DIGITIZATION OF AQUILINA'S MALTESE-ENGLISH DICTIONARY: TOWARDS A
COMPREHENSIVE ONLINE LEXICON FOR MALTESE SPEAKERS

By

MITCHELL ROBERT CRAVENS

A Thesis Submitted to the Honors College
In Partial Fulfillment of the Bachelors degree
With Honors in
Linguistics

THE UNIVERSITY OF ARIZONA

December 2013

Approved by:

Dr. Adam Ussishkin
Department of Linguistics
The University of Arizona supports the dissemination and preservation of scholarship produced by University of Arizona faculty, researchers, and students. The University Library, in collaboration with the Honors College, has established a collection in the UA Campus Repository to share, archive, and preserve undergraduate Honors theses.

Theses that are submitted to the UA Campus Repository are available for public view. Submission of your thesis to the Repository provides an opportunity for you to showcase your work to graduate schools and future employers. It also allows for your work to be accessed by others in your discipline, enabling you to contribute to the knowledge base in your field. Your signature on this consent form will determine whether your thesis is included in the repository.

Name (Last, First, Middle)
Craven, Mitchell

Degree title (eg BA, BS, BSE, BSB, BFA):
BA Linguistics, BA Spanish

Honors area (eg Molecular and Cellular Biology, English, Studio Art):
Linguistics

Date thesis submitted to Honors College:
11 Dec 13

Title of Honors thesis:
Digitization of Aquila's Maltese-English Dictionary: Towards a Comprehensive Online Lexicon for Maltese Speakers

The University of Arizona Library Release Agreement
I hereby grant to the University of Arizona Library the nonexclusive worldwide right to reproduce and distribute my dissertation or thesis and abstract (herein, the "licensed materials”), in whole or in part, in any and all media of distribution and in any format in existence now or developed in the future. I represent and warrant to the University of Arizona that the licensed materials are my original work, that I am the sole owner of all rights in and to the licensed materials, and that none of the licensed materials infringe or violate the rights of others. I further represent that I have obtained all necessary rights to permit the University of Arizona Library to reproduce and distribute any nonpublic third party software necessary to access, display, run or print my dissertation or thesis. I acknowledge that University of Arizona Library may elect not to distribute my dissertation or thesis in digital format if, in its reasonable judgment, it believes all such rights have not been secured.

☐ Yes, make my thesis available in the UA Campus Repository!
Student signature: ___________________________ Date: 11 Dec 2013
Thesis advisor signature: ___________________________ Date: 11 Dec 2013

☐ No, do not release my thesis to the UA Campus Repository.
Student signature: ___________________________ Date: ___________________________
Abstract

For over ten years, the Psycholinguistics and Computational Linguistics Lab at the University of Arizona has attempted to digitize Joseph Aquilina’s *Maltese-English Dictionary* (first published in 1996) in the interest of providing a comprehensive online dictionary and searchable lexicon for Maltese speakers. The initial premise of this project was to use Optical Character Recognition Technology (OCR) to automatically scan all pages in Aquilina’s dictionary into a human and machine-readable .rtf format, but problems with the OCR technology led to a long process of manual error correction of the mistakes found in every page-file. The author, as the main corrector of these systematic errors, spent several years fixing them by hand, then realized that the employment of automation and the greater use of technology could make this process significantly easier and more efficient. This paper details the techniques the author, in conjunction with several colleagues, has developed to speed up the correction process and ensure a more accurate, complete final product. In addition, it gives guidelines and code for the creation of the searchable database referenced above, which will serve as an important tool to online Maltese speakers as they seek to express themselves both online in and everyday life.
Introduction

The Maltese language has a long history of resilience despite continual invasion and linguistic dominance from foreigners. The Maltese islands, Malta, Gozo, and Comino are located approximately halfway between Sicily and Tunisia, which has made them a hotbed of occupation for forces from both sides of the Mediterranean (Brincat 1). Among these invaders are the Romans (218 BC - 535 AD), the Arabs (870 AD - 1048 AD), and the Normans (1090 - 1530) (ibid & Marsh 718). The Arab and Norman occupations had especially important consequences for the language of Malta, as their long cultural and political dominance resulted in the imposition of their native languages on the Maltese people (Brincat 1). This led to the integration of many Arab and Sicilian grammatical structures and vocabules into Maltese, which essentially created a Semitic language with a majority-Romance lexicon (Aquilina 1958, p. 58). In addition, Malta was a British colony from 1800-1964, which brought with it the influx of English words into Maltese (Mitchell 381). The language today consists of approximately of 32% etymologically Arabic words, 52% Sicilian and Italian words, and 6% English words (Brincat 1). It is written in a Latin script with several non-standard characters; these are the letters ċ (pronounced ʧ̆), ġ (ʤ̆), and ż (z), and the digraph gh (ʍ) (“New Maltese Orthography” 468).

Maltese today shares the title of Official Language of the Maltese government with English (Mitchell 382). The country is officially bilingual, with Maltese being used primarily at home and sometimes in official functions, and English being used in the school system and in many administrative contexts (Brincat 1). Over 3/4 of all Maltese are bilingual, and everyday speech, especially from the country’s younger generations, often consists of code-switching between Maltese and English (ibid). Despite this, Maltese’s presence in digital media, especially
on the Internet, is not strong. Currently there are no comprehensive dictionary resources on the
web, which leaves Maltese speakers with no means to quickly and efficiently clarify their
linguistic questions when creating online resources exclusively in Maltese or when translating
phrases between Maltese and English. A comprehensive print resource does exist, however, in
the form of Joseph Aquilina’s *Maltese-English Dictionary*, which was published in several
volumes from 1987-2000.

Thus this project, performed by the Psycholinguistics and Computational Linguistics Lab
(PsyCol) at the University of Arizona, in conjunction with The University of Malta and Midsea
Books Limited, set out to digitize Aquilina’s *Dictionary* in the interest of providing Maltese
speakers with electronic resources that would allow them to employ the Maltese language in
their interactions online. The goal of our project is two-fold: First, to convert every dictionary
page into electronic format, and secondly, to create a searchable, free database that allows
Maltese speakers to search individual words and phrases in both Maltese and English and returns
relevant lexicographical information for these queries. This paper will explain the process in
which previous work was done on these undertakings, both by the author’s and others’ hands,
and demonstrate several automation techniques the author has devised in the interest of making
this yet-to-be-completed project more efficient and rapid. Finally, the author will explain the
methods by which the aforementioned searchable online database will be created, and highlight
some of the important functionalities it makes available to Maltese users of the Internet.

