PAN
ARCHIVES OF ACOUSTICS
Vol. **46**, No. 2, pp. 237–247 (2021)
DOI: 10.24425/aoa.2021.136574

Research Paper

Conditional Random Fields Applied to Arabic Orthographic-Phonetic Transcription

El-Hadi CHERIFI*, Mhania GUERTI

Department of Electronics, Signal and Communications Laboratory
National Polytechnic School
El-Harrach 16200, Algiers, Algeria; e-mail: mhania.guerti@enp.edu.dz
*Corresponding Author e-mail: el_hadi.cherifi@g.enp.edu.dz

(*received March 22, 2020; accepted January 28, 2021*)

Orthographic-To-Phonetic (O2P) Transcription is the process of learning the relationship between the written word and its phonetic transcription. It is a necessary part of Text-To-Speech (TTS) systems and it plays an important role in handling Out-Of-Vocabulary (OOV) words in Automatic Speech Recognition systems. The O2P is a complex task, because for many languages, the correspondence between the orthography and its phonetic transcription is not completely consistent. Over time, the techniques used to tackle this problem have evolved, from earlier rules based systems to the current more sophisticated machine learning approaches. In this paper, we propose an approach for Arabic O2P Conversion based on a probabilistic method: Conditional Random Fields (CRF). We discuss the results and experiments of this method apply on a pronunciation dictionary of the Most Commonly used Arabic Words, a database that we called (MCAW-Dic). MCAW-Dic contains over 35 000 words in Modern Standard Arabic (MSA) and their pronunciation, a database that we have developed by ourselves assisted by phoneticians and linguists from the University of Tlemcen. The results achieved are very satisfactory and point the way towards future innovations. Indeed, in all our tests, the score was between 11 and 15% error rate on the transcription of phonemes (Phoneme Error Rate). We could improve this result by including a large context, but in this case, we encountered memory limitations and calculation difficulties.

Keywords: Orthographic-To-Phonetic Transcription; Conditional Random Fields; text-to-speech; Arabic speech synthesis; Modern Standard Arabic.

1. Introduction

Speech and text (written words) are the two main ways in human communication. The fundamental units composing a written text, in any language, are called graphemes, while phonemes are descriptors of how a word – in a given language – is pronounced. The object of the *Character-To-Sound* CTS (or *Orthographic-To-Phonetic*: O2P) Transcription is to find the most appropriate pronunciation for any given written word. This is by no means a trivial task for most languages. The O2P Transcription plays a central role in many applications. It is part of the speech synthesis applications TTS, and is also exploited in the processing of the *Out-Of-Vocabulary* words (OOV) pronunciations in the Automatic Speech Recognition systems (ASR). The basic step in the speech synthesis process, after many pre-processing operations, is firstly to convert the text to its phonetic form. Then, the phoneme sequences are used to synthesize the appropriate sounds. The difficulty of the operation is the fact that the mapping between graphemes and phonemes is not bijective. Indeed, for the same sequence of orthographic characters (graphemes) can correspond to different pronunciations according to several considerations, such as phonetic pharyngalization or emphatisation and their influences on the selected synthesis units, and according to other phonological considerations, such as co-articulation and assimilation phenomena that modify the phonetic content expected for the same orthographic transcription. These phenomena are amply presented and discussed in this study.

In this work, we introduce an advanced Data-driven probabilistic approach – *Conditional Random Fields* (CRF) – to automate the process of Arabic *Orthographic-To-Phonetic* Transcription (O2P), also

called Grapheme-To-Phoneme (G2P) Conversion. In Sec. 2, we analyze the problem posed by the automatic phonetization of Arabic language texts at the orthographic, phonetic and phonological levels. Our phonetization system which will be implemented based on (CRF) approach should implicitly take into account these alterations thanks to the training phase using our *MCAW-Dic* phonetic dictionary. We present at Sec. 3 the most replied strategies for implementation of *Text-To-Speech* (TTS) systems. The Sec. 4 provides a literature review of O2P transcription process for Arabic language. The *Conditional Random Fields* (CRF) method will be presented in Sec. 5, as well as the training and test Database, and we give then the experimental conditions and results. We present also some conclusions and ideas for perspectives in Sec. 6.

