Open-Source Boundary-Annotated Corpus for Arabic Speech and Language Processing

Claire Brierley[1], Majdi Sawalha[2], Eric Atwell[1]
University of Leeds[1] and University of Jordan[2]
[1] School of Computing, University of Leeds, LS2 9JT, UK
[2] Computer Information Systems Dept., King Abdullah II School of IT, University of Jordan, Amman, Jordan
E-mail: scscb@leeds.ac.uk, sawalha.majdi@gmail.com, eric@comp.leeds.ac.uk

Abstract

A boundary-annotated and part-of-speech tagged corpus is a prerequisite for developing phrase break classifiers. Boundary annotations in English speech corpora are descriptive, delimiting intonation units perceived by the listener. We take a novel approach to phrase break prediction for Arabic, deriving our prosodic annotation scheme from *Tajwīd* (recitation) mark-up in the Qur'an which we then interpret as additional text-based data for computational analysis. This mark-up is prescriptive, and signifies a widely-used recitation style, and one of seven original styles of transmission. Here we report on version 1.0 of our Boundary-Annotated Qur'an dataset of 77430 words and 8230 sentences, where each word is tagged with prosodic and syntactic information at two coarse-grained levels. In (Sawalha *et al.*, 2012), we use the dataset in phrase break prediction experiments. This research is part of a larger-scale project to produce annotation schemes, language resources, algorithms, and applications for Classical and Modern Standard Arabic.

Keywords: prosodic annotation; psycholinguistic chunking; phrase break prediction

1. Introduction

It is universally recognised that whatever the language, people process speech (and text) in chunks (Ladd, 1996), which in turn can be interpreted *syntactically* as function word groups (Liberman and Church, 1992) and *prosodically* as tone units (Croft, 1995; Roach, 2000). Phrase break prediction is a classification task within the Text-to-Speech synthesis pipeline that attempts to simulate human chunking strategies by assigning prosodic-syntactic boundaries to input text. A boundary-annotated and part-of-speech (PoS) tagged corpus (§2) is therefore an essential language resource for training such classifiers. Our research applies techniques honed on English (Brierley, 2011) to another stress-timed language, Arabic, and to the entire text of the Qur'an. One novelty is that we derive a coarse-grained boundary annotation scheme for Arabic from traditional recitation mark-up (*Tajwīd*) in the Qur'an; this is then compared with existing schemes for British and American English speech corpora (Taylor and Knowles, 1988; Beckman and Hirschberg, 1994). We then merge a PoS-tagged version of the text (Dukes, 2010) with our prosodic Qur'an, where each of the 77430 words is classified in terms of a finite set of boundary categories **`{major, minor, none}`**. An additional novelty is that we use compulsory and recommended **{** (٦٥) **,** ◌ۘ**,** ◌ۚ**}** and prohibited stops **{** ۙ **}** in *Tajwīd* mark-up (*cf.* Al-'ashmuni, 1973) to segment the text into 8230 sentences. Finally, we plan to evaluate the applicability of our Qur'an dataset as a training corpus for predicting boundaries in Modern Standard Arabic (MSA) text. This entails the creation of a second (smaller) boundary-annotated corpus for MSA, which is also segmented into sentences. We thus offer two unique language resources for exploring the prosody-syntax interface in Arabic, intended for open-source distribution. A related LREC submission (Sawalha *et al.*, 2012) uses our corpora to develop and evaluate the performance of several probabilistic, syntax-based phrase break classifiers.

2. Boundary Annotation Schemes for English

The Lancaster/IBM Spoken English Corpus or SEC (Taylor and Knowles, 1988) established a tripartite boundary annotation scheme **`{major, minor, none}`** for British English. Theoretically, major boundary markers **(||)** in this scheme denote pauses, and minor boundary markers **(|)** define tone units (Roach, 2000). Tone units (*i.e.* intonational phrases or chunks) are sequences that contain at least one accented word, namely: a word realised with pitch fluctuation on the syllable carrying primary stress (Croft, 1995). In practice, major boundaries do not *only* denote sentence segmental pauses, as in the following example from SEC A06 (informal news commentary on housing) annotated by Bryony Williams:

> '…For the thousand Turkish workers and their families | who lived in them | have left || taking advantage of a double pay offer || a cash grant from the government | and money from Mannesmann | to return home ||…'

In the above sentence, major boundary markers correspond to a *comma*, a *colon*, and a *full stop* respectively in the orthographic transcription of this utterance.