**Work Completed Prior to Automation**

The work completed prior to the implementation of automation techniques, which was
performed from 2002-2013, consisted of 6 principal phases: First, the physical dictionary was
deconstructed and each page was photocopied and enlarged to a 8.5 x 11 inch reprinted size.
Then, these pages were scanned into .rtf files using OCR technology. Next, an initial, brief sweep of the scanned pages was made to ensure they had rendered in the proper format, in which any extremely salient OCR errors were corrected. Then, pages with incorrect scanning of non-Roman characters (including Arabic, Greek, Hebrew, and Aramaic glyphs) were marked and archived for future reconciliation with their hard copies. Later, these non-Roman characters were inputted manually into the problematic pages using custom keyboards for their corresponding languages and Unicode character codes. Finally, every page in the digital dictionary was checked for remaining OCR and non-standard character problems. Today, due to human error and the differing degrees of focus each research participant put into our pursuit, over 100 pages with OCR errors, formatting issues, or non-Roman character mistakes remain.

I will now chronicle how each step in the pre-automation process was performed, and display what the digital pages looked like so that the reader has an idea of how they have evolved and been refined over time. Each image below will chronicle the progression of page 315 of the dictionary over its various iterations. I will provide several comments about the state of this page at each step for the reader to have a more concrete idea of what changed over time.
1. Deconstruction of the dictionary and photocopying into 8.5 x 11" size:

**FEDA** [Sol 159v/121r]: , v.t. (imperf. b-i, pp. m-i/ti). 1. To redeem (Krisu miet biex jifídina, Christ died to redeem us). 2. To ransom, purchase the freedom of a slave, etc. (~ l-i-isera u takom il-helsem, he redeemed the slaves and set them free). 3. To buy back, recover or take out of pawn (~ d-deheb li keli mihran, he has taken the gold out of pawn: ~ c-eens, he redeemed the emphyteusis by purchasing the ground rent payable on it). 4. To recover the expense incurred in the purchase of s.th. (il-parata bhiggata imma ma fæghix-ix, I sold the potatoes at a loss, did not recover the cost). [Car] FIDÌ/FIDU, v.n. Redeem, redemption, recovery of the cost ([Aq] m-inixx ser taphitini haqq il-~ taphi-compounds, you are not giving me enough to recover the cost). [Sol 159r] FIDWÀ, v.n. Redemption (~ b- t-bnieddal midmab, man's redemption from sin; [GMA] siwet ~ twa Lhoud, said of s.th. for which one had to spend a great deal of money). [Id. 159v] FEEDEJ, n.m. (~ i. & pl. ~ jo) Redeemer; il-~, the Redeemer Jesus Christ. [Car] TIFĐIJA, n.f. (morphol. v.n. of tadda II) Redemption. [VMB] TIFĐIJA, v.i. To barely cover expenses (in a transaction) (ki x marru mas-sw? Kemmm kemm wished ji~, how are you doing in business? Just well enough not to lose money). [Sol 150r] NIFĐA, v.i. (pass. of i.) To be redeemed, ransom, taken out of pawn. (~ Ar. fidda, feda ~ fadd, to give a ransom for a captive (Hava), fida~: lor M. form cp. M. faddalit. ~ faddalit, the Redeemer (id.); nifda~ nifadda VII, to be redeemed (id.).)

**FEDEJ** Redeemer; n.m.ag. of feda, q.v.

**FEDDEJ** To tame, domesticate. I, see fiddi.

**FEDEKOMMISS** [Mag], n.f. (~ jiet) Federation, [Bus] ~ ALIST, n.m. (~ a. pl. ~ i) ~ ALISTIA, n.zc. (~ a. pl. ~ i) Federalist, [Id.] ~ ALIZMU, n.m. Federalism. [Id. ~ a. v.t. To imperate (~ a. pl. ~ jo) Fidescommi. trust. [Id.] ~ ARJU, n.m. (~ a. pl. ~ i) (leg.) Fidescommunic. trust. (< lt. fidescommunic-o -arjo)

**FEDEIZZONJI** [Mag], n.f. (~ jiet) Federate, [Bus] ~ ALIST, n.m. (~ a. pl. ~ i) ~ ALISTIA, n.zc. (~ a. pl. ~ i) Federalist, [Id.] ~ ALIZMU, n.m. Federalism. [Id. ~ a. v.t. To imperate (~ a. pl. ~ jo) federation, < federation, ~ arifederat) To federate. (< federation, ~ arifederat) To federate. (< federation, ~ arifederat) To federate. (< federation, ~ arifederat) To federate. (< federation, ~ arifederat) To federate.

**FEZZ [Vass] (~ FEGG), v.i. & n. 1. To emerge, come into view (ix-xemex ~ et, the sun has appeared, emerged). 2. To burst or break open, to break through (beda j~ l-i-biset, the sprouts have begun to break through [the soil]). 3. [Fai] To expand (womb, in giving birth). 4. vn. Emerging, appearance ([Car] mal~ tax-xemex, with the rising of the sun [on the horizon]); [FEGGA, v.n. & ~ a. 1. Sudden and brief appearance. 2. A craving for (tariqjeu ~ , he felt a sudden craving for). 3. a.f. [VMB] kwehba, shooting star, meteor. 4. [Sol 166r] Impetus. [Car] FEZGIEHN, vn. Emerging, appearance. [For poss. origin cp. It. affacc-tare ~ faetté (fàgger) ~ fegg ~ cp. also Ar. b-wa fagg'a, to fall upon s.o. unexpectedly, surprise (Havat)].

**FEZMIA** [KM] 14, p. 41], n.f. (pl. ~ jo) Physical appearance such as creates a personal impression of one's character (m'genx bireeja ~ sejed tatt miyom, imjarqwa wigj, he doesn't look crazy at all, he is just very nervous). [etym. 7]

**Fezattet**. One's favourite person; one's blue-eyed boy.

**FEZATON** [q.v. (Chickens', pigeon's, etc.) liver as prepared for cooking. (~ t. fegarini, pl.)

**FEHEMI** [Sol 159v] (~ FHMM), v.t. & i. (imperf. b-e; pp. miiefham). 1. To understand (qad tìfhem x~ qad ngaddik, are you following what I am saying? ~ miiefham, are you following me; qed niiefham, I am with you). ~ bi, (i) he meant, understood (x'tifhem biha din il-kema), what do you mean (or understand) by this word?, (ii) he understood: a language (ma jifhemx bil-Franki, he does not understand French; ma jifhem b'seijn, he doesn't understand any language; (ii) he is obsolete, pig-headed; taw x'tfhem li ~, he gave him to understand that ... niiefham jien li ~, I think, have a feeling that ... 2. To hold an opinion about s.th. (kulhadd jishtem kif jifhem, everyone follows his own opinion; kif tiiefhemh~, what is your opinion? ma niiefhamich hekk, I am not of that opinion). 3. To have knowledge of (of, b) or ability in (ma jifhemx, he has no knowledge of the thing; jifhem al-hif, he is a man of great knowledge, well-versed in the matter, an expert; ma jifhemx fil-nigawja, he has no experience or ability in business; jifhem fil-medićina, mutika, etc., he is an expert in medicine, music, etc.); dak fex jifhem~ that man has no knowledge of any kind; iss, ghass x'tfhem, kifj, you think you know but you don't; (vulg.) ma jifhemx fil-hurra~, he is an arrant dunce, an ignoramus). 4. To intend (~b-hek niifham nafighe biema, a.m. (~ b ~ a. pl. ~ jo) Fidescommi. trust. [Id.] ~ ARJU, n.m. (~ a. pl. ~ i) (leg.) Fidescommunic. trust. (< lt. fidescommunic-o -arjo)

