The Making of a Large English-Arabic/Arabic-English Dictionary: the Oxford Arabic Dictionary

Tressy Arts
Oxford Arabic Dictionary
tressy.arts@gmail.com

Abstract

In this presentation, we will illustrate the process of making our brand-new Arabic-English/English-Arabic dictionary, which is due for publication in print and online in August 2014. It is intended for speakers of both English and Arabic. It contains over 26,000 entries on each side, including many up-to-the-minute words and expressions. Collocations and examples are an important feature. The dictionary has been compiled using dictionary writing software that enables editors to work and communicate with one another regardless of their location. It will be available in both print and online.

We show the entire process of making an Arabic dictionary, from finding a reliable framework in both languages, to developing a unique online functionality. We show the difficulties lexicographers face when compiling an Arabic dictionary, and the ways in which we dealt with those. In addition, the Oxford Arabic Dictionary has quite a few features that are entirely new to Arabic dictionaries, and we illustrate how we went about developing those.

Keywords: Arabic; bilingual; dictionary-making process

1 Arabic Lexicography

Arabic lexicography has a centuries-old tradition, starting with the *Kitāb al-'Ayn*, a large monolingual dictionary, in the eighth century. This impressive legacy is both a blessing and a curse, since conservatism rules, and the mediaeval tomes like the 13th century *Lisān al-'Arab* still count as *the* standard in lexicography. Modern monolingual dictionaries often do little more than repeat what has been said before, regardless of whether the senses, examples, or even the headwords mentioned are still in actual use. Many Arabic lexicographers, linguists and laymen see Classical Arabic, the language of the Koran and the pre-Islamic poetry, as the perfect standard, and any deviations from that are seen as corrupting the language. Therefore many loanwords and words derived from colloquial Arabic are not included in modern dictionaries, despite them being commonplace in Standard Arabic texts. Haywood says in his "Arabic Lexicography": "The lexicographers helped to keep the written language static, and to aid the understanding of it, as the spoken dialects diverged more and more from it."

(Haywood 1960: 116). There are several Arabic Language Academies which aim to find proper Arabic words (that is, words which confirm to the Arabic root-and-pattern system for word forming, rather than words which are simply English words written in Arabic letters) for new concepts, but they are slow-moving and don't always communicate clearly, so that loanwords for new concepts often are commonplace long before an approved version is released. Sometimes the two continue to exist side by side (*tilifūn* and *hātif* (telephone)), sometimes the loanword remains strong (*tilifizyūn* (television)), sometimes the approved version gains prominence (*ḥāsūb* (computer) is more common than *kumbyūtir* these days).

Still, it is rare to find modern words in monolingual dictionaries, either loanwords, or ones formed according to Arabic morphology. One will be hard put to find *ḥāsūb* (computer) in most, and a common modern word like *nazzala* (to download) I haven't been able to find in any, not even online.

Bilingual lexicography with Arabic as source or target language also has a long history. Already in the 11th century al-Zamak̲šarī published an Arabic-Persian dictionary, and al-Kāšḡarī a Turkish-Arabic one. For Europeans, the first bilingual dictionary was Golius' *Lexicon Arabico-Latinum*, published in Leiden in 1653. The first Arabic-English one was produced by Edward William Lane in the late 19th century.

Looking at current-day Arabic-English/English-Arabic lexicography, however, one cannot help but be a bit disappointed with what's there, especially considering the status of both languages as major world languages. The standard Arabic-English work, Hans Wehr's *Dictionary of Modern Written Arabic*, was translated from German and not much updated since it was first published in 1961. Though groundbreaking at the time, it is nowhere near comparable to modern bilingual dictionaries for most other languages. Obviously the word list is outdated, but also the presentation of the entries is not really what one would expect from a modern dictionary: long lists of possible English translations are given for most Arabic words, without any guidance for the user on which translation to choose in which context; no word senses are distinguished. No examples are given and only very few collocations, many of which don't actually exist in modern Arabic. Yet, this is still the dictionary that everyone translating from Arabic into English will use.

The situation for English-Arabic is even direr. Arguably the best option is still Oxford's English-Arabic (hand-written!) dictionary from 1972 (cf. Benzehra 2012); the most popular work is *al-Mawrid* (Baalbaki & Baalbaki 2013), which is updated regularly, but has certain major flaws. The English used seems mostly derived from very old texts, and its policy for including new words in updated editions is not clear: the most recent edition, from 2013, has "fuscous" and "silvern", featured new entries "naughties" [sic] and "smirt", but not "blog" or "text message". Examples given seem to be taken from very old English dictionaries/texts ("swallows that affect chimneys", "the hes would quarrel and fight with the females"). As to the Arabic translations, many of the senses given cannot be attested to exist in modern English, senses are ordered from oldest to most modern, leading to the most common sense often being the last one, and the editors even create new Arabic "words" to serve as translations:

> compilers of bilingual dictionaries are not only entitled to coin words which may or may not gain currency, but that coinage becomes an essential duty of theirs, especially with a language like Arabic, where a huge number of terms, particularly scientific ones, are lacking. (Baalbaki 2004: 68)

This statement is disputable: Arabic does not lack terms for scientific or any other areas, but some of the terms will be loanwords or multi-word constructions, so it depends on what is understood by "Arabic terms". We believe that it's more useful for an English-Arabic dictionary to give those loanwords or multi-word phrases, than to coin new words which are useful for neither decoding or encoding users.

Other European languages fare slightly better, but not much, with the exception of Dutch, which has some good and modern dictionaries, published at the beginning of this century (*Woordenboek Nederlands-Arabisch/Arabisch-Nederlands* (Hoogland et al. 2003) and *Leerwoordenboek Arabisch* (Van Mol & Bergman 2001)).

There are plenty of online Arabic dictionaries, but the accuracy there usually leaves much to be desired.

So in developing a large bilingual Arabic-English/English-Arabic dictionary that was corpus-based and could live up to modern expectations, our team was truly taking on a unique endeavour. Not being able to rely on any previous works, either bilingual or monolingual, all progress had to be made by painstaking research.