2. Problems of O2P transcription for Standard Arabic

The O2P transcription of the text (or phonetization) consists in determining the sequence of phonemes corresponding to the pronunciation of this text. The difficulty of the operation lies in the fact that this transformation is not one-to-one: a same sequence of orthographic characters, or graphemes, may correspond to a different pronunciation. These difficulties are not due only to problems at a lexical level; a simple dictionary would then suffice to provide the phonetic transcription of a word. However, pronouncing a word, i.e. a sequence of sounds, or a phrase, i.e. a sequence of words, is not just a simple sound-to-sound transcription. Phenomena of orthographic, phonetic and phonological order modify the expected phonetic content of a simple lexical transcription. We can cite, among the main phenomena effectively acting on transcription, the example of co-articulation, which means that the sequence of phonemes (phonemic chain) is different from the sequence of phones (phonetic chain). A principle of least effort can be drawn as an articulation justification for these phenomena (Priva, 2012). The Arabic language contains these phenomena that alter the theoretical pronunciation of phonemes, we can cite: germination, emphatisation, glottal stop, the partial and total assimilation... etc.

2.1. Orthographic difficulties of phonetic transcription of Standard Arabic

The Arabic alphabet has Semitic origins derived from the Aramaic writing system, and is among the oldest alphabets in the world. The Arabic writing system contains a regular alphabet for consonants, diacritics for vowels and other signs used:

- twenty-eight graphemes representing the sounds of consonants;
- three diacritical vowels which appear above the graphemes representing the consonants; the diacritical vowels indicate that the consonants on which these signs appear are vocalized;
- the shedda or gemination sign; this sign normally appears on a consonant to indicate that it is geminate, that is, the corresponding sound is repeated;
- three symbols of "*Tanwin*": Tanwin fatha, Tanwin Kasra and Tanwin Damma; they appear on any consonant to indicate certain sequences of phonemes, for example in the following words: "كبيرًا" /kabi:r**an**/ (huge), "صغيرٍ" /s'aGi:r**in**/ (small) and "جميلٌ" /ʒami:l**un**/ (beautiful) have the symbol of Tanwin;
- some ligature symbols like Alif-Lam, Lam Alif, etc., as for example in the words "البيتِ" /? albajti/ (the house) and "لأنّ" /la?anna/ (because) where the ligature symbols are Alif-lam and Lam-alif, respectively.

The Arabic writing system also uses some special symbols and some punctuation marks. An exclusive feature of writing in the Arabic language is that the graphemes are connected. An Arabic letter changes geometric shape depending on its position in the word. There are, in general, three forms for each grapheme and these forms vary depending on whether the grapheme appears at the beginning, in the middle or at the end of the word.

Three of the 28 letters of the Arabic alphabet show variations in writing: a variant of the tā' ("ة" the *tā' marbuta*), a variant of the alif called "brief or twisted" ("ى" the alif *maqsura*), five variants of the hamza and the alif mamduda "آ" which is a replacement character for either the sequence "voweled hamza + hamza quiescent" or the sequence "hamza with alif as support + fatha + /a:/" (Table 1).

Table 1. Variants of consonants.

tâ' marbuta	ة
Alif maqsura	ى
Alif + *hamza* Above	أ
wâw + *hamza* Above	ؤ
yâ' + *hamza* Above	ئ
Alif + *hamza* Below	إ
Alif + + Hamzat Wasl	ٱ
Alif + Maddah	آ

The numeral one hundred (مائة) can be written with a spelling alif in the singular, in the dual (مائتان) and agglutinated with other numerals, as in "ثلاثمائة" (three hundred). There are other possible spellings of this word such as: مئة (this script is in fact its original script) or, very rarely, مأة. The alif of the attention particle "ها" is elided when it is prefixed to demonstratives:

- هذا but pronounced "هاذا /ha:Da:/" (this, this, this one);

- هذه but pronounced "هاذه /ha:Dihi/" (this, this one – *feminine*);
- هؤلاء but pronounced "هاؤلاء /ha:?ula:?/" (These).

Or when it is prefixed to personal pronouns beginning with hamza:

- هأنا but pronounced "هاأنا /ha:?ana:/" (I'm here);
- هأنتم but pronounced "هاأنتم /ha:?antum/" (You are there);
- هأنتما but pronounced "هاأنتما /ha:?antuma:/" (You are there – "in dual form").