[Car] FEMHI, n.m. (~ jiet) Opinion, understanding (ma nafighe x~ ghandu, I can't understand his mind); [Vass] ~ et id-drawwa, skill, expert skill; ghandu ~ ghollih, he is a queer man, you never know what he thinks; blaghinghair ~ unreasonable (niiejemd bbl~ unreasonable/pig-headed man); ~ x~ xAd., in s.o.'s opinion; salt~ li, of the opinion that (m'genx bi xa din il~ , I am not of this opinion; biiejemd tatt ~ ikebha, a man who does not easily change his opinion, who means what he says or thinks; aheb tatt ~ wadda, we are of the same opinion); biiejemd tatt hejemux, an opinionated man. 2. Intention; b~ i, with the intention of (ghandu bbl~ li iftalq minn Malta, he still intends to leave Malta; ~ tajba, with a good intention). [Fai: II] FUMI, a.m. (~ o. ~ n) Intelligible. [H] (Pant GLM, p. 118)
The image above is simply a .pdf copy of the 8.5 x 11” scans that were performed at the outset of the project. As the Automatic Character Recognition (OCR) software has not been applied yet, the page appears exactly as it does in the physical dictionary. Therefore, none of the problems due to OCR are present yet.

2. Initial OCR:

feda
FEDA [Sol 158v/121rJ, (i)FDJ, v.t. (imperf. i-i, pp. mi[di] i). To redeem (Kristu miet biex jifidina, Christ died to redeem us). 2. To ransom, purchase the freedom of (a slave, etc.) (- i-ilsiera u tahom il-helsien, he redeemed the slaves and set them free). 3. To buy back, recover or take out of pawn (- d-deheb li kella mirhun, he has taken the gold out of pawn; i-cens, he redeemed the emphyteusis by purchasing the ground rent payable on it). 4. To recover the expense incurred in the purchase of s.th. (il-patata bhgetha imma ma idejtheix, I sold the potatoes at a loss, did not recover the cost). [Car] FIDI/FIDU, vn. Redeeming, redemption, recovery of the cost ((Ar) m‘intix ser taghtini haqq il-- taghha, you are not giving me enough to recover the cost). [Sol 159rJ FIDWA, v.n.u. Redemption (il-- tal-bniedem midstub, man's redemption from sin; (GMAJ sweiet _ ta`Lhudi, said of s.th. for which one had to spend a great deal of money). [id. 158v] FEDDEJ, n.m.ag. ("n. & pl. - ja) Redeemer; il-, the Redeemer Jesus Christ. [Car] TIFDIJA, n.f. (morphol. the vn. of *fedda 11) Redemption. VI [VMBJ TFIEDA., v.i. To barely cover expenses (in a transaction) (kif tmorru mas-saq? Kemm kemm which jied ji -, how are you doing in business? Just well enough not to lose money). VII [Sol 130r] NFEDA, v.i. (pass. of I) To be redeemed, ransomed, taken out of pawn. (Ar.I.S--y’FDJ, feda < l.S-- fada, to give a ransom for (a captive) (Hava); fidwa: for M. form cp. "","LJi-- fadiwiit (= M. .fidwiet), pl. of "",-fidad, ransom (id.): feddej < l.S...lJi...fali fididi, the Redeemer (id.); rifeda < l.S...l4;1 nfada VII, to be redeemed (id.)]

feddej. Redeemer; n.m.ag. of feda, q.v.

fedda. To tame, domesticate. II, see ,fidil. FEDEKOMMESS [Fal 11 M], a.m. (f. ]a, pl. - i) & n. (leg.) 1. Entrusted. 2. n. (pl. - i) Fideicommisum, trust. [id.J - ARJIU, n.m. (pl. -i) (leg.) Entrusted, fiduciary. [<!t. jidecommiss-o; -arja]
- i) Federalist. [id.] - ALIZMU, n.m.


FEGG [Vass] (y’FGG), v.i. & n. 1. To emerge, come in to view (ix-xemx - et, the sun has appeared, emerged). 2. To burst or break open, to break through (beda j - l-inbit, the sprouts have begun to break through (the soil)). 3. [Fal] To expand (womb, in giving birth). 4. vn. Emerging, appearance (/[Car] mal- - tax-xemx, with the rising of the sun (on the horizon)). FEGGA, vn.u. & a. 1. Sudden and brief appearance. 2. A craving for (tatu/qieta -, he felt a sudden craving for). 3. a.f. [VMBJ kewkba -, shooting star, meteor. 4. [Sol116r] Impetus. [Car] FEGGIEN, vn. Emerg.

315

fehem

ing, appearance. [For poss. ongm cp. !t. affaċċiare-faċċé (faggìfegg - cp. also Ar.-fàga’à, to fall upon s.o. unexpectedly, surprise (Hava)]

FEGHIA [KMZ 114, p. 41] J, n.f. (pl. -iet) Physical appearance such as creates a personal impression of one's character (m’ghandux - zejjed ta’ m’inqun, imqarqar’ wsq, he doesn't look crazy at all, he is just very naughty). [etym. ?]

fegattell. One's favourite person; one's blue-eyed boy.

Var of jìgelitell, q.v.

FEGATINI, n.pl. (cul.) (Chicken's, pigeon's, etc.) liver as prepared for cooking. [<!t. fegatini, pl.]

