Are you sure your tool does what it is supposed to do? Validating Arabic root extraction

Janneke van der Zwaan
Netherlands eScience Center, Amsterdam, the Netherlands

Maksim Abdul Latif
Department of Philosophy and Religion Studies, Utrecht University, Utrecht, the Netherlands

Dafne van Kuppevelt
Netherlands eScience Center, Amsterdam, the Netherlands

Melle Lyklema
Department of History and Art History, Utrecht University, Utrecht, the Netherlands

Christian Lange
Department of Philosophy and Religion Studies, Utrecht University, Utrecht, the Netherlands

Abstract

Although in the digital humanities, researchers use software tools to conduct their research, and often apply these tools to data the software was not developed for, there has been little attention for investigating tool performance on this data. This is strange because in order to be able to appraise the results of digital humanities research, it is important to understand to what extent the tool output is correct. To illustrate the importance of the validation of tools, this article presents a case study of validating Arabic root extraction tools. Arabic words are based on root letters; three root letters usually demarcate a semantic field. Thus, roots can be used for studying semantic fields. For example, researchers can gain insight into the relative importance of the different senses (i.e. seeing, hearing, touching, smelling, and tasting) in Arabic jurisprudence (*fiqh*) by extracting and counting roots. A problem is that there are only a few usable tools available. We take three root extraction tools, Khoja (Khoja and Garside, 1999, *Stemming Arabic Text.* Lancaster, England: Lancaster University), ISRI (Taghva *et al.*, 2005, Arabic stemming without a root dictionary. In *International Conference on Information Technology: Coding and Computing (ITCC'05).* Vol. 2. Las Vegas, NV, April 2005 pp. 152–57), and AlKhalil (Boudlal *et al.*, 2010, Alkhalil morpho sys1: a morphosyntactic analysis system for Arabic texts. In *International Arab Conference on Information Technology*. New York, NY: Elsevier Science Inc., April 2017, pp. 1–6),

Correspondence:
Janneke van der Zwaan,
Science Park 140, 1098 XG
Amsterdam, the Netherlands.
E-mail:
j.vanderzwaan@
esciencecenter.nl

doi:10.1093/llc/fqz045 Advance Access published on 9 August 2019

and create manually annotated gold standard data consisting of three samples of approximately 1,000 words from important books of Islamic jurisprudence. We show that Khoja is the best root extraction tool for our data. We also demonstrate that the relative counts of individual roots differ among tools, which leads to a different interpretation depending on which tool is chosen. This means that findings based on automatically extracted roots should always be interpreted with care.

1 Introduction

The 'instruments' used by Digital Humanities researchers consist of software tools. Some of the most commonly used tools for research involving (historic) text are tokenizers, part of speech taggers, and named entity extractors. Often, these kinds of tools have been developed and validated for particular types of data, e.g. newspaper articles (Vossen *et al.*, 2016) or social media text (Hutto and Gilbert, 2014). In the field of Digital Humanities, however, it is not uncommon to apply tools to other types of data than they were originally developed for. One of the consequences of applying these tools to other kinds of data, is that tool performance might significantly decrease compared with the performance originally reported for these tools. For example, when applying a part of speech tagger developed for contemporary Dutch to 17th century Dutch, accuracy drops from an estimated 96.5–70.9% (Van den Bosch et al. 2007; Tjong Kim Sang et al., 2017).

In the case of Digital Humanities research based on Arabic text, this problem is compounded by the relative inattention within the natural language processing community for Arabic language processing tools as compared with English and other Latin script-based languages. As a result, there are fewer tools available for Arabic, which means that Digital Humanities scholars studying Arabic texts have fewer options to choose from and are thus more likely to use tools that have not been developed specifically for their type of data. As a consequence, undetected performance loss represents a serious risk for Digital Humanities research on Arabic text.

In order to provide insight into the issues surrounding the use and validation of tools and to raise awareness of the importance of validation, this article presents a validation of tools for extracting roots from Arabic text. Automatic root extraction allows researchers to gain insight into much broader concepts than when searching for and counting individual words. For example, a researcher may be interested in studying the 'ratio of the senses' (McLuhan, 1962) in a given body of literature, by determining the relative importance of words referring to seeing, hearing, touching, smelling, and tasting. One way of doing that would be to count the number of words that refer to each sense (and divide by the total number of words in the (sub)corpus compensate for (sub)corpus size). However, enumerating all words that have to do with seeing, hearing, touching, smelling, and tasting is a lot of work. Even if researchers came up with tentative comprehensive lists, they might miss variants that are unique to a specific work or genre. In addition, in order to study other semantic fields, such as 'writing' and 'telling', the researcher would have to come up with more new word lists.

Root extraction provides a much faster approach to counting the number of words that refer to concepts. Arabic is based on a system of (usually three) 'root letters' that define a semantic field, and from which—by adding enclitics, proclitics, prefixes, suffixes, and infixes, by doubling a root letter, and by filling in various short and long vowels—all verbs, nouns, and adjectives pertaining to the semantic field are derived. For example, the root k-t-b denotes the semantic field of 'writing'. Its verbal variants, like in the case of most other Arabic verbs, include several so-called forms (usually up to ten are counted). Form 1 (k-t-b) means 'to write', form 2 (k-t-t-b) 'to make someone write', form 3 (k-[long ā]-t-b) 'to write to someone', form 4 ([short a]-k-t-b) 'to dictate a text', etc. Nominal variants include m-k-t-b-t ('library'), k-t-[long ā]-b ('book'), and many more. There are many more intricacies that are beyond the scope of this article, but the basic idea is this: we want to gain insight into semantic fields such as the senses and this is much easier by counting roots than by counting words. However, in order to be able to do this, reliable root extractors are required.

The field of Digital Humanities generally pays little attention to validation, although having insight into the performance of tools on the data to which they are applied is critical for appraising the output of Digital Humanities research. We argue that by paying more attention to validation of performance, the development of tools for Arabic text can be improved, although Arabic Digital Humanities continues to be affected by a range of other issues as well.