2 The Arabic Language

The lack of reliable predecessors is far from the only complicating factor in Arabic lexicography. To highlight a few more:

2.1 Diglossia

In the Arabic world a diglossia exists, with the "high" variant, Modern Standard Arabic, being the accepted language for any written and official spoken discourse. Meanwhile, a multitude of "dialects" or "colloquials" are the languages people actually speak. These are officially known as dialects, but are often mutually unintelligible, and can be considered different languages on purely linguistic grounds. It is hard to say how many exist, as there is a dialect continuum, but in many countries the dialect of the capital has the major status (so if you pick up a text book or travel guide in "Egyptian Arabic", this will be the dialect of Cairo, etc.). The "dialects" have no agreed orthography, though they are used more and more in written communication, mostly on the internet, and even a few novels have appeared.

Since we wanted to make a broadly useful dictionary, not focusing on one or a few dialects, and since it is hard to write dialects because of the lack of standard orthography, we chose to use (almost) exclu-

sively Standard Arabic for the Arabic in the dictionary. Even for English expressions which are very informal or chatty, we have tried to give Standard Arabic equivalents, or if needed descriptions. Both languages have level markers so the user is alerted if there is a difference in level.
Then we had to devise a strategy for deciding which words are indeed Modern Standard Arabic, and should be included in the dictionary. We could not simply include all written words in general texts, since these days many dialect words are written. We could also not rely on dictionaries to verify existence of words, since many common modern words are not listed in monolingual or bilingual dictionaries. So the criterion we decided on was that words which are used without quotation marks in a reasonable number of non-specialist, otherwise Standard Arabic texts, could themselves be considered Standard Arabic, and therefore deserved a place in the dictionary. On the other hand, if a word is only found in dictionaries, we did not include it.

2.2 No Native Speakers

It is standard practice these days in writing bilingual dictionaries to have them compiled by native speakers of the target language. However, there are no native speakers of Modern Standard Arabic; children in Arabic countries grow up with one of the dialects or even a non-Arabic language (Berber, Kurdish) as their mother tongue. Standard Arabic is learnt, like a foreign language, at school and in the mosque. Research has shown that indeed Arabic people process Standard Arabic in the brain in the same way as foreign languages (Ibrahim & Aharon-Peretz 2005). So although the editors hired were highly educated linguists, they lacked that native speaker sense that lexicographers for most other languages have.

2.3 Geographical Spread

In addition, the Arab world has a very large geographical spread, and regional language preferences exist even in Standard Arabic, so editors need to constantly be wary of using words and phrases that are limited to their home country. Very little research has been done in this area, partly because of the prevailing ideal of Arabic being one uniform language.
If we take all these factors together (no reliable antecedents, no native speaker sense, uncharted geographical differences), we can understand that it's very hard for the "native speaker"[1] editors to feel secure in their translation of a word or phrase, necessitating elaborate checking in a corpus and on the internet, as well as discussion with native speakers from other locations, to find the right terms. This meant that the project took much longer, and was more expensive, than originally envisaged.

1 We will use this to refer to native speakers of one of the Arabic dialects with good knowledge of Standard Arabic, for want of a better term.

3 The Bases

For the Arabic-English side, the data of Hoogland's Arabic-Dutch *Woordenboek* mentioned above, which had got very good press, was licensed.
For the English-Arabic side, we used an English framework that had been developed for use as a basis for Oxford unabridged bilingual dictionaries, expanded with words that are especially relevant for Arabic, like the English names of the Islamic months.

4 Vocalization

The Arabic script is a consonant script. Vowels are indicated by diacritic signs over and under the letter (see figure 1), but are not commonly written, so a word like *al-maḡrib* (Morocco) will be written as *al-mḡrb*. In addition, the three cases Arabic has are most often only indicated by vowels, and hence not visible in most texts.

أُحِبُّ اللُّغةَ العَرَبيّةَ أحب اللغة العربية

Figure 1: Vocalized (left) and unvocalized (right) Arabic.

For learners of Arabic it is useful to be able to find all vowels both in the Arabic-English side (if one doesn't know a word, one probably doesn't know how to pronounce it), and in the English-Arabic side (when one finds a new Arabic word, it's useful to find the vowels). Similarly, case endings are useful to understand the syntax and word combinations. So in order to be most useful for non-Arab users, as well as clearest for Arab users, we wanted all Arabic on both sides of the dictionary, headwords, translations, examples and descriptions, to have full vocalization, including the case endings. This placed an extra burden on our editors, since most people are not used to writing vocalized Arabic, and there is no standard vocalization system, so we had to devise our own rules (most were taken over from the system used in the Hoogland dictionaries). It also made the use of an adapted font necessary, since most Arabic fonts are not designed with the vowels in mind, and they can "disappear" in certain letter combinations.
For the English-Arabic side, not only did we want the translations to be written in the above-mentioned vocalization system, we also wanted to provide the unpredictable grammatical information for single-word Arabic translations (the plurals of nouns and adjectives, and conjugational information and infinitives of many verbs are unpredictable in Arabic, so it's useful when this is provided with a word functioning as a translation). This information was already present for all the headwords in the Arabic framework, so a "translation picker" tool was developed within Oxford University Press: the

English-Arabic translator could enter an unvowelled word in this tool in the dictionary database, and was presented with all possible vocalized headwords in the Arabic data, with their grammatical information (see figure 2).

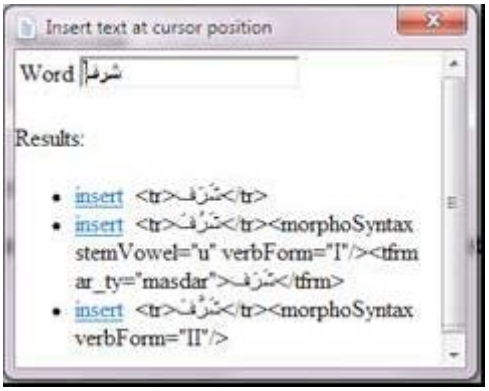


Figure 2: The Translation Picker in use.