FEHEM [Sol 159rJ (i)FHIM, V.t. & i. (imperf. i-e; pp. mi[hum]. 1. To understand (qet fidem x’qed nghidekk? are you following what I am saying?: jihimnti?, are you following me?; qed nijhem, I am with you); - bi, (i) he meant, understood (x’tijhem biha din il-kelma?, what do you mean (or understand by this word?)), (ii) he understood: a language (ma jifhimx bil-Franċiż, he does not understand French); ma njhem b’xejn, (i) he doesn't understand any language; (ii) he is obstinate, pig-headed; tah x’jifhem li ... , he gave him to un-
(kulhadd jaghmel kif jifhem, everyone follows his own opinion; kif tifhimha?, what is your opinion?; ma nifhimhix hekk, I am not of that opinion). 3. To have knowledge (of, ji) or ability in (ma jifhemx, he has no knowledge of the thing; jifhem Ħafna, he is a man of great knowledge, wellversed in the matter, an expert; ma jifhemx fn-negozju, he has no experience of or ability in business; jifhem jil-medicina, muzika, etc., he is an expert in medicine, music, etc.); dak jiex jifhem?, that man has no knowledge of any kind; iss, għax tifhem, ħej!, you think you know but you don’t; (vulg.) ma jifhemx ħa’, he is an arrant dunce, an ignoramus). 4. To intend (b’hekk nifhmu negħa gieħ u qima, with this we intend to show her honour and respect; jifhmu li din is-sena r-reforma għandha tgħadd, they have in mind to carry out the reform this year). 5. Intellect ([Aq Taħti Saltniet] irrat tal-ħajja la għandha - u anqas dehen, the wheel of life has neither intellect nor understanding). [Carj FEHIM, vn. Understanding. [Sol 159r] FEHMIA, vn.u. (pI. -iet) Opinion, understanding (ma nafx x’ għandu, I can’t understand his mind); [Vass] - et id-drawwa, skill, expert skill; għandu - għalik, he is a queer man, you never know what he thinks; bla/mingiżjair -, unreasonable (bniedem bla -, unreasonable/pig-headed man); ji li - ta’ x’d. in s.o.’s opinion; tal- li, of the opinion that (m’iniex ta’ din il- -, I am not of this opinion; bniedem ta’ iebxa, a man who does not easily change his opinion, who means what he says or thinks; ahna ta’ - waħda, we are of the same opinion); bniedem ta’ fehemtu, an opinionated man. 2. Intention; bil- li, with the intention of (għadu bil-xdi jiltaq minn Malta, he still intends to leave Malta; b’ - taħba, with a good intention). [Fal IIMJ FHUMI, a.m. (f. -ja, pI. -n) Intelligible. 11 [Panz GLM, p. 118J

Several structural issues exist with this OCR product. First, the Arabic included in many entries’ etymologies is not represented. Instead of rendering the Arabic characters as they appear on the scanned dictionary page, the OCR uses members of the Roman character set to represent them. Clear examples of this problem are represented in blue on the above page 315. To show a specific instance, the highlighted phrase Ar.l.Ş̣y should actually appear as Ar. فد١. Secondly, the OCR software incorrectly renders many symbols and letters. I have compiled an extensive list of these errors that I will address with a find-and-replace script in the following section of this thesis. In page 315, these errors are highlighted in blue. These types of errors are split into two different categories: First, the incorrect representation of individual letters, and secondly, the unfaithful replacement of grammatical symbols. Individual letters are often replaced by a similar-looking letter or at times a combination of two different letters that when juxtaposed are similar in appearance to a single letter. In page 315, a common and repetitive error is the representation of the letter l with an uppercase I. This can be seen in a few examples highlighted in blue on the page; one particular instance is the text pl. (abbreviation for “plural”) being written as pI. In mistakes with symbols, two common errors emerge. To start, the closing bracket ] is frequently transcribed as J. This is evident in almost all the etymologies on the page; in one case, what should be [<lt. fegatini, pl.] is written as [<lt. fegatini, pl.J. Finally,
the character ~ is consistently represented as - in page 315. This is significant because ~ is used throughout the dictionary as a placeholder for the headword, both in determining variations of the word and in creating example sentences.

It is important to note that the text above is not fully representative of how each page appeared after the initial OCR process was performed. In fact, each page was originally scanned into a two-column form so as to emulate the form of the physical copies of the dictionary pages. This form only appears when each .rtf document is opened in Microsoft Word. I have foregone this form at this step because even more errors appear when each .rtf page is opened in Word. As a result of the additional markup Microsoft Word performs (over a traditional text editor, for example), many additional problems arise when each document is displayed in this program. These include jumbled text, unfaithful positions of text boxes, and superfluous images. Unfortunately, much of the work done on the dictionary has been in Microsoft Word, which caused us to have to occupy ourselves with many more errors than necessary. As the final product of this project is going to be in .rtf form, I have decided to only include .docx versions of page 315 where absolutely necessary, as the additional errors simply make the OCR’d text more difficult to read. I will, however, address some of the problems caused by working in Microsoft Word in the next section, which concerns the initial correction of highly visible OCR problems.

Looking at page 315 above, it is obvious that a significant amount of work is required to bring this page towards being a carbon copy of the .pdf version seen in the previous section. Almost all of the pages on which we initially performed OCR suffer from the same problems delineated above. However, it is important to keep in mind that despite the problems in accurately depicting the Arabic, letters, and symbols, the majority of the text is in fact intact. Many entire definitions and sample sentences, such as “To be redeemed, ransomed, taken out of
pawn” are completely correct, and faithful to the original dictionary page. This is inextricably the case for the grand majority of dictionary pages. That is to say, the problems that exist are not so overwhelming that OCR is not worth performing. Although the error correction process that followed, as described below, was tedious and long, it was certainly less time consuming than typing up all the pages from scratch.

3. Initial Correction Process:

During this step, an initial sweep of each page was made to seek out and correct any extremely salient OCR errors. As I was not a part of the project at this time, I am unable to determine exactly what was done, but based on seeing various previous versions of the pages I have worked on, I can infer that several important types of corrections were made. These corrections concern the problematic formatting caused by viewing each .rtf dictionary file in Microsoft Word. I will include a .docx copy of page 315 below to demonstrate the issues caused by Word.
As one can see by comparing the look of the page between its .rtf form and Microsoft Word rendering, Word creates additional formatting by attempting to render the page in its original form. Word uses text boxes to attempt to emulate the template of the hard copy dictionary pages. Two large text boxes columns are placed adjacent to each other to represent the two-column format present in the actual dictionary, and additional text boxes are created to correctly place the headwords and the page numbers. These text boxes, along with the additional text formatting added by Word, are the crux of the problems that were seemingly worked on during the initial correction phase of our project. In the following paragraph, I will explain how the above version of page 315 is representative of some of these errors, and demonstrate what additional errors it lacks.

Page 315 is a typical example of how Microsoft Word adds superfluous text formatting to each .rtf dictionary file. There are three principal text-related errors caused by Word: First, extra spaces and tabs are added within the large columns; secondly, fonts and font sizes are changed by Word to try to accommodate the apparent size of text on the page; and lastly, non-existent numbers and symbols are added to many pages as a result of incomplete formatting transfers from the scanned .rtf files. In the above page 315, extra spaces and tabs are represented in blue highlighting, while font issues are highlighted in blue and formatting-related numbers in yellow. In terms of unnecessary spacing, note how the phrase “To be redeemed, ransomed, taken out of pawn” contains a large amount of tabs between “be” and “redeemed.” When editing in Word, it is impossible (at least from the Word visual interface) to delete these additional tabs, as the interface does not let the user highlight them or place the cursor in the white space they create. Much ado was made about these uneditable white spaces, and a significant amount of time was spent to correct them. After an inordinate amount of time looking for a solution, I found out that
copying and pasting the relevant text into a text editor, then re-pasting it back into the Word document often worked, but not always. Essentially, like many of the other issues caused by Word, this is a cosmetic one, and we should have realized earlier that it was irrelevant to editing the actual text of the dictionary.

