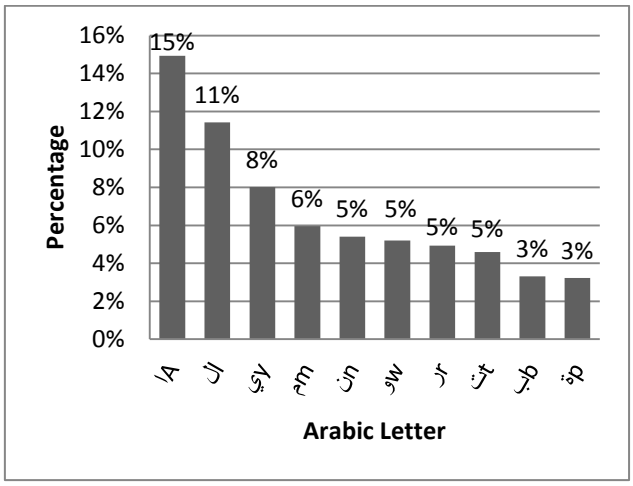

Figure 3: Most Frequent 10 Letters

7 UNIGRAMS IN ARABIC AND ENGLISH

Unigrams represent how frequent a certain token has been written in a corpus. Arabic has a larger number of unigrams because Arabic has a very rich and complex morphology than English [7]. Moreover, the concatenation of affixes (prefixes and suffixes) with stems generates new unigrams. Figure 4 shows the distribution of unique words (unigrams).

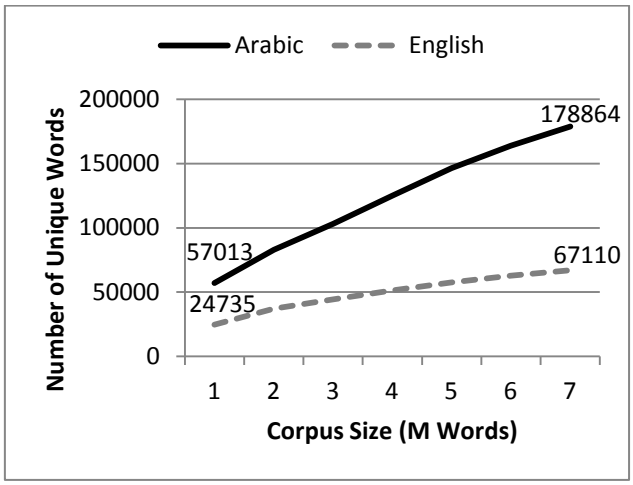

Figure 4: Number of Unique Words in Arabic and English

Table I shows the most frequent 20 words in Arabic and English corpora in addition to the percentage of appearance. It is observed that the majority of these words is function words (prepositions represent ~9%) and have no direct relation with the idea of the document. However, they play a significant role in binding words together.

TABLE 1: MOST FREQUENT 20 WORDS IN ARABIC AND ENGLISH

Arabic		English	
word	**%**	**word**	**%**
fy في	3.55	the	5.1
mn من	2.09	of	2.59
>n أن	1.4	in	2.36
ElY على	1.4	to	2.18
<lY إلى	1.06	and	1.9
<n إن	0.61	a	1.38
En عن	0.58	that	1.25
Alty التي	0.54	for	0.73
wqAl وقال	0.41	on	0.73
mE مع	0.4	The	0.57
*Al*y* الذي	0.36	is	0.57
bEd بعد	0.29	with	0.51
*h*h* هذه	0.28	said	0.47
byn بين	0.26	by	0.42
qd قد	0.25	as	0.4
*h*A* هذا	0.24	was	0.38
lA لا	0.24	it	0.36
mA ما	0.23	from	0.35
lm لم	0.18	an	0.31
>nh أنه	0.18	not	0.31

8 PUNCTUATION MARKS

One of the most useful features in detecting sentences boundaries and tokens is punctuation marks. Unfortunately, writers do not pay attention to punctuation marks usage in Arabic, and they are considered by some as redundant

cosmetic marks [7]. Figure 5 shows punctuation marks distribution in Arabic and English. It is remarkable that Arabic documents are full of inconsistent styles of punctuation marks like two consecutive commas, mixing of single and double quotations, two consecutive question marks, and incorrect representation of period as a zero digit.

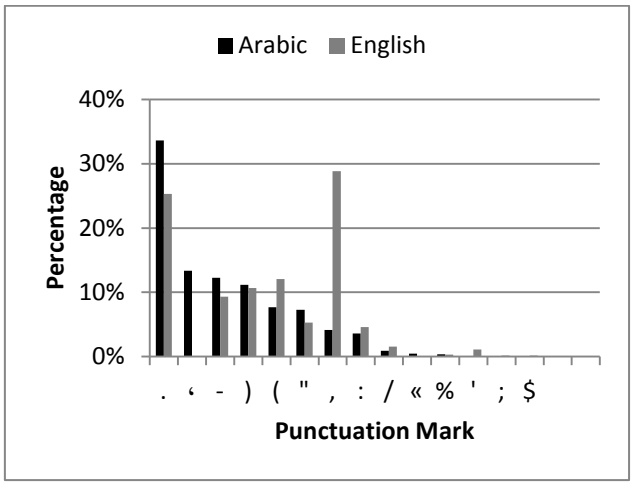

Figure 5: Punctuation Marks in Arabic and English

9 MSA AMBIGUITY

Short vowels are indicated by diacritics and are very often omitted from the modern writing style. It can be easily observed that MSA tends to be simpler than the Classical Arabic in grammar usage, syntax structure, morphological and semantic ambiguity. This will help normal Arabic readers to understand the written text. For example, 69% of words in the Arabic corpus have only 1 identified morphological analysis (one morphological interpretation), and 19% have 2 analyses, while high ambiguous words (3+ analyses) represent 12% only as shown in Figure 6.