By simply selecting the one they wanted, a correctly vocalized word with all relevant grammatical information was entered as translation.

This had an added advantage: if the translator could not find the word they wanted as translation in the Translation Picker, it meant that that word was not in the Arabic lemma list for the Arabic-English dictionary. The translator then made a note for the chief editor, who would decide if the Arabic word in question was a valuable addition to the Arabic-nglish lemma list, and add it to the latter if it was.

5 The Database Used

The dictionary writing software used was DPS (Digital Publishing System) produced by French company IDM, a database system specialized in creating and developing dictionaries, which is used by Oxford University Press and many other publishers. Its application, the Entry Editor, lets users connect to the central database in Oxford and download and upload entries to work on. It lets users edit entries uploaded by others, keeps track of who made which changes, lets you revert to any earlier uploaded version, and allows users to communicate with one another through a system of searchable annotations that can be made on every level (entry, sense, translation, etc.), a very valuable tool for a dictionary where much discussion of terms needs to take place. Via its search engine, editors could make highly specific searches, for example every translation that has a “region: Egypt” attached, and for the project managers its workflow manager allowed us to distribute the workload and keep track of progress.

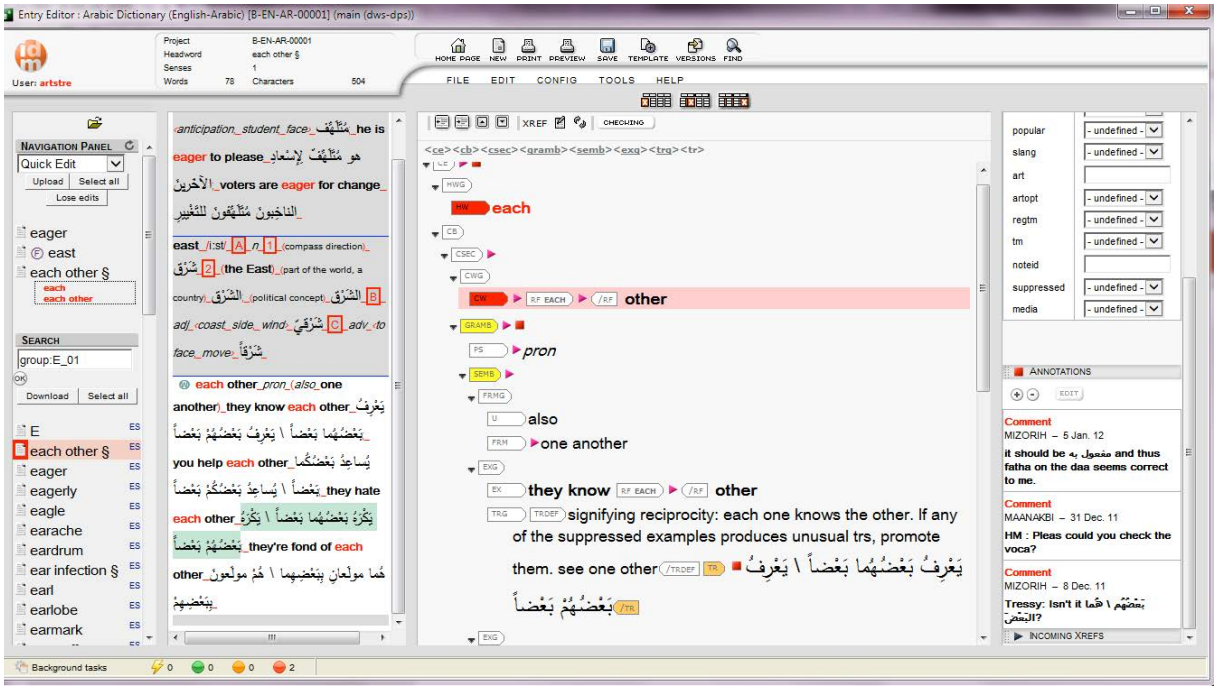


Figure 3: The Entry Editor showing entries in a group, preview, tag tree, and attributes and annotations.