The font size and formatting problems created by Word are visible in the text jumble of ❉.$jL..ǐJl, which is highlighted in blue above. The size of this text is smaller than the rest of the text on the page, which is by default set to size 9; the above phrase is size 8. It is also italicized, which is unfaithful to the original page. Finally, the symbols [४;] and ……. are added into page 315, and do not exist in the page when opened by a text editor. Overall, the additional problems presented by Microsoft Word created surplus complications that our team sought to correct in the initial OCR correction phase. However, we would have done better to do all our editing in a text editor, as we essentially were making corrections that were only in the text due to Word’s additional formatting.

4. Non-Latin Character Markup:

The next step in the editing process was inputting ‘(@)’ symbols wherever jumbled non-roman characters were found. These characters included Arabic, Greek, Hebrew, and Aramaic text, among others. The purpose of this endeavor was to mark where these characters occurred so that someone who was literate in Arabic and Greek (these are the two principal non-roman languages that occur throughout the dictionary) could go through the pages later and input the characters manually. These pages were found by doing a mass search of the entire set of dictionary files for the letters ‘Ar.’, ‘Gr.’, and ‘Gk’ (for some reason, Greek is abbreviated in both these ways). Using this method, over 800 pages were identified. Then, several other
researchers and I went through the pages one-by-one, searched for ‘Ar.’, ‘Gr.’, and ‘Gk.’ again in every page, and replaced the Arabic, Greek, and other non-Roman characters we found with @ symbols. As this step is fairly simple and self-explanatory, I will not include a copy of page 315 at this stage, and move on to the next step of our project.

4. Correction of non-roman characters and phrases:

At this point, the @ symbols replaced for the non-roman text were finally edited to make the pages more true to their original, hard-copy form. For several months, a colleague and I went through the set of approximately 800 pages and replaced the @ symbols with the correct non-Western text. We used foreign keyboards, especially Greek and Arabic configurations, to accomplish this. Inputting the non-Western text into Microsoft Word directly was difficult due to formatting problems (unfortunately, we were still working in Word at this point), so often we compiled the Arabic and Greek text in a text editor, then copied and pasted it into word. Below I will display a copy of page 315 with corrected Arabic.

fedda
315
femhem

FEDA [Sol 158v/121r], (√FDJ), v.t. (imperf. i-i, pp. mifdi) 1. To redeem (Kristu miet biex jifdina, Christ died to redeem us). 2. To ransom, purchase the freedom of (a slave, etc.) (~ l-ilsiera u tahom il-helsien, he redeemed the slaves and set them free). 3. To buy back, recover or take out of pawn (~ d-deheb li kellu mirhun, he has taken the gold out of pawn; ~ ċ-ċens, he redeemed the emphyteusis by purchasing the ground rent payable on it). 4. To recover the expense incurred in the purchase of s.th. (il-patata bghettha imma ma jdejthieix, I sold the potatoes at a loss, did not recover the cost). [Car] FIDI/FIDU, vn. Redeeming, redemption, recovery of the cost ([Aq] m'intix ser taghtini haqq il-~ taghha, you are not giving me enough to recover the cost). [Sol 159r] FIDWA, vn.u. Redemption (il-~ tal-bniemed middnub, man's redemption from sin; [ĠMA] swiet ~ ta'Lhudi, said of s.th. for which one had to spend a great deal of money). [id. 158v] FEDDEJ, n.m.ag. (f. & pl. ~ ja) Redeemer; il-~, the Redeemer Jesus Christ. [Car] TIFDIJA. n.f. (morphol. the vn. of *fedda II) Redemption. VI [VMB] TFIEDA,. v.i. To barely cover expenses (in a transaction) (kif tmorru mas-suq? Kemm kemm wiehed ji ~, how are you doing in business? Just well enough not to lose money). VII [Sol 130r] NFEDA, v.i. (pass. of I) To be redeemed, ransomed, taken out of pawn. [(Ar. دمـ دمـيد) feda < YI fada:, to give a ransom for (a captive) (Hava); fidwa: for M. form cp. @ YI fada: wa:t (= M. fidwiet), pl. of fidja, ransom (id.); feddej < ٌفـدـدـي, the Redeemer (id.); nfeda < نـفـدـدـي nfeda: VII, to be redeemed (id.)]

feddej. Redeemer; n.m.ag. of fedda, q.v.
feddel. To tame, domesticate. II, see fidil.

FEDEKOMMESS [Fal IIM], a.m. (f. ~a, pl. ~i) & n. (leg.). 1. Entrusted. 2. n. (pl. ~i) Fideicommissum, trust. [id.] ~ ARJU, n.m. (pl. ~) (leg.) Entrusted, fiduciary. [<It. fidecommiss-o; -aria]

fedeltà. Fidelity, faithfulness, loyalty. See fidil.


FEGG [Vass] (~FGG), v.i. & n. 1. To emerge, come in to view (ix-xemx ~et, the sun has appeared, emerged). 2. To burst or break open, to break through (beda j~l-inbiet, the sprouts have begun to break through (the soil)). 3. [Fajl] To expand (womb, in giving birth). 4. vn. Emerging, appearance ([Car] mal~ tax-xemx, with the rising of the sun (on the horizon)). FEGGA, vn.u. & a. 1. Sudden and brief appearance. 2. A craving for (tatu/ġietet-, he felt a sudden craving for). 3. a.f. [VMB] kewkba ~, shooting star, meteor. 4. [Sol 166r] Impetus. [Car] FEGGIEN, vn. Emerging, appearance. [For poss. origin cp. It. affacciare > faċċi > fegg ~ cp. also Ar. ICO fagaḍa, to fall upon s.o. unexpectedly, surprise (Hava)]

FEĦHA [KMŻ 114, p. 41], n.f. (pl. ~iet) Physical appearance such as creates a personal impression of one's character (m'għandux ~żejjed ta' mignun, imqarqaċ wiqs, he doesn't look crazy at all, he is just very naughty). [etym. ?]

fegatell. One's favourite person; one's blue-eyed boy. Var of jigatell, q.v.

FEGATINI, n.pl. (cul.) (Chicken's, pigeon's, etc.) liver as prepared for cooking. [<It. fegatini, pl.]