Or when it is also prefixed with the pronoun "أنا" (the pronoun of the subject "I") followed by the demonstrative "ذا", it is written هأنذا but pronounced "هاأنذا /ha:?anaDa:/" (Here I am again!). The alif of the demonstrative "ذا" is elided when it is suffixed with a lâm as in: "ذلك /Da:lika/" (that one), "ذلكما /Da: likuma:/", "ذلكم /Da:likum/", "ذلكن /Da:likunna:/" (same meaning: that, that one; but in dual and plural, respectively). Some words are always pronounced with an alif (long vowel /a:/) that is not written, namely:

- الله "/?*alla:h*/" (Allah);
- إله "/?illa:h/" (God);
- لكنْ "/la:kin/" ("but", before a sentence);
- لكنَّ "/la:kinna/" ("but", before a name or a pronoun).

The relative pronouns of the third person masculine singular الذي "/?allaDi:/" and feminine التي "/?allati:/", and the third person masculine plural الذين, are written with a single letter lâm "/l/" (that of article) but are pronounced with two lâm. It should be noted that the same one relating to the dual and the feminine plural, is not in this case, the adequacy between graphemes and phonemes is respected there: اللَّذَانِ - اللَّذَيْنِ - اللَّتَيْنِ - اللَّتَانِ, as well as اللَّوَاتِي - اللَّاتِي. The second Waw /w/ "و" (the long vowel /u:/) in (داوود) /dawu:d/ and (طاووس) /tawu:s/ is elided in writing but still pronounced.

2.2. Phonetic and phonological difficulties of O2P of standard Arabic

Depending on the geometric configuration of the vocal tract, different sounds are produced. Each of the articulators of the vocal tract can indeed take a considerable number of positions. The passage from one position to another does not happen abruptly but on a continuum. Only a limited number of configurations of the phonatory apparatus are used linguistically for the production of the speech sounds: *the phones*.

As in English and French, the transcription of Arabic graphemes may depend on the preceding and/or following words. In Arabic, this type of contextual dependency is encountered with any word beginning with the prefix "ال" (the equivalent of "la" or "le" in French, or "the" in English) which is followed by what is normally called in Arabic a "solar consonants" (Table 2). When a word begins with "ال" followed by a letter "solar consonant" and the word is preceded by a vowel, the prefix "ال" is pronounced as /?a/, the following solar character is geminated (مشدّد), and the word that contains the prefix "ال" is confused (linked) with its predecessor. For example, "بدأ اللَعبُ" /bada?alla?'ibu/ (the game has started). If the word prefixed with "ال" is not preceded by a vowel, the "ال" is pronounced as /?a/ and the following "solar letter" is geminated but the two words will not be linked in pronunciation. Indeed, this situation is called "supporting vowels" or "connecting vowels". The supporting vowels are the three short vowels (damma, fatha and kasra) used, in a phonological (syllabic) conditioning framework, to replace a sukun "ْ" at the end of a preposition or a verb followed by a word starting with an unstable hamza, in order to solve a phonological problem. However, it is impossible, in Arabic, to have two sukuns which follow each other immediately, then, when we are faced with the case where a word ending in a sukoun (مِنْ /min/ "of", مَنْ /man/ "who?, the one who", هُمْ, /hum/ "them", قَالَتْ /qa:lat/ "she said") and followed by a word beginning with an unstable hamza, we replace the sukun of the first word by a supporting vowel; this vowel is generally a kasra except in the following cases:

- a fatha when it is about مِنْ followed by the article ال;
- a damma when it is about the masculine plural pronouns هُمْ, كُمْ or the verbal suffix of the masculine plural تُمْ.

Table 2. Classification of consonants taking into account the transcription constraints.

Solar	Lunar
ت ث د ذ ر ز س ش ص ض ط ظ ل ن	ء ا ب ج ح خ ع غ ف ق ك ه م و ي

The following examples clearly show the replacement of the sukun in each of the cases listed (this represents the general case, but there are some exceptions):

مِنَ ٱلْكتاب → مِنْ + الْكتاب of the book

قالتِ ٱلْبنت → قالتْ + الْبنت The girl said

هُمُ ٱلناس → هُمْ + الناس These are the people

The phonetic system of MSA contains certain phonological rules governing the articulation of the final silent (unvocalized) "nun" sound of the letter Nun Sakina "نْ" /n/, and the inflectional "tanwin". It is also a question of studying the influence of the final Nun Sakina and the tanwin on the following word, from where several phonological phenomena appear within the same word, or at boundaries between words – which alter the theoretical pronunciation of phonemes

– and which the phonetization system must take into account. These rules are: clear pronunciation إظهار, assimilation (Idgham) إدغام, concealment إخفاء, and substitution إقلاب. Indeed, from a phonological point of view, a *Nun Sakina* and a *tanwin* are considered to be the same because at the level of their sound, the tanwin sounds like a Nun Sakina (Table 6). Thus, they will follow the same pronunciation rules, we will give a brief overview here, and for more details, the reader can refer to the exhaustive study by Alduais (2013).