6 The Corpus Used

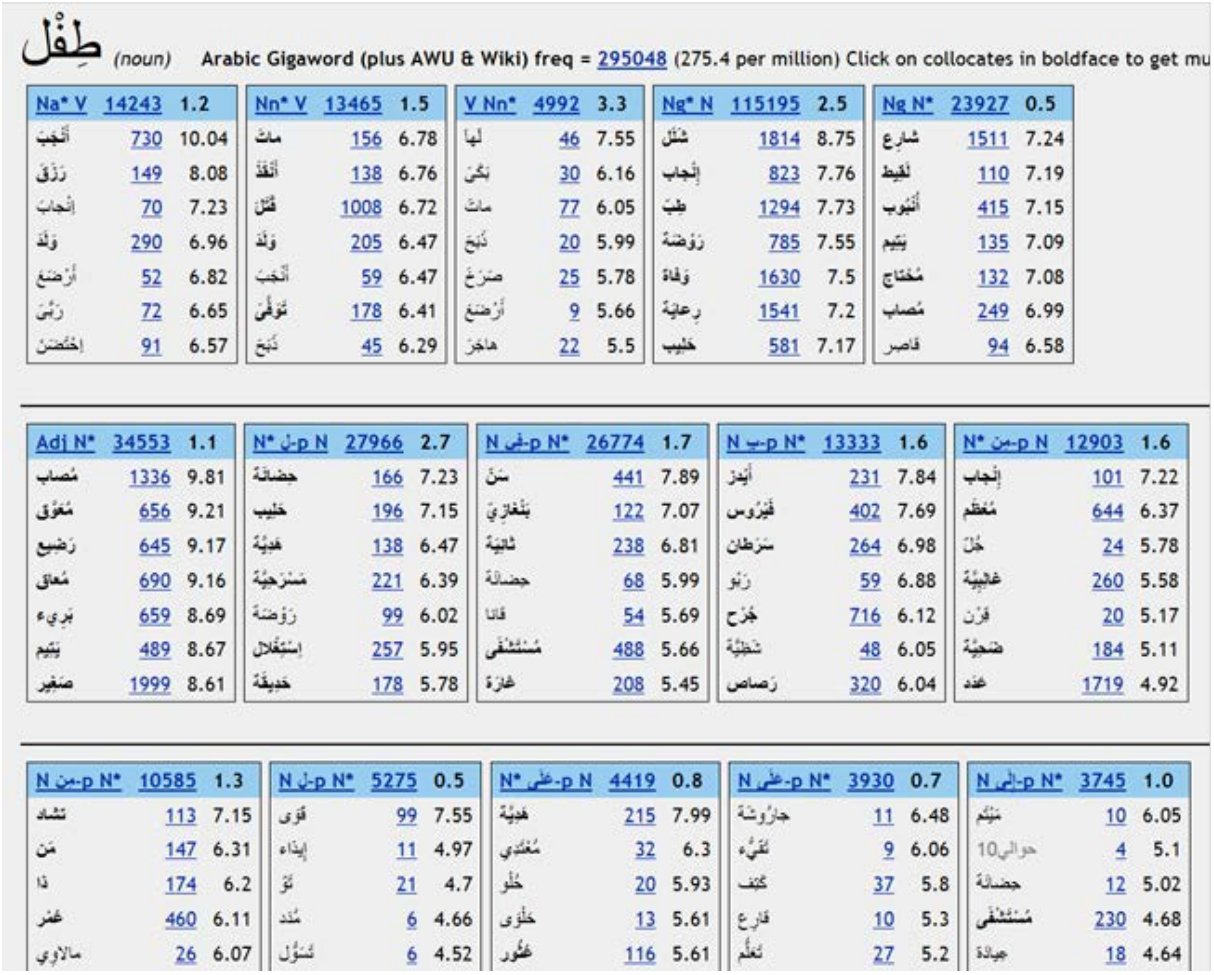
The Oxford English Corpus contains 2.5 billion words of modern English from all over the globe, and as such was a very valuable resource for our editors. A unique new resource, however, was the Oxford Arabic Corpus, made searchable with the Sketch Engine software, developed by Lexical Computing Limited of Brighton. This enabled us to do truly unprecedented work.

The Oxford Arabic Corpus comprises the Arabic Gigaword Corpus Fourth Edition from Linguistic Data Consortium: 840 million words of news text from nine publications covering the period 1996-2008, plus 10 million words of fiction from the Arabic Writers Union of Damascus, and 30 million words from Arabic Wikipedia.

After the corpus was assembled, the raw data was processed using MADA (Habash, Rambow & Roth 2009). MADA uses the Buckwalter morphological analyzer (by Linguistic Data Consortium) to provide alternative analyses (part-of-speech, vowelled form, and lemma) of each input token, after which a Support Vector Machine classifier ranks the competing analyses in context. The highest ranked analysis is loaded into the Sketch Engine corpus software (Kilgariff et al. 2004), allowing items to be searched and concordanced by word form or lemma, and collocate phrases to be displayed by part of speech structure and grammatical class, as illustrated in Figure 4.

Here we see the Word Sketch for the word *ṭifl* (child). The abbreviation with the asterisk is the search term, and the sequence is from right to left, so the top left column is verbs (V) followed by the search

term (N for noun) in the accusative (a). The top collocate here, with 730 results, is *anjaba* (to give birth to). The results give us collocates in the form of verbs, nouns in so-called genitive constructs, adjectives, and prepositions with other nouns.


طِفْل (noun) Arabic Gigaword (plus AWU & Wiki) freq = 295048 (275.4 per million) Click on collocates in boldface to get mu

Na* V	14243	1.2
أَنْجَبَ	730	10.04
رَزَقَ	149	8.08
إِنْجَابَ	70	7.23
وَلَدَ	290	6.96
أَرْضَعَ	52	6.82
رَبَّى	72	6.65
إِخْتَطَفَ	91	6.57

Nn* V	13465	1.5
مات	156	6.78
أَنْقَذَ	138	6.76
قُتِلَ	1008	6.72
وَلَدَ	205	6.47
أَنْجَبَ	59	6.47
تُوُفِّيَ	178	6.41
ذَبَحَ	45	6.29

V Nn*	4992	3.3
لها	46	7.55
بكى	30	6.16
مات	77	6.05
ذَبَحَ	20	5.99
صَرَعَ	25	5.78
أَرْضَعَ	9	5.66
هاجَرَ	22	5.5

Ng* N	115195	2.5
شَلَلِ	1814	8.75
إِنْجاب	823	7.76
طِبِّ	1294	7.73
رَوْضَةِ	785	7.55
وَفاةِ	1630	7.5
رِعايَةِ	1541	7.2
حَلِيب	581	7.17

Ng N*	23927	0.5
شارِع	1511	7.24
لَقِيط	110	7.19
أَنْبُوب	415	7.15
يَتِيم	135	7.09
مُخْتاج	132	7.08
مُصاب	249	6.99
قاصِر	94	6.58

Adj N*	34553	1.1
مُصاب	1336	9.81
مُعَوَّق	656	9.21
رَضِيع	645	9.17
مُعاق	690	9.16
بَرِيء	659	8.69
يَتِيم	489	8.67
صَغِير	1999	8.61

N* لِ-p N	27966	2.7
حَضانَة	166	7.23
حَلِيب	196	7.15
هَدِيَّة	138	6.47
مَسْرَحِيَّة	221	6.39
رَوْضَة	99	6.02
إِسْتِقْلال	257	5.95
حَدِيقَة	178	5.78

N فِي-p N*	26774	1.7
سَنّ	441	7.89
بِلَغارِيّ	122	7.07
ثانِيَة	238	6.81
حضانة	68	5.99
قانا	54	5.69
مُسْتَشْفَى	488	5.66
غارَة	208	5.45