FEHEM [Sol 159r] (~FHM), v.t. & i. (imperf. i-e: pp. niżhum). 1. To understand (qed tifhem x'qed nghilkew?, are you following what I am saying?: jilmìni?, are you following me?: qed niżhum, I am with you); ~ bi, (i) he meant, understood (x'tifhem biha din il-kelma?, what do you mean (or understand) by this word?),(ii) he understood: a language (ma jifhem bil-Franċez, he does not understand French); ma jifhem b'xewjin, (i) he doesn't understand any language; (ii) he is obstinate, pig-headed; tah x'tifhem li ..., he gave him to understand that ...; niżhum jien li ..., I think, have a feeling that ... 2. To hold an opinion about s.th.(kalhadd jaqhnell kif jifhem, everyone follows his own opinion; kif tifhimha?, what is your opinion?; ma niżhimmie hekk, I am not of that opinion). 3. To have knowledge (of, fi) or ability in (ma jifhemx, he has no knowledge of the thing; jifhemx hafna, he is a man of great knowledge, well-versed in the matter, an expert: ma jifhemx fi-n-zeigt, he has no experience of or ability in business; jifhem fil-medicina, muzika, ečč., he is an expert in medicine, music, etc.); dak jexx jifhem?, that man has no knowledge of any kind; iss, ghax tifhem, hej!, you think you know but you don't; (vulg.) ma jifhimx jil-hara', he is an arrant dunce, an ignomorous. 4. To intend (b'hekk niżhmu naghtuha gieħ u qima, with this we intend to show her honour and respect; jifhmu li din is-xena riforma ghandha tghaddi, they have in mind to carry out the reform this year). 5. Intellect ([AQ Taht Tiet Salmiet] irrota tal-hajja la ghandha ~ u angas dehen, the wheel of life has neither intellect nor understanding). [Car] FEHIM, vn. Understanding. [Sol 159r] FEHMI/A, vn.u. (pl. ~iet) Opinion, understanding (ma nafl x' ~ ghandu, I can't understand his mind); [Vass]~ et id-drawwa, skill, expert skill; ghandu ~ ghalih, he is a queer man, you never know what he thinks; bla/minghajr ~, unreasonable (bniemed bla~, unreasonable/pig-headed man); fil-~ ta' x. hd., in s.o.'s opinion; tal-~ li, of the opinion that (m'niexx ta' din il~ ~, I am not of this opinion; bniemed ta'~ iehsa, a man who does not easily change his opinion, who means what he says or thinks; ahna ta' ~ wahda, we are of the same opinion); bniemed ta'fehmutu, an opinionated man. 2. Intention; bil- li, with the intention of (għadu bil~ li jillaq minn Malta, he still intends to leave Malta; b'~ tajba, with a good intention). [Fal IIM] FHUMI, a.m. (f. ~ja, pl. ~n) Intelligible. II [Panz GLM, p. 118]

Unfortunately, even at this point in the process, not all of the Arabic was
completely finished. This is due to encoding problems between .rtfs worked on via PC and Mac. We are still not sure of the source of this problem, but often files that had been worked on using both types of operating systems contained jumbled Arabic text. The several roman character-represented Arabic entries on page 315 above are a result of this issue. For example, what should be written as فصداوات is represented as ÝїÇæÇÊ; this type of error occurs in one other location on the page and seemingly was present in about 1/5 of pages after this step of editing. This was an unfortunate setback for our project, but not devastating, as we were already planning to go back through every single page of the dictionary to look for any additional errors we may have missed in our previous edits. This final step is described below.

5. Final modifications:

In this iteration, all 1,673 pages of the dictionary were reviewed for further issues related to OCR, formatting, and non-roman characters. I performed this task alone, taking several months to do so. Once again, this process was performed in Microsoft Word, which implied correcting the additional formatting complications discussed above. While correcting each page, I was principally looking for missing Arabic and Greek etymologies, formatting inconsistencies, and incorrect IPA transcriptions. I spent about 3 minutes on each page, first searching for ‘Ar.’, ‘Gr.’, ‘Gk.’ and ‘@’, then doing a visual sweep of each entry and the guide words at the top of the page. After completing this process on every digitized page, all 1673 files were sent to Malta for further editing. As the only significant modifications to page 315 were correcting the jumbled Arabic text, I will not include the current copy of that page here.
Despite the ‘finished’ pages being out of our hands, we realized that the dictionary still had significant issues, especially in terms of OCR, Arabic, and formatting errors. We determined that the potential for human error was just too great in this project for future work to be done entirely manually, and decided that future edits would have to come in an automated, code-driven form. The problem that we decided needed the most attention was the large amount of systematic OCR errors still present in the dictionary files. Many of these problems were solvable via a simple mass find-and-replace operation, but at the time our team lacked the technical expertise to write a script to do this. After gaining programming experience during the subsequent summer, I decided that I would attempt to tackle the problem using Python. The following section of this thesis describes this process and the script that resulted.

**Mass Find-and-Replace of Systematic OCR Errors**

Throughout my entire work on the dictionary to this point, I kept a comprehensive list of errors that systematically occurred in the pages I worked on. This error list consisted of over 80 items. Upon learning how to code over the past year, I realized that I was capable of creating a mass find-and-replace script for these errors. The script is written in Python and uses a dictionary with key-value pairs to substitute incorrect characters for the correct ones. It iterates over a set of .txt files within a specific directory, edits each individual file, and saves over it.

The dictionary of key-value pairs seeks to correct four principal types of errors that were consistently found throughout the dictionary. First, a few punctuation marks are doubled by the OCR; for example, the string `;;` often
appeared in the digitized pages and was replaced by a single semicolon. Secondly, both roman and non-roman letters were incorrectly represented by OCR. The Greek letter θ, for example, was frequently written as fÆ in the scanned pages. Next, symbols, particularly the radical symbol √, were sometimes incorrectly displayed. √ has over ten variations throughout the dictionary, and the find-and-replace script seeks to resolve this. Lastly, a few characters, phonetic transcriptions of the Arabic and Greek text found in various etymologies, needed to be changed to IPA notation, so the find-and-replace script accounted for this as well. The pharyngeal character D, for example, needed to be changed to D⁸ to accommodate the modern IPA character set.