Speech corpora for American English, such as the Boston University Radio News Corpus (Ostendorf *et al.*, 1995) use ToBI or the ***To***nes and ***B***reak ***I***ndices annotation scheme (Beckman and Hirschberg, 1994) which identifies *five* theoretical levels of juncture between words: **`{0,1,2,3,4}`**. Break index **`{0}`** denotes no separation or *cliticization* (Ananthakrishnan and Narayanan, 2008), while index **`{1}`** applies to most phrase medial junctures between words. The 'correct' labelling of coarticulation is debateable, as in this SAMPA phonetic transcription **`/Di:jA:mi:/`** where *the army* (*i.e.* two consecutive words) is realised as one unit via the y-glide **`/j/`**. Index **`{2}`** is a special (and somewhat ambiguous) case, denoting either a hesitation that does not affect the tonal contour, or a disjuncture that is less strong than expected (Grabe, 2001). Indices `{3}` and `{4}` correspond to `minor` and `major` boundaries in the British system.

Both SEC and the Boston University Radio News corpus are widely-used resources for Text-to-Speech Synthesis, Automatic Speech Recognition, and Machine Translation applications but are largely representative of read speech, namely: speech delivered in a natural but controlled manner (Hasegawa-Johnson *et al.*, 2005). Therefore, the above boundary annotation schemes, and their implementation in English speech corpora, do not identify the disfluencies (*i.e.* filled pauses, repetitions, and false starts – *cf.* Stolcke and Shriberg, 1996) characteristic of spontaneous speech. These are outside the scope of our work, since we are interested in optimised (*i.e.* intelligible and naturalistic) chunking of text to maximise communication effectiveness.

3. Pause Markers in the Qur'an

Qur'anic verses are meant to be recited aloud from memory at least as much as they are meant for silent reading:

> '…The Arabic word *qur'an* means "recitation"...While the words have…been available in written form, equal prominence has been given to the continuing oral tradition…' (Denny, 1976).

The art of *Tajwīd* has developed over time to help believers achieve "clearly articulated recitation", and one aspect of this is the system of stops and starts وَقْفْ وَ ٱبْتِدَاء or *waqf wa ibtidā* defining intelligible and naturalistic phrasing *within* and *between* verses (Denny, 1989). We have derived a coarse-grained boundary annotation scheme for Arabic (Brierley *et al.*, 2011) from *Tajwīd* stops and starts mark-up in a reputable edition of the Qur'an[1], and in a widely-used recitation style: *ḥafṣ bin 'āṣim* (*cf.* Sharaf, 2004). This uses the *Qurayshi* or Meccan dialect, and, according to a 'strong' *hadīth*, is one of seven original styles of transmission:

> '…The Qur'an has been revealed to be recited in seven different ways, so recite of it that which is easier for you…' (*Sahih al-Bukhari* in Gilchrist, 2011)

Our annotation scheme is coarse-grained because, for our immediate purposes (Sawalha *et al.*, 2012), we have collapsed eight degrees of boundary strength (*i.e.* three major boundary types, four minor boundary types, and one *prohibited* stop) into the familiar `{major, minor, none}` set. Future work will implement the full fine-grained boundary annotation scheme for text analytic investigation and experimentation with an updated version of the corpus. For the present, we note that in addition to its specificity, boundary mark-up in the Qur'an is *prescriptive and proactive* rather than descriptive and reactive, as in existing systems for English. Figure 1 displays Verse 45 from Chapter 29 of the Qur'an (*Al-Ankabūt* or The Spider) in decorative *othmāni* script, followed by the same verse as it appears in our corpus, in MSA script and with `major/minor` boundary mark-up. It also displays a transliteration and an English translation of the text.

