Journal of Speech Sciences 2(2):175-191.2012.
Available at: http://www.journalofspeechsciences.org

Prosody Prediction for Arabic via the Open-Source Boundary-Annotated Qur'an Corpus

SAWALHA, M.S.[1]*, BRIERLEY, C.[2], ATWELL, E.[2]

[1] University of Jordan

[2] University of Leeds

Abstract

A phrase break classifier is needed to predict natural prosodic pauses in text to be read out loud by humans or machines. To develop phrase break classifiers, we need a boundary-annotated and part-of-speech tagged corpus. Boundary annotations in English speech corpora are descriptive, delimiting intonation units perceived by the listener; manual annotation must be done by an expert linguist. For Arabic, there are no existing suitable resources. We take a novel approach to phrase break prediction for Arabic, deriving our prosodic annotation scheme from Tajwid (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. We then use this dataset to train, test, and compare two probabilistic taggers (trigram and HMM) for Arabic phrase break prediction, where the task is to predict boundary locations in an unseen test set stripped of boundary annotations by classifying words as breaks or non-breaks. The preponderance of non-breaks in the training data sets a challenging baseline success rate: 85.56%. However, we achieve significant gains in accuracy with a trigram tagger, and significant gains in performance recognition of minority class instances with both taggers via the Balanced Classification Rate metric. This is initial work on a long-term research project to produce annotation schemes, language resources, algorithms, and applications for Classical and Modern Standard Arabic.

Keywords: phrase break prediction, prosodic annotation, Tajwid recitation, N-gram and HMM taggers, boundary-annotated and PoS-tagged Qur'an

*Corresponding author: sawalha.majdi@gmail.com

1. Introduction

An accepted Universal of language is that people process speech (and text) in chunks (1), which in turn can be interpreted *syntactically* as function word groups (2) and *prosodically* as tone units (3, 4). A phrase break classifier is needed to predict natural chunks in text to be read out loud by humans or machines. 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. To develop phrase break classifiers, we need a boundary-annotated and part-of-speech tagged corpus. Boundary annotations in English speech corpora are descriptive, delimiting intonation units perceived by the listener; manual annotation must be done by an expert English linguist. Our research applies techniques honed on English (5) to another stress-timed language, Arabic, and to the entire text of the Qur'an (§4). For Modern Arabic, there are no existing suitable resources with prosodic phrase boundaries annotated by Arabic linguistics experts. However, the Qur'an can be used as a reputable "gold standard" for phrasing in Arabic, because traditional editions include boundary mark-up to aid correct recitation, based on long-established traditions of Quranic Arabic linguistics developed to help believers read and understand the Quran. We can harness the recitation markup in traditional Quran editions, to use these as phrase-break markup in a Boundary-Annotated Quran Corpus.

Chunking text via automatic assignment of sentence-medial and sentence-terminal prosodic-syntactic boundaries is a Natural Language Processing (NLP) and machine learning task which attempts to simulate human parsing and phrasing strategies. The latter are represented by "gold standard" boundary annotations in a speech corpus. Phrase break classifiers are typically trained and tested on such datasets, and assume prior sentence segmentation and part-of-speech (PoS) tagging for input text. Here, we utilize our boundary-annotated *Qur'an* corpus of Classical Arabic (6) to develop and evaluate two probabilistic taggers (n-gram and HMM) for the phrase break prediction task, using two different feature sets. We regard the Qur'an as a reputable 'gold standard' for phrasing in Arabic because *recitation* is integral to this text, and many editions (§4) already carry prescriptive boundary mark-up representative of the long-established traditions of Arabic linguistics. Hence we plan to assess the naturalness and intelligibility of outputs from our best-performing tagger over a sample of Modern Standard Arabic (MSA) text (6).

2. Phrase Break Prediction

Automated phrase break prediction is a natural language processing (NLP) task within the Text-to-Speech (TTS) synthesis pipeline, and sub-divides input sentences into meaningful chunks to copy the way in which a native speaker might parse or phrase the utterance. This equates to classifying junctures between words, or the words themselves, in terms of a finite set of boundary types, for example **`breaks`** or **`non-breaks`**. Establishing these delimiters is an essential component of the symbolic linguistic representation of text as output to a speech synthesizer.