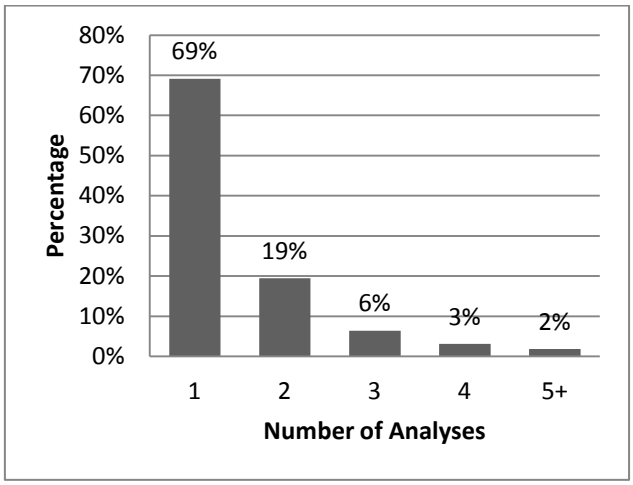

Figure 6: Distribution of Number of Word Analyses

Because Sakhr’s Morphological Analyzer provides an ordered list of analyses according to usage frequency, it was discovered that 92% of words occupy the first position in analyses, and 5% occupy the second one as shown in Figure 7, which means that MSA in most cases is not so ambiguous, and words occupy the “trivial” analysis. For example, the word “للحاكم *llHAkm*” has more than one analysis (لِلْحَاكِمِ *liloHaAkimi*, to/of/for the ruler, لِلِحَاكُمْ *liliHaAkumo*, to/of/for your beards, etc), but the first one is usually recognized.

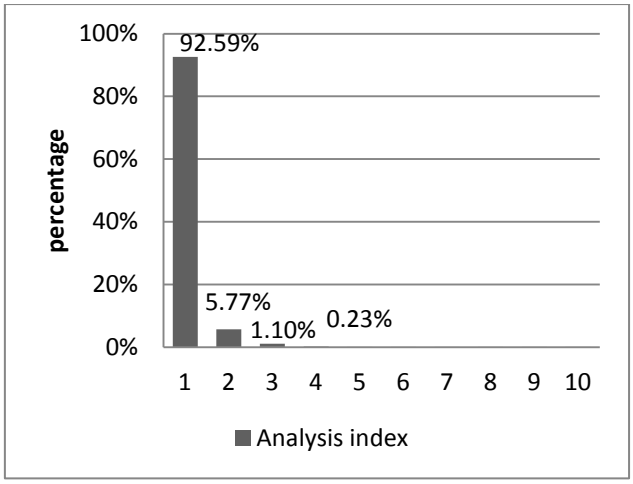

Figure 7: Distribution of the Selected Analysis Index

Figure 8 shows the relation between the word length and its morphological ambiguity (number of analyses). On the average, an Arabic word has 1.5 analyses, and in the extreme cases when length of word is too short (1 character) or too long (15+ characters), it tends to have only one analysis.

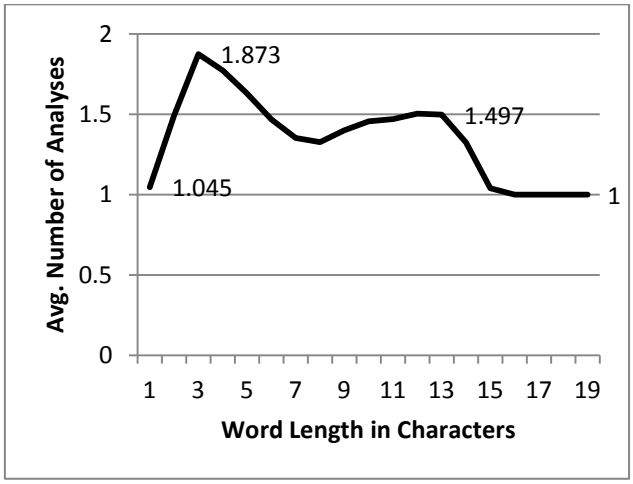

Figure 8: Morphological Ambiguity and Word Length

10 POS DISTRIBUTION

Arabic grammarians traditionally analyze all Arabic words into three main parts of speech or categories, which are further sub-categorized and collectively cover the whole Arabic language [6]. These parts are: **Noun** (a name or a word that describes a person, thing, or idea), **Verb** (a word that denotes an action), and **Particle** (anything else, includes prepositions, adverbs, conjunctions, interrogative particles, exceptions, and interjections). Figure 9 shows the POS distribution after manual POS disambiguation of the Arabic Corpus.

It is notable that nouns represent 62% of POS, verbs represent 10%, while particles represent 28%. In addition, the usage of imperative verbs and passive voice of past and present verbs is rare in MSA (less than 1%), and they are usually replaced by less ambiguous words and structures. For example, instead of writing the ambiguous passive verb in the sentence "اُفتُتح المشروع *AfttH Alm$rwE*" (was-inaugurated the-project), another simple structure is used " تم افتتاح المشروع *tm AfttAH Alm$rwE*" (has-been inaugurating the-project).

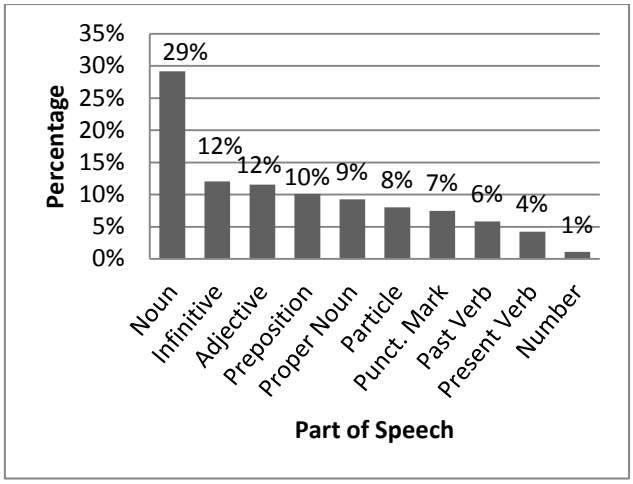

Figure 9: Most Frequent 10 POS's

In the following sections, we will describe some of the lexical and morphological statistics that are derived from POST after assigning each word in a sentence to its appropriate morphological analysis based on its context. The morphological analysis includes information about stem (which is divided more into root and morphological pattern), affixes (prefixes and suffixes), and morphosyntactic features (like the gender, number, person, case ending, etc.)

