A Lexicon of Connected Components for Arabic Optical Text Recognition

Yousef S. Elarian
King Fahd University of Petroleum & Minerals
yarian@kfupm.edu.sa

Fayez M. Idris
German-Jordanian University
fayez.idris@gju.edu.jo

Abstract

Arabic is a cursive script that lacks the ease of character segmentation. Hence, a unit for Arabic text recognition that is discrete in nature was suggested, viz. the connected component. A lexicon listing valid connected components of Arabic is necessary to any system that is to use such unit. Here, we produce and analyze a comprehensive lexicon of connected components in two ways. The resulting lexicon contains 684,743 entries, showing a percent decrease of 97.17% from the word-lexicon.

1. Introduction

It has been said that "the personal computer has grown in many directions since its birth, but one feature remains the same: The keyboard [1]." The substitute for the standard input unit is to recognize humans' communication forms, mainly, speech and images. Text recognition systems have the advantage of keeping the interaction between the human and the computer quiet and private [2]. The aim of text recognition is to transform written text into a computer comprehensible representation [3-5].

Text recognition systems need to be presented with the list of units they are to learn. Such list of allowed vocabulary is referred to as the *lexicon* [6]. The unit of a lexicon may range from character shapes, as in optical character recognition systems (OCR's), to complete words, as in holistic systems. The advantage of bigger units is that they require less effort for segmentation. The advantage of smaller units is in ease of learning and compact lexicons [7].

When running an OCR, the input images need to be segmented into the units that constitute the lexicons used to train it. Arabic script is cursive in both printing and handwriting.

Character segmentation in cursive scripts suffers from the classic 'hen and egg' dilemma: To recognize a character, segmentation is needed; and to segment a character, it needs to be recognized. Even the trained Arabic reader may need to backtrack between the segmentation and recognition steps [8]. Therefore, holistic word recognition, that doesn't segment characters, is gaining attraction [9].

An alternative unit for Arabic optical text recognition (AOTR) which is readily segmentable is the *connected component* (CC). CCs can range in length between single letters and complete words. A thorough listing of all possible CCs explodes exponentially with the CC size. For connected components to be used as a semi-holistic unit for training and testing recognition systems, a lexicon of CCs with a tractable size needs to be present. In this work, we produce and assess a mere lexicon of connected components from Arabic words.

The rest of this paper is organized as follows: related Arabic-script characteristics are exposed in Section 2. Section 3 presents literature briefly. Section 4 details the steps followed to obtain the lexicon. In Section 5, we present and analyze results. Finally we conclude in Section 6.

2. Characteristics of the Arabic scripting system

Arabic script has some aspects that can make it peculiar. It enjoys well-defined rules governing the connection and separation of characters. Some letters never connect to subsequent letters in the same word. These are shown in Figure 1. The leftmost character in particular, *Hamzah*, is never connected from the right side either. Other letters always connect in both printing and handwriting [10]. See Table 1 for examples of the Arabic printed and handwritten matching in their cursive script.
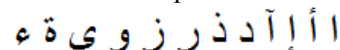
ا أ إ آ د ذ ر ز و ى ة ء

Figure 1. Non-connectable Arabic characters

In modern Latin scripting, as a contrasting example, a writer is free to connect or separate characters, as exemplified in Table I. Such freedom may form a source of ambiguity. This helps interpreting the following quote:

"*Arabic language is the easiest and clearest language in the world. It is useless to try to find new ways to make it easier and clearer. If you receive any letter you will not face any difficulty to read it even if it is written with the worst Arabic font* [11]."

Table 1. Printed and handwritten Arabic and Latin script samples

Samples	Arabic	Latin
Printed	هذه الصفحة	Paris
Handwritten	هذه الصفحة	PARIS Paris

Another characteristic of Arabic script is that some characters share glyph shapes and differ only in points (dots and *Hamzah*). Typically, two to three letters share a glyph. Some letters share the glyph shape of others in some but not all positions. Figure 2 shows examples of the above rule.