N بِ-p N*	13333	1.6
أَيدز	231	7.84
فَيْرُوس	402	7.69
سَرَطان	264	6.98
رَبْو	59	6.88
جُرْح	716	6.12
شَظِيَّة	48	6.05
رَصاص	320	6.04

N* مِن-p N	12903	1.6
إِنْجاب	101	7.22
مُعْظَم	644	6.37
خُلْ	24	5.78
غالِبِيَّة	260	5.58
فَرْن	20	5.17
ضَحِيَّة	184	5.11
عَدَد	1719	4.92

N مِن-p N*	10585	1.3
تَشاد	113	7.15
مَن	147	6.31
ذا	174	6.2
عُمْر	460	6.11
مالاوي	26	6.07

N لِ-p N*	5275	0.5
قَوِيّ	99	7.55
إِيذاء	11	4.97
لَوْ	21	4.7
مُذ	6	4.66
تَسَوُّل	6	4.52

N* عَلى-p N	4419	0.8
هَدِيَّة	215	7.99
مُعْتَدٍ	32	6.3
حُلْو	20	5.93
حَلْوى	13	5.61
عُثُور	116	5.61

N عَلى-p N*	3930	0.7
جازُوشَة	11	6.48
لَقِيء	9	6.06
كَتِف	37	5.8
قارِع	10	5.3
تَعَلَّم	27	5.2

N إِلى-p N*	3745	1.0
مَيْلَم	10	6.05
حوالي10	4	5.1
حضانة	12	5.02
مُسْتَشْفَى	230	4.68
عِيادَة	18	4.64

Figure 4: Word Sketch for ṭifl (child).

This allowed us to fine-tune translations, add relevant collocations and examples, and discover new words and senses.

Using the corpus wasn't without its pitfalls though. Correctly tagging Arabic is much more complex than English or Italian. Due to the complexity of Arabic's morphology and the surface-form ambiguity, mistagging in the corpus is inevitable – *wīlz* (Wales) is consistently analysed as a form of *lazza* (to be compressed), because *wīlz* is not in the lemma list and *wa-yaluzzu*, a conjugated form of *lazza*, has the same surface form: WiYLZ and WaYaLuZzu (the capitals are the letters visible in the surface form). Also, because the case endings aren't usually visible, and Arabic syntax has a VSO structure, differentiation between verb + nominative noun and verb + accusative noun is nearly impossible. In Figure 4, the biggest result in the verb + nominative noun table is *qatala* (to kill). We can be reasonably certain that in most of those cases the child is the unfortunate object, rather than the subject, and so is in fact an accusative noun.

Editors therefore needed a sharp critical eye when analysing the results.

7 Finding New Words and Examples in the Corpus

The English framework was developed by a skilled team of Oxford lexicographers, and continued to be expanded, so we could reliably assume that all relevant words and phrases were included. However, for the Arabic, this was a different story: the Arabic-Dutch dictionary was developed in the nineties, when Arabic corpus linguistics was still very primitive. The editors could only search the corpus on surface forms, which is remarkably difficult in a language where a simple verb like "to eat" can have no fewer than 46 surface forms, most of which also can mean "to feed". They found ways to work around those problems though (Hoogland 2004), and the chief editor calculated that the dictionary covered 99.95% of (non-lemmatized) word forms in Modern Arabic texts (Hoogland 2003). However, with the new technology at our disposal in the form of the Oxford Arabic Corpus and Google, we could certainly see many possibilities for improving and expanding the Arabic framework.

We started on the level of the vocabulary, comparing the lemma list of the corpus to the Arabic-Dutch lemma list, and added any lemmas we thought warranted inclusion, like *rawwasa* (to sharpen; to supply with a header). However, this only discovered words that were already listed in the Buckwalter lexicon. Because of Arabic's complex morphology, it's often not possible to automatically distil a lemma from a string if the lemma is unknown, so many surface forms were still unidentified. Mohammed Attia, one of our translators and a computational linguist, had developed a system to distil potential lemmas from the corpus by checking frequency and compliance with Arabic morphological patterns (Attia et al. 2014). These potential lemmas were again compared to our lemma list, and thus we managed to find genuine new words, many of which not only weren't listed in the Arabic-Dutch data, but had not been previously listed in any Arabic dictionaries, like *tamāhā* (to be congruent) and *ṭawā'ifī* (sectarian).

So altogether we had four ways to expand the Arabic lemma list to contain more, currently relevant, entries: listing Arabic translations for English words that weren't in the Arabic data, comparing the lemma list of the Gigaword corpus with the Hoogland lemma list, using Attia's potential lemma extractor, and old-fashioned handwork: critical reading of web sites and being alert to any newly developing words and collocations, like *al-rabī' al-'arabī* (Arab spring) and *taġrīda* (tweet). In these ways we managed to expand the lemma list by over 2,000 entries. At the same time, we pruned entries that were deemed obsolete or archaic, thus removing about four hundred entries from the original Arabic data. This latter was not a priority however, since, from a user's point of view, obsolete entries are not in the way: if a user doesn't encounter the word, they are not going to look it up; so we focused on improving the relevant entries rather than on pruning the less relevant ones. All in all we raised the number of Arabic entries from 24,682 to 26,316.

On the level of individual entries, the editors used the corpus, as well as Google, to check the existing examples for relevance, to find new or better examples and collocations, and on occasion find entirely new senses for existing words (Arts & McNeil 2013). Especially on this microstructural level, the Arabic-English dictionary has been greatly expanded compared to its Arabic-Dutch predecessor.