For example, over 10 years ago, it has already been noted that many of the tools that are developed for Arabic are hard to use by others than the original developers, because they are closed source, lack documentation, and/or use custom input formats (Atwell *et al.*, 2004). More recently, Alosaimy and Atwell (2015) conclude that it has not improved since then. Also, there is a general shortage of data that can be used for the development and/or validation of tools, mainly because researchers that create and/or validate tools do not make this data available for others. These issues not only hamper Digital Humanities research, but also restrict progress on existing tools as well. Because this study only uses existing tools, we make no source code available. We do provide the gold standard under a cc-by 4.0 license.[1]

This article is organized as follows. In Section 2, we provide some context and elaborate the senses use case to illustrate the kinds of research questions that scholars are able to pursue using root extraction in Arabic and Islamic studies. Section 3 introduces tools for Arabic root extraction. Section 4 describes our selection of tools for the validation, the validation data, and how performance was measured. Section 5 presents the results, and our conclusions follow in Section 6.

2 The Ratio of the Senses in Islamic Jurisprudence

Root extraction in Arabic, in theory if not in practice, provides researchers with a powerful basis for conceptually parsing large corpora of Arabic digitized texts. The currently available Arabic digital corpus is impressive by any standard; it is estimated to be at least ten times as large as the entire classical Latin and Greek corpus taken together (∼1.1 billion words versus ∼150 million words). More and more Arabic texts, both of the premodern and the modern period, are in the process of being digitized. The corpus includes texts from genres as diverse as Islamic theology, Qur'an commentaries, Islamic law, poetry and prose (both secular and religious), and Arabic newspapers. As digitization of Arabic texts progresses, the field of Arabic and Middle East Studies Digital Humanities is gaining momentum, a development that has recently resulted in several joint publications, a series of international conferences, and an emerging network of Islamicate Digital Humanists (Muhanna, 2016; Miller *et al.*, 2018).

To illustrate the great potential of root extraction for the advancement of knowledge about the Arabic literary heritage, consider the example of Islamic law and ethics, which are jointly referred to by Muslims worldwide as the Sharia. Since its inception in the 7th century of the Common Era, Sharia jurisprudence has consistently been written in Arabic, in a fairly unchanging linguistic form, such that a legal text written in, say, 10th century Northern Persia is still commensurable, in grammar and lexicography, with a text written in 18th century Morocco. Though divided into several schools (there are five major ones, four Sunni schools, and one Shi'i school), the Sharia tradition, next to the Qur'an, is arguably the single most important shared point of reference connecting Muslim believers living as far apart as in Morocco, Indonesia, and China.

A carefully assembled corpus of Sharia jurisprudence (*fiqh*) spanning the entire history of Islam has been created and made publicly available by the Utrecht-based 'Bridging the Gap'—project.[2] The corpus offers researchers the ability to read Islamic legal literature 'from a distance' (Moretti, 2013), discerning basic patterns of legal thought and tracing longue-durée developments in Sharia discourse. A reliable computational mechanism to extract roots from words would constitute a sharp knife to cut through this corpus. Take, for example, the history of perception, or sensory history, a budding field in cultural studies (Howes, 2004). It has been a commonplace in the Western study of world history that the modern West's rise to scientific and industrial preeminence owes much to the privileged position Western culture accords the sense of sight, declaring sight to be detached, objective, rational, and hence conducive toward scientific and industrial progress. In contrast, oriental cultures, according to the classic hypothesis by media theorist Marshall McLuhan (1962), would have remained mired in an aural/oral culture, or worse, privileged the 'lower' senses of smell, taste, and touch.

Now imagine root searching the 14 centuries of Islamic legal and ethical discourse along the lines of the five senses. The primary root connected to the semantic field of seeing is b-ṣ-r. This gives us a plethora of verbs, nouns, and adjectives, all related to perception by the eye: baṣar 'sight', istibṣār 'insight', abṣār 'eyes', baṣīra 'mental perception', abṣara 'to observe', istabṣara 'to be able to see', etc. Similarly, the root s-m-ʿgives us samʿ 'hearing', sammāʿa 'stethoscope', mismaʿ 'ear', sāmiʿ 'listener', samiʿa 'to hear', tasammaʿa 'to eavesdrop', etc. From the root sh-m-m, we can derive shāmma 'sense of smell', mashmūm 'musk, spoiled food', shammām '(tobacco) snuffer', shamma 'to smell', ishtamma 'to sniff', etc. The root dh-w-q forms the words dhawq 'sense of taste', dhawwāq 'gourmet', tadhawwuq 'gustatory delight', madhāq 'taste', tadhawwaqa 'to relish', etc. Finally, from the root l-m-s, we encounter the following derivations: lams 'touch, touching', lamsa 'stroke', malmas 'point of contact', mulāmasa 'sexual intercourse', talammasa 'to grope', talāmasa 'to touch each other', etc. A reliable root extractor would be able to catch all the various derivations of a single root, and thus save researchers a lot of time: instead of running searches for all the different derivations, it would suffice to search for a root. As a welcome side, effect, unexpected, or previously unknown derivations might come to researchers' attention.

On a more conceptual level, a reliable root extractor would enable researchers to gage the respective weight of the five senses in Islamic legal discourse, thus enabling us to determine the ratio of the senses (McLuhan, 1962) in Sharia law. Shifts in this ratio over the centuries might be studied. The legal schools in Islamic law are based in different regions of the Islamic world. This could possibly result in different, regionally defined ratios of the senses (e.g. according to Shi'i law in Iran and Iraq as opposed to Maliki law in North Africa). Or the ratio of the senses of Islamic law could be compared with the ratio of the senses obtaining in, say, Arabic poetry.

To illustrate such a use of root extraction in Islamic legal literature, we have used the Khoja root extractor[3] to measure the ratio of the senses in the entire corpus and to compare the ratio of the senses emerging from Shi'i jurisprudence to the ratio of the senses emerging from the legal literature of the Sunni schools. To enrich results, we have included secondary roots in this search. Next to b-ṣ-r, s-m-ʿ, sh-m-m, dh-w-q and l-m-s, there are several secondary roots that also refer to acts of seeing, hearing, smelling, tasting, and touching. The two most important secondary roots for each sense are n-ẓ-r, r-ʾ-y (seeing); ṣ-w-t, ḍ-j-j (hearing); r-y-ḥ, ʿ-ṭ-r (smelling); ṭ-ʿ-m, l-dh-dh (tasting); m-s-s, l-ṣ-q (touching).