Isolated		Beginning
ن	ي ئ	ب ت ث	بـ تـ ثـ يـ نـ
Ending		**Middle**
ن	ـي ـئ	ـب ـت ـث	ـبـ ـتـ ـثـ ـيـ ـنـ

Figure 2. Examples of letter-shapes that differ only in dots/Hamzah

A connected component (CC) refers to whatever can be written before the pen must be lifted and translated [12]. Hence, CCs seem to be the easiest unit for the task of segmentation from the script images. In Arabic, a CC appears when a non-connectable character occurs, or, otherwise, at the end of the word. Besides, mere CCs don't include points and other diacritic marks which appear abundantly in Arabic script [9]. Figure 3 shows the CCs of an Arabic text with distinctive black and white tiles.

الأشكال المنصلة

Figure 3. Connected components

3. Literature survey

According to the level of interaction between segmentation and recognition, optical text recognition systems are associated to one of three strategies [13]:

1. Segmentation-based: where attempts to dissect images to classifiable units are done before passing the results to the classifier.

2. Recognition-based segmentation: where components of images which match with classes of the system's alphabet are looked for and decided on by aid and feedback from the recognition stage. Segmentation and recognition of letters are accomplished at the same time [14]. A popular family of this strategy is Hidden Markov Models (HMM).

3. The holistic segmentation-free: where words are recognized as a whole.

Segmentation-based systems segment images into lines, words or characters [15]. Alternatively, over-segmentation techniques tend to break the images down into small strokes and then group these into characters [16]. However, it is reported that there exists no segmentation algorithm which is likely to separate the characters of Arabic script with reasonable accuracy [11]. On the other hand, holistic techniques have achieved high word recognition rates for cursive scripts [17]. Their disadvantage is that their lexicons are always limited to a manageable count of words.

The CC unit was used, in limited form, by Allam [18] and has been recently declared as the basic pictorial block for AOTR [10,19]. Allam has located groups of connected characters by contour tracing. Khorshed [9] has inspected the vertical projection to determine whether a white column is an inter-word or intra-word spacing. Special treatment has been required to separate sub-words when an overlap exists [20]. We are aware of no previous collection of a lexicon of CCs.

4. Lexicon production and reduction

Production of a lexicon of CCs encompasses several steps. A lexicon that reflects valid surface-words in the Arabic language is first obtained. Then two reduction steps reduce it into lexical sub-words (pointed CCs) and mere CCs (point-normalized (PN) CCs). These steps, along with their input data and their outcomes, are illustrated in Figure 4 and described in the following subsections.
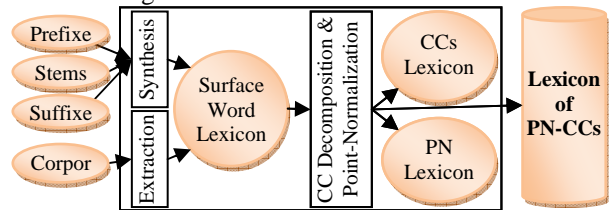

Figure 4. The block diagram.

4.1 Producing the surface-word lexicon

The surface-word lexicon is obtained through two different approaches: analysis and synthesis. The analytical approach processes large amounts of representative texts, known as *corpora*, by parsing, normalizing, and discarding redundant entries. The synthesis approach starts from the smallest linguistic meaningful parts, known as *morphemes*, and assembles them into valid words. Both approaches, along with their input data, are detailed below.

4.1.1. The Analytical Approach. The input data for this approach consists of two corpora and a dictionary. Corpora are ideally expected to reflect natural language statistics. The dictionary plays a different role: it asserts seeing a complete list of words, regardless of the frequency of their usage. The input data is tokenized into words, which are then normalized and stored without repetitions.

The first corpus, provided by the Dustour newspaper [21], consists of local Jordanian news for 53 months. The second input source is the Corpus of Contemporary Arabic (CCA) [22] consisting of internet texts of distinct subjects. The dictionary used is the Salmone Arabic-to-English dictionary, encoded as part of the Perseus project at Tufts University [23]. The dictionary provides rich vocabulary and phrasal expressions, as well.

The tokenization step parses files of mixed Arabic and Latin characters into lists of Arabic words. It filters out tags and non-Arabic alphanumeric characters, tokenizes words based on spaces and punctuation marks, and combines the output of each source into a single text file. Table 2 lists the counts of the result of word tokenization. It also reports the counts of CCs in the sources before processing.