2.1. General Procedure for Phrase Break Prediction

Phrase break prediction assumes prior sentence segmentation and part-of-speech tagging for input text, and therefore punctuation and syntax are traditionally used as classificatory features. Another prerequisite is a boundary-annotated and part-of-speech (PoS) tagged corpus (6) as 'gold standard' for developing phrase break classifiers. The classifier is trained on a substantive sample of 'gold-standard' boundary-annotated text, and tested on a smaller, unseen sample from the same source *minus* the boundary annotations.

2.2. Machine Learning Approaches to Phrase Break Prediction

There are two generic approaches to machine learning: rule-based or probabilistic. Phrase break models exemplifying these two approaches are: (i) Liberman and Church's *chinks 'n' chunks* algorithm (1992) (*cf.*2); and (ii) Taylor and Black's Markov model (1998) (*cf.*7) used in Edinburgh's *Festival*[1] Speech Synthesis system. In the former, chinks are closed-class function words, while chunks are open-class content words; the algorithm inserts a phrase break at every punctuation mark, and whenever a content word is immediately followed by a function word. Taylor and Black's statistical model conditions the probability of juncture type (*i.e.* $P(j_i)$ in Equation 1) on: (i) the *prior probability* of each class given the immediate context (*i.e.* the PoS trigram in which that juncture is embedded or $P(C_i \mid j_i)$ in Equation 1); and (ii) the *likelihood* of each class given the previous sequence of N juncture types, where in this case, $N = 6$ (Equation 1).

$$P(j_i) \propto P(j_i \mid J_{i-1}^{N}).P(C_i \mid j_i) \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \text{(1)}$$

2.3. Metrics for Phrase Break Prediction

Performance is primarily evaluated in terms of *accuracy*, namely: the number of correct predictions – or the sum of *true positives* and *true negatives* (TP + TN) – made during test. There are also other relevant metrics such as *f-score* and balanced classification rate (BCR). The former is the trade-off between, or weighted mean, of *recall* (*i.e.* TP total / total number of boundaries in the sample), and *precision* (*i.e.* TP total / total number of boundaries retrieved). The latter (*i.e.* BCR) mitigates against high accuracy scores arising from class imbalance, a typical scenario for phrase break prediction since instances of the majority

[1] http://www.cstr.ed.ac.uk/projects/festival/

class **`(non-breaks)`** greatly outnumber minority class instances **`(breaks)`** in the corpus. BCR is computed as the average of *breaks-correct* and *non-breaks-correct* and thus considers relative class distributions (Equation 2).

$$BCR = 0.5 * \left(\frac{\text{TP}}{\text{total true positives}} + \frac{\text{TN}}{\text{total true negatives}}\right) \qquad \ldots\ldots\ldots\ldots\ldots\ldots..(2)$$

3. Boundary Annotation Schemes for English

The Lancaster/IBM Spoken English Corpus or SEC (8) 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 (4). 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 (3). 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 (9) use ToBI or the ***To****nes and* ***B****reak* ***I****ndices* annotation scheme (10) which identifies *five* theoretical levels of juncture between words: **`{0,1,2,3,4}`**. Break index **`{0}`**denotes no separation or *cliticization*(11), 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 (11). 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. 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) 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.

4. Building the Open-Source Boundary Annotated Qur’an Corpus

We derive a coarse-grained boundary annotation scheme for Arabic from traditional recitation mark-up (*Tajwid*) in the Qur’an; this is then compared with existing schemes for British and American English speech corpora (8, 10). We then merge a PoS-tagged version of the text (12) 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 *Tajwid* mark-up (*cf.* 13) to segment the text into 8230 sentences.

A prerequisite for developing and evaluating phrase break classifiers is a “gold standard” boundary-annotated and PoS-tagged corpus. We regard the Qur’an as a reputable “gold standard” for phrasing in Arabic because recitation is integral to this text, and many editions already carry prescriptive boundary mark-up representative of the long-established traditions of Arabic linguistics.

4.1. 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…’ (14).