11 STEM DISTRIBUTION

Most Arabic words are morphologically derived from a list of roots; it can be tri-, quad-, or pent-literal. Most of these roots are tri-literal. Arabic words may have no root (for the majority of function words, some of proper nouns and borrowed words). Figure 10 shows the distribution of root types. This figure shows that quad-literal roots are rarely used in MSA.

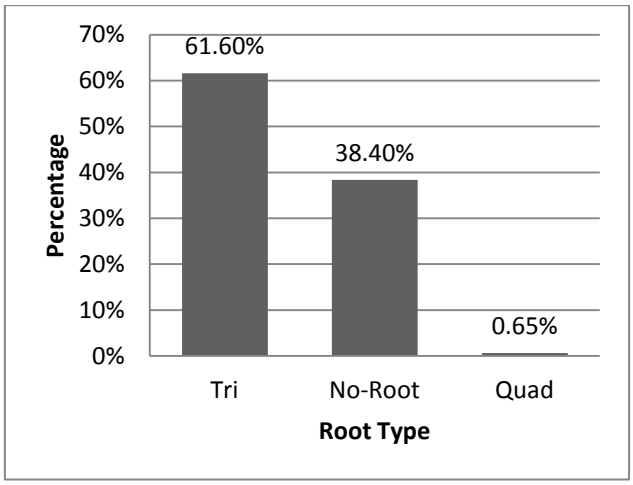

Figure 10: Root Type Distribution

Figures 11 and 12 show the most frequent roots, and morphological patterns, respectively. The most frequent roots used during this period of time were "رءس *r's*" and "عرق *Erq*" because of the events that were happening in " العراق *AlErAq*, Iraq" and their effect on most of the publications and media. The most frequent morphological patterns are both "فَعْل *faEol*" which represents the noun, and infinitive, and "فَاعِل *faAEil*" which represents the adjective in most cases.

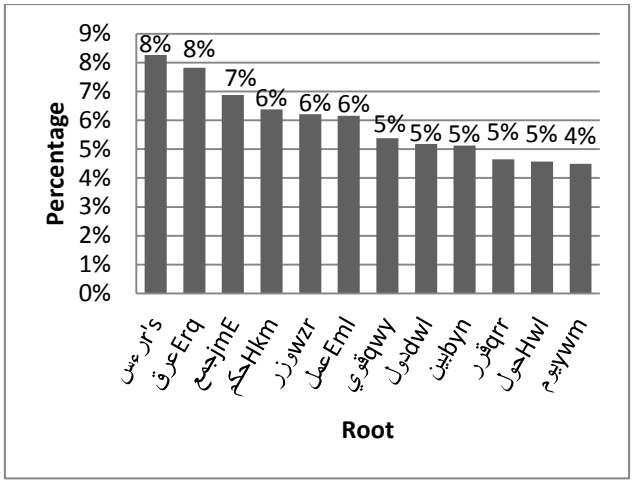

Figure 11: Most Frequent Roots

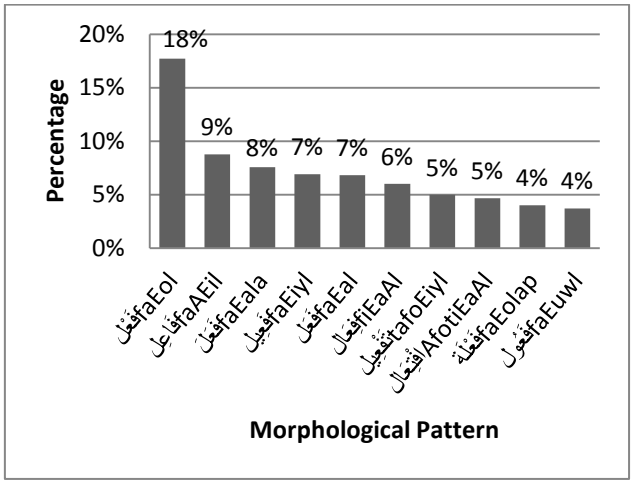

Figure 12: Most Frequent 10 Morphological Patterns

12 AFFIXES DISTRIBUTION

Affixes (prefixes and suffixes) are agglutinated to the beginning and the end of Arabic words. Prefixes are generally conjunctions, prepositions, and determiners (and include also the person conjugation of verbs in the present tense "أنيت" حروف المضارعة). Suffixes are the conjugation terminations of verbs and they are the dual/plural/feminine marks for nouns, and pronouns attached at the end of words [5].

Figures 13 and 14 show the distribution of prefixes and the conjugation person of present verbs.

We can observe that most words have no prefixes (87%), and 12% have only 1 prefix ("و *w*", "ب *b*", or "ل *l*"), while other prefixes are rarely used.