2.2.1. Clear pronunciation (Al-Idh-har) **الإظهار**

It consists to get out each sound from its articulation point without resorting to nasalization. The articulation point (makhraj) of the /n/ sound is the tip of the tongue against the palate near the upper incisors. This rule concerns six phonemes called "the guttural sounds" (Halqiya): ء هـ ع ح غ خ (/h/, /?/, /?'/, /x/, /G/, /X/). Example: مِنْ خَوْفٍ (/min Xawfin/) and مِنْ إِلَهٍ (/min?ila:hin/).

2.2.2. Idgham (a form of assimilation)

Appears when the assimilation of Nun Sakina or tanwin takes place in the presence of the following sounds: ي - ر - م - ل - و - ن (/j/, /r/, /m/, /l/, /w/, and /n/). Idgham refers to the suppression of the alveolar nasal sound of Nun Sakina or tanwin whenever it occurs in the final position of a word and is followed by a word beginning with one of the above six sounds. As a result, these sounds will be geminate مشدّدة. In writing, gemination is represented by the doubling of the geminate letter (Ferrat, Guerti, 2016). But it is interesting to note that no assimilation will take place if these sounds occur within a word. There are two types of Idgham: Idgham without nasalization and Idgham with nasalization.

2.2.3. Idgham with nasalisation

It is also called incomplete assimilation. The sounds of idgham with nasalization: ي - ن - م - و (/j/, /n/, /m/, and /w/). Assimilation with nasalization means that the sound of the Nun Sakina or the tanwin will be heard in an incomplete or partially assimilated way. The nasalization remains as an indicator of the suppression of the /n/ sound. It is a regressive but not progressive assimilation that takes place here because there is a complete absorption of one final phoneme into another initial phoneme (Abu-Salim, 1988). Example: مِنْ نعمةٍ is pronounced مِنّعمة /min^{n}ni?'matin/, or أُمَةً واحدة is pronounced أُمَّتواحدة /ummataw^{n}wa:hida/.

2.2.4. Idgham without nasalization

Idgham without nasalization is when the letter Nun Sakina or tanwin at the end of a word is followed by the letters ر /r/ or ل /l/ of another word. The pronunciation will be done without nasalization. In writing, the Arabic letters Lam - ل and Ra - ر will then bear the diacritical sign of gemination "*Shedda* ّ". We speak here about a complete assimilation because we will not find any sound trace of the phoneme /n/. We are also talking here, about a regressive and not progressive assimilation because there is a complete fusion of the final sound into the initial sound with the cancellation of any phonetic feature of the suppressed sound. Example: مِنْ ربّك which is pronounced مرّبك /mirrabika/ or ويلٌ لِكل which is pronounced ويلّكلِ /wajlullikulli/.

2.2.5. Ikhfaa (concealment)

In the phonetic system of MSA, there are 15 sounds known to be "weak sounds" because they are treated differently when preceded by the unvocalized nasal alveolar sound /n/, whether at the within the same word, or between borders of words in a sentence.

Whenever the alveolar nasal sound /n/ of "unvocalized n" or tanwin is followed by one of the 15 sounds mentioned above (Table 3), a concealment process will take place. In such a case, the /n/ will change its articulation point, but retain its nasalization trait. The sound /n/ is maintained, not assimilated, but it is not explicitly spoken. We are faced with a phenomenon of partial regressive assimilation (Roach, 1987). We recall that in the case of tanwin, this phenomenon will only take place between words, but not within the word. Examples: كنتم /kun^{0}tum/, or الإنسان /al?in^{0}san/ or مِنْ طِينٍ /min^{0}t'i:nin/, and in the case of tanwin, عينٌ جارية /?'ajnun^{0}ʒa:rija/. With [n^{0}] denoting the phenomenon of *nun Sakina concealment*.