The art of *Tajwid* 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 (15). We have derived a coarse-grained boundary annotation scheme for Arabic (16) from*Tajwid*stops and starts mark-up in a reputable edition of the Qur’an[2], and in a widely-used recitation style: *ḥafṣ bin ‘āṣim*(*cf.*17). 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 (18))

Our annotation scheme is coarse-grained because, for our immediate purposes (19), we have

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

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 (figure 1). 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 2 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.

Arabic	**BAC**	**Explanation of Tajwid Symbol**
﴿٦٥﴾	||	End of verse is a compulsory major break.
ـۘ	||	Major and compulsory verse-medial break which completes the meaning of a phrase.
ـۚ	||	Minor break: a break is allowed and preferable.
ـۗ	|	Minor break (continuation mark): the reader can continue without pausing, but a pause is preferable.
ـۖ	|	Minor break permitted: readers can pause if they wish, but it is better not to.
ـۜ	|	Minor break for a shorter time without breathing, where last pronounced letter before break is pronounced without its short vowel.
ـۛ ـۛ	|	Alternative boundaries in the same phrase: if the reader breaks in one position, they must not break in the other, and vice versa.
ـۙ	**non-break**	Non-break: pausing is not permitted as it would change the meaning of the verse.

Figure 1: Mapping from Tajwid symbols to coarse-grained tripartite boundary annotation scheme for Arabic. The majority of words do not carry Tajwid boundary markup and these are thus tagged as non-breaks in our corpus

وَلَا يَحۡزُنكَ قَوۡلُهُمۡۘ إِنَّ ٱلۡعِزَّةَ لِلَّهِ جَمِيعًاۚ هُوَ ٱلسَّمِيعُ ٱلۡعَلِيمُ ٦٥	ٱتۡلُ مَآ أُوحِيَ إِلَيۡكَ مِنَ ٱلۡكِتَٰبِ وَأَقِمِ ٱلصَّلَوٰةَۖ إِنَّ ٱلصَّلَوٰةَ تَنۡهَىٰ عَنِ ٱلۡفَحۡشَآءِ وَٱلۡمُنكَرِۗ وَلَذِكۡرُ ٱللَّهِ أَكۡبَرُۗ وَٱللَّهُ يَعۡلَمُ مَا تَصۡنَعُونَ ٤٥	فَوَيۡلٞ لِّلۡمُصَلِّينَ ٤ ٱلَّذِينَ هُمۡ عَن صَلَاتِهِمۡ 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 ğamīʿ^{an}* ‖ *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* | *waallāhu 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 wrong doing, 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 2: 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)

An additional novelty is that we use compulsory and recommended, plus prohibited stops in *Tajwid*mark-up to segment the text into sentences (*cf.* Figure 3). Such ‘sentences’ may constitute the grammatical units of common parlance but may also be realised as sequences of intonation units or *extended sentences*(3) which resemble mainstream sentences in their ‘feeling of closure’ (3). Novelty aside, our taggers (20) require sentence segmentation (20 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 (7).

﴿٦٥﴾	ۘ	ۚ	ۙ
Compulsory break	Compulsory break	Recommended break	Prohibited stop

Figure 3: Compulsory, recommended and prohibited stops in *Tajwid*mark-up

4.2. Course-Grained Syntactic Annotation

Traditional Arabic grammar (21-23) 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 (19). Qur’anic Arabic is fully vowelised, unlike MSA; and this facilitates syntactic analysis via this ostensibly straightforward scheme which, without vowelisation, becomes problematic (24). 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 (25). Subsequently, we extend our sparse tag set to differentiate a limited selection of subcategories extracted from fully parsed sections of QAC, the *Qur’anic Arabic Corpus*[3](12). Morpho-syntactic analysis in QAC is fine-grained. For example, in an earlier version of the corpus (v.2.0), the word الرَّحِيمِ *ar-raḥīm* in Chapter 1:3 (*the Most Merciful*) is tagged as follows (*cf.* Figure 4).

```
r~aHiymi  Al+ POS:ADJLEM:r~aHiymROOT:rHm MS GEN
```

Figure 4: QAC sample of part-of-speech tags for an Arabic word

An explanation of this tagging scheme can be found in (26). However, items in bold in Figure4 indicate that each *token* carries an over-arching PoS tag derived from the *stem* of the word. Thus the token الرَّحِيمِin 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 extract this information from QAC to tag each token with its main part-of-speech; we also map these categories to the tripartite notation of traditional Arabic grammar: **{noun, verb, particle}**.