We consider MSA script as preferable for speech and language processing, and for boosting the currency of this corpus for the wider research community. An additional novelty is that we use compulsory and recommended, plus prohibited stops in *Tajwīd* mark-up to segment the text into sentences (*cf.* Figure 2). Such 'sentences' may constitute the grammatical units of common parlance but may also be realised as sequences of intonation units or *extended sentences* (Chafe in Croft, 1995) which resemble mainstream sentences in their 'feeling of closure' (Croft, 1995). Novelty aside, our taggers (Sawalha *et al.*, 2012) require sentence segmentation (Bird *et al.*, 2009, p.198), and classifying words (*e.g.* as `breaks` or `non-breaks`) in situ within a sentence is the usual approach to phrase break prediction (Taylor and Black, 1998).

4. Course-Grained Syntactic Annotation

Traditional Arabic grammar (Wright, 1996; Ryding, 2005; Al-Ghalayyni, 2005) classifies words into one of three syntactic categories `{noun, verb, particle}`, and we therefore retain this coarse-grained feature set as the default in our initial experiments (Sawalha *et al.*, 2012). Qur'anic Arabic is fully vowelised, unlike MSA; and this facilitates syntactic analysis via this ostensibly straightforward scheme which, without vowelisation, becomes problematic (Sawalha, 2011b). For example, native Arabic speakers will use context to disambiguate the non-vowelised form ورد ***wrd***, which could either be the *noun* وَرْدٌ ***ward***^{un} (*roses*), or the *verb* وَرَدَ ***warada*** (*to come*). A further problem is the mismatch between descriptive frameworks for Arabic and English (*aka* 'Western') grammar; Arabic nouns subsume adjectives, adverbs, and some prepositions, while particles also subsume some prepositions, as well as conjunctions and negatives (Maamouri *et al.*, 2004). Subsequently, we extend our sparse tagset to differentiate a limited selection of subcategories extracted from fully parsed sections of QAC, the *Qur'anic Arabic Corpus* (Dukes, 2010). Morpho-syntactic analysis in QAC is fine-grained. For example, in an earlier version of the corpus (v.2.0), the word ٱلرَّحِيمِ *ar-raḥīm* in Chapter 1:3 (*the Most Merciful*) is tagged as follows (*cf.* Figure 3).

An explanation of this tagging scheme can be found in Dukes and Habash (2010). However, items in bold in Fig. 3 indicate that each *token* carries an over-arching PoS tag derived from the *stem* of the word. Thus the token ٱلرَّحِيمِ in this verse is an *adjective*. QAC defines 10 major syntactic categories: `{nouns; pronouns; nominals; adverbs; verbs; prepositions; 'lām prefixes; conjunctions; particles; disconnected letters}`. We therefore tag each token via the QAC PoS schema, plus the tripartite notation of traditional Arabic grammar: `{noun, verb, particle}`.

[1] http://tanzil.net/download

وَلَا يَحْزُنكَ قَوْلُهُمْ ۘ إِنَّ ٱلْعِزَّةَ لِلَّهِ جَمِيعًا ۚ هُوَ ٱلسَّمِيعُ ٱلْعَلِيمُ ﴿٦٥﴾	ٱتْلُ مَآ أُوحِىَ إِلَيْكَ مِنَ ٱلْكِتَٰبِ وَأَقِمِ ٱلصَّلَوٰةَ ۖ إِنَّ ٱلصَّلَوٰةَ تَنْهَىٰ عَنِ ٱلْفَحْشَآءِ وَٱلْمُنكَرِ ۗ وَلَذِكْرُ ٱللَّهِ أَكْبَرُ ۗ وَٱللَّهُ يَعْلَمُ مَا تَصْنَعُونَ ﴿٤٥﴾	فَوَيْلٌ لِّلْمُصَلِّينَ ۙ ﴿٤﴾ ٱلَّذِينَ هُمْ عَن صَلَاتِهِمْ سَاهُونَ ﴿٥﴾
وَلَا يَحْزُنْكَ قَوْلُهُمْ || إِنَّ الْعِزَّةَ لِلَّهِ جَمِيعًا || هُوَ السَّمِيعُ الْعَلِيمُ ||	اتْلُ مَا أُوحِيَ إِلَيْكَ مِنَ الْكِتَابِ وَأَقِمِ الصَّلَاةَ | إِنَّ الصَّلَاةَ تَنْهَى عَنِ الْفَحْشَاءِ وَالْمُنْكَرِ | وَلَذِكْرُ اللَّهِ أَكْبَرُ | وَاللَّهُ يَعْلَمُ مَا تَصْنَعُونَ ||	فَوَيْلٌ لِلْمُصَلِّينَ الَّذِينَ هُمْ عَنْ صَلَاتِهِمْ سَاهُونَ ||
walā yaḥzunka qawluhum || *inna al-ʿiza^{ta} lillāhi ǧamī^{ʿan}* || *huwa as-samīʿu al-ʿalīmu* ||	*'utlu mā ūḥiya 'ilayka mina al-kitābi wa 'aqimi aṣ-ṣala^{ta}* | *inna aṣ-ṣala^{ta} tanhā ʿani al-faḥshā'i wa al-munkari* |*walaḏikru allāhi 'akbaru* | *wa allāhu yaʿlamu mā taṣnaʿūna* ||	*fawayl^{un} lilmuṣallīna al-laḏīna hum ʿan ṣalātihim sāhūna* ||
And let not their speech grieve you. Indeed, honor [due to power] belongs to Allah entirely. He is the Hearing, the Knowing.	Recite, [O Muhammad], what has been revealed to you of the Book and establish prayer. Indeed, prayer prohibits immorality and wrongdoing, and the remembrance of Allah is greater. And Allah knows that which you do.	So woe to those who pray, [But] who are heedless of their prayer –