The normalization step removes characters causing linguistically acceptable variations of the same surface words. It removes optional diacritic marks and a calligraphic elongation character called *Tatweel*. After this step, repeated instances are removed. Counts of the entries of the word lexicon are reported in Table II. The low percent decrease[1] of words in the dictionary is due to the low repetition there.

Table 2. Counts of Arabic words and CCs in the analytical inputs

Corpus	With repetition	Without repetition	% decrease
Dustour Words	11,386,925	235,973	97.93
CCA Words	594,119	85,482	85.61
Salmone Words	85,640	44,171	48.42
Dustour CCs	28,519,734	47,676	99.83
CCA CCs	1,348,019	24,027	98.22
Salmone CCs	166,597	15,031	90.98

4.1.2 The synthesis approach. The synthetic approach starts with dictionaries of morphemes to systematically produce variations of words. It depends on Buckwalter's dictionaries of: prefixes (containing Arabic prefixes and their concatenations), stems (containing roots and their inflections from patterns) and suffixes (containing Arabic suffixes and their concatenations) and three compatibility tables listing the allowed combinations of entries from these dictionaries. The surface-word lexicon results from the Cartesian product of the three dictionaries filtering out entries containing incompatible parts. This results in 39,399,206 words with repetition, 24,122,954 unique words and 2,162,960 unique CCs. The former behavior is depicted in the pseudo-code of Figure 5.

```
For every prefix p
  For every stem s
    Filter out p+s if p,s incompatible
    For every suffix f
      Filter out p+s+f if p,f incompatible
      Filter out p+s+f if s,f incompatible
      store p+s+f
```

Figure 5. Pseudo-code of the synthetic lexicon generation approach.

4.2 Point-normalization and CC-tokenization

Characters that share the same glyph, except for points, are mapped together. This lossy mapping is intended to ignore all dots and points. Table 3 shows the necessary replacements to achieve an encoding that doesn't differentiate between characters sharing the same primary glyph.

Characters in the thick cells act (and hence map) differently in the cases of their final and non-final positions. Characters that are not mentioned in Table 3 have no similarities with other character glyphs. These remain untouched.
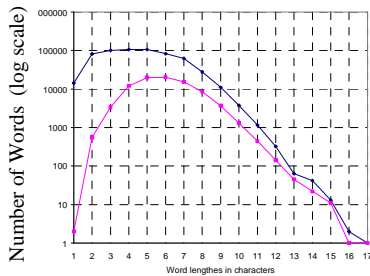
Table 3. Replacements made to point-normalize entries

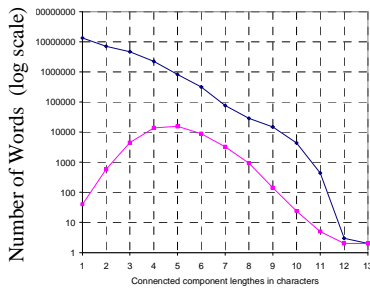| Final		Non final	
ا ←	أ إ آ	ا ←	أ إ آ
ب←	ت ث	ب ←	ﺘﺗ
ح←	خ ج	ﺤ ←	ﺠ ﺨ
د←	ذ	د ←	ذ
ر←	ز	ر ←	ز
س←	ش	ﺴ ←	ﺸ
ص←	ض	ﺼ ←	ﺾ
ط ←	ظ	ﻄ ←	ﻆ
ع ←	غ	ﻌ ←	ﻐ
ق ←	ق	ﻔ ←	ﻗ
ن ←	ن	ﺒ ←	ﻧ
ه ←	ة	---	---
و←	ؤ	و ←	ؤ
ى ←	ئ	ﺒ ←	ﻧ
ى ←	ي	ﺒ ←	ﻧ