8 Microstructure and Translations

On both sides of the dictionary, entries are divided into one or more senses, which have disambiguators in the source language indicating the meanings. Many senses have examples to illustrate typical uses of the headword in that sense. Idioms are given separately, outside the scope of the senses. Several types of translations are given:

- the direct translation, which is an equivalent of the headword in that sense and can be used as its translation in most contexts, or which is the equivalent of the example given, e.g. "computer" and *ḥāsūb.*
- the approximate translation, which is a nearly equivalent translation, or a translation in certain contexts (which is then specified), e.g. *Allāhu yuḵallīka*, literally "God bless you", which is used to thank someone, gets the approximate translation "thank you". Approximate translations are indicated with an approximation sign (≈).
- the translation or approximate translation followed by an explanation in brackets. Often the translation is useful for the encoding user, and the explanation for the decoding user, with the explanation further specifying the translation, e.g. "muezzin (*mosque official who recites the call to prayer*)". The explanation is in brackets in the English-Arabic and in brackets and italics in the Arabic-English.
- the definition: where a headword or example doesn't have an equivalent in the target language (for English e.g. "haggis", for Arabic for example many Islamic terms), a description of what it means is given in the target language. In the English-Arabic these definitions are in square brackets, in the Arabic-English in italics.

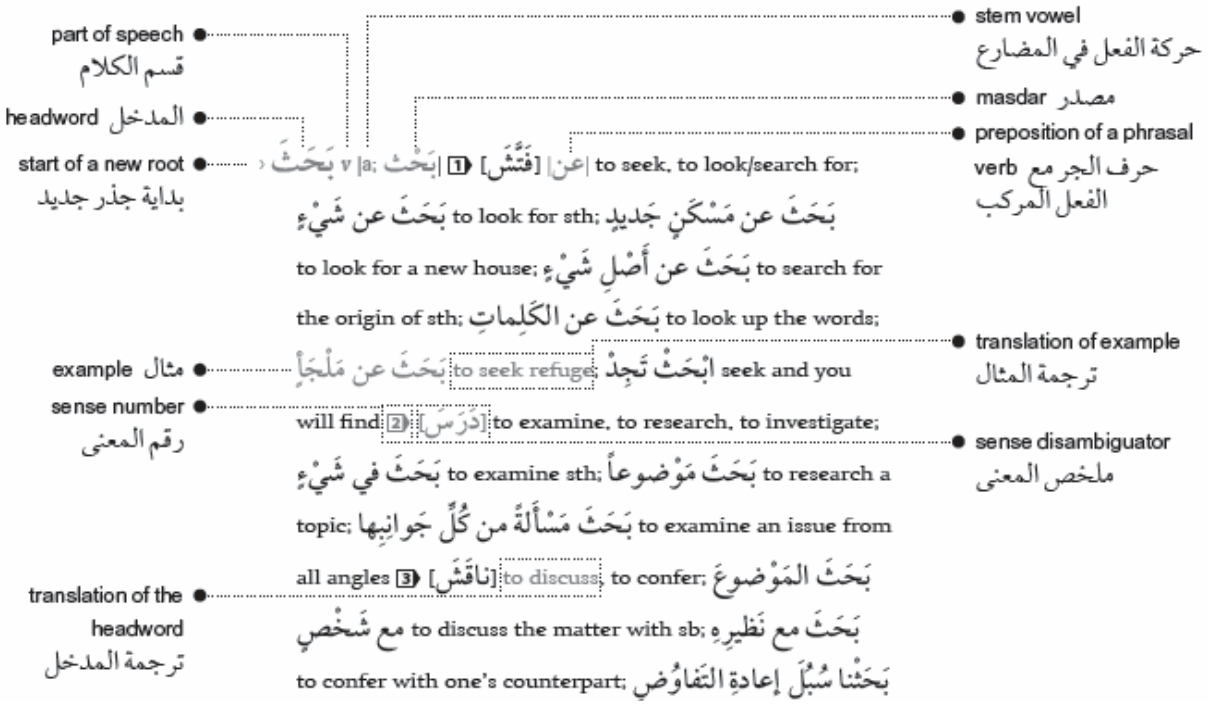


Figure 5: An Arabic-English entry with some aspects of its structure illustrated.

8.1 Arabic-English Translations

The existing Hoogland Arabic-Dutch dictionary was translated by Dutch Arabists with a good command of English. They made sure to keep the translation relation between the Arabic and the English in mind, rather than translating from Dutch to English, but the existing Dutch translations were too valuable a tool to disregard. Also, most of the editors had previously worked on the Arabic-Dutch dictionary, so they were familiar with Arabic lexicography and with this specific dictionary.
Then all entries were reviewed by native English-speaking Arabists, correcting the English where necessary and further improving the entries.
Since the original was made in a time when corpus research was hardly possible, the existing entries and examples were checked in the corpus and on Google: are the senses and examples representative of current Arabic, or are they so-called "dictionary words/examples", copied from monolingual dictionaries that are not in actual use any more? The latter were weeded out and where needed we replaced the examples by new ones distilled from modern language found in the corpus and Google. We checked if all senses were attested, and checked senses we couldn't find evidence for with native speakers. Sometimes the corpus showed senses that were not yet listed, and we added those. During the entire process, we could ask native speakers' or other editors' opinion via annotations.
See figure 5 for an example of an Arabic entry and its structure.

8.2 Arabic-English Translations

The English-Arabic side of the dictionary was formed by having the English framework translated into Arabic by translators and lexicographers from several Arabic countries: Algeria, Tunisia, Egypt, Palestine, Lebanon, and Iraq. For certain fields, namely Medicine, Technology, and Law and Business, bilingual experts were found to advise on the terminology. The English framework contains field markers for many entries, so exporting all entries in a specific field was easily done.
Since, as I stated above, no one is a true native speaker of Arabic, it was important for the translators and reviewers to be able to check the translations not just against their own language sense, but also against a corpus. Even a 900-million-word corpus is not quite reliable enough to state that a certain word sense/construction is never used, so for verification of usage of Arabic translations of English terms, Google proved invaluable. Using quotation marks makes it possible to search for exact constructions, enabling translators and reviewers to verify that the translations are actually in use in modern Arabic.
Annotations were used to communicate with other native speakers verifying translations or asking for suggestions ("in Iraq we say this, but do you also use this in North Africa?").
All entries were, after translation, reviewed one or more times by revising editors, at least one of which, for every entry, was a native speaker. Often the reviewer would discuss with the original translator to find the best solution for tricky aspects.