Figure 1: Original boundary annotations in Qu'ranic verses (top row) mapped to major/minor boundary symbols as in SEC (second row), plus transliteration and translation views of the text (third and fourth rows)

﴿٦٥﴾	ۘ	ج	لا
Compulsory break	Compulsory break	Recommended break	Prohibited stop

Figure 2: Compulsory, recommended and prohibited stops in *Tajwīd* mark-up

```
r~aHiymi    Al+ POS:ADJ LEM:r~aHiym ROOT:rHm MS GEN
```

Figure 3: QAC sample of part-of-speech tags for an Arabic word

5. Building the Dataset

To build the Boundary-Annotated Qur'an Corpus we have extracted, processed, and merged information from two online sources: the Tanzil Qur'an project (Zarabi-Zadeh, 2012) and an earlier version of QAC, the Qur'anic Arabic Corpus (Dukes, 2010). A full account of dataset build is intended for a future publication, but outline processing steps involved: (i) gathering and tracking boundary stops from Tanzil; (ii) extracting PoS tags from QAC; (iii) collapsing boundary stops into two alternative coarse-grained schema; (iv) collapsing PoS tags into two alternative coarse-grained schema; (v) merging these two data streams; (vi) segmenting long paragraphs into sentences.

The constructed boundary annotated corpus of 77430 words and 8230 sentences is stored in a tab separated column file, with each word also stored in a separate file (*cf.* Figure 4). The first four columns contain tracking information, including Sura (*i.e.* chapter) number, and Aya (*i.e.* verse) number, (the first two columns). The Arabic word in Othmani and then MSA script occupy the fifth and sixth columns respectively. Part-of-speech information is given in the next two columns, with tripartite coarse-grained tags in column seven, and more detailed syntactic annotation in column eight. Column nine stores the Tajwīd boundary symbol (if present); and the next two columns show each word classified in terms of boundary type: boundary types stored as **`{major, minor, none}`**, and then as **`{breaks, non-breaks}`**. The penultimate column identifies sentence terminals, and the last column gives the word-for-word English translation.

5.1 Preparing the Dataset as Input to the Tagger

Both taggers used in our experiments take input text segmented into sentences. Since we have classified compulsory and recommended stops in recitation mark-up as major breaks, these are used to identify sentence terminals. Then for our series of experiments, we prepare different permutations of the data to include/exclude words mapped to coarse and slightly finer-grained PoS and either two or three boundary classes. Figure 5 shows sample training input to the tagger as nested lists of tuples with: (i) Arabic words as they appear encoded in UTF-8; and (ii) a facsimile view of Arabic words in their normal orthographic form.