2.2.6. Iqlab (substitution)

The Iqlab occurs when the "sound" changes its initial exit point, as in the case of Nun Sakina or tanwin, if followed by the phoneme [b] - ب, changes to Mîm - م /m/. Example: سميعٌ بصير is pronounced سميعُمْبصير /sami:?'u**m^{n}**basi:r/. Note that this phenomenon ap-

Table 3. The 15 "weak sounds" of the Arabic language (Alduais, 2013).

Occlusive	/t/ = /ت/, /d/ = /د/, /k/ = /ك/
Fricative	/s/ = /س/, /z/ = /ز/, /f/ = /ف/, /T/ = /ث/, /D/= /ذ/, /S/ = /ش/
Occlusive-fricative	/ʒ/ = /ج/
Emphatic	/s'/= /ص/, /d'/ = /ض/, /t'/ = /ط/, /D'/ = /ظ/, /q/ = /ق/

Table 4. Example of homophonic words in Arabic.

	Example 1	Example 2	Example 3
Pronunciation	[?ila:huna]	[?'ala:]	[?inSa:?a]
Possible writing	إلى هنا (So far)	على (On)	إن شاء (If he wants)
Orthographic variant	إلهنا (Our Lord)	علا (He rose)	إنشاء (Construction)

Table 5. Arabic consonant and vowels and their IPA (International Alphabet Phonetic) and SAMPA (Speech Assessment Methods Phonetic Alphabet) transcriptions (Wells, 2002).

Type	Consonant																
Arabic	ء	ب	ت	ث	ج	ح	خ	د	ذ	ر	ز	س	ش	ص	ض	ط	ظ
IPA	ʔ	b	t	θ	ʒ	ħ	X	d	ð	r	z	s	ʃ	ṣ	ḍ	ṭ	ðˤ
SAMPA	?	b	t	T	Z	x	X	d	D	r	z	s	S	s'	d'	t'	D'

Type	Consonant											Vowel					
Arabic	ع	غ	ف	ق	ك	ل	م	ن	هـ	و	ي	ـَ	ـُ	ـِ	ا	و	ي
IPA	ʕ	ɣ	f	q	k	l	m	n	h	w	j	a	u	i	a :	u :	i:
SAMPA	?'	G	f	q	k	l	m	n	h	w	j	a	u	i	a:	u:	i:

Table 6. Other orthographic symbols (tanwin, shadda and sukun) with SAMPA transcription.

Doubled case endings (tanwin)		
ـً	tanwin fatha	/an/
ـٌ	Tanwin damma	/un/
ـٍ	Tanwin kasra	/in/

Syllabification marks	
ـّ	*Shadda* (gemination mark denotes the consonant doubling)
ـْ	*Sukun* (to indicate that the letter doesn't contain any vowel)

pears in the boundaries between words, as it may appear inside the word as in "عنبر" /?'anbar/ which is pronounced "عمبر" /?'am^{n}bar/. The sign (^{n}) is used to represent the phenomenon of nasalization (الغنّة) in the representation of pronunciation. In addition, there are in Arabic language what we call *homophones* such as in Table 4. We have adopted for the phonetic transcriptions the SAMPA code (Tables 5 and 6). The SAMPA phonemic transcription system was a choice for practical reasons, and its use is widely supported in this field for the Arabic language. We could have used the IPA system; the result would have been the same. An additional reason that prompted us to choose SAMPA notation rather than IPA is that SAMPA is a user-friendly system for computer coding. There are a number of alternative Unicode-oriented versions of the IPA, but SAMPA suits us the most when it comes to the possibility of integrating parts of other software into the final TTS system: such as the toolbox of natural language processing (NLTK), the hidden Markov model toolkit (HTK), or the speech synthesis engine MBROLA among others, which adopt this notation.

3. Strategies for implementation of text-to-speech systems

Technically, we can first distinguish *Lexicon-Based* O2P systems. Each lexical entry, written in graphemic form, matches a phonemic form. The disadvantage of this approach resides in its practical implementation. It requires indeed a storage capacity proportional to the size of the lexicon and the search time of a word can become prohibitive for *real-time* systems or those placed in *embedded situations*. Because of these functional constraints, the use of a lexicon is usually reserved for the transcription of morphemes. However, rules are needed to determine the pronunciation of a word then constituted by the juxtaposition of morphemes. The *Rules-Based* transcription systems involve O2P transcription rules, but also lexicons of exceptions to these rules.