Figure 1 shows the relative root counts for the senses roots in the Fiqh corpus. The results provide interesting insights into Islamic legal thought, showing it to (1) clearly emphasize sight over hearing (against McLuhan's assumption of an Oriental oral/aural bias), (2) almost completely ignore issues connected to the semantic field of smell (as opposed to, e.g., modern Western anxieties about smelliness), and (3) be more concerned with issues of taste (think of halal food) than of touch (e.g. between members of the opposite sex). The Sunni and the Shi'i legal ratio of the senses seem remarkably similar overall. Shi'i jurists appear to focus more than their Sunni counterparts on aspects of vision, while Sunni jurists seem to be marginally more concerned than Shi'i jurists with issues of tasting and food. However, to be able to state such differences with more nuance and confidence, i.e. for the Sunni–Shi'i comparison to become viable, an accurate root predictor is necessary. Do such tools for root extraction exist?

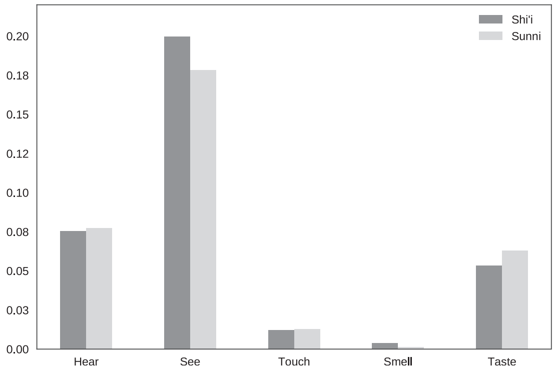

Fig. 1. Relative root counts for the senses roots in the Fiqh corpus extracted with Khoja

3 Tools for Root Extraction

There are two types of tools that can be used for root extraction: stemmers and morphological analyzers. Stemming is the process of removing enclitics, proclitics, prefixes, suffixes, and infixes, while morphological analyzers try to automatically extract a wide range of linguistic information about each word in a text, such as the root, suffixes, prefixes, infixes, and the pattern used to create the specific word form. Stemmers that perform root extraction are called ‘heavy’ stemmers, in contrast to ‘light’ stemmers that remove less morphological variance. Examples of heavy stemmers are the Khoja stemmer (Khoja and Garside, 1999), ISRI (Taghva *et al.*, 2005), and the stemmer proposed by Al-Kabi *et al.* (2015) . There also are multiple morphological analyzers for Arabic, but most are not available or difficult to use (e.g. requiring transliterated input) (Atwell *et al.*, 2004). Tools that are available include BAMA (Buckwalter, 2002), Elixir-FM (Smrž, 2007), and AlKhalil (Boudlal *et al.* 2010).

Even when stemmers and analyzers are available, it can be challenging to use them because of limited documentation. In addition, there is no standard output format, which makes it difficult to compare results of different analyzers. Recently, however, tools that provide uniform interfaces to multiple analysis tools and standardized, comparable output have been introduced. SAWAREF is able to run seven morphological analyzers, and combines the results with the aim of improving the results of individual tools (Alosaimy and Atwell, 2016). Software Architecture for Arabic language pRocessing (SAFAR) provides access to three analyzers and five stemmers (Jaafar and Bouzoubaa, 2015).

Both SAWAREF and SAFAR are closed source software. SAWAREF is only available through a web interface, which make it cumbersome to use it in text processing pipelines.[4] In addition to a web interface,[5] SAFAR has binaries available that can be used to analyze text directly. Because of this convenience, we decided to use SAFAR for the validation.

Research presenting stemmers or morphological analyzers do not always include performance results. For example, Boudlal *et al.* (2010) argue that it is hard to evaluate their morphological analyzer because there are no suitable test corpora. In other cases, tools are evaluated on other tasks than root extraction, e.g. full text search (Taghva *et al.*, 2005; Larkey *et al.*, 2007), or part of speech tagging (Alosaimy and Atwell, 2016).

For evaluating root extraction performance, two strategies can be distinguished: using gold standard data, or lexicons with generated words. Ghwanmeh *et al.* (2009) claim their stemmer produces correct roots for up to 95% on a manually annotated corpus of 242 abstracts from the Proceedings of the Saudi Arabian National Computer conferences. However, they do not provide details about how the evaluation data were created and did not make the data available. Khoja (2001) reports a root extraction performance of up to 97%, but does not specify how performance was measured or what data were used.

Al-Shawakfa *et al.* (2010) present a comparison of six root finding algorithms. They reimplemented the tools based on descriptions in papers. To evaluate the performance, an evaluation lexicon containing 27.6 million words was generated from a dictionary of 3,823 triliteral roots. Reported performance ranged from 14 to 39%. These numbers are much lower than what is reported by the tools themselves. An explanation for this difference is the fact that the lexicon of generated words very likely contains words that do not occur in natural language texts.

Table 1. Overview of the root extraction tools used in this study

Name	Type	Publication	Root extraction performance
Khoja	Stemmer	Khoja and Garside (1999)	Up to 97% (Khoja, 2001), but no details about evaluation
ISRI	Stemmer	Taghva *et al.* (2005)	Less than Khoja (Al-Shawakfa *et al.*, 2010)
AlKhalil	Analyzer	Boudlal *et al.* (2010)	Not evaluated

Table 2. Overview of the annotated text samples

Author	Title	School of law	Century (CE)	# words in sample
Al-Mawardi	Hawi	Sunni	1,058	985
Al-Tusi	Mabsut	Shi'i	1,067	981
Al-Sarakhsi	Mabsut	Sunni	1,090	996

4 Method

After trying out the different stemmers and analyzers provided by SAFAR, the Khoja and ISRI stemmers and AlKhalil analyzer were selected for the validation study. The Khoja and ISRI stemmers were selected because they are heavy stemmers, i.e. stemmers that try to reduce words to their roots. All other stemmers in SAFAR are light stemmers. Furthermore, only one of the analyzers in SAFAR could be used for the purpose of root extraction. Two analyzers did not work as expected: MADAMIRA (Pasha *et* al., 2014) gives an error message when it is called from SAFAR, and the BAMA output does not contain roots.