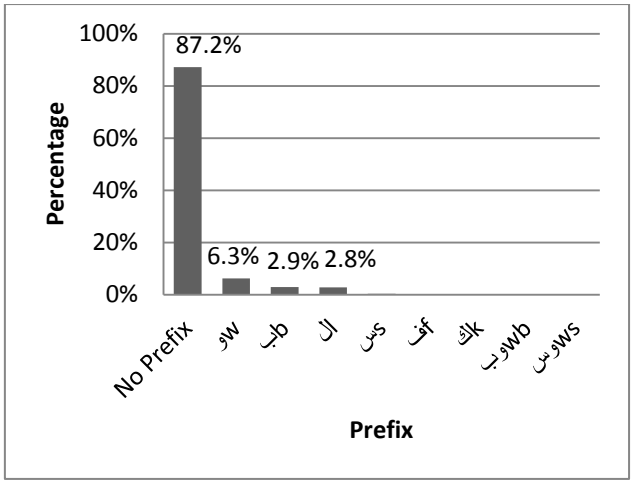

Figure 13: Most Frequent Prefixes

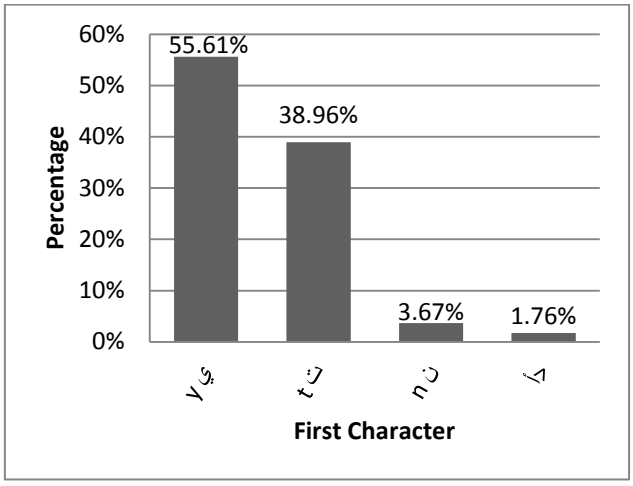

Figure 14: Person Conjugation of Present Tense

On the other hand, Figure 15 shows the suffixes distribution, and it is notable that 76% of words have no suffixes, and 17% have simple ones, while other suffixes are rarely used.

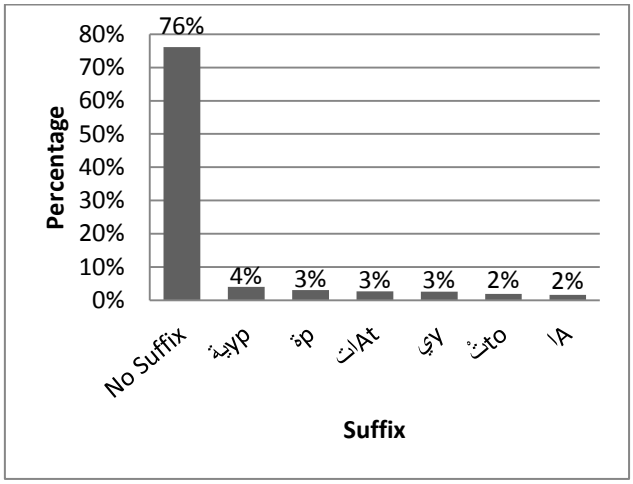

Figure 15: Most Frequent Suffixes

13 MORPHOSYNTACTIC FEATURES

In this section we show the distribution of *gender*, *number*, *person*, *case ending*, and *definiteness*.

Gender النوع in Arabic can be masculine, feminine, or neuter (like function words). Figure 16 shows the distribution of gender. It is notable that masculine words are more frequent than feminine words (1.5 times).

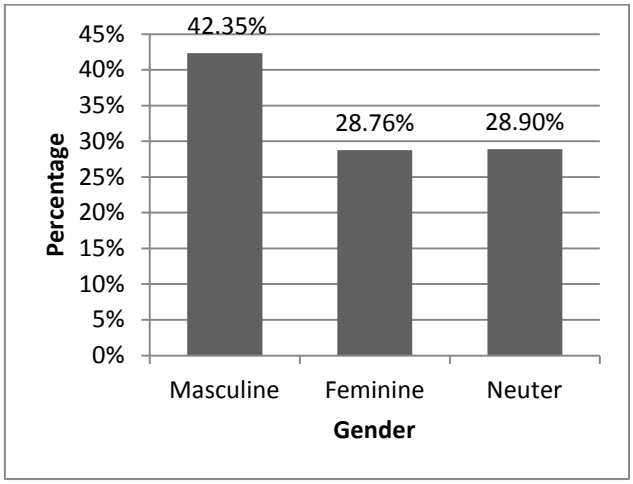

Figure 16: Gender Distribution

Number العدد in Arabic can be singular, dual, or plural (plural is divided more into regular plural and broken plural). Figure 17 shows the distribution of number. It is notable that singular words are more frequent than plural words, while using dual number is very limited (~5%).

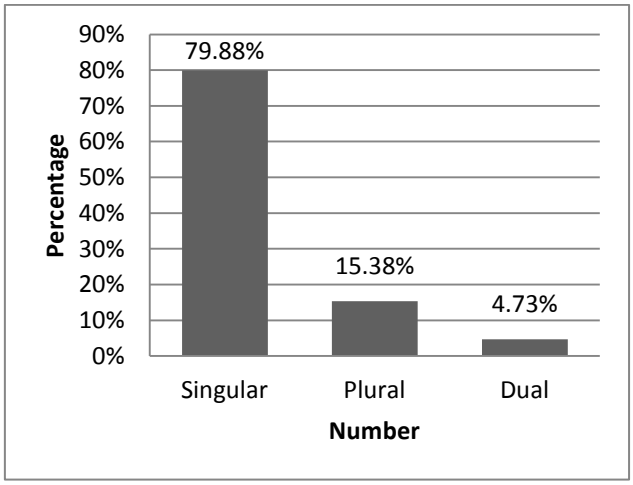

Figure 17: Number Distribution

Person الشخص in Arabic can be first person (narrator متكلم), second person (interlocutor مخاطب), or third person (absent غائب). Because of the narrative nature of most of Arabic publications (especially newswire and media), the third person is dominant (~97%) while second and first persons are almost equal as shown in Figure 18.

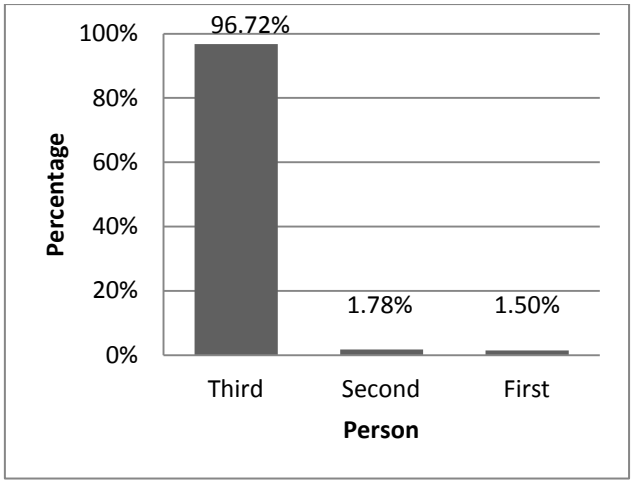

Figure 18: Person Distribution

Case Ending الحالة الإعرابية for nouns can be nominative مرفوع, accusative منصوب, genitive مجرور, or given مبني (fully diacritized without considering the case ending mark), while the case ending for verbs can be indicative مرفوع, subjunctive منصوب, jussiveمجزوم , or given مبني . Examples for given nouns are particles, and pronouns, and for given verbs are past verbs, imperative verbs, and present verbs with some suffixes.

