Automatic Splitting of Lexical Entries from Classical Dictionaries

GEORGE A. KIRAZ

Nineteenth and early twentieth-century scholarship arguably produced the foundational lexica for most disciplines.[1] For Greek, we have Liddell and Scott's *A Greek-English Lexicon,* first published in 1891 and later revised, and Sophocles' *Greek Lexicon of the Roman and Byzantine Periods* (1990). For Arabic, Lane's multi-volume An *Arabic-English Lexicon*, published between 1863 and 1893, is an essential tool for any serious Arabist, as well as Biberstein-Kazimirski's *Dictionnaire arabe-français*, published in 1860. The Armenian scholar is still in need of Bedrossian's *New Dictionary*. *Armenian-English* (1875-1879) and the scholar of Christian Palestinian Aramaic has only Schulthess' *Lexicon Syropalaestinum* (1903) to rely on. The same can be said for other ancient languages, and Syriac is no exception.

Until recently, access to these dictionaries was difficult to obtain as only elite universities would have them in their collections. The situation has changed drastically in recent years with digitization projects (and print-on-demand reproductions as all of these dictionaries are now in the public domain) such as the efforts of archive.org. In some cases, Obtical Character Recognition (OCR) was applied on scanned dictionaries to make them machine-readable and therefore, more searchable. This, however, is not possible for many languages—such as Syriac—where there does not exist a reliable OCR technology despite some scientific research in the field, most notably by Clocksin.[2]

A step further would be creating lexical databases that would allow users to search lexemes and present the results from various existing dictionaries. One such tool exists for Syriac: dukhrana.org allows multi-lexical search where one would type a lexeme—either in the Syriac script or in transliteration—and the web site will display pages from various dictionaries. The user is then given the opportunity to navigate through pages of a particular dictionary.

The Syriac Electronic Data Research Archive (SEDRA) hosted at sedra.bethmardutho.org, attempts to take this further by displaying only the lexical entry the user is seeking, allowing the user to compare the descriptions of a lexeme in various dictionaries side-by-side. To achieve this, one has to take an image of a dictionary page and then split that image into multiple images, each

30

pl. = rt. m. *a sprinkling; dew, fine rain, moisture; drops of dew.*

barley-water.

fut. part. *to meet, encounter; a)* or *with the sword* or *with shields = to attack,* also *to sustain* or *meet an attack, resist; no one could resist him; to resist* disease or cold; met. *to meet in argument, refute, confute. b) to happen, befall, come upon,* usually of misfortune; *evil will happen unto you,* Deut. xxxi. 29. DERIVATIVES.

pl. f. *the earth,* opp. *the heavens* and *the seas,* Gen. i. 1–10; *a country, land, a piece of land, field, ground, soil, the floor* of a house; *the land of Egypt; good ground; the floors of the house; an earthworm.* DERIVATIVES.

pl. from adj. and subst. *earthly, terrestrial; an earthly being, a dweller on the earth; land winds.*

f. pl. m. f. from adj. *earthly, terrestrial; earthly wisdom.*

for Ar. *a nest, a flock* of birds, *shoal.*

APHEL of; *to beat out thin.*

rt. m. *the expanse of heaven, the firmament.*

fut. inf. *to strike, beat, hammer* as a blacksmith.

or f. adj. from *Arsaces,* the name or title of the founder of the Parthian empire. *Seleucia* and *Ctesiphon,* the chief cities of the Arsacian kings are called hence *royal, chief, principal; the chief monastery.*

f. pl. &c. Gr. = Syr. adj. *orthodox, holding the right faith.*

adv. *orthodoxly.*

pl. or m. ὀρθόδοξος, *orthodox;* see above.