Text of connected components can easily be tokenized based on the non-connectable characters of Figure 1 and on the word-end delimiter.
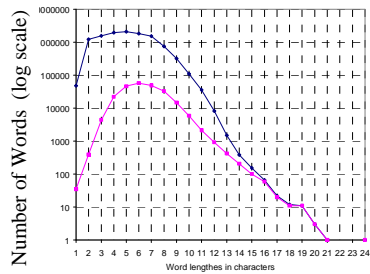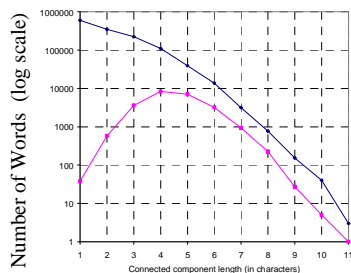

(a)

(b)

(c)

(d)

Figure 6. Reduction in the count of units corpora due to redundancy removal for (a) Dustour words, (b) CCA words, (c) Dustour CCs, and (d) CCA CCs. (Higher curve is for the corpus and lower curve is for the lexicon.)

5. Results and analysis

This section presents graphs that show counts of entries (words or CCs) per entry length (in characters). The impact of different levels of reduction on the counts and distribution of the entries of the several word-lexicons is observed.

5.1 Reduction due to redundancy removal

The reduction in counts per object size for words and CCs of the two corpora we have, viz. Dustour and CCA, due to the surface-word lexicon extraction are shown in Figure 6. The Salmone dictionary is not studied here for it doesn't represent natural frequencies of entry counts. Notice that the y-axis is logarithmic.

These graphs show up phenomena that appear quite frequently in linguistics. The lexicon curves take the rough shape of a bell. In their ascending sides, the lexicon curves are governed by the maximum number of combinations that a small number of characters can produce. The difference of lexicon in counts from corpus curve is to its maximum in this part. This reflects the trend of languages to concentrate on shorter vocabulary for common use, which is a doctrine in data compression. The collapsing part in both, the lexicon and corpus curves, are due to the limited number of longer words that are actually used in a language. This phenomenon is stronger in CCs due to their reusability in many words.

5.2 Lexicon-wise reduction

Figure 7 shows the reduction on each of the four surface lexicons. It's clear how the synthesized lexicon is by orders of magnitude larger than the others. The object lexicon and its PN version share the same range on x.

5.2 Object-wise reduction

Figure 8 shows almost the same information of Figure 7, but allowing the comparison of the performance of lexicons. We display one more category, the size of the set of all combinations of characters allowed in each category, given by:

$$\text{Combinations} = |ending\ form\ characters| \times \left(\sum_{n}^{N-1} |\text{non} - \text{connectable characters}|^{n}\right)$$
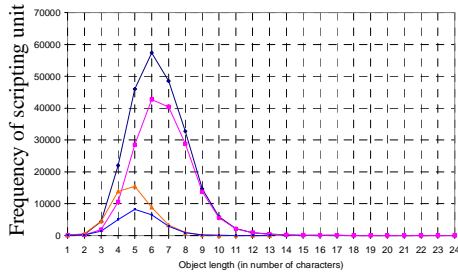
where *N* is the object size in number of characters, the magnitude operator refers to size in number of

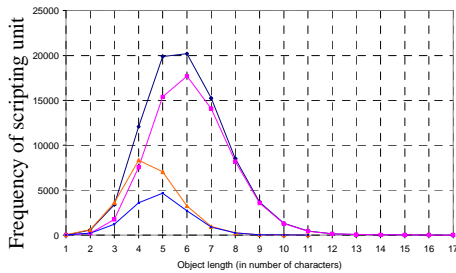characters, and *ending form characters* are all characters except *Hamzah*.