See figure 6 for an example of an English-Arabic entry and its structure.

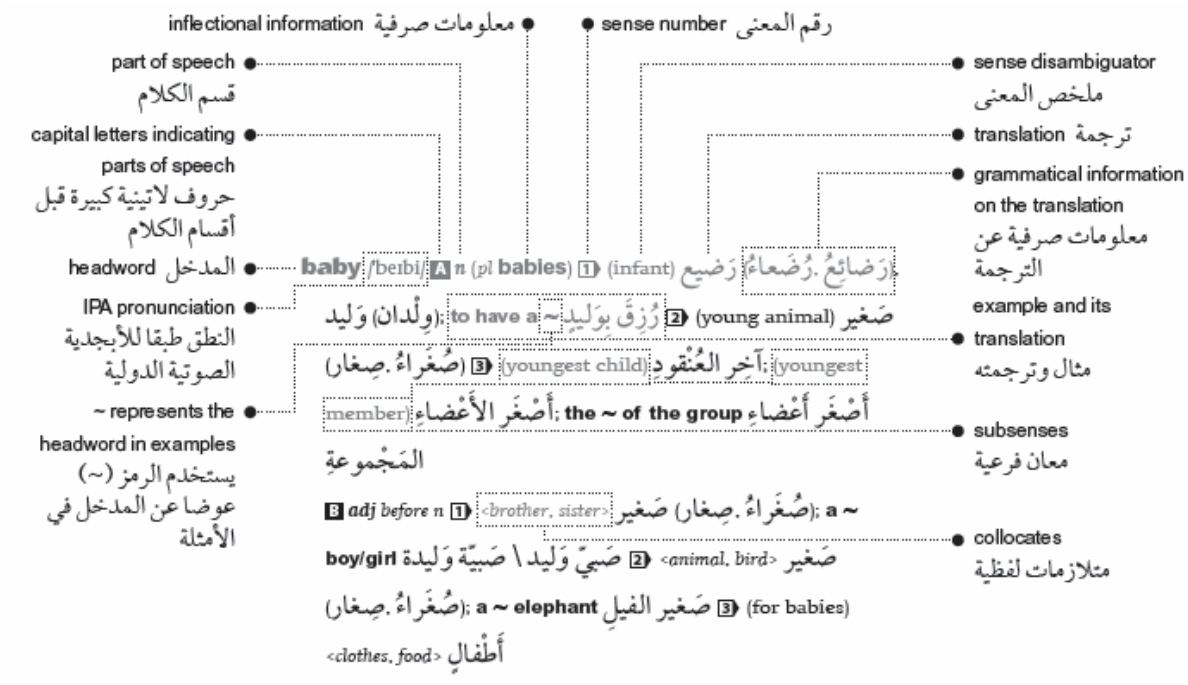


Figure 6: An English-Arabic entry with some of its structure illustrated.

9 Technical Challenges

A number of technical challenges had to be overcome in getting the dictionary writing software to work in harmony with the Arabic. The first challenge was the simple fact that Arabic is written from right to left and the letters within a word are joined up. This causes problems whenever Arabic and computers come into contact with each other, and made the news when Arabic and Persian visitors to the London Olympics were told the equivalent of N O D N O L OT E M O C L E W.

We had to make sure that the software could deal with English and Arabic, and even with English words inside Arabic tags, and vice versa, which gave the OUP dictionary technology staff no end of work. One of the largest problems turned out to be any numbers in examples, which in Arabic go in the same direction as in English, but which consistently were turned back to front in Arabic tags, since we had given the instruction that Arabic tags needed to output from right to left! The dictionary technologists then developed "bidirectional-override" tags, in which data that needed to output in a different direction than the default direction could be entered. This worked inasmuch as the numbers came out correctly... however, the rest of the data switched place, so the start of the sentence jumped to after the number and the back to before it. After months of tinkering we eventually had to instruct the typesetters to set all numerals back to front.

An additional complication is that our text is not just Arabic, but half Arabic and half English. Due to the large volume of text, we had to have English and Arabic constantly interplay, rather than having them in separate columns (see Figure 7).

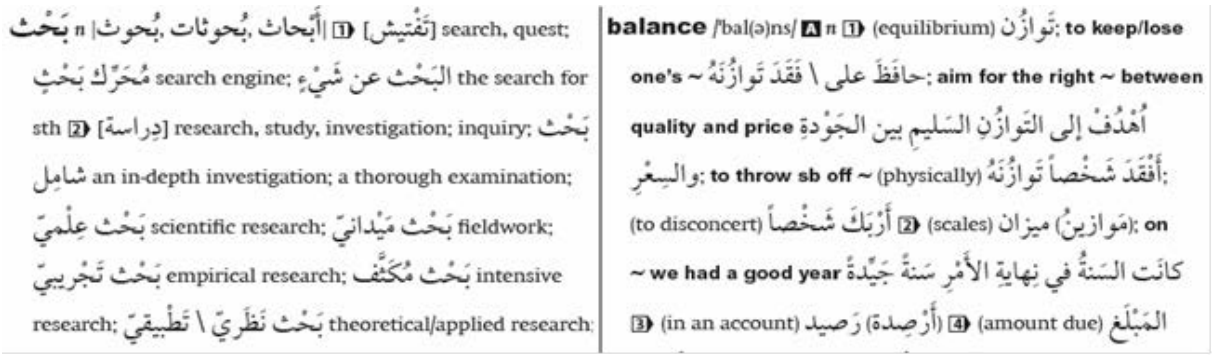


Figure 7: Both sides of the dictionary have text running on.

The direction of the dictionary is left-to-right, so fitting right-to-left phrases in there was a major challenge, not in the least because some information that is to the right (e.g. the start of an Arabic example sentence) needs to go on the top line in the case of a line break, whereas other information on the right (e.g. the grammatical information with an Arabic translation) needs to go on the bottom line in case of a line break. It took many tries and elaborate diagrams until we had all the kinks sorted out.