Table 1 lists the tools used in this study. In addition to the name, type, and a reference to a publication about the tool, the table summarizes what is known about the root extraction performance. The best performing root extractor seems to be Khoja although it is unclear on what kind of text a performance of 97% was obtained. In a comparison of root extraction algorithms, Khoja outperformed the other algorithms, including ISRI (Al-Shawakfa *et al.*, 2010). Because the algorithms were tested on artificial instead of natural text, we do not include the actual numbers, but note Khoja was better than ISRI. For AlKhalil, no root extraction performance was found.

To determine to what extent the roots extracted by the tools are correct, we created gold standard data. The gold standard data were created by taking approximately 1,000 words from the beginning of three major 'classical' books of Islamic jurisprudence from the 11th century Common Era. The roots of every word were then manually extracted by a trained native speaker. The three books were chosen because they are seminal, authoritative works in the jurisprudence of both Sunni and Shi'i Islam. Islamic jurisprudence is characterized by a relatively stable style and vocabulary over the centuries, and is arguably a core genre of literature in Islamic civilization as a whole. This amply justifies using our three books to create gold standard data. Table 2 shows the title, author, and number of words in the sample of each book. Because tools may tokenize a text in a different way than a person does, the annotator was provided with a list of tokens to extract roots from. The list of tokens was obtained by running Khoja on the texts. To be able to properly take into account the context of words, the annotator also had access to the complete texts. The gold standard data are available under a cc-by 4.0 license.

In Arabic, not all words are derived from a root. These so-called letters are annotated in the gold standard with #. For convenience, we consider # to be a 'special root' and treat the label as such in the remainder of this article. Khoja and AlKhalil mark letters explicitly, ISRI ignores them. Additionally, a word can be derived from multiple roots. Usually, the context determines which root is the correct one, but, even in context, words can be ambiguous. This means that root extraction is a multilabel classification problem; given a word in a text, the task is to predict the correct roots for this word. In practice, the number of correct roots is usually one; in the gold standard data 576 of 2,865 words were annotated with two roots and all other words were assigned a single root. The stemmers always assign a single root to a word, but AlKhalil can produce multiple analyses containing different roots. For the evaluation texts, the average number of roots

Table 3. Performance of the Khoja stemmer

Text	No. of words	Precision	Recall	F1
Al-Mawardi, Hawi	985	55.9	55.4	54.8
Al-Tusi, Mabsut	981	53.0	52.5	51.6
Al-Sarakhsi, Mabsut	996	62.3	59.8	59.1
Combined	2,962	61.5	55.9	55.5
Combined (NLTK stopwords removed)	2,230	74.1	72.7	71.7
Combined (custom stopwords removed)	1,793	83.1	81.8	80.9

Table 4. Performance of the ISRI stemmer

Text	No. of words	Precision	Recall	F1
Al-Mawardi, Hawi	985	36.4	38.8	37.0
Al-Tusi, Mabsut	981	35.5	38.7	36.5
Al-Sarakhsi, Mabsut	996	36.6	38.2	36.6
Combined	2,962	37.2	38.6	36.5
Combined (NLTK stopwords removed)	2,230	46.3	49.9	46.7
Combined (custom stopwords removed)	1,793	54.7	59.5	55.9

Table 5. Performance of the AlKhalil analyzer

Text	# words	Precision	Recall	F1
Al-Mawardi, Hawi	955	33.4	29.5	29.4
Al-Tusi, Mabsut	944	34.7	33.0	31.9
Al-Sarakhsi, Mabsut	955	30.9	28.1	28.4
Combined	2,854	37.0	30.1	30.3
Combined (NLTK stopwords removed)	2,177	44.5	36.3	37.8
Combined (custom stopwords removed)	1,770	52.9	45.2	46.7

predicted by AlKhalil is 2.24 (minimum: 1; maximum: 9); 1,422 of 2,854 words are assigned a single root.[6]

Tool performance is measured by calculating precision, recall, and F1-measure for each root in the gold standard separately and then taking the average weighted by the root frequencies in the gold standard. Given the number of true positives (tp), true negative (tn), false positives (fp), and false negatives (fn) for a specific root, precision and recall are defined as:

$$\text{Precision} = \text{tp}/(\text{tp} + \text{fp})$$
$$\text{Recall} = \text{tp}/(\text{tp} + \text{fn})$$

And F1 measure is the harmonic mean of precision and recall:

$$\text{F1} = 2 \cdot (\text{precision} \cdot \text{recall})/(\text{precision} + \text{recall}).$$

Because the tools use different tokenizers, performance is calculated only on the tokens that are returned by a tool. As not all words in a text convey meaning, and we are less interested in root extraction performance for irrelevant words, we also calculate performance for the texts with stopwords removed. Two stopword lists are used; a standard list from the NLTK (Bird *et al.*, 2009), which contains 248 words, and a custom list created by the authors that contains 544 words.[7]

For answering our research question, we are interested in accurate roots counts, and this does not necessarily require accurate root extraction for individual words (cf. Forman, 2005). So, in addition to root extraction performance, we also explore the counts for the different roots in the gold standard and the ones extracted by the tools. For this analysis too, we filter stopwords to get an approximation of the performance on relevant versus irrelevant words. We repeat counting the senses roots in the gold standard data and the complete Fiqh corpus. Finally, we investigate relative root counts for roots that do not occur in the gold standard.

5 Results

Tables 3–5 contain root extraction performance for Khoja, ISRI, and AlKhalil, respectively. Without filtering, stopwords from the text performance is between ∼30 and ∼55%. This is much lower than 97% that was reported for Khoja (see Table 1). When filtering stopwords, performance increases to ∼80%. Unsurprisingly, performance gets better if more stopwords are removed. The performance is reasonable at least for Khoja and when ignoring stopwords.