4.3. 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 (27) and an earlier version of QAC, the Qur’anic Arabic Corpus (12). 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 5). 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

[3] http://corpus.quran.com/

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 Tajwid 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.

1	1	1	1	بِسۡمِ	بِسْمِ	N	NOUN	-	-	non-break	-	in-(the)-name
1	1	1	2	ٱللَّهِ	اللَّهِ	N	NOUN	-	-	non-break	-	(of)-allah
1	1	1	3	ٱلرَّحۡمَٰنِ	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	ٱلۡحَمۡدُ	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	ٱلۡعَٰلَمِينَ	الْعَالَمِينَ	N	NOUN	۝	||	break	terminal	of-the-universe
1	3	1	1	ٱلرَّحۡمَٰنِ	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	يَوۡمِ	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	نَعۡبُدُ	نَعْبُدُ	V	VERB	-	-	non-break	-	we-worship
1	5	1	3	وَإِيَّاكَ	وَإِيَّاكَ	N	PRONOUN	-	-	non-break	-	and-you-alone
1	5	1	4	نَسۡتَعِينُ	نَسْتَعِينُ	V	VERB	۝	||	break	terminal	we-ask-for-help
1	6	1	1	ٱهۡدِنَا	اهْدِنَا	V	VERB	-	-	non-break	-	guide-us
1	6	1	2	ٱلصِّرَٰطَ	الصِّرَاطَ	N	NOUN	-	-	non-break	-	(to)-the-path
1	6	1	3	ٱلۡمُسۡتَقِيمَ	الْمُسْتَقِيمَ	N	NOMINAL	۝	||	break	terminal	the-straight

Figure 5: Sample of the tab separated column file for our boundary-annotated Arabic corpus

5. Taggers

We implement a trigram tagger based on the ***N****atural* ***L****anguage* ***T****ool****K****it*'s (20) `Ngram Tagger` class to assign boundaries to a corpus of Qur'anic Arabic which is segmented into sentences and PoS-tagged, and where outputs from the tagger can be evaluated against 'gold standard' boundary annotations in the dataset (6). We also implement an HMM or sequence model based on NLTK's `HiddenMarkovModelTagger` class. Input to the tagger is the same in both cases: our purpose-built Qur'an dataset (6) is segmented into 8230 sentence tokens, and each sentence token is represented as a list of tuples from which we specify permutations of features that match our research questions. A sample *Qur'anic sentence* is given in Figure 6.

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 6 shows sample training input to the tagger as nested lists of tuples.

```
[((ذَلِكَ, N), non-break), ((الْكِتَابُ , N), non-break), ((لَا, P), non-break),
((رَيْبَ, N), break), ((فِيهِ, P), non-break), ((هُدًى, N), non-break),
((لِلْمُتَّقِينَ , N), break)]
```

Figure 6: 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

5.1. The Trigram Tagger

Our trigram tagger is coded in Python and trained on Qur'an text represented as **`(PoS, boundary-type)`** or **`((word, PoS), boundary-type)`** pairings. For the former, it assigns the most likely boundary type (*e.g.* **`break`** or **`non-break`**) based on the current PoS, plus the two preceding boundary types as context. Figure 7 is an adaptation from (20 p.204): shaded areas denote context, and the target for prediction is *italicised.*

Readers will note that this trigram tagger is based on Python dictionaries: a look-up table is consulted to determine an appropriate tag for each instance; and the tagger backs off to a majority class tagger (*i.e.* tags the instance as `non-break`) if look-up fails.