Gr. = Syr. *orthodoxy, holding the right faith; the orthodox faith.*

= rt. f. *trembling, fear.*

or ἄριθμος, *a troop of soldiers.*

or pl. m. *a large spoon.*

part. and APHEL of; *to blow, cause the wind to blow.*

fut. inf. imper. act. part. pass. part. (cognate roots in Heb. and cf.) *to shed, pour out* or *down* water, rain, blood, tears; *bloodshed, manslaughter; a manslayer, homicide; bloodthirsty; breathing out poison;* metaph. *to throw up* a mound; *to upset; to pour out* the heart or soul in prayer; wrath or evil; gifts, the grace of God, the Holy Spirit; *he gave himself up of his own will unto death; to shed forth mercy;* pass. part. metaph. *shed* or *scattered abroad, dissipated, diffuse, fluid; diffused, scattered light; a distracted mind; a fluid body.* ETHPE. imper. *to be shed, poured out,* esp. of blood, but also of dry things, corn, ashes, stones; metaph. *to give oneself up to; to be spread abroad, diffused; as a torrent rushing impetuously; the hope of the wicked is poured out,* i.e. *flows away like water; the divine word is spread abroad.* DERIVATIVES.

m. *the shedding* of blood.

pl. rt. *a user of charms and incantations,* esp. *a snake-charmer.*

rt. f. *snake-charming, enchantment.*

= *a writing, document, bond; a bill, bond.*

rt. *a) a being shed, spread far and wide, diffused* as liquids, sand, light; *bloodshed;* metaph. *the diffusion of knowledge. b) fluidity, liquidness.*

Figure 1

image containing only one lexical entry. This process was performed manually on T. Audo's[3] Syriac-Syriac dictionary from 1897–1901 and proved to be extremely laborious and prone to error.

The present paper describes an algorithm that takes as input scanned images of a printed dictionary and outputs images of the lexical entries contained in the dictionary. The algorithm looks at the image as a matrix of pixels and is therefore language-independent. It assumes, however, that dictionaries are typeset in a systematic way in one or more columns. It further assumes that each lexical entry begins in a new line and is either indented, out-indented, or begins in a larger font size if indentation is not used. This seems to be the case with all the dictionaries which were examined for the project.

The recognition of lexical entries is performed in the following steps. First, the location of the columns in a page is recognized. Next, lines are recognized. Finally, the first line of a lexical entry is recognized. The recognition of columns and lines is achieved by computing image histograms. These steps are described in the following sections with an introduction to notion of image histograms.

Historgrams

Histograms, introduced by Karl Pearson in 1891,[4] are a graphical representation of the density of given data. A bitonal image of a printed page gives the printed letters in a foreground color (usually black) and the background color (usually white). If one wanted to know the density of black pixels in a particular column (that is, a column of an image, not a text column), one simply counts the number of black pixels in that particular column. Consider Figure 1.

The text is typeset in two text columns, but the image itself is a matrix of pixels—5,175 columns and 6,642 rows—the width and height, respectively, of the image in pixels which one can obtain from any image editing software. If we were to count for each column in the matrix the number of black pixels, and then draw a line reflecting the total counted, we end up with the histogram in Fig. 2, where the histogram is shown in red. In this particular case, the histogram was normalized by dividing the totals by two so that the drawing of the histogram does not go well into the text. If the count of black pixels of a particular column was 100, a line length of 50 was drawn. One notices that the black line separating the columns is the densest column. Another way to visualize a histogram is to imagine one putting their hand at the bottom of a page and pushing all the black pixels upward so that there are no white pixels in-between.

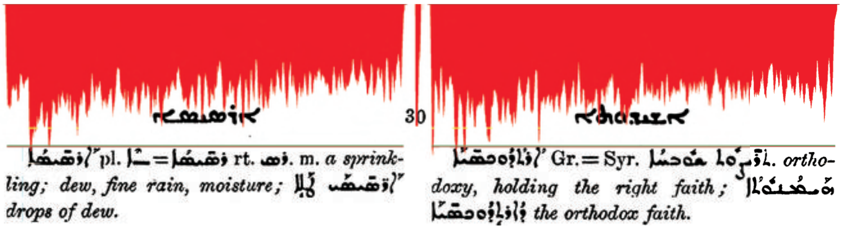


Figure 2

Detecting Text Columns

Most printed dictionaries are typeset in two columns, while others are set in smaller, one-column trim-size. I have not seen dictionaries in three columns, but the general principles discussed here will still apply. The first step in the current algorithm is to detect the beginning and end of each column by means of computing the histogram of image columns, as shown in Fig. 2. As a scanned image may not be perfect and may include speckles or small smudges, the algorithm defines a threshold that needs to be reached for each count of black pixels in an image column to be considered. If the count is below that threshold, the algorithm considers the image column to contain zero black pixels. The threshold is defined by trial-and-error, differs from one dictionary to another, and depends on the quality of the scans. Once the boundaries of the two text columns are identified, a box is drawn (in yellow in our case) around them to help debugging when issues arise. The rest of the algorithm only looks at pixels within the text columns.