Below is the find-and-replace script, which was applied to the remaining pages with errors. Note: The escaped characters beginning with \u are Unicode code points for non-ASCII characters that are represented differently in memory than standard characters. These include é, è, and the superscript pharyngeal character ª discussed in the previous paragraph.

```python
import os, glob, re

def findandreplace(replacement_text, keypairsdic):
    for i, j in keypairsdic.items():
        replacement_text = replacement_text.replace(i, j)
    return replacement_text

path = '/Users/mitchellcravens/Thesis/Yes/'
for infile in glob.glob(os.path.join(path, '*.txt')):
    with open(infile, "r+", encoding = "UTF-8") as f:
        text = f.read()
        f.seek(0)
        dictionary = {'>>':'))',',j':'√',;;':';','iI':'II','ã':'√','v':'

\u1E0D':'d\u02C1','\u1E0C':'D\u02C1','\u1E63':'\u00B9\u0302\u0301','\u1E62':'S\u02C1','\u1E6D':'Tu\u02C1','\u1E6C':'Tu\u02C1',
        'aI':'a\u0301','eI':'e\u0301','iI':'i\u0301','oI':'o\u0301','uI':u0398':u03B8','/E':u03B8',':',':',':@f\u2019,

u1E0D':\u00C1,'u1E0C':\u00C1,'u1E63':u00B9,'u1E62':S,'u1E6D':Tu,'u1E6C':Tu,
        'aI':a\u0301,'eI':e\u0301,'iI':i\u0301,'oI':o\u0301,'uI':\u0398',\u03B8',',

\u1E62':S,'\u1E6D':Tu,'\u1E6C':Tu,
        'aI':a\u0301,'eI':e\u0301,'iI':i\u0301,'oI':o\u0301,'uI':u\u0398',\u03B8',',

\u1E6C':Tu,'\u1E6D':Tu,
        'aI':a\u0301,'eI':e\u0301,'iI':i\u0301,'oI':o\u0301,'uI':u\u0398',\u03B8',',
```

```
In addition to the pages that still need to be revised, this script should be applied to the rest of the pages in the dictionary. Instead of relying on manual editing, this is a simple way to replace many of the mistaken characters in a quick, automated form. However, despite its efficiency, this script currently lacks two basic characteristics. First, it is unable to correct errors that rely on context, as it only does character-for-character replacements and does not use any regular expressions. Therefore, errors like the letter O appearing instead of the number 0, which cannot be solved by a simple find-and-replace, still occur rampantly throughout the dictionary. Secondly, as will also be seen in the following section, this script is written only to process text files. Text files do not maintain the formatting included in .rtf files, so as a result important characteristics like italicization and bolding are lost. Italicization is especially significant because it delineates Maltese sample sentences. In this case, it will be necessary to either write a script that is able to parse .rtf files, or use regular expressions to enumerate the contexts in which various formatting characteristics occur. Regardless, this script is a step in the right direction, as automating find-and-replace tasks allows future workers on this project to spend more time on items that actually require human analysis.

**Database Creation Through XML Scripting of Dictionary Pages**

After the editing of the digitized dictionary pages is complete, each page
will be transposed into XML format in the interest of creating a searchable online database containing the entirety of Aquilina’s *Maltese-English Dictionary*. Ideally, this database will allow Maltese speakers to access each dictionary entry, including the definition of each headword, its inflections, its etymology, and other important information. The imagined interface of this database will consist of a search bar, where the reader can enter a Maltese word and be presented with the English entry corresponding to it. From there, he or she will be taken to the entry’s webpage, where it will be possible to select and view those individual aspects of the entry mentioned above (root, sample sentences, etc.). The page will then load the selected aspects individually when the user clicks on them, and he or she will be able to copy and paste the data required for their online or offline activity.

As previously stated, the data from each entry of the dictionary will be stored in XML format, which uses a hierarchy of tags to store information. Jeff Berry, a colleague tasked with the transposition to XML process, has devised a Python script that stores all the relevant information from each dictionary page in .xml files. This script, displayed below, can be divided into three principal sections. First, the class `Parser` is defined to create methods to identify the relevant parts (roots, definition, etc.) of each dictionary entry through the use of regular expressions. Then, the function `parseEntry` is used to write each of these parts into XML format and compile them into a single entry under the `<entry>` tag. Lastly, the `demo` function is defined in order to compile all the entries for each page of the dictionary under a solitary `<page>` tag. Dr. Berry’s script is shown below.
import re, codecs

class Parser:
    def __init__(self):
        self.malteseChars = u'\u010a\u010b\u0120\u0121\u0126\u0127\u017b\u017c'
        self.malteseCaps = u'\u010a\u0120\u0126\u017b'
        self.arabicChars = []
        for i in range(1536, 1792):
            self.arabicChars.append(unichr(i))

    def fillTilde(self, instring, stem):
        x = instring.split('-
')
        return stem.lower().join(x)

    def findStem(self, entry):
        stemP = re.compile(u'([A-Z] + self.malteseCaps + u"]*[ \|,])"
        stem = stemP.match(entry)
        try:
            stem.group(1)
            return stem.group(1)
        except AttributeError:
            stemP = re.compile(u'([a-z] + self.malteseChars + u"]")"
            stem = stemP.match(entry)
            return stem.group(1)

    def findSource(self, entry):
        sourceP = re.compile(u'([a-zA-Z0-9. ]*)')
        source = sourceP.search(entry)
        try:
            source.group(1)
            return source.group(1)
        except AttributeError:
            return ''

    def findRoot(self, entry):
        rootP = re.compile(u'([A-Z] + self.malteseCaps + u"]")")
        root = rootP.search(entry)
        try:
            root.group(1)
            return root.group(1)
        except AttributeError:
            return ''

    def findPos(self, entry):
        posP = re.compile(u', ([a-z.]*) \(')
        pos = posP.search(entry)
        try:
            pos.group(1)
            return pos.group(1)
        except AttributeError:
            return ''

    def findInfl(self, entry):
        inflP = re.compile(u'(\([a-z.-]\))'')
        infl = inflP.search(entry)
        try:
            infl.group(1)
            return infl.group(1)
        except AttributeError:
return ''

def findDefs(self, entry):
    defsP = re.compile(u'|(l\.*\).+\[A-Z]\{2,\})
    defs = defsP.search(entry)
    try:
        defs.group(1)
        return defs.group(1)
    except AttributeError:
        return 'no match'

def findEtym(self, entry):
    etymP = re.compile(u'\([\]')
    etym = etymP.search(entry[::-1]) #[::-1] = string reversing
    try:
        etym.group(1)
        return etym.group(1)[::-1]
    except AttributeError:
        return 'no match'

def findNote(self, entry):
    noteP = re.compile(u'\+ (.*)$')
    note = noteP.search(entry)
    try:
        note.group(1)
        return note.group(1)
    except AttributeError:
        return ''

def processDefs(self, entry):
    sensesP = re.compile(u'([0-9].\D*)')
    senses = sensesP.findall(entry)
    output = ''
    for i in range(len(senses)):
        output += '\n\t\t\t<sense>'
        defi = senses[i].split('. (')
        output += defi[0]
        if len(defi) > 1:
            multipleusages = defi[1][:-3].split(' ; ')
            for j in multipleusages:
                src = self.findSource(j)
                if len(src) > 0:
                    nosrc = ''.join(j.split((src + ' ')))
                    gloss = nosrc.split(', ')
                else:
                    gloss = j.split(',')[ ]
                output += '\n\t\t\t<usage>' + gloss[0]
                if len(gloss) > 1:
                    output += '\n\t\t\t<gloss>' + gloss[1] + '</gloss>'
                if len(src) > 0:
                    output += '\n\t\t\t<source>' + src + '</source>'
            output += '</usage>'
        else:
            output += '</sense>'
    return output

def parseEntry(entry):
    output = []
    p = Parser()
Along with this script, Dr. Berry has provided a sample page that is outputted from his program. It is not apparent at this point if the script in its current form is applicable to the dictionary files we have worked on, as the program contains several syntactical errors that keep it from fully compiling in Python. However, once the script is updated, it seems that the results are exactly what we are looking for. The below document is an XML version of page 111 of the dictionary created by Dr. Berry.