Both *Lexicon-Based* and *Rule-Based* approaches use sets of rules as well as lexicons. They can be distinguished by the large number of rules relative to the size of the lexicon and by the fact that the lexicon solution begins his treatment by a lexical morphological decomposition. Most commercialized speech synthesis systems are based on both a morphological decomposition, transcription rules and access to exceptional lexicons (Polyakova, Bonafonte, 2005).

Many research studies have focused on the automatic inference of transcription rules from examples. The objective is to identify transcription rules and exceptions from transcripts examples. One can quote some work using Hidden Markov Models (HMM) (Van Coile, 1991), inference techniques by analogy for French (Yvon, 1996) or for English (Bagshaw, 1998), or the stochastic approaches (Luk, Damper, 1996).

One must add to previous approaches, the purely functional approaches where one does not try to discover a set of transcription rules but to link a graphemic input to phonetic output. One can quote the

work of Sejnowsky and Rosenberg (1987) the NETtalk system – that models the transcription function by a Neural Network and our recent framework on the Arabic language (Cherifi, Guerti, 2017) based on finite-state transducer approach. The Data-driven techniques such as Conditional Random Fields (CRFs) or Joint Multigram Model (JMM) were successfully exploited in O2P Transcription to speech synthesis for many languages (Lafferty *et al.*, 2001; Casacuberta, Vidal, 2007). We recall that all these existing methods for O2P Transcription are based on a dictionary constituted by pairs *Word-Pronunciation* (i.e. *Word-Phonetic Transcription*) as the only source of data. In the next section, we review the main approaches to Arabic O2P Transcription in literature.

4. Literature review of O2P transcription process for MSA

Limited researches have been carried out on Arabic Language phonetization in comparison to other languages. We disclosed related work in this area such as Al-Ghamdi *et al.* (2004). This is mainly related to more than one factor. Arabic text is usually written without diacritic, this causes a shortage of a comprehensive Arabic phonetic corpus. Selim and Anbar (1987) developed a rule based phonetic transcription system for Arabic text; a non-diacritic 291 words used from newspapers. Their system showed a moderate ratio of accuracy although they used a limited numbers of tested words.

El-Imam (1989; 2004) proposed a system to phonetize Arabic text and addressed the problems related to transcription of Arabic O2P by studying the properties of Arabic phonology including phonetic rules. However, these rules were not clearly prioritized, so they might contradict each other and produce inappropriate output, since some of these rules have to be visited before others and the output of the previous rule will be an input to the next one. Ahmed (1991) utilized about 150 allophones, vowels/constants combinations. He applied a set of *Letter-To-Sound* rules to simplify computer voice production. His results clarified that the rules were the main part and is considered as the backbone of any Arabic Text-To-Speech application. Al-Ghamdi *et al.* (2004) used the Arabic phonology rules to convert text to sound symbols by listing a number of phonetic and phonemic rules with some exceptional words, but they did not implement a system in order to test the performance of these rules.

In the other hand, many researchers have realised that modelling the pronunciation variation can enhance the performance of the O2P process. They have proposed different approaches, of which the classical approach involves generating an Arabic multi-pronunciation dictionary. For instance, Biadsy *et al.* (2009) have generated a multi-pronunciation dictionary using pronunciation rules and then MADA (*Morphological Analysis and Disambiguation for Dialectal Arabic*), as a morphological disambiguation tool, to determine the most likely pronunciation of a given word in its context. The proposed method reported a significant improvement of 4.1% in accuracy compared to the baseline system (Habash *et al.*, 2009). Elshafei *et al.* (2008) have provided a limited set of phonetic rules for automatic generation of an Arabic phonetic dictionary. The rules were mainly direct O2P with a few rules for the assimilation of "lam" [*l*] with solar letters, the conversion of [n] to [m] when followed by [b], and emphatics with pharyngeal vowels. The effectiveness of using the generated dictionary was tested using a large vocabulary speaker-independent Arabic ASR system and achieved a comparable accuracy with the same vocabulary-size English ASR system. This work was then implemented in many other publications such as in Al-Ghamdi *et al.* (2009) and Abuzeina *et al.* (2012).