1	1	1	1	بِسْمِ	بِسْمِ	N	NOUN	-	-	non-break	-	in-(the)-name
1	1	1	2	ٱللَّهِ	اللَّهِ	N	NOUN	-	-	non-break	-	(of)-allah
1	1	1	3	ٱلرَّحْمَٰنِ	الرَّحْمَٰنِ	N	NOMINAL	-	-	non-break	-	the-most-gracious
1	1	1	4	ٱلرَّحِيمِ	الرَّحِيمِ	N	NOMINAL	۞	||	break	terminal	the-most-merciful
1	2	1	1	ٱلْحَمْدُ	الْحَمْدُ	N	NOUN	-	-	non-break	-	all-praises-and-thanks
1	2	1	2	لِلَّهِ	لِلَّهِ	N	NOUN	-	-	non-break	-	(be)-to-allah
1	2	1	3	رَبِّ	رَبِّ	N	NOUN	-	-	non-break	-	the-lord
1	2	1	4	ٱلْعَٰلَمِينَ	الْعَالَمِينَ	N	NOUN	۞	||	break	terminal	of-the-universe
1	3	1	1	ٱلرَّحْمَٰنِ	الرَّحْمَٰنِ	N	NOMINAL	-	-	non-break	-	the-most-gracious
1	3	1	2	ٱلرَّحِيمِ	الرَّحِيمِ	N	NOMINAL	۞	||	break	terminal	the-most-merciful
1	4	1	1	مَٰلِكِ	مَالِكِ	N	NOUN	-	-	non-break	-	(the)-master
1	4	1	2	يَوْمِ	يَوْمِ	N	NOUN	-	-	non-break	-	(of-the)-day
1	4	1	3	ٱلدِّينِ	الدِّينِ	N	NOUN	۞	||	break	terminal	(of-the)-judgment
1	5	1	1	إِيَّاكَ	إِيَّاكَ	N	PRONOUN	-	-	non-break	-	you-alone
1	5	1	2	نَعْبُدُ	نَعْبُدُ	V	VERB	-	-	non-break	-	we-worship
1	5	1	3	وَإِيَّاكَ	وَإِيَّاكَ	N	PRONOUN	-	-	non-break	-	and-you-alone
1	5	1	4	نَسْتَعِينُ	نَسْتَعِينُ	V	VERB	۞	||	break	terminal	we-ask-for-help
1	6	1	1	ٱهْدِنَا	اهْدِنَا	V	VERB	-	-	non-break	-	guide-us
1	6	1	2	ٱلصِّرَٰطَ	الصِّرَاطَ	N	NOUN	-	-	non-break	-	(to)-the-path
1	6	1	3	ٱلْمُسْتَقِيمَ	الْمُسْتَقِيمَ	N	NOMINAL	۞	||	break	terminal	the-straight

Figure 4: Sample of the tab separated column file for our boundary-annotated Arabic corpus

```
[((u'\u0630\u064e\u0644\u0650\u0643\u064e',             u'N'),             u'non-break'),
((u'\u0627\u0644\u0652\u0643\u0650\u062a\u064e\u0627\u0628\u064f',  u'N'),  u'non-break'),
((u'\u0644\u064e\u0627',  u'P'),  u'non-break'),  ((u'\u0631\u064e\u064a\u0652\u0628\u064e',
u'N'),     u'break'),     ((u'\u0641\u0650\u064a\u0647\u0650',     u'P'),     u'non-break'),
((u'\u0647\u064f\u062f\u064b\u0649',                  u'N'),                  u'non-break'),
((u'\u0644\u0650\u0644\u0652\u0645\u064f\u062a\u0651\u064e\u0642\u0650\u064a\u0646\u064e',
u'N'), u'break')]
```

```
[((ذَٰلِكَ , N), non-break), ((ٱلْكِتَٰبُ , N), non-break), ((لَا , P), non-break), ((رَيْبَ , N), break),
((فِيهِ , P), non-break), ((هُدًى , N), non-break), ((لِّلْمُتَّقِينَ , N), break)]
```

Figure 5: A single Qur'anic "sentence" as training input to the tagger: words are PoS-tagged via the set of {N, V, P} for binary classification