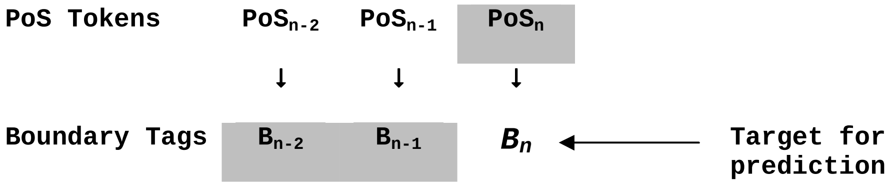

Figure 7: Abstract representation of trigram context used for predicting breaks or non-breaks

5.2. The HMM Tagger

One drawback of this method is that there is no way to revise previously assigned boundaries as the algorithm iterates through the list (*i.e.* the sentence). To resolve this, we also implement NLTK's HMM tagger for comparative evaluation (§6). For these initial experiments, we have simply used the `train()` and `evaluate()` methods with default parameter settings, plus the `train()` method with labeled and unlabeled sequences (*i.e.* training and test set splits), to determine the optimal/most probable combination of break types for each sentence via the Maximum Likelihood Estimate (MLE), which maximizes the joint probability of symbol/state sequences. The HMM tagger generates a probability distribution over all possible boundary types - either **`break`** versus **`non-break`** (the two-class problem), or **`major/minor/non-break`** (the three-class problem). The product of these probabilities then gives a probability score for each boundary sequence, and the highest-scoring sequence is then chosen.

6. Evaluation

The immediate research question pertaining to this study is: Can we successfully recapture prosodic boundaries authenticated by Tajwid recitation markup using probabilistic taggers trained and tested on our *Boundary-Annotated Qur'an Corpus*?

6.1. Methodology

To address this question, we comparatively evaluate the performance of a trigram tagger and an HMM tagger firstly on our Qur'an dataset, with different permutations of features. The first round of experiments uses tripartite PoS categories **`{noun, verb, particle}`** to predict: (i) **`breaks`** versus **`non-breaks`**; and (ii) boundaries of type: **`major, minor, none`**. The second round uses ten PoS features to resolve both tasks: binary classification, and the 3-class problem. The Qur'an dataset is split into the same partitions for training and test in both cases; the training set comprises 70112 words and 7381 sentences, and the test set comprises 7318 words and 849 sentences. The number of sentences in the test set also equates to the number of major breaks in the test set. Non-breaks in the test set total 6469, and this total sub-divides into 6261 non-breaks and 208 minor breaks for the 3-class problem. These are supervised machine learning experiments that assume the classes are mutually exclusive, such that each Arabic word will be resolved as an instance of one, *and only one*, specific boundary type.

6.2. Test Set Selection

Test set sentences were not randomly selected. There is agreement on the provenance of most Qur'anic verses in terms of whether they originate from the Prophet's period of residence in Mecca or Medina. However, there are 21 (out of 114) chapters where Mecca/Medina verse associations are in doubt (*cf.*28). Meccan and Medinan verses differ stylistically (28), and therefore the 21 disputed chapters were used as our test set, since they constitute a representative sample of both styles and a fair test for a tagger trained on the rest of the corpus.

6.3. Confusion Matrices

Tagger accuracy for each classification task can be expressed as an overall percentage calculated by summing the number of correct predictions for each boundary type, and dividing this total by the total word count (*i.e.* the total number of items to be classified). Output predictions are presented as a confusion matrix where false positives and false negatives (FPs and FNs) are used to infer basic issues in performance. Table 1 is an example of the confusion matrix for the two-class problem, where shaded area counts constitute the proportion of correct predictions (*true positives* and *true negatives*) retrieved during test for our trigram tagger using very coarse-grained PoS. Readers will note that class distributions in the test set are highly skewed: 6261 non-breaks versus 1057 breaks.

Table 1: Example confusion matrix for binary classification with the trigram tagger using (word, PoS) pairings

		Predicted +ve	**Predicted -ve**
Breaks	1057	**380**	677
Non-breaks	6261	167	**6094**

6.4. The Two-Class Problem

Table 2 displays results for binary classification experiments with both taggers, and feature set permutations which include/exclude words PoS-tagged at two different levels of granularity. What is immediately obvious is that data skew (*i.e.* the over-preponderance of non-breaks) sets a high baseline accuracy of 85.56%. Nevertheless, the trigram tagger in Runs 1 and 5 significantly outperforms the baseline for both syntactic feature sets: 88.47% for 3 PoS categories, and 88.44% for 10 PoS categories.

Table 2: Experimental results for binary phrase break classification on the Qur'an test set of 7318 words.