Figure 19 shows the distribution of case ending for nouns and verbs. We can observe that the case ending for verbs (if not given) tends to be indicative (~81% of the cases), and for nouns (if not given) it tends to be genitive (~56% of the cases).

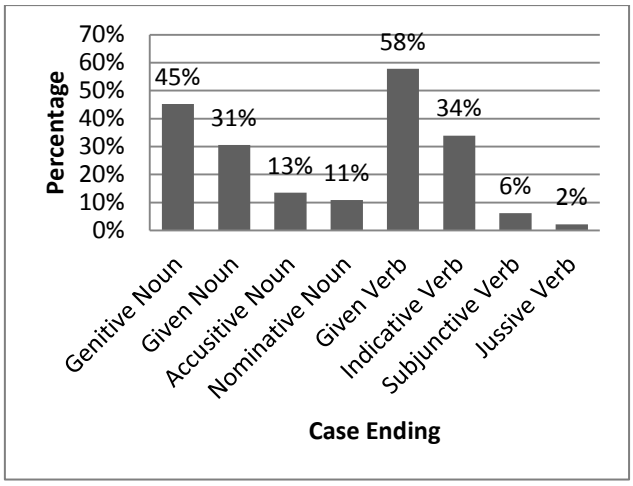

Figure 19: Case Ending Distribution

Figure 20 shows the distribution of diacritics extracted from the fully diacritized corpus. It is notable that “Fatha” is the most frequent diacritic and forms with “Kasra”, “Sukun” and “Damma” ~97% of the whole diacritics.

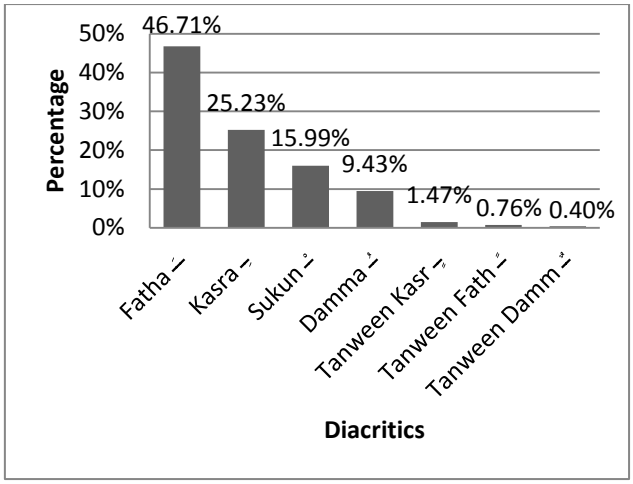

Figure 20: Diacritics Distribution

Definiteness التعريف in Arabic can be definite with the definite article AL معرف بال, definite without AL معرف بغير ال (like proper nouns, pronouns, and in possessive pronouns suffixes cases), or indefinite نكرة as in Figure 21.

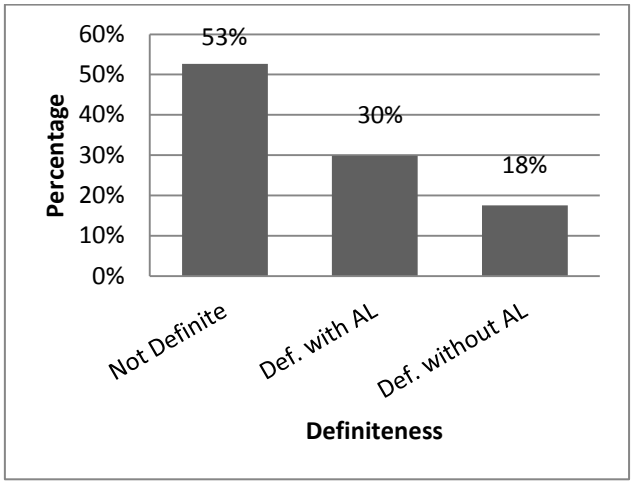

Figure 21: Definiteness Distribution

14 CORPUS COVERAGE

In this section we discuss the coverage of existing unique words in POST and compare it with an arbitrary recent corpus that is crawled by Sakhr's news gathering service (Johaina http://johaina.sakhr.com) which gathers Arabic text from more than 400 Arabic sources. The objective of this comparison is to answer the following question: If we have an arbitrary recent corpus, what are the differences between our "old" tagged corpus and this new one in terms of new unique words, new stems, and new proper nouns?

To study the unique words coverage, we gathered a recent corpus from Johaina with a size of 14M words (double POST size), and normalized tokens in both corpora (to exclude mismatches due to spelling mistakes in the crawled corpus and POST corrected corpus). Out of 172K normalized unique words in POST and 298K normalized unique words in Johaina, there was an intersection of 124K words which represents 73% of POST and 42% of Johaina as shown in Figure 22.