However, because these performance results are weighted by the frequencies of roots that occur in the gold standard, incorrect roots extracted by the

Table 6. The overlap between the roots in the gold standard and the ones extracted by the tools

	Roots	Khoja		ISRI		AlKhalil	
Text	**in gs**	**Roots**	**Overlap (%)**	**Roots**	**Overlap**	**Roots**	**Overlap (%)**
Al-Mawardi, Hawi	255	332	216 (84.7)	405	162 (63.5)	504	196 (76.9)
Al-Tusi, Mabsut	282	343	233 (82.6)	433	179 (63.5)	499	219 (77.7)
Al-Sarakhsi, Mabsut	287	343	249 (89.6)	421	190 (66.2)	531	209 (72.8)
Combined	541	627	461 (85.2)	883	363 (67.1)	883	407 (75.2)

Percentages denote the amount of overlap with the gold standard.

Table 7. The top 15 roots in the gold standard and absolute and relative deviations for the different tools

Root	Stop-	Count	Khoja			ISRI			AlKhalil		
	word	**in gs**	**Count**	**Abs. dev.**	**Rel. dev. (%)**	**Count**	**Abs. dev.**	**Rel. dev. (%)**	**Count**	**Abs. dev.**	**Rel. dev. (%)**
#		898	76	822	91.5	0	898	100.0	561	337	37.5
???		86	11	75	87.2	0	86	100.0	0	86	100.0
???	c	65	55	10	15.4	13	52	80.0	66	1	1.5
???		46	46	0	0.0	43	3	6.5	44	2	4.3
???		45	2	43	95.6	0	45	100.0	42	3	6.7
???		38	36	2	5.3	30	8	21.1	33	5	13.2
???	c	34	31	3	8.8	0	34	100.0	22	12	35.3
???		34	18	16	47.1	10	24	70.6	0	34	100.0
???		31	30	1	3.2	0	31	100.0	19	12	38.7
???		30	29	1	3.3	29	1	3.3	30	0	0.0
???		28	28	0	0.0	28	0	0.0	27	1	3.6
???		26	26	0	0.0	25	1	3.8	20	6	23.1
???		26	29	3	11.5	0	26	100.0	0	26	100.0
???	c	26	26	0	0.0	23	3	11.5	17	9	34.6
???		23	22	1	4.3	22	1	4.3	24	1	4.3

Table 8. Overall comparison of the root counts in the gold standard and the ones extracted by the tools

Measure	Khoja	ISRI	AlKhalil
Mean absolute error	2.9	3.8	2.8
Median absolute error	0	1	0
Max. absolute error (#)	822	898	337
Max. absolute error (not #)	75	88	96
Mean relative error (%)	63.5	77.6	76.2
Median relative error (%)	0.0	33.3	0.0
Relative error <2.5%, *n* (%)	327 (60.4)	234 (43.3)	277 (51.2)

tools (i.e. roots that are extracted by the tools, but do not occur in the gold standard) do not affect the performance. Additionally, it is unclear what these performance results mean for the numbers we need to answer our research question, i.e. the counts of individual roots. Table 6 shows the number of different roots in the gold standard and extracted by the tools. The tools consistently extract more different roots; whereas the gold standard contains 541 different roots (∼250–290 per text fragment), the numbers of roots extracted by the tools vary from 627 (with 85.2% overlap with the gold standard roots) to 883 roots (with 67.1–75.2% overlap). The roots extracted by Khoja have the most overlap with the gold standard root set, both ISRI and AlKhalil extract more different roots. Because AlKhalil often extracts multiple roots for a word, it is not surprising that the root set extracted by this tool is large. Because ISRI only extracts a single root for each word, extracting many roots not in the gold standard will affect the root counts of individual roots. For answering our type of research question, we need accurate root counts. Table 7 compares the root counts for the 15 roots that occur most frequently in the gold standard data to the root counts according to the three different tools. The table also specifies the absolute and relative difference between

Table 9. Top 15 most frequent roots extracted by the tools that do not occur in the gold standard (sw = stopwords: c = custom list, b = custom list and nltk list, # = number of times the root extracted should have been a letter (#))

Khoja				ISRI				AlKhalil			
Root	**sw**	**Count**	**#**	**Root**	**sw**	**Count**	**#**	**Root**	**sw**	**Count**	**#**
??	b	96	11	??	b	80	80	NOANALYSIS		164	72
??	b	87	86	???	–	79	0	???	–	122	93
??	b	63	63	??	b	60	60	???	–	100	93
??	b	35	35	??	c	42	42	???	–	89	89
???	b	28	28	??	c	30	0	???	–	75	75
??	b	25	24	???	b	30	30	???	c	75	75
???	b	20	20	???	c	28	28	???	–	74	74
???	b	18	18	??	b	24	24	???	–	74	74
??	b	16	16	???	c	23	0	???	–	58	58
??	b	16	16	???	b	18	18	???	–	41	32
???	–	15	14	???	–	18	0	???	c	41	0
??	b	13	13	??	b	16	16	???	–	40	40
??	b	13	13	???	c	16	16	???	–	37	32
???	–	12	0	???	–	16	0	???	–	36	0
???	b	12	12	?????	c	13	13	???	–	35	0

Table 10. Statistics about root counts for roots not in the gold standard

Measure	Khoja	ISRI	AlKhalil
Total no. of roots not in gs	166	520	476
No. of NLTK stopwords (%)	42 (25.3)	52 (10.0)	8 (1.7)
No. of custom stopwords (%)	71 (42.8)	102 (19.6)	23 (4.8)
Mean frequency	5.3	2.9	6.4
Median frequency	2	1	2

the gold standard root counts and the ones extracted by the tools. The results vary between roots and tools. Although Khoja and AlKhalil attempt to mark letters (#), the total counts are underestimated by both tools. As ISRI does not extract letters, it has a count of 0 for #. For the other roots, some root counts are quite accurate for some tools (e.g. Khoja extracts the correct root count for ??? (book)), while others are completely off the mark (e.g. ISRI extracts a root count of 0 for ??? (universe)). Table 8 summarizes the similarities and differences between the root counts in the gold standard and the ones extracted by the tools. While the top absolute difference is substantial (337–898), the second difference is much smaller (75–96). The mean absolute error is 3–4 for all tools, and the median is 1 to 0. The mean relative error seems substantial, but is dominated by some outliers; the median relative difference is 0–33%. Finally, the table reports the number of roots for which the difference with the gold standard root counts is small (i.e. <2.5%), which is 40–60% of the gold standard roots. We can conclude the root counts for individual roots that occur in the gold standard are quite accurate, except the root counts for #. Khoja again appears to be the superior tool: the mean absolute deviation of AlKhalil is only slightly lower (2.8 versus 2.9), it has the lowest relative deviation (63.5%), and it produces the most root counts that are close to the gold standard root count (327 of 541 roots).