RUN	**TAGGER**	**INCLUDE WORD?**	**NUMBER OF POSTAGS**	**NUMBER OF CLASSES**	**ACCURACY**	**BCR**	**TPs**	**TNs**	**FPs**	**FNs**
Base	Baseline	✓	3 or 10	2	85.56%	0.50	0	6261	0	1057
Base	Baseline		3 or 10	2	85.56%	0.50	0	6261	0	1057
1	Trigram	✓	3	2	88.47%	0.67	380	6094	167	677
2	HMM	✓	3	2	82.63%	0.72	601	5446	815	456
3	Trigram		3	2	85.56%	0.50	0	6261	0	1057
4	HMM		3	2	85.56%	0.50	0	6261	0	1057
5	Trigram	✓	10	2	88.44%	0.66	372	6100	161	685
6	HMM	✓	10	2	82.66%	0.72	600	5449	812	457
7	Trigram		10	2	86.31%	0.55	108	6208	53	949
8	HMM		10	2	86.32%	0.55	114	6203	58	943

Success rate for the HMM tagger is below par, but its superior true positive hit rate (*i.e.* 600 TPs) and BCR statistic of 0.72 suggest that this tagger has learnt the concept better than the others. Brierley (5) recommends consideration of more than one evaluation metric when comparing phrase break classifiers. She also recommends further significance testing to verify apparent gains in accuracy, and to explore conflicting results: in this case, accuracy versus BCR scores for the HMM tagger in Runs 2 and 6. We therefore consider the alternative metric of BCR or Balanced Classification Rate (*cf.* Equation 2) to assess how well each model has learnt the concept. The *trigram* tagger in Run 1 correctly predicts 380 breaks set against the baseline prediction of zero. Hence BCR for Run 1 represents a significant gain in performance. What is additionally interesting is that the *HMM taggers* in Runs 2 and 6 also represent a *statistically* significant gain in performance in terms of BCR even when set against Run 1. This claim is verified by applying McNemar's significance test to compare the performance of both classifiers (29). In this case, the focus for comparison is success in minority class recognition. For example, we assembled counts for concordant and discordant output predictions for the trigram and HMM taggers in Runs 5 and 6 in a 2 x 2 contingency table (Table 3).

Table 3: McNemar's significance test is performed on discordant pairs in shaded areas in table.

	Run 5:breaks	**Run 5:**non-breaks
Run 6:breaks	489	923
Run 6: non-breaks	44	5862

There are 923 instances where the HMM tagger identifies a `break` and the trigram tagger a `non-break`. Similarly, there are 44 instances where the HMM tagger identifies a `non-break` and the trigram tagger a `break`. In all, the HMM tagger predicts 1412 `breaks` (hence 5906 `non-breaks`), and the trigram tagger 533 `breaks` (hence 6785 `non-breaks`). Clearly, these marginal probabilities are different. McNemar's test determines whether differences in the proportions of discordant predictions for the minority class (923 versus 44) are significant. Here, it turns out that they are: the two-tailed p-value is <0.000001, and the odds ratio is 20.98 with a 95% confidence interval[4]. Thus, the while the HMM tagger over predicts, it captures many more TPs and achieves a better average positive hit rate.

6.5. The Three-Class Problem

Table 4 records results for tripartite classification.

[4] Calculated via http://vassarstats.net/propcorr.html

Table 4: Experimental results for tripartite phrase break classification on the Qur'an test set of 7318 words.

RUN	TAGGER	INCLUDE WORD?	NUMBER OF POSTAGS	NUMBER OF CLASSES	ACCURACY	BCR	TPs	TNs	FPs	FNs
Base	Baseline	✓	3 or 10	3	85.56%	0.50	0	6261	0	1057
Base	Baseline		3 or 10	3	85.56%	0.50	0	6261	0	1057
9	Trigram	✓	3	3	88.69%	0.65	333	6157	128	700
10	HMM	✓	3	3	81.46%	0.64	371	5590	821	536
11	Trigram		3	3	85.56%	0.50	0	6261	0	1057
12	HMM		3	3	85.56%	0.50	0	6261	0	1057
13	Trigram	✓	10	3	88.62%	0.65	323	6162	122	711
14	HMM	✓	10	3	81.18%	0.63	361	5580	834	543
15	Trigram		10	3	86.17%	0.54	98	6208	63	949
16	HMM		10	3	86.62%	0.55	117	6222	45	934

Significant gains in both accuracy and BCR over baseline performance were achieved by the trigram tagger for the 3-class problem using both feature sets in Runs 9 and 13: 88.69% and 88.62% respectively. The HMM tagger also achieved significant gains in terms of BCR (Runs 10 and 14), and in one experiment (Run 16), where words were disabled as a feature, improved on baseline success rate, albeit at the expense of BCR.