6. Scheme Ratification on Modern Standard Arabic

In a related LREC submission (Sawalha *et al.*, 2012), we construe our boundary-annotated and PoS-tagged Qur'an as a 'gold standard' for supervised learning of the phrase break prediction task. The Qur'an is a rich dataset, despite its size, and has previously been used as an evaluative 'gold-standard' for machine learning (*e.g.* for Arabic morphological analysers in Morpho Challenge 2009)[2]. The general procedure is to train the classifier on a substantive sample of 'gold-standard' boundary-annotated text, and to hold out a smaller sample from the same source for testing. Although target boundary sites in the test set are available to the researcher for comparative evaluation, they are missing from test data presented to the classifier. Classifier accuracy therefore equates to the number of correct boundaries retrieved during test.

6.1 Evaluation Metrics

Accuracy is a standard metric in machine learning. However, in the case of phrase break prediction, success rate (*i.e.* the number of breaks and non-breaks correctly

[2] http://research.ics.tkk.fi/events/morphochallenge2009/datasets.shtml

predicted) needs to be complemented with other metrics since the baseline set by the majority class (*i.e.* the skewed proliferation of non-breaks in the data) is guaranteed to be challenging (Brierley, 2011). The Information Retrieval metrics of *precision*, *recall*, and *f-score* have previously been used (*e.g.* Koehn *et al.*, 2000; Atterer and Klein, 2002). In our related LREC 2012 submission (Sawalha *et al.*, 2012), we adopt Balanced Classification Rate given by the formula:

$$0.5 * \left(\frac{\text{TP}}{\text{total positive instances}} + \frac{\text{TN}}{\text{total negative instances}}\right)$$

This places equal emphasis on model capture of true positives as well as true negatives.

وَلَا يَحْزُنكَ قَوْلُهُمْ ◌ۘ إِنَّ ٱلْعِزَّةَ لِلَّهِ جَمِيعًا ◌ۚ هُوَ ٱلسَّمِيعُ ٱلْعَلِيمُ
Let not their speech grieve thee: for all power and honour belong to Allah: it is He Who heareth and knoweth (all things).

Figure 6: Arabic chunk boundary symbols mirrored by punctuation in the corresponding English translation

6.2 Delimiting Sentences in the MSA Corpus

Our MSA corpus replicates our Qur'an dataset classification of each word in terms of two levels of syntactic plus prosodic information. For the latter, "sentences" within longer paragraphs are readily identified via major breaks as sentence terminals, whereas for MSA text we segment on punctuation.

Working with MSA text is not straightforward. First, it is not fully vowelised, and restoring full vowelisation is an essential preliminary step to morphological analysis, POS-tagging and parsing. In our "gold-standard" excerpt[3] from the Corpus of Contemporary Arabic (Al-Sulaiti, 2006), full vowelisation has been restored automatically by the SALMA Tagger (Sawalha, 2011a; Sawalha, 2011b). Another problem is that sentences in Arabic can be very long, and punctuation is sparse at best. For this study, sentence segmentation was done manually. A longer term goal is to develop reliable chunking algorithms for Arabic such that MSA text can be chunked automatically and extra intelligible and naturalistic boundaries inserted which meet with human approval.

6.3 Long-term Goals

Our over-arching research objectives are: (i) to determine whether Qur'anic Arabic speech rhythms still inform native speaker intuitions, and parsing and phrasing strategies, for Modern Standard Arabic; and (ii) to analyse and leverage prosodic-syntactic boundary correlates in the Qur'an for Arabic speech and language applications. This will eventually entail use of subjective human judgment to scrutinise output predictions from our best-performing tagger which is first evaluated on the boundary-annotated Qur'an (*ibid*), and then tested on unseen 'gold standard' PoS-tagged MSA text[4].

We take a novel approach to phrase break prediction for Arabic, deriving our prosodic annotation scheme from *Tajwīd* (recitation) mark-up in the Qur'an; as previously stated (§3), this prescribes intelligible and naturalistic phrasing *within* and *between* verses (Denny, 1989). For example, in Figure 6 compulsory and highly recommended verse-medial breaks in Chapter 10.65 chunk the text into meaningful units which are retained via punctuation in Yusuf Ali's acclaimed English translation (2000).