As shown in Table 6, the set of roots extracted by all three tools is much larger than the set of roots in the gold standard. Table 9 shows the top 15 root counts for roots extracted by the different tools that do not occur in the gold standard. In addition to the roots and the counts, the table lists whether the root is a stopword (sw), and how many of the occurrences should be letters (#) according to the gold standard. The root frequencies are in the same order of magnitude as the top 15 counts of roots in the gold standard (164-12). However, for Khoja and ISRI, most of the words in the top 15 are stopwords, and should have been a letter (#). This means that, as long as researchers

Table 11. Results for the senses roots in the gold standard

			Khoja			ISRI			AlKhalil		
Sense	**Root**	**gs count**	**Count**	**Abs. diff.**	**Rel. diff.** (%)	**Count**	**Abs. diff.**	**Rel. diff.** (%)	**Count**	**Abs. diff.**	**Rel. diff.** (%)
Hear	???	3	3	0	0	3	0	0	3	0	0
	???	0	0	0	0	0	0	0	0	0	0
	???	0	0	0	0	0	0	0	0	0	0
See	???	1	1	0	0	1	0	0	1	0	0
	???	7	6	1	14.3	5	2	28.6	6	1	14.3
	???	7	0	7	100	0	7	100	0	7	100
Touch	???	0	0	0	0	0	0	0	0	0	0
	???	1	0	1	100	0	1	100	1	0	0
	???	0	0	0	0	0	0	0	0	0	0
Smell	???	0	0	0	0	0	0	0	0	0	0
	???	0	0	0	0	0	0	0	0	0	0
	???	0	0	0	0	0	0	0	0	0	0
Taste	???	0	0	0	0	0	0	0	0	0	0
	???	0	0	0	0	0	0	0	0	0	0
	???	0	0	0	0	0	0	0	0	0	0

do not want to know frequencies of stopwords or letters, these counts will be ignored. For AlKhalil, the situation is different because AlKhalil usually extracts multiple roots for a word. This results many non-stopwords in the top 15, many of which should have been a letter according to the gold standard. If researchers want to count these roots, then the root counts can be distorted. Moreover, for a total of 164 words, AlKhalil returns no analysis (NOANALYSIS). Table 10 shows some more information about the root frequencies of roots that do not occur in the gold standard. In addition to the proportion of roots extracted that are stopwords, the mean and median frequencies are reported. The results show that while the top frequencies are of the same order of magnitude as the top 15 roots that do occur in the gold standard, the mean and median frequencies are much lower (mean $\leq$ 6, and median $\leq$ 2). In practice, the root counts of these roots will be only slightly affected. With regard to these results, Khoja and ISRI are the better tools; Khoja because it has the smallest set of roots not in the gold standard and ISRI because it has the lowest mean and median frequencies for roots not in the gold standard.

Next, we present results about the senses roots. Table 11 contains data about the numbers of senses roots in the gold standard, and how many were counted by the different tools. Only five of the nine senses roots occur in the gold standard. The counts of two of these five are correctly extracted by all three tools (i.e. ??? (hear), and ??? (see)), one root is missed by AlKhalil only (i.e. ??? (touch)). Furthermore, the counts of one root are off by 1 or 2 for one root (i.e. ??? (see)), and all tools fail to extract one root (i.e. ??? (see)). Although not all counts are correct, the overall picture painted by the different tools is quite comparable. When we look at the relative root counts in the Fiqh corpus, we come to the same conclusion. Figures 1–3 contain the relative roots counts for the senses roots as extracted by Khoja, ISRI, and AlKhalil, respectively. The differences between the Shi'i versus Sunni counts are small, and the trends between the five senses for each tool are comparable.

Based on these results, it is tempting to conclude that it does not really matter which tool is used for root extraction. However, upon closer inspection of the relative root counts for the gold standard roots in the complete Fiqh corpus, interesting differences between tools come to light. From Table 6, we know that the tools do not extract all gold standard roots. So, this is the first type of difference between relative root counts that occurs. For 136 of 541 gold standard roots (25.1%) at least one of the tools does not extract the root from the Fiqh corpus (Khoja: 55, ISRI 31, and AlKhalil 110). Because the gold standard is a sample of texts from the Fiqh corpus, these

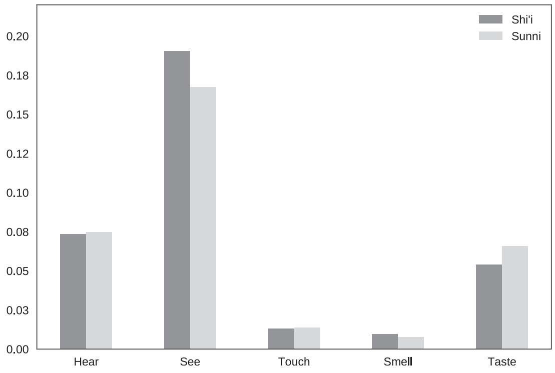

Fig. 2. Relative root counts for the senses roots in the Fiqh corpus extracted with ISRI

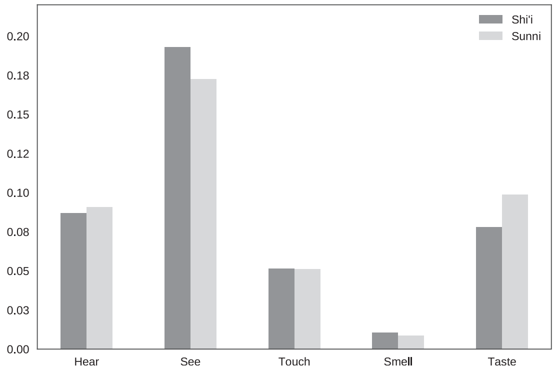

Fig. 3. Relative root counts for the senses roots in the Fiqh corpus extracted with AlKhalil

roots should have a non-zero count. Eighteen of these (3.3%) are stopwords; so, mostly these can be considered meaningful words, that researchers, given their particular research questions, might want to count. For these roots using the wrong tool lead to zero counts and subsequently erroneous conclusions about the relative importance of roots e.g. schools of law. Figure 4 shows the relative root counts of the three tools for ??? (foundation). AlKhalil does not extract this root at all.