7. Scheme Ratification on Modern Standard Arabic

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)[5]. 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.

7.1. 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

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

"gold-standard" excerpt[6] from the Corpus of Contemporary Arabic (30), full vowelisation has been restored automatically by the SALMA Tagger (31, 24). 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.

7.2. 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 (6), and then tested on unseen 'gold standard' PoS-tagged MSA text[7].

We take a novel approach to phrase break prediction for Arabic, deriving our prosodic annotation scheme from *Tajwid* (recitation) mark-up in the Qur'an; as previously stated (§4), this prescribes intelligible and naturalistic phrasing *within* and *between* verses (15). For example, in Figure 8 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 Tajwid system of chunk boundary delimiters, and other features extracted from the orthographic form (5, 6, 35) as additional sources of text-based data for computational analysis. Text analytics techniques honed on English (5) 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.

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

8. Conclusion

Our boundary-annotated Qur'an corpus is a unique, open-source dataset for Arabic phrase break prediction and for Arabic speech and language processing in general. Boundary annotations in this corpus differ from similar corpora for English in that they are proactive, not reactive, and provide detailed and

[6] http://www.comp.leeds.ac.uk/cgi-bin/scmss/cca_gs_color_coded.py
[7] http://www.comp.leeds.ac.uk/sawalha/goldstandard.html

corroborated guidance for the reader/speaker on optimal parsing and phrasing strategies for interpreting and conveying meaning. Thus, in the longer term, we are interested in the possibility of leveraging this received wisdom for Modern Standard Arabic language engineering applications. This will entail enriching the dataset with morpho-syntactic analyses via the SALMA tagger (32, 31, 24), and with symbolic prosodic information (5, 33), prior to exploratory text data mining and feature extraction of prospective boundary correlates.

This paper constitutes initial work and compares the performance of sequence models for Qur'anic Arabic phrase break prediction. The trigram and HMM taggers in these experiments are prototypes, and use coarse-grained syntactic features only. Nevertheless, sharable experience and insights of interest to fellow corpus linguists are to be gained from the present implementation and evaluation. As with English (2, 11, 34), syntactic information proves a reliable feature, but what is especially interesting is that our highest accuracy scores have been achieved with a very coarse-grained feature set with a long-established history: the tripartite classification of Arabic words as **{noun, verb, particle}** in traditional Arabic grammar (*cf.*9).

What also emerges is the vexed question of class imbalance, potentially compounded by the problem of sparse data: our Qur'an corpus is only 77430 words long, and it is one of a kind. The morphological complexity of Arabic increases the likelihood of data sparseness. We will ascertain whether data sparseness is affecting classification results and if so, how this can best be addressed as part of future work.

Another recommendation (*cf.*5) for *understanding* as well as evaluating classifier performance in this task is to use a combination of performance metrics (not just accuracy) to determine how well the classifier has learnt the concept: selective use of one or other metric, and inconsistency of metrics used across studies in phrase break prediction is counter-productive – and prosodic-syntactic chunking is already inherently variable.

This is original research in that: (i) our goal is to derive chunking algorithms for Arabic speech and language applications from traditional prosodic mark-up in the Qur'an; and (ii) our underpinning question is whether Qur'anic Arabic speech rhythms still inform native speaker intuition and judgment when processing Modern Standard Arabic. This, along with our other recent publications (10, 16, 19), represents groundwork for a larger-scale project to produce annotation schemes, language resources, algorithms, and applications for Classical and Modern Standard Arabic.

9. Acknowledgements

The research presented in this paper is an extension of our LREC conference papers (6, 19). To build the Boundary-Annotated Quran Corpus, we extracted, processed and extended data from two online

sources: the Tanzil Quran project (27) and the Quranic Arabic Corpus (12, 26, 36). For our experiments, we used Python tools from the open-source NLTK Natural Language Tool Kit (20).