Detecting Lines within Columns

Once the text columns are detected, the next step is to detect typeset text lines. The algorithm navigates through each row of a given text column and computes a histogram by counting black pixels in each given row. In the case of nicely-spaced text lines, there is always white space that separates the lines. In some cases, letter descenders and very tall ascenders may cause two or more text lines to be recognized as one unit. However, this does not hinder the process, because the ultimate objective is not recognizing text lines, it is recognizing the beginning of a lexical entry. An example is given in Fig. 3 where the histograms for each column are given on the left and right edges of the image.

Detecting the Beginning of a Lexical Entry

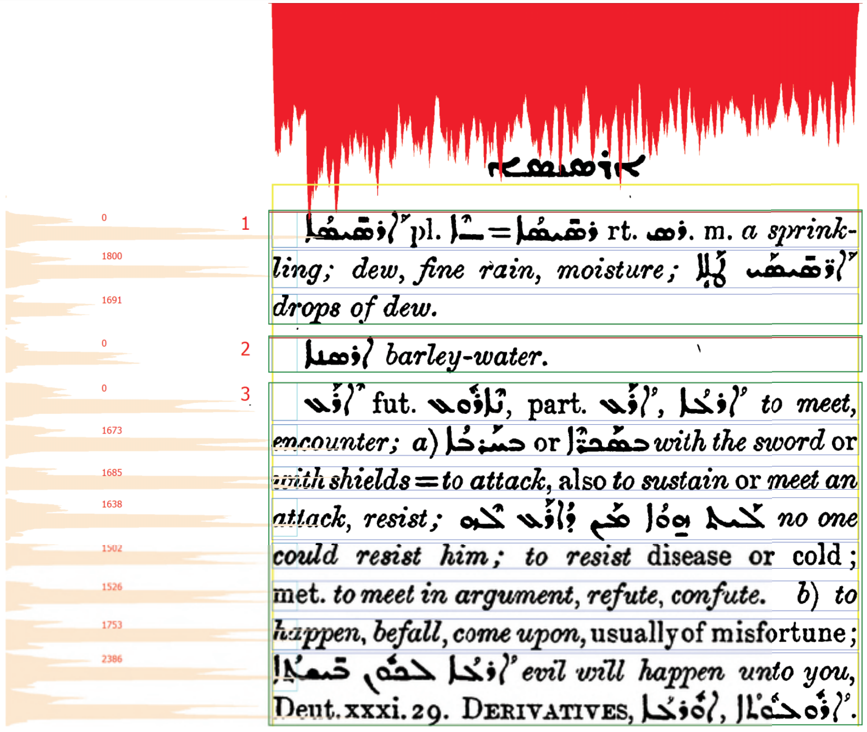
As mentioned above, a lexical entry in a printed dictionary is either indented (as in the examples shown here) or out-indented. The algorithm was initially designed for indented texts. To detect if a particular line begins a lexical entry, a box of a certain fixed width and fixed height (shown in light blue at the beginning of each line in Fig. 3) is set as the target of a black pixel search, where a count of all the black pixels in that box is computed. If the count falls below a certain threshold, we assume that the line is indented, and is the beginning of a new lexical entry. Programmatically, a Boolean variable named *IsIndented* is set to "true" when this condition is met. In the case of hanging indented lexica, the algorithm works as before, but the interpretation of the Boolean *IsIndented* variable is changed. If set to "false," then we know that there is a new lexical entry.