A second type of difference that occurs is that the relative root count between the schools is reversed, or so small that there does not seem to be a difference between schools. This is problematic because it changes the interpretation of the results. For example, in Fig. 5, according to Khoja and AlKhalil, the root for determine (???) is more prominent in the Sunni school, while for ISRI, the numbers suggest it is more prominent in the Shi'i school. In Fig. 4, Khoja suggests the root of foundation is more important for the Shi'i school, while ISRI suggests there is not much of a difference. As this type of difference is hard to detect automatically, we visually inspected the graphs for all 500 non-stopword gold standard roots, and counted 51 (11.3%) roots with reversed bars for schools (cf. Fig. 5) and 81 (15.0%) that are close for one tool, but have bigger differences for one or more of the other tools (cf. Fig. 4). These numbers should be seen as estimates rather than exact statistics.

A third type of difference is the relative difference between root counts as extracted by a tool. For example, given two roots, a researcher wants to relate to each other, one tool may overestimate the root count of root 1 and underestimate the count of root 2, while a second tool may do the reverse. This may lead to erroneous conclusions about the relative importance of one root versus another. Because whether this type of difference occurs depends on the semantic fields a researcher is interested in, we did not try to estimate how often this happens for the gold standard roots.

Because Khoja has the highest root extraction performance, the root set extracted by Khoja has most overlap with the gold standard, Khoja has the most accurate root counts for the gold standard, and extracts the least roots that do not occur in the gold standard, we conclude Khoja is the best root count extractor for our data. For the senses roots, the differences between the three tools are small, both when looking at results for the gold standard (Table 11) and the complete Fiqh corpus (Figs. 1–3). When studying root counts in the complete Fiqh corpus, we see differences between tools that lead to different interpretations about the importance of a root for a school of law. This means that researchers have to be very careful when interpreting relative root counts.

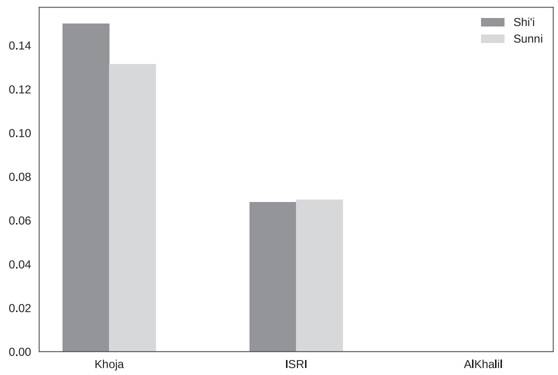

Fig. 4. Relative root counts of ??? (foundation) for the two main schools of law (Shi'i and Sunni) extracted from the Fiqh corpus by the three root extractors

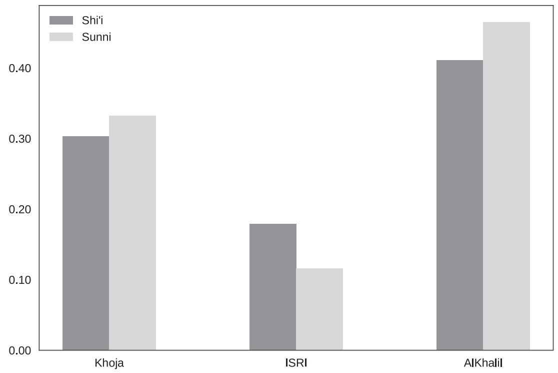

Fig. 5. Relative root counts of ??? (determine) for the two

6 Conclusion

Automatically extracting roots in Arabic texts allow researchers to quantify the relative importance of semantic fields without having to rely on possibly incomplete, manually created word lists. As an example, we presented a case study about understanding the importance of the different senses for different schools of law in Arabic legal texts. The reliability of (relative) root counts depends on the performance of the tool used to extract the roots. In this article, we presented a validation of three Arabic root extraction tools. The results show that, for a gold standard consisting of three texts of approximately 1,000 words, tool performance is ∼30–55%. This is much lower than the performance reported by the tool developers themselves. When removing stopwords, performance increases to ∼80%. Based on the different analyses of the roots extracted by the tools and the roots in the gold standard, we conclude that, for our data, Khoja is the best root (count) extraction tool.

All three tools show similar results for the senses use case. Although it is tempting to conclude any of these tools can be used to study semantic fields in Arabic, when comparing root counts between the tools in the complete Fiqh corpus, we found noteworthy differences for individual roots. In addition to the fact that all tools extract roots that do not occur in the gold standard, we saw roots for which one tool shows a big difference between relative root counts of different schools, while for another tool only a marginal difference was found, and other cases in which the interpretation of the relative root counts was reversed. We estimated that for ∼25% of the non-stopword roots in the gold standard, the root counts for one tool have a different interpretation than for other tools.

If researchers are interested in exploring the semantic fields that these roots represent, then it is unclear what the correct relative importance in different subcorpora (e.g. different schools of law) is. Because it is unknown for which roots the relative root counts are reliable, all results based on counts of automatically extracted roots should be interpreted with care and other types of evidence should be provided to substantiate findings. For future work, we propose to investigate what roots tools can extract reliable (relative) roots counts for. Furthermore, there is work to do on improving root extraction tools, and investigating alternative methods for studying root-based semantic fields in Arabic.

This case study of validating root extraction tools was presented to illustrate the importance of the validation of tools for Digital Humanities research. The second contribution of this article is a dataset for evaluating root extraction that is made available under a cc-by 4.0 license.

Acknowledgments

This work was supported by the Netherlands eScience Center.

References

Al-Kabi, M. N., Kazakzeh, S. A., and Abu Ata, B. M. (2015). A novel root based Arabic stemmer. *Journal of King Saud University – Computer and Information Sciences*, **27**(2): 94–103.