10. References

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

2.Liberman, M.Y., Church, K.W. Text Analysis and Word Pronunciation in Text-to-Speech Synthesis. In: Advances in Speech Signal Processing.Furui S., Sondhi, M.M., editors. New York. Marcel Dekker Inc; 1992.

3. Croft, W. Intonation Units and Grammatical Structure. Linguistics.1995; 33: 839-882.

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

5. Brierley, C. 2011. Prosody Resources and Symbolic Prosodic Features for Automated Phrase Break Prediction.[PhD Thesis]. Leeds: School of Computing. University of Leeds; 2011.

6. Brierley, C., Sawalha, M., Atwell, E. Open-Source Boundary-Annotated Corpus for Arabic Speech and Language Processing. In: Proceedings of LREC 2012: Language Resources and Evaluation Conference. May 2012; Istanbul, Turkey. 2012.

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

8.Taylor, L.J., Knowles, G. Manual of Information to Accompany the SEC Corpus: The machine readable corpus of spoken English. 1988.**[**Accessed: January 2010]. Available from:

http://khnt.hit.uib.no/icame/manuals/sec/INDEX.HTM

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

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

11. Grabe, E. 2001. Prosodic Annotation.PowerPoint. 9th ELSNET European Summer School on Language and Speech Communication, Prague.[Accessed: 2006].

12.Dukes, K. The Quranic Arabic Corpus (v. 2.0); 2010.[Accessed: August 2011]. Available from:http://corpus.quran.com

13. Al-'ashmuni, Ahmad bin Muhammad Abdul-Kareem. منار الهدى في بيان الوقف والإبتدا، ومعه المقصد لتلخيص ما في المرشد في الوقف والابتداء manar al-huda fi bayan al-waqfwa al-'ibtida' Mustafa Al-baabi Al-halabi.1973.

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

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

16. Brierley, C., Sawalha, M.; Atwell, E. Arabic Phonetics and Phonology for Text Analytics and Natural Language Processing Applications. PowerPoint presentation for Arabic Phonetics and Phonology PG Workshop. York.2011.

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

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

19. Sawalha, M., Brierley, C., Atwell, E. Predicting Phrase Breaks in Classical and Modern Standard Arabic Text.In: Proceedings of LREC 2012: Language Resources and Evaluation Conference.Istanbul, Turkey. May 2012.2012.

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

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

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

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

24. Sawalha, M. Open-Source Resources and Standards for Arabic Word Structure Analysis: Fine Grained Morphological Analysis of Arabic Text Corpora. [PhD. Thesis].Leeds:School of Computing. University of Leeds.2011.

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

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

27. Zarabi-Zadeh. Tanzil Quran Project. 2012. [Accessed: April 2012]. Available from: http://www.tanzil.net

28.Sharaf, A.M. Macci, MadaniShurahs. 2011. [Accessed: October 2011]. Available from: http://www.textminingthequran.com/wiki/Makki_and_Madani_Surahs

29. Dietterich,T.G. Approximate Statistical Tests for comparing supervised classification learning algorithms. In: Neural Computation. 1998; 10:1895-1924.

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

31.Sawalha, M. The SALMA – Gold Standard. 2011. [Accessed: September 2011]. Available from:http://www.comp.leeds.ac.uk/sawalha/goldstandard.html

32. Sawalha, M. , Atwell, E. 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 2010.2010.

33. Brierley, C.; Atwell, E. ProPOSEC: a Prosody and PoS Spoken English Corpus. In: Proceedings of LREC 2010: Language Resources and Evaluation Conference.Valetta, Malta. May 2010.2010.

34.Ingulfsen, T., Burrows, T.; Buchholz, S. Influence of Syntax on Prosodic Boundary Prediction. In: Proceedings, INTERSPEECH 2005.2005;1817-1820.

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

36. Dukes, K., Atwell, E. LAMP: A Multimodal Web Platform for Collaborative Linguistic Analysis. In: Proceedings of LREC 2012: Language Resources and Evaluation Conference.Istanbul, Turkey. May 2012.2012.