For some users it may seem a bit odd at first to have to "jump" when reading a translation, but we have found that one gets used to it very quickly, and indeed this is the way the entries are presented in most dictionaries. The alternative, writing the English on the left and the Arabic on the right, leaves too much white space and isn't feasible for a print dictionary of this size.

10 Finding an Arabic Word in the Print Dictionary

An important factor of Arabic is that it is root-based, that is, every word has a root of (usually three) consonants carrying the basic meaning of a word (e.g. *ktb* with basic meaning "writing", which is modified by adding vowels and affixes, making *kātib* (writer), *kitāb* (book), *kataba* (he wrote), *kitāba* (writing), *maktaba* (library; bookshop), etc. Arabic dictionaries, with the exception of learners' dictionaries, are usually ordered by these roots, rather than in "proper" alphabetical order. We chose to do this as well, since the advantage of having all words of one root in one outweighs the difficulty a beginner may have in finding the root of a word. This means the Arabic-English side has a kind of double ordering, first the roots in alphabetical order, then a logical order for the words within one root.

Loanwords do not have an Arabic root. For them we listed each written letter as a root letter, and fitted them in into the root system like that.

11 Finding an Arabic Word in the Electronic Dictionary

Though all words are fully vocalized, we cannot expect the user of the electronic dictionary to enter a word with all its vowels, especially not if it's a word they don't know – not knowing a word means you could at best make an educated guess at its vowels, and there is no standard system of vocalization. So one of the challenges for the electronic dictionary was to make it so that the user is able to enter the unvowelled or partially vowelled form of a word, and be redirected to the entry or entries that correspond to that form.

But that is not the biggest challenge. I mentioned before the many possible surface forms of one lemma. Also, words are often joined together: the article *al-* is prefixed to the noun or adjective, object and possessive pronouns are suffixed, and some grammatical words like *wa-* (and) and *li-* (for) are joined to the word following them as well. All this can be combined, so for example *wali'uxtihi* (and for his sister) is one string. This can make strings of letters highly ambiguous – a common string like *l'nh* can be interpreted in at least seven different ways! We want the user to be able to enter any string they don't understand, without having to distinguish the different morphemes themselves, which can be a challenge for even quite advanced Arabic learners. For this, we again use the Buckwalter Arabic Morphological Analyzer integrated with our own headword list. Thus the strings are analysed into the appropriate morpheme(s) and the user will be redirected to the relevant entry. For example if a user enters the string وكتبه (and his books), they will be redirected to كِتاب (book), with the information that it's preceded by وَ (and) and followed by هُ (his). In case of multiple possible analyses, like كتب, which can mean "he wrote", "writing", or "books", it gives the possible entries with a summary of the part of speech and meaning, and allows the user to choose which entry to display fully.

12 Conclusion

The world of Arabic lexicography is a very challenging field, where little can be relied on, and expected and unexpected pitfalls abound. Despite, or maybe even because of this, it is a fascinating area, where truly significant achievements can be made. With English and Arabic being world languages, and being two of the six official languages of the UN, it is amazing that so few resources exist for translation between the two, with no reasonably modern ones.

We feel that this dictionary fills that hole, and with the opportunity to constantly update that online dictionaries offer, will continue to fill the gap for many years to come.

We hope this look behind the scenes has been interesting for lexicographers and other linguists.

13 References

Arts, T. & McNeil, K. (2013). Corpus-based lexicography in a language with a long lexicographical tradition: The case of Arabic. In *Proceedings of WACL'2, Second Workshop on Arabic Corpus Linguistics, 22 July 2013*. Lancaster University, UK.

Arts, T. et al. (Forthcoming 2014). *Oxford Arabic Dictionary*. Oxford: Oxford University Press.

Attia, M., Pecina, P., Toral, A. & Van Genabith, J. (2014). A corpus-based finite-state morphological toolkit for contemporary Arabic. In *Journal of Logic and Computation*, 24 (2), pp. 455-472.

Baalbaki, R.M. (2004). Coinage in Modern English-Arabic Lexicography. In *Zeitschrift für Arabische Linguistik*, 43, pp. 67-71.

Baalbaki, M. & Baalbaki, R.M. (2013) *Al-Mawrid Al-Hadeeth*. Beirut: Dar El-Ilm Lilmalayin.

Benzehra, R. (2012). Modern English-Arabic Lexicography: Issues and Challenges. In *Dictionaries: Journal of the Dictionary Society of North America*, 33, pp. 83-102.

Doniach, N.S. (ed.) (1972). *Oxford English-Arabic Dictionary*. Oxford: Oxford University Press.

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

Haywood, J.A. (1965). *Arabic Lexicography*. Leiden: Brill.

Hoogland, J. (2003). Coverage. In *The Nijmegen Arabic/Dutch Dictionary Project*. Accessed at: http://wba.ruhosting.nl/Content1/1.4_Coverage.htm [10/04/2014].

Hoogland, J. (2004). Working Methods. In *The Nijmegen Arabic/Dutch Dictionary Project*. Accessed at: http://wba.ruhosting.nl/Content1/1.4_Working_Methods.htm [10/04/2014].

Hoogland, J. et al. (2003). *Woordenboek Arabisch-Nederlands*. Amsterdam: Bulaaq.

Ibrahim, R. & Aharon-Peretz J. (2005). Is Literary Arabic a Second Language for Native Arab Speakers?: Evidence from Semantic Priming Study. In *Journal of Psycholinguistic Research*, 34 (1), pp. 51-70.

Kilgarriff, A., Rychly, P., Smrz, P. & Tugwell, D.A. (2004). The Sketch Engine. In *EURALEX Lorient Proceedings*. Lorient, France.

Van Mol, M. & Bergman, K. (2001). Leerwoordenboek Arabisch. Amsterdam: Bulaaq.

Wehr, H. (1979) *A Dictionary of Modern Written Arabic*. Rev. ed. Urbana: Spoken Language Services.