- ◆ Words
- ▲ Connencted components
- ■ Point-normalized words
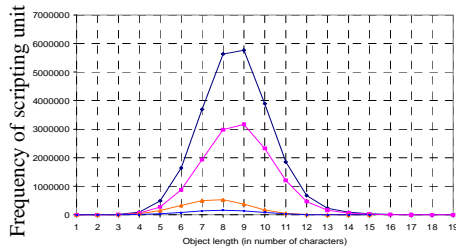- ＋ Point-normalized connected components
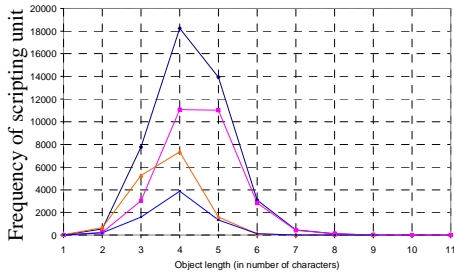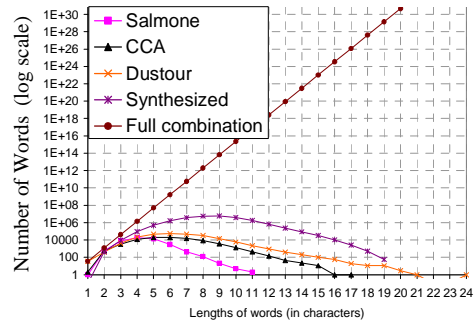

(a)

(b)

(c)

(d)

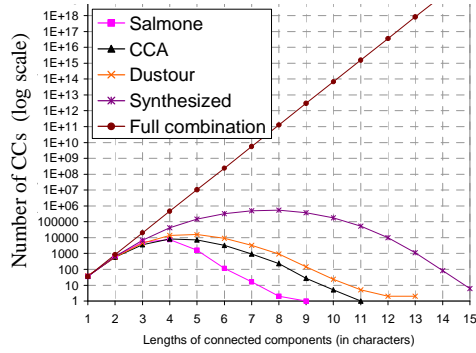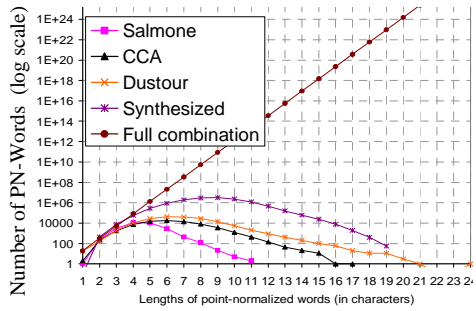**Figure 7. Lexicon scripting unit distribution over length of object for the (a) Salmone (b) Dustour (c) CCA and (d) Synthesized lexicons.**

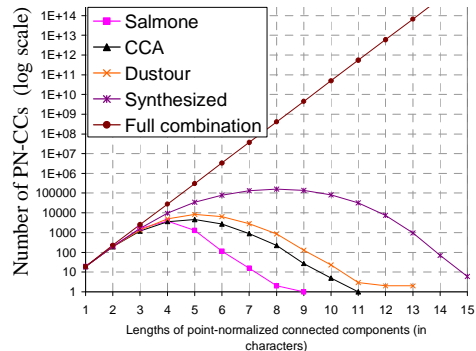
(a)

(b)

(c)

(d)

Figure 8. Comparisons between the counts of unique units per character length for the (a) word, (b) CCs, (c) PN-word, and (d) PN-CCs lexicons plus the full combinations curve.

Table 4 presents a summary of some statistics of the lexicons obtained.

Table 4. Statistical summary of objects in the lexicons

Lexicon	Category	Longest Word	PN-words	Longest CCs	PN-CCs
Salmone	*Length*	11		9	
	Count	44171	28788	15031	7129
	decrease	83.86%	75.23	52.57	--
CCA	*Length*	17		12	
	Count	85482	70333	24027	13527
	decrease	84.18%	80.77	43.70	--
Dustour	*Length*	24		13	
	Count	235973	176122	47676	25157
	decrease	89.34%	85.71	47.23	--
Synthesiz-ed	*Length*	19		6	
	Count	24122954	9925052	2162960	674583
	decrease	97.20%	93.20	68.812	--
Union	*Length*	24		13	
	Count	24166215	9965531	2173121	684743
	decrease	97.17%	93.13	68.49	--

6. Conclusion

Arabic connected components have a level of diversity between that of single characters and that of words (inclusively). Besides, being bare of points reduces their number further. To be used, these units must be comprehensively listed in a lexicon of reasonable size. We address the problem of production of the lexicon analytically and synthetically. Reduction of the size of the lexicon comes inherently in the concept of mere CCs. The resulting lexicon contains 684,743 entries, having a percent decrease of 97.17% from the corresponding word-lexicon.