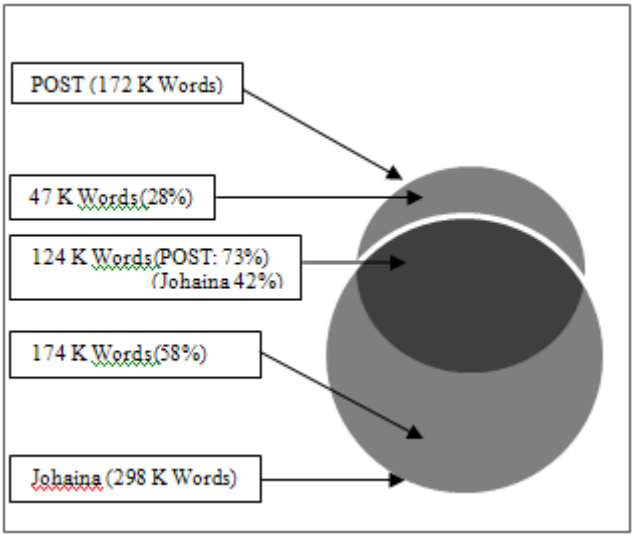

Figure 22: Unique Words Coverage

When we analyzed the words that are found in POST but not found in Johaina and vice versa, we observed the following:

- Missing stems in POST (with affix expansion) represent 11% of these words which indicate new stems in MSA or uncovered ones in POST like: "حلحلة *HlHlp*, حوكمة *Hwkmp*, and تمدرس *tmdrs*", while missing stems in Johaina represent 2% of these words like: "فدائي *fdA}y*, مستنسخ *mstnsx*, and تسلحية *tslHyp*" that are no longer mentioned extensively in modern writings as obtained from Johaina corpus.
- Stems with different affixes and obsolete/new proper nouns represent 98% and 87% of POST and Johaina stems in order, as shown in Figure 23.

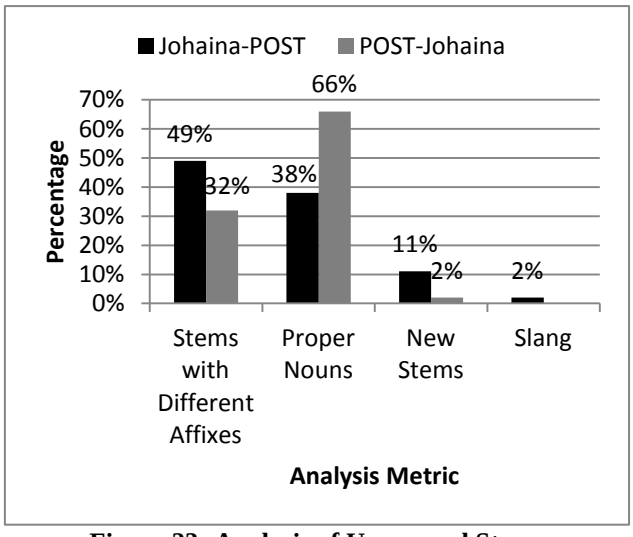

Figure 23: Analysis of Uncovered Stems

15 MSA MORPHOLOGICAL COVERAGE

The morphological analyzer uses the lexical database (LDB) to analyze and synthesize Arabic words. LDB contains lists of stems, roots, morphological patterns, prefixes, and suffixes, etc., as mentioned in common Arabic lexicons and resources (like المعجم الوسيط and المعجم العربي الأساسي).

In this section we study the coverage of these morphological data that appeared in our tagged corpus with respect to the corpus size. For any of the next information, we consider a single existence of any morphological data value as covered, otherwise, we consider this value uncovered (unused).

For **stem coverage**: Figure 24 shows the relation between the corpus size and existing stems. LDB contains 38,500 tri-literal stems, 1,200 quad-literal stems and 6,500 stems with no-root. For the whole corpus size (7M words), the coverage percentages of stems reached 52%, 39% and 57%, respectively. Examples of uncovered stems are: ميعاس myEAs, قاووق qAwwq، and تيهور tyhwr.

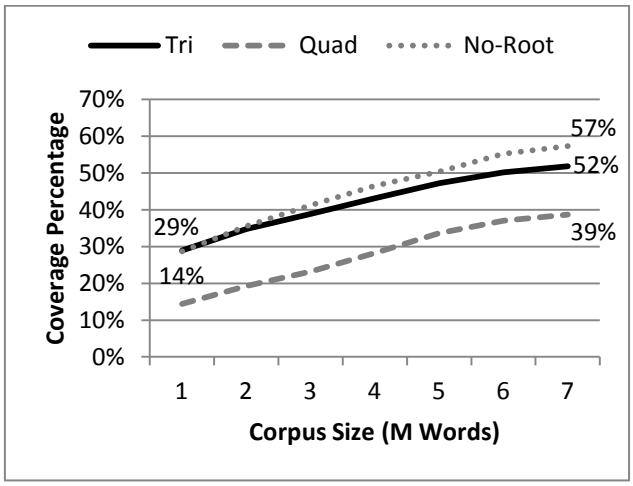

Figure 24: Stems Coverage Distribution

For **root coverage**: Figure 25 shows the relation between the corpus size and existing roots. LDB contains 5,000 tri-literal roots, and 800 quad-literal roots. For the whole corpus size (7M words), the coverage percentages reached 86% and 34%, respectively. Examples of unused roots are: kdn كدن, جدح jdH, and يفخ yfx.

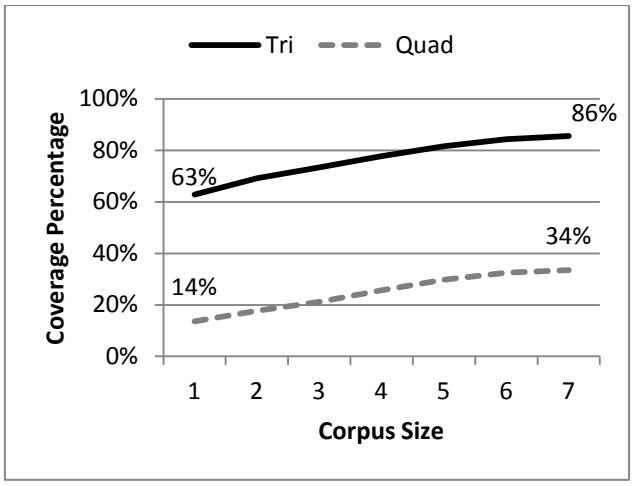

Figure 25: Roots Coverage Distribution

For **morphological patterns coverage**: Figure 26 shows the relation between the corpus size and the existing morphological patterns. LDB contains 540 tri-literal morphological patterns, and 110 quad-literal morphological patterns. For the whole corpus size (7M words), the coverage percentages were 55% and 46%, respectively. Examples of unused morphological patterns are: تَفَيْعَلَ tafayoEala, فِعْوَال fiEowaAl, and يُنْفَعَل yunofaEal.