Producing the Output

Once we know where a lexical entry begins, it becomes easy to draw a box (in Green in Fig. 3) around each lexical entry and extract its image pixels to copy into a separate image file. There is, however, one case that needs to be considered: the spanning of a lexical entry to multiple columns within the same page or onto the next page. The algorithm pays special attention to the first line of each text column. In the case of indented dictionaries, it expects the first line of a text column to be indented. If that is not the case, the algorithm then rightly assumes that the column continues the definition of a lexeme as connected to the entry started in the previous column or page. In the case of the SEDRA database, the various parts of the lexical entry from each column is still extracted into a separate image, but all of the images of the one lexical entry are given the same file name suffixed with A, B, C, etc. One can, of course, option to merge all of these images into one image programmatically.

The SEDRA Workflow

The input to our workflow is a scanned dictionary in PDF format. We save the PDF in PNG format where each page becomes a separate file. Before pro-



ܐܪܣܣܐ pl. ــ̈ܐ = ܪܣܣܐ rt. ܪܣ. m. *a sprinkling; dew, fine rain, moisture;* ܐܪܣܣܝ ܛܠܐ *drops of dew.*

ܐܪܣܢܐ *barley-water.*

ܐܪܥ fut. ܢܐܪܘܥ, part. ܐܪܥ, ܐܪܥܐ *to meet, encounter; a)* ܣܝܦܐ or ܒܣܟܪ̈ܐ *with the sword* or *with shields = to attack,* also *to sustain* or *meet an attack, resist;* ܠܗ ܕܐܪܥ ܡܢ ܗܘܐ ܠܐ *no one could resist him; to resist* disease or cold; met. *to meet in argument, refute, confute. b) to happen, befall, come upon,* usually of misfortune; ܒܝܫܬܐ ܠܟܘܢ ܐܪܥܐ *evil will happen unto you,* Deut. xxxi. 29. DERIVATIVES, ܐܘܪܥܐ, ܐܘܪܥܘܬܐ.

Figure 3

cessing the PNG images, we make sure that the pages are de-skewed and despeckled, otherwise, the algorithm does not produce good results. Once fed into the above algorithm, the output will be given as follows:

1. A debug image is produced for each PNG page image. An example is given in Fig. 4. The debug image is mostly of use to the programmer for debugging issues that may arise (e.g. a text column was not detected). This debug image gives the histograms described above as well as black pixel counts of the indentation box. In addition, lines are drawn around columns, lines of text, and lexical entries.
2. A proof image is produced for each PNG. The proof image is similar to the debug image, but it only contains a box around each lexical entry and a sequential number for the lexical entries. It also marks when a column continues a lexical entry form a previous page. An example is given in Fig. 5. These images are used by lexicographers, usually students, who check if the algorithm has failed in detecting a lexical entry.

31

1 f. = rt. . *borrowed, assumed, presumed;* *the supposed father.*

2 = rt. . *apparently, in appearance, by supposition, falsely, in pretence.*

3 , Heb. *a desert* or *waste.*

4 rt. . f. *assuaging, ease.*

5 pl. m. pl. f. *a testicle.*

6 or pl. m. Ar. *a cobbler, shoemaker.*

7 from the above; f. *cobbling, shoemaking.*

8 , f. pl. f. *a yoke of land = about two-thirds of an acre; a field, a piece of land, a farm;* *acres of vineyard;* *Church lands* or *endowments;* metaph. *the Church's field of labour;* *the plot of faith;* *the Apostles laboured in this track.*

9 m. perh. *the wood of the acacia,* but the word is variously used to translate *box, beech, teak,* and *almug = sandal-wood;* *beechen planks.*

10 Aphel of and of .

11 (cognate roots and) fut. , part. , *to stroke, soothe, charm serpents,* *a snake-charmer, user of charms.* Derivatives, , , .

12 rt. . m. *a charm, incantation.*

13 or m. *a lodging, inn, house of entertainment for travellers.*

14 rt. found in Heb., Chald., Ar., verb only used in Heb. Hithpoal and in Chald. Ithpael; root-meaning *to make firm, support, found,* hence is derived

15 , pl. f. rt. . *the bottom, the deepest part;* *the bottom of the well;* *the bottom of the pit;* *the depth of the sea;* *foundations of the walls,* Is. xvi. 7.