Our original insight is then to view the Tajwīd system of chunk boundary delimiters, and other features extracted from the orthographic form (Islamic Bulletin, 2012) as additional sources of text-based data for computational analysis. Text analytics techniques honed on English (Brierley, 2011) will then be used to discover significant linguistic patterns in the vicinity of these benchmark phrase break annotations, to be evaluated as classificatory features in machine learning experiments. The best-performing feature set will then be evaluated on and adapted for Modern Standard Arabic.

7. Conclusion

The Qur'an may already be construed as a multimodal dataset by virtue of its prosodic (*Tajwīd*) annotations and *salutation* marks {۩}. We have compiled a unique, open-source, boundary-annotated Qur'an for Arabic speech and language processing, which we utilise immediately in phrase break prediction experiments for Classical and Modern Standard Arabic (Sawalha *et al.*, 2012). Our longer term aim is to unravel the linguistic wisdom of prosodic-syntactic chunking inherent in recitation mark-up in the Qur'an, and to leverage this knowledge in Arabic Natural Language Engineering applications. We therefore plan to enrich our dataset with: (i) very fine-grained morpho-syntactic analyses using the SALMA tagger (Sawalha and Atwell, 2010); (ii) more fine-grained boundary annotations; and (iii) projected prosody (Brierley, 2011; Brierley and Atwell, 2010), potentially as part of an ongoing project (Atwell *et al.*, 2011).

8. References

Ali, A.Y. 2000. *The Holy Quran* (translated by Abdullah Yusuf Ali). Hertfordshire. Wordsworth Editions Ltd.

Al-'ashmuni, Ahmad bin Muhammad Abdul-Kareem.

[3] http://www.comp.leeds.ac.uk/cgi-bin/scmss/cca_gs_color_coded.py

[4] http://www.comp.leeds.ac.uk/sawalha/goldstandard.html

1973. منار الهدى في بيان الوقف والإبتدا، ومعه المقصد لتلخيص ما في المرشد في الوقف والابتداء manar al-huda fi bayan al-waqf wa al-'ibtida' Mustafa Al-baabi Al-halabi.

Al-Ghalayyni. 2005. جامع الدروس العربية "Jami' Al-Duroos Al-Arabia" Saida - Lebanon: Al-Maktaba Al-Asriyiah "المكتبة العصرية".

Al-Sulaiti, L., Atwell, E. 2006. ‘The design of a corpus of contemporary Arabic.’ *In International Journal of Corpus Linguistics*. Vol. 11, pp. 135-171.

Ananthakrishnan, S. and Narayanan, S.S. 2008. ‘Automatic Prosodic Event Detection Using Acoustic, Lexical, and Syntactic Evidence.’ *IEEE Transactions on Audio, Speech, and Language Processing, TASLP 2008*. Vol. 16.1: 216-228.

Atwell, E., Brierley, C., Dukes, K., Sawalha, M. and Sharaf, A.M. 2011. ‘An Artificial Intelligence Approach to Arabic and Islamic Content on the Internet.’ *National Information Technology Symposium (NITS)*. Riyadh, Saudi Arabia.

Beckman, M. and Hirschberg, J. 1994. *The ToBI annotation conventions*. The Ohio State University and AT&T Bell Laboratories, unpublished manuscript. Online. Accessed September 2011. ftp://ftp.ling.ohio-state.edu/pub/phonetics/TOBI/ToBI/ToBI.6.html.

Bird, S., Klein, E. and Loper, E. 2009. *Natural Language Processing with Python.* Sebastopol, CA. O’Reilly Media, Inc.

Brierley, C. 2011. Prosody Resources and Symbolic Prosodic Feartures for Automated Phrase Break Prediction. PhD Thesis. School of Computing. University of Leeds.

Brierley, C. and Atwell, E. 2010. ‘ProPOSEC: a Prosody and PoS Spoken English Corpus.’ In *Proceedings of LREC'10: Language Resources and Evaluation Conference*, Valetta, Malta. May 2010.

Brierley, C. and Atwell, E. ‘ProPOSEL: A prosody and POS English lexicon for language engineering.’ In *Proceedings of LREC'08: Language Resources and Evaluation Conference*, Marrakech, Morocco. May 2008.