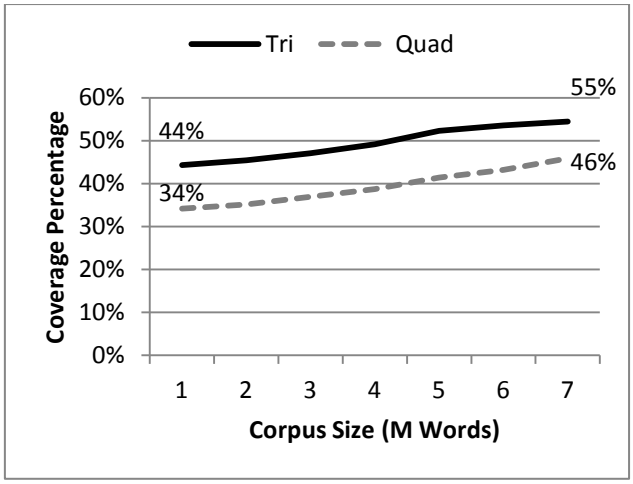

Figure 26: Morphological Patterns Coverage Distribution

For **prefixes coverage**: Figure 27 shows the relation between the corpus size and existing prefixes. LDB contains 140 Prefixes. For the whole corpus size (7M words), the coverage percentage was only 15%. Examples of unused prefixes are: أو >w, أس >s, and أوب >wb.

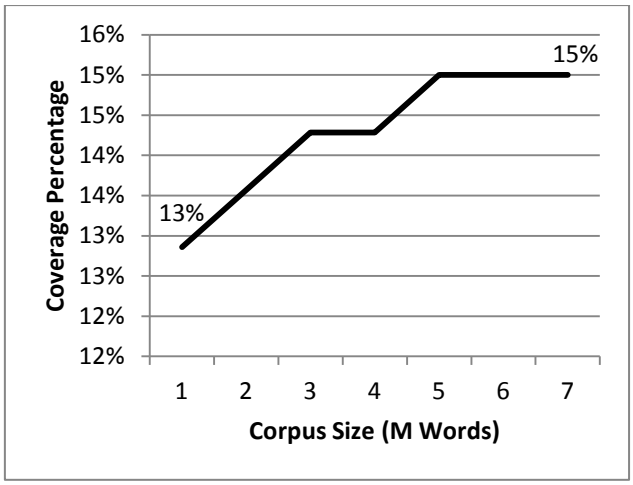

Figure 27: Prefixes Coverage Distribution

For **suffixes coverage**: Figure 28 shows the relation between the corpus size and existing suffixes. LDB contains 700 suffixes. The coverage percentage was 32%. Examples of unused suffixes are: كهن khn ،كهاء khA, and اكما AkmA.

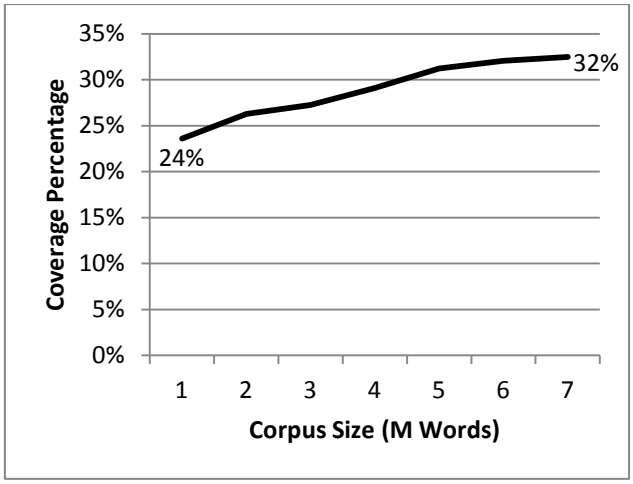

Figure 28: Suffixes Coverage Distribution

16 OTHER ANNOTATED CORPORA

Some previous attempts of Arabic corpora analysis are discussed in this section.

The Penn Arabic Treebank (PATB): Treebank is designed to support the development of data-driven approaches to NLP, human language technologies, automatic content extraction (topic extraction and/or grammar extraction), cross-lingual information retrieval, information detection, and other forms of linguistic research on MSA in general [8]. (http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2004T11)

NEMLAR Arabic Written Corpus: aims to achieve a well-balanced corpus that offers a representation of the variety in syntactic, semantic and pragmatic features of modern Arabic language. The time span of the data included goes from late 1990's to 2005. The corpus is provided in 4 different versions: a) raw text, b) fully vowelized text, c) text with Arabic lexical analysis, and d) Arabic POS-tagged. (http://catalog.elra.info/product_info.php?products_id=873)

Prague Arabic Dependency Treebank (PADT): is a project of analyzing large amounts of linguistic data in Modern Written Arabic in terms of the formal representation of language that originates in the Functional Generative Description [9]. PADT does not only consist of multi-level linguistic annotations of the MSA, but it even has a variety of unique software implementations, designed for general use in NLP.
(http://ufal.mff.cuni.cz/padt/PADT_1.0/docs/index.html)

CLARA (Corpus Linguae Arabicae): The ultimate goal of this project is building a balanced and annotated corpus. The annotation is done for morphological boundaries and Part Of Speech (POS) [10].
(http://enlil.ff.cuni.cz/veda/projekty/clara.htm)

Table II shows some information about these corpora.

TABLE 2: ANNOTATED CORPORA INFORMATION

Corpus	Size (Words)	Years	Sources	Annotation
Sakhr	7 M	2002-2004	Different sources	POS+Morph
PATB	340 K	2000-2002	AFP, Al-Hayat, An Nahar	POS+Morph +Syntax
NEMLAR	500 K	1990-2005	Islamonline, RDI, An Nahar	POS+Morph
PADT	113 K	2000-2003	AFP,Ummah, An Nahar, Al-Hayat Xinhua	POS+Morph +Syntax
CLARA	100 K	1997-1999	Different sources	POS+Morph

These annotated corpora use different morphological analyzers. At many levels, there are no standards. There are none for basic Arabic linguistic terms and their definitions, none for terms and their translation into English, and none for test collections and performance evaluations [11]. (Sakhr uses Sakhr's morphological analyzer, PATB and PADT use Buckwalter Arabic morphological analyzer (BAMA), while NEMLAR uses ArabMorpho© morphological analyzer).