16 , pl. from a Heb. and Chald. word for fire, f. *fever;* *quotidian fever;* *quartan fever;* metaph. *the fever of schism.*

= , *six;* with *the sixth;* *the sixth day;* *the six days* of the Creation; = *sixteen;* = *sixty.*

generally followed by . interj. expressing desire and longing, *O that! would that!*

and from and . adv. of time; *last year;* *a year ago, for a year past.*

, pl. f. *a) a sign, mark, pledge, token;* *the sign* or *token of the covenant,* Gen. ix. 12; *circumcision,* Gen. xvii. 11. Used in John ii. 11–18 and throughout this Gospel of the signs whereby our Lord manifested Himself; *signs and wonders;* *whosoever has not my token let him not enter;* *a sign of approaching death;* *the sign of the cross;* cf. *signing, making* the sign of the cross. *b) a standard;* *standard-bearers.* *c)* pl. *a figure, description* of a pers. or thing; *he learnt the design* i.e. of a garden *from her.* *d) a constellation;* *the signs of the heavens, the figures* or *constellations of the heavens;* *an astronomical figure* or *symbol.* Cognate, , .

fut. 1 pers. , inf. , imper. , pl. , act. part. , pl. m. f. , pass. part. f. or *to come, arrive,* opp. *to go;* *my hour is come;* *his letter has reached us.* With preps. of the manner, *they came by ship;* *with a great army;* *to the neighbourhood of = near to;* with *to follow,* *follow me;* *thou hast followed my teaching;* with *to surround.* With *a) of place* or *result;* *to Galilee;* *to his place;* metaph. *it was apparent;* , , *to come to be, come to pass, come to something, result in;* *to turn out*

Figure 4

3. Output images. Each lexical entry is saved in a separate PNG file, or multiple files when it spans columns. These are the files that are fed to the SEDRA database.

Once uploaded to the SEDRA database, the images await tagging by lexicographers (recall that they are merely unsearchable images). A tool is provided on the SEDRA website which is accessible to users of type "tagger" or "lexicographer." The tagger is shown an image and can type the headword in a textbox. If the headword already exists in the SEDRA lexical database, the user picks that lexeme and tags the image with it. If the lexeme does not exist in the SEDRA database, the user can create a new lexeme entry and tag the image with the newly created lexeme.

Conclusion

This article described an algorithm that takes a scanned dictionary and saves each of its lexical entries in one or more files. The algorithm makes use of histogram computation and does not require any linguistic knowledge. It can be applied to scanned images of any dictionary that is typeset systematically. A caveat, hinted at above, needs to be stressed. The algorithm fails if the input images are skewed or full of specks or smudges. There are software libraries that can be used to de-skew and de-speckle images programmatically, but we have chosen to do these tasks manually as the number of dictionaries that we needed to process is small.

NOTES

[1] For a comprehensive history of the contribution of philology to the humanities, see James Turner's *Philology: The Forgotten Origins of the Modern Humanities* (Princeton: Princeton University Press, 2014), especially Chapters 5 & 9.

[2] William Clocksin, "Towards Automatic Transcription of Estrangelo Script" in *Hugoye: Journal of Syriac Studies* (2003) 6.2: 249–268.

[3] T. Audo, *Sīmtā ḏ-lešānā suryāyā (Dictionnaire de la langue chaldéenne)*. 2 vols. Mosul: Imprimerie des pères dominicains, 1897–1901. Reprint in 1 vol., Chicago: Assyrian Language and Culture Classes Incorporated, 1978; Stockholm: The Assyrian-Federation in Sweden, 1979; Glane/Losser: Bar Hebraeus Verlag, 1985. Reprint in 2 vols. titled *Treasure of the Syriac Language, A Dictionary of Classical Syriac,* with a new introduction by G. A. Kiraz and abbreviation list by Y. Unval, Piscataway: Gorgias Press, 2008. [in Syriac].

[4] K. Pearson, "Contributions to the Mathematical Theory of Evolution. II. Skew Variation in Homogeneous Material." *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 186 (1895): 343–414.