```xml
<?xml version="1.0" encoding="UTF-8"?>
<page>
    <entry>
        <stem>BEŻBEŻ</stem>
        <source>[Sol 53v]</source>
        <root>BŻBŻ</root>
        <pos>v.t.</pos>
        <inflect>(pp. mbeżbeż)</inflect>
        <definition>
            <sense>1. To seize s.o. by the forelock (gen. by way of reprimand). </sense>
            <sense>2. To reproach, to reprimand, to rebuke, to chide</sense>
            <usage>beżbiżtu biex ma jergax jindahal fejn ma jesghux</usage>
            <gloss>I reprimanded him to make sure that he won't interfere again in matters that don't concern him</gloss>
        </definition>
        <source>[Vass]</source>
    </entry>
    <entry>
        <stem>bżiebeżta l-anical</stem>
        <source>[Mer]</source>
    </entry>
    <entry>
        <stem>bżiebeżta l-an ġħa</stem>
        <source>[Mer]</source>
    </entry>
    
    <sense>1. Lock of hair, tuft (of cotton, etc.); forelock; tuft of hair that stands out from the rest, bżiebeżta l-angħ, (cul.) lit. angels' forelocks, name of very thin wheaten pasta having the shape of locks of hair. </sense>
</page>
```
<sense>2
  <usage>child's word) Peni</usage>
</sense>

<subentry>I
  TBEŻBEŻ, v.i. To be admonished; to have had one's lesson (warning). [Vass]
  TBEŻBIŻ, vn. (~ A, vn.u.) Reprimanding; warning. [loc. form. - but for the nom. form cp. Ar. bzbu:z cannelle (robinet mobile) (Dozy)]
</subentry>

<entry>
  <stem>beżbież</stem>
  <definition>no match</definition>
</entry>

<entry>
  <stem>beżbuża</stem>
  <definition>no match</definition>
</entry>

<entry>
  <stem>beżina</stem>
  <inflect>(food)</inflect>
  <definition>no match</definition>
</entry>

<entry>
  <stem>BEŻIK</stem>
  <source>[ESIJ]</source>
  <definition>no match</definition>
  <etymology>[DER Fr. besique, through Eng. bezique] </etymology>
</entry>

<entry>
  <stem>beżlaq</stem>
  <definition>no match</definition>
</entry>

<entry>
  <stem>BEŻLEK</stem>
  <source>[Vass]</source>
</entry>
<root>BŻLK</root>
<pos>v.i.</pos>
<inflect>({pp. mbeżlek)}</inflect>
<definition>
  <sense>1. To suck milk frequently but in little doses
      <usage>ilu jbeżlek f'sidri nofs sigieha</usage>
      <gloss>the baby has been at my breast (off and on) for the last half hour</gloss>
      <source>[Car]/</source>
  </sense>
  <sense>2. To produce a smacking sound while drinking. </sense>
  <sense>3. To make a mess of s.th., to cook badly
      <usage>ikel imbeżlek</usage>
      <gloss>badly-cooked food (such as overcooked pasta)</gloss>
      <usage>gbejniet imbeżila</usage>
      <gloss>too soft country cheese</gloss>
  </sense>
</definition>

<subentry>BEŻLIEK, n.m.ag. (f. ~a, pl. in) One who slobbers. [Car] ~I, a.m.ag. (r. ~ija, pl. ~in). That makes a much of s.th. That sucks milk smacking his lips as he does so. II TBEŻLAK, v.i. To become very soft as a result of being overdone/over-cooked. [Vass] TBEŻLIK, vn. (~A, vn.u.) Messy, overcooked food. [Bus] Light sucking of milk.
</subentry>
<etymology>[Barb. derives this v. from N. Afr. Ar. بزولة M. beżżula, a woman's breast. But this explanation leaves the existence of the 4th radical unexplained. My explanation is that beżlek is denom. from the noun beżeq, spittle, Ar. بزق buza:q with inserted l as 3rd radical to denote frequency of action. Note that beżlek or beżłaq is not only used in the sense of 'to suck frequently' but also in the sense of 'to spit frequently'. In this sense beżlek is a syn. of beżzaq, 2nd form of bezq, to spit. Ar. بزق bazaqa. Another liquid consonant forming quadriliterals from original triliteral roots is r which in Ar. frequently interchanges with l; cp. M. xeblek, to twine about, from vb. root XBK, to twine, interweave, and Ar. شربكة DER [fabaka] DER جربكة</etymology>

<note>[VG Gozo] mbeżlek/mbeżłaq is said also of cement, mortar, etc. which needs more kneading because it is still too watery. </note>
</entry>
</page>

As referenced earlier, the above document contains the information pertinent to each entry of page 111 in a hierarchical XML format. Each entry is clearly separated from other entries by the <entry> tag, and items like <root>, <sense>, <usage>, and <gloss> are organized with respect to one another. Given the organized nature of this output from Dr. Berry’s script, it is indubitable that this program provides the solution our project needs to begin the database creation process. However, a few problems do exist with the script. First, not all entries are effectively captured into the XML tree, especially those entries that lack some of the property information (root, source, inflection) typically listed at the beginning.
of the entry. These items often appear as `<definition> no match` in the XML text, despite the fact that a definition is in fact present on the page. For example, the word beţlaq on page 111 is defined as “To make a mess of, make slurry,” but has a “no match” value in the `<definition>` tag in the above XML document. This problem needs to be corrected so as to incorporate as many entries as possible into the final database.

Secondly, it is unclear in Berry’s script how each entry will be inputted into the XML processing source code. The script references a variable `entry` from the outset, but does not provide information on how to programatically obtain the text of each entry to be processed by the program. This will need to be clarified before the script can be applied to the whole dictionary. Additionally, the script does not contain a way to loop through all the files in the dictionary, and is only executable for one file at a time. Finally, the program was only created to process .txt files, which poses a problem if any formatting is going to be maintained in the online database. Overall, however, the script provides a viable method by which to transpose the digital pages into a searchable, archivable format, and simply needs a small amount of syntactical and functional editing to become directly implementable.

**Conclusion**

While the Maltese Dictionary Project is still not completed, the shift towards automation implied by the use of a mass find-and-replace script and an XML parser brings us closer to achieving the ultimate goal of creating a searchable database for Maltese speakers to use. Despite the inefficiencies that have plagued this project in the past as a result of a lack of technical expertise, its prognosis is optimistic and we are closer than ever to being finished. As the find-and-replace and XML scripts are refined, our pages will come closer and closer to being faithful to the original dictionary. Whether this project is completed in a year or five years, when our
searchable database is finally available it will provide a comprehensive Maltese-English
dictionary that will allow Maltese speakers to better express themselves online.

The most important conclusion that can be drawn from our work up to now on Aquilina’s
dictionary is that automation has incredible value in this type of project. The hours previously
poured into this project are innumerable, and after learning how to automate many of the tasks
we performed by hand, many of them seem somewhat wasted. This project, which has now
lasted over ten years, should serve as a precedent to future users of OCR technology, be it for the
purpose of language preservation or other reasons, to make use of automation techniques from
the beginning. Regardless of the extra labor required to learn these skills, the amount of time that
can be saved over the long run is simply amazing. In the field of language preservation, this is
especially significant because time is often of the essence, especially when dealing with
endangered languages. The author hopes that the techniques delimited in this paper will be useful
to future projects in language preservation and inspire more use of automation in the field of
linguistics in general.
Bibliography