17 CONCLUSIONS

In this paper, we presented lexical and morphological statistics of an Arabic POS-Tagged corpus and basic statistical differences between Arabic and English languages. Some useful statistics about the general characteristics (ambiguity,

usage and coverage) of MSA were also obtained. In NLP applications, there is a new tendency to make use of statistical methods. The idea underlying this approach is observing how the language is actually used and drawing conclusions, instead of trying to formalize the language. The results given in this paper can be extended on this line. They are useful for statistical NLP approaches and different applications like Optical Character Recognition (OCR), spelling correction, POS disambiguation and diacritization, MT, IR, and IE.

REFERENCES

[1] Jurafsky, D. and Martin, J.H. (2008). Speech and Language Processing: An Introduction to Speech Recognition, Computational Linguistics and Natural Language Processing. 2nd Edn., Prentice Hall, ISBN: 10: 0131873210, pp: 1024.

[2] Atwell, E., Al-Sulaiti, L., Al-Osaimi, S. & AbuShwar, B. (2004). A review of Arabic corpus analysis tools. Proc. JEP-TALN'04 Arabic Language Processing.

[3] Elhadj, Y. (2009). Statistical Part-of-Speech Tagger for Traditional Arabic Texts. Journal of Computer Science 5 (11). Imam Muhammad Bin Saud University, KSA

[4] Seikaly, Z.A. (2007). The Arabic Language: The Glue that Binds the Arab World. AMIDEAST, America-Mideast Educational and Training Services, Inc. http://www.amideast.org/publications/arabic- language.pdf

[5] Kadri, Y., & Nie, J. Y. (2006). Effective Stemming for Arabic Information Retrieval. In Proceedings of the challenge of Arabic for NLP/MT Conference. The British Computer Society. London, UK.

[6] Khoja, S. (2001). APT: Arabic part-of-speech tagger. Proceeding of the Student Workshop at the 2nd Meeting of the NAACL, (NAACL'01), Carnegie Mellon University, Pennsylvania, pp: 1-6.

[7] Alotaiby, F., Alkharashi, I., & Foda, S. (2008). Processing Large Arabic Text Corpora: Preliminary Analysis and Results. King Saud University.

[8] Maamouri, M. & Bies, A. (2004). Developing an Arabic Treebank: Methods, Guidelines, Procedures, and Tools. In Proceedings of COLING 2004. Geneva, Switzerland.

[9] Hajič O. & et al (2006), The Challenge of Arabic For NLP/MT, Tips and Tricks of the Prague Arabic Dependency Treebank, International Conference at The British Computer Society (BCS), 23 October, London.

[10] Zemanek, P. (2001), CLARA (Corpus Linguae Arabicae): An Overview . In ELSNET (Ed.), Proceedings of ACL/EACL 2001 Workshop on Arabic Language Processing. Toulouse, France.

[11] Al-Sughaiyer, I., & Al-Kharashi, I. (2004), Arabic morphological analysis techniques: A comprehensive survey. Journal of the American Society for Information Science and Technology, 55(3):189–213.

BIOGRAPHY

Hamdy Mubarak received his BSc degree in computer engineering from Alexandria University in 1992 with degree "Distinct with honor". Hamdy is currently the Arabic NLP Research and Development manager at Sakhr Software (www.sakhr.com). He also worked as a Software Manager at the Cairo Engineering center for Ellipsis Digital Systems (USA) (2001-2003). His main research interests are Natural Language Processing, machine translation, Information Extraction, and Information Retrieval.

إحصائيات معجمية وصرفية للمكنز العربي المُرمّز

حمدي مبارك، كريم شعبان، فرات عادل
شركة صخر لبرامج الحاسب، القاهرة، مصر

خلاصة:

عملية ترميز أقسام الكلام للنصوص تعتبر من المكونات الأساسية للعديد من تطبيقات معالجة اللغة الطبيعية، وعملية بناء مكنز مرمز يدويا تساعد على دراسة إحصائيات رئيسية للغة ما، والذي يشكل أساسا هاما لبناء نظام يقوم بالترميز الآلي لهذه اللغة.

في هذا البحث نقدم الكثير من الإحصائيات المعجمية والصرفية للغة العربية تم استخلاصها من مكنز صخر المرمز يدويا. هذا المكنز يتكون من ٧ ملايين كلمة مختارة من عدد كبير من الدول العربية وتغطي مواضيع مختلفة خلال الأعوام من ٢٠٠٣م إلى ٢٠٠٤م. تم استخدام هذه الإحصائيات الهامة أثناء بناء مشكل آلي إحصائي للنصوص العربية يتميز بدقة عالية في تشكيل بنية الكلمة واختيار قسم الكلم المناسب لكل كلمة داخل النص.

هذه الإحصائيات تشمل معلومات خاصة بأطوال الجمل والكلمات، علامات الترقيم، تكرار الحروف العربية وأيضا حركات التشكيل، بالإضافة إلى معلومات معجمية وصرفية لتكرار أقسام الكلام المختلفة، تكرار جذوع الكلمات، تكرار السوابق واللواحق، تكرار الجذور، وتكرار الموازين الصرفية. تشمل أيضا هذه الإحصائيات سمات صرف–نحوية مثل تكرار النوع (مذكر، مؤنث)، تكرار العدد (مفرد، مثنى، جمع)، تكرار الشخص (متكلم، مخاطب، غائب)، تكرار الحالات الإعرابية المختلفة (مرفوع، منصوب، مجرور، مجزوم) وغيرها.

تم دراسة مدى تغطية هذا المكنز المرمز للغة العربية المعاصرة من حيث تغطية كل من جذوع الكلمات، الجذور، الموازين الصرفية، السوابق واللواحق.

تم مقارنة هذه الإحصائيات بمثيلتها في اللغة الإنجليزية عن طريق دراسة مكنز كبير للغة الإنجليزية كلما أمكن ذلك.