Acknowledgment

References

[1] Powalka RK. An algorithm toolbox for on-line cursive script recognition [dissertation]. Nottingham Trent University; 1995.

[2] Homayoon SMB, Nathan K, Clary GJ, Subrahmonia J. Challenges of handwriting recognition in Farsi, Arabic and other languages with similar writing styles an on-line digit recognizer. Proceedings of the 2nd Annual Conference on Technological Advancements in Developing Countries, Columbia University, New York, July 23-24, 1994.

[3] Erlandson EJ, Trenkle JM, Vogt RC. Word-level recognition of multifont Arabic text using a feature-vector matching approach. Proceedings of the International Society for Optical Engineers, SPIE 1996; 2660: 63–70.

[4] Hamami L, Berkani D. Recognition system for printed multi-font and multi-size Arabic characters. The Arabian Journal for Science and Engineering; 2002 April; 27(1B): 57-72.

[5] Schurmann J, Bartneck N, Bayer T, Franke J, Mandler E, Oberlander M. Document analysis from pixels to contents; Proc. IEEE, 1992, July. 1101-19.

[6] Lexicon. (2010, May 8). In *Wikipedia, The Free Encyclopedia*. http://en.wikipedia.org/w/index.php?title=Lexicon&oldid=360937035

[7] Vinciarrelli A. A survey on off-line cursive word recognition. Pattern recognition 2002; 35:1433-1446.

[8] Timar G, Karacs K, Rekeczky C. Analogic preprocessing and segmentation algorithms for offline handwriting recognition. Journal of Circuits, Systems, and Computers 2003; 12(6):783-804.

[9] Khorsheed MS. Off-line Arabic character recognition –a review. Pattern analysis & applications 2002; 5:31–45.

[10] Khedher M, Abandah G. Arabic character recognition using approximate stroke sequence. Third Int'l Conf. on Language Resources and Evaluation (LREC 2002), Arabic Language Resources and Evaluation –status and prospects workshop; 2002, June.

[11] Abuhaiba ISI. A discrete Arabic script for better automatic document understanding. The Arabian Journal for Science and Engineering 2003; 28(1B): 77-94.

[12] Abuhaiba ISI, Holt MJJ, Datta S. Recognition of off-line cursive handwriting. Computer Vision and Image Understanding 1998; 71: 19-38.

[13] Safabakhsh R, Adibi P. Nastaaligh handwritten word recognition using a continuous-density variable-duration HMM. The Arabian Journal for Science and Engineering 2005; 30:95-118.

[14] Cheung A, Bennamoun M. An Arabic optical character recognition system using recognition-based segmentation. Pattern recognition 2001; 34:215-233.

[15] Amin A. Recognition of printed Arabic text based on global features and decision tree learning techniques. Pattern recognition 2000; 33: 1309-23.

[16] Romeo-Pakker K, Miled H, Lecourtier Y. A new approach for Latin/Arabic character segmentation. IEEE 1995; 874-7.

[17] Khorsheed MS. Automatic recognition of words in Arabic manuscripts [dissertation]. University Of Cambridge; 2000, June.

[18] Allam M. Segmentation vs. segmentation-free for recognising Arabic text. Proceedings of the International Society for Optical Engineers, SPIE 1995; 2422: 228–35.

[19] Abandah G, Khedher M. Printed and handwritten Arabic optical character recognition –initial study. A report on research supported by the Higher Council of Science and Technology. Amman, Jordan, 2004, August.

[20] Timsari B, Fahimi H. Morphological approach to character recognition in machine-printed Persian words Proc. SPIE Vol. 2660, p. 184-191, Document Recognition III, Luc M. Vincent; Jonathan J. Hull; Eds.

[21] Website at URL: http:\\www.Dustour.com.jo

[22] Al-Sulaiti L. Designing and developing a corpus of contemporary Arabic [dissertation]. The University of Leeds; 2004, March.

[23] Smith D. An advanced learner's Arabic-English dictionary encoded by the Perseus Project, Tufts University. [Online]. Available from URL: http://www.tei-c.org/P4X/DTD/tei2.dtd.