Croft, W. 1995. ‘Intonation Units and Grammatical Structure.’ *Linguistics*. 33: 839-882

Denny, F.M. 1976. Review [untitled]. *Journal for the Scientific Study of Religion*. 15.3: 287-289

Denny, F.M. 1989. ‘Qur’an Recitation: A Tradition of Oral Performance and Transmission.’ *Oral Tradition*. 4/1-2: 5-26

Dukes, K. 2011. *The Qur’anic Arabic Corpus*. Online. Accessed: August 2011. http://corpus.quran.com

Dukes, K. and Habash, N. 2010. ‘Morphological Annotation of Qur’anic Arabic.’ In *Proceedings of Language Resources and Evaluation Conference (LREC 2010)*, Valletta, Malta.

Gilchrist, J. 2011. ‘Jam’ Al-Qur’an: The Codification of the Qur’an Text.’ Online. Accessed September 2011. http://www.answering-islam.org/Gilchrist/Jam/index.html

Grabe, E. 2001. ‘Prosodic Annotation.’ PowerPoint. *9th ELSNET European Summer School on Language and Speech Communication, Prague*. Accessed: 2006.

Islamic Bulletin. 2012. The Holy Quran Color Coded with Tajweed Rules. Online. Accessed: Feb. 2012. http://www.islamicbulletin.com/services/details.aspx?id=260

Ladd, R. 1996. *Intonational Phonology* Cambridge. Cambridge University Press.

Liberman, M.Y. and Church, K.W. 1992. ‘Text Analysis and Word Pronunciation in Text-to-Speech Synthesis.’ In *Advances in Speech Signal Processing*. Furui S. and Sondhi, M.M. (eds.). New York. Marcel Dekker Inc.

Maamouri, M., Bies, A., Buckwalter, T. and Mekki, W. 2004. *The Penn Arabic Treebank: Building a Large-Scale Annotated Corpus*. Philadelphia. Linguistic Data Consortium.

Ostendorf, M., Price, P. and Shattuck-Hufnagel, S. 1996. *Boston University Radio Speech Corpus*. Philadelphia. Linguistic Data Consortium.

Roach, P. 2000. *English Phonetics and Phonology: A Practical Course* (3rd. edition). Cambridge. Cambridge University Press

Ryding, Karin C. 2005. *A Reference Grammar of Modern Standard Arabic*. Cambridge. Cambridge University Press.

Sawalha, M., Brierley, C., and Atwell, E. 2012. ‘Predicting Phrase Breaks in Classical and Modern Standard Arabic Text.’ In *Proceedings of LREC 2012: Language Resources and Evaluation Conference*. Istanbul, Turkey. May 2012.

Sawalha, M. 2011a. *The SALMA – Gold Standard*. Accessed: September 2011 http://www.comp.leeds.ac.uk/sawalha/goldstandard.html

Sawalha, Majdi. 2011b. Open-source Resources and Standards for Arabic Word Structure Analysis. Leeds: University of Leeds PhD.

Sawalha, M. and Atwell, E. 2010. ‘Fine-Grain Morphological Analyzer and Part-of-Speech Tagger for Arabic Text.’ In *Proceedings of LREC'10: Language Resources and Evaluation Conference*, Valetta, Malta. May 2010.

Sharaf, Jamal Ad-Deen Muhammad. 2004. مصحف الصحابة في القراءات العشر المتواترة من طريق الشاطبية والدرة mushaf as-sahabah fi al-qira'at al-'ashr al-mutawatirah min tariq ash-shatibyyah wa al-durrah Tanta: Dar As-Sahaba lil-Turath.

Taylor, P. and Black, A.W. 1998. ‘Assigning Phrase-Breaks from Part-of-Speech Sequences.’ In *Computer Speech and Language*. 12.2: 99-117.

Taylor, L.J. and Knowles, G. 1988. ‘Manual of Information to Accompany the SEC Corpus: The machine readable corpus of spoken English.’ Accessed: January 2010. http://khnt.hit.uib.no/icame/manuals/sec/INDEX.HTM

Wright, W. 1996. A Grammar of the Arabic Language, Translated from the German of Caspari, and Editted with Numerous Additions and Corrections Beirut: Librairie du Liban.