Al-Shawakfa, E., Al-Badarneh, A., Shatnawi, S., Al-Rabab'ah, K., and Bani-Ismail, B. (2010). A comparison study of some Arabic root finding algorithms. *Journal of the American Society for Information Science and Technology*, **61**(5): 1015–24.

Alosaimy, A. and Atwell, E. (2015). A review of morphosyntactic analysers and tag-sets for Arabic corpus linguistics. In *Corpus Linguistics*, Lancaster, UK, July 2015, pp. 16–19.

Alosaimy, A. and Atwell, E. (2016). Ensemble morphosyntactic analyser for classical Arabic. In *Proceedings 2nd International Conference on Arabic Computational Linguistics*. Konya, Turkey, April 2016.

Atwell, E., Al-Sulaiti, E., Al-Osaimi, S., and Abu Shawar, B. (2004). A review of Arabic corpus analysis tools. In *Proceedings of TALN04: XI Conference Sur le Traitement Automatique des Langues Naturelles*. Fes, Morocco, April 2004, pp. 229–34.

Bird, S., Loper, E., and Klein, E. (2009). *Natural Language Processing with Python*. Newton, MA: O'Reilly Media Inc.

Boudlal, A., Lakhouaja, A., Mazroui, A., Meziane, A., Bebah, M., and Shoul, M. (2010) Alkhalil morpho sys1: a morphosyntactic analysis system for Arabic texts. In *International Arab Conference on Information Technology*. New York, NY: Elsevier Science Inc., April 2017, pp. 1–6.

Buckwalter, T. (2002). Buckwalter Arabic morphological analyzer version 1.0. https://catalog.ldc.upenn.edu/LDC2002L49 (accessed 11 October 2018).

Forman, G. (2005). Counting positives accurately despite inaccurate classification. In *European Conference on Machine Learning*. Heidelberg, Berlin: Springer, pp. 564–75.

Ghwanmeh, S., Kanaan, G., Al-Shalabi, R., and Rabab'ah, S. (2009). 'Enhanced algorithm for extracting the root of Arabic words' In *2009 Sixth International Conference on Computer Graphics, Imaging and Visualization*. Piscataway, NJ: IEEE, pp. 388–91.

Howes, D. (2004). *Sensual Relations: Engaging the Senses in Culture & Social Theory*. Ann Arbor, MI: University of Michigan Press.

Hutto, C. J. and Gilbert, E. (2014). Vader: a parsimonious rule-based model for sentiment analysis of social media text. In *Eighth International Conference on Weblogs and Social Media (ICWSM-14)*, Ann Arbor, MI, pp. 216–25.

Jaafar, Y. and Bouzoubaa, K. (2015). Arabic natural language processing from software engineering to complex pipeline. In *2015 First International Conference on Arabic Computational Linguistics (ACLing)*. Piscataway, NJ: IEEE, pp. 29–36.

Khoja, S. (2001). APT: Arabic part-of-speech tagger. In *Proceedings of the Student Workshop at the Second Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL2001)*. Pittsburgh, PA, pp. 20–5.

Khoja, S. and Garside, R. (1999). *Stemming Arabic Text*. Lancaster, England: Computing Department, Lancaster University.

Larkey, L. S., Ballesteros, L., and Connell, M. E. (2007). Light stemming for Arabic information retrieval. In *Arabic Computational Morphology*. Berlin, Germany: Springer, pp. 221–43.

McLuhan, M. (1962). *The Gutenberg Galaxy: The Making of Typographic Man*. Abingdon: Routledge & Kegan Paul.

Miller, M., Romanov, M., and Bowen Savant, S. (2018). Roundtable: digital humanities for middle east studies. *International Journal of Middle East Studies*, **50**: 103–9.

Moretti, F. (2013). *Distant Reading*. Brooklyn, NY: Verso.

Muhanna, E. (ed.) (2016). *Digital Humanities and Islamic & Middle East Studies*. Berlin, Germany: De Gruyter.

Pasha, A., Al-Badrashiny, M., Diab, M., *et al.* (2014). MADAMIRA: a fast, comprehensive tool for morphological analysis and disambiguation of Arabic. *LREC*, **14**: 1094–101.

Smrž, O. (2007). *Functional Arabic Morphology: Formal System and Implementation*. Ph.D. thesis, Charles University.

Taghva, K., Elkhoury, R., and Coombs, J. (2005). Arabic stemming without a root dictionary. In *International Conference on Information Technology: Coding and Computing (ITCC'05)*. Vol. 2. Las Vegas, NV, April 2005, pp. 152–7.

Tjong Kim Sang, E., Bollman, M., Boschker, R., *et al.* (2017). The CLIN27 shared task: translating historical text to contemporary language for improving automatic linguistic annotation. *Computational Linguistics in the Netherlands Journal*, **7**: 53–64.

Van den Bosch, A., Busser, B., Daelemans, W., and Canisius, S. (2007). An efficient memory-based morphosyntactic tagger and parser for Dutch. *Computational Linguistics in the Netherlands 2006: Selected Papers of the 17th Computational Linguistics in the Netherlands Meeting*. Utrecht, Netherlands: LOT, pp. 99–114.

Vossen, P., Agerri, R., Aldabe, I., *et al.* (2016). NewsReader: using knowledge resources in a cross-lingual reading machine to generate more knowledge from massive streams of news. *Knowledge-Based Systems*, **110**: 60–85.

Notes

[1] https://github.com/arabic-digital-humanities/root-extraction-validation-data

[2] https://github.com/arabic-digital-humanities/fiqh

[3] More details about the root extraction tools can be found in section 3.

[4] SAWAREF can be found at http://sawaref.al-osaimy.com. Although there a box where text can be pasted, it did not work for our text. We also tried contacting the researchers involved, but they did not respond.

[5] http://arabic.emi.ac.ma/safar/

[6] There is a difference between the total number of words for the gold standard and AlKhalil, because the gold standard is based on a different tokenization of the text. The gold standard was tokenized using Khoja instead of AlKhalil.

[7] The custom stopword list is available in the dataset (see note 1).