Ramsay *et al.* (2014) developed a comprehensive knowledge-based model for automatically generating a phonetic transcription of a given Arabic text. This model is based on a set of language-dependent pronunciation rules that works on converting fully diacriticised Arabic text into the actual sounds, along with a lexicon for exceptional words. Al-Daradkah and Al-Diri (2015) have developed an automated O2P Transcription process by using Arabic language phonology rules supported by a dictionary of exceptional words. The system was tested on a publicly dataset that contains 620 fully diacritics Arabic sentences formed from 3440 words and consists of 27 030 graphemes which were manually segmented. The system showed a high precision of 99.19%. Sindran *et al.* (2016) presented an O2P transcription system for MSA at five levels: phoneme, allophone, syllable, word and sentence. The accuracy of the system was better than 99% for Arabic texts without some type of named entities like: dates, numbers, acronyms, abbreviations and special symbols.

5. Conditional random fields for Arabic O2P

In recent years, speech synthesis techniques – and in particular the O2P transcription phase – have achieved excellent performance levels thanks to the use of discriminating probabilistic models such as maximum entropy models (Toutanova *et al.*, 2003), or CRFs (Tsuruoka et al., 2009). However, for the Arabic language, this potential is not yet exploited. CRFs are discriminating probabilistic models introduced by (Lafferty et al., 2001) for sequential annotation. They have been used in many Natural Language Processing tasks, where they give excellent results (McCallum, Li, 2003; Sha, Pereira, 2003).

CRFs allow an observation x to be associated to an annotation y based on a set of labelled examples, i.e. a set of pairs (x, y). Most of the time (and it is the case in this paper), x is a sequence of units (here, a sequence of orthographic graphemes) and y is the sequence of corresponding labels (here, the sequence of their phonemes). CRFs are discriminating models which belong to the family of non-oriented graph models. They are defined by X and Y, two random fields respectively describing each unit of the observation x and its annotation y, and by a graph $G = (V, E)$ such as $V = X \cup Y$ is the set of nodes and $E \subseteq V \times V$ the set of edges. Two variables are related in the graph if they depend on each other. The graph on the Y field of linear CRFs, drawn in Fig. 5, reflects the fact that each label is supposed depending on the previous label and the next one and, implicitly, on the complete x data.

5.1. The CRF linear chain

In our study, we consider a specific CRF form, namely the CRF linear chain. An example of a CRF linear chain is shown in (Fig. 1).

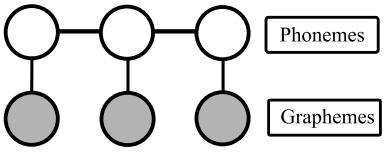

Fig. 1. The CRF linear chain.

We can take, as a first application example of this chain, the case of the homophone word [?'ala:] which has orthographically two writings, so two meanings (Table 4). Figure 2 shows the first orthographic possibility, and Fig. 3 the second one.

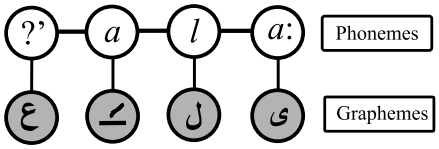

Fig. 2. The linear chain of the word [?'ala:] with the first orthographic possibility.

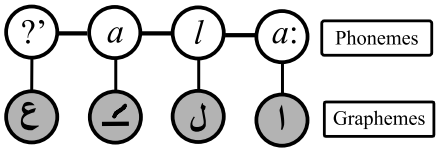

Fig. 3. The linear chain of the word [?'ala:] with the second orthographic possibility.

In this example, the phonetic transcription is the same although the corresponding orthographic is in two forms. The following example, shown by Fig. 4, we have the CRF chain transcribing the word [?'ambar]

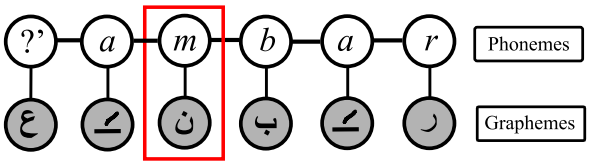

Fig. 4. The linear chain of the word [?'ambar] with an assimilation Idgham.

which shows a phonological phenomenon (a complete consonant assimilation process, see Subsec. 2.2).

As shown in the Fig. 1, each node represents, generally, a random variable. Assuming we make the first order Markov assumption, all the nodes in the graph form a linear chain. Using the definition from (Lafferty *et al.*, 2001), a linear chain CRF applied to any labelling problem, is specified by the following conditional probability:

$$P(Y/X) = \frac{1}{Z(X)} \exp\left\{\sum_{k=1}^{m} \alpha_k F_k(Y, X)\right\}. \qquad (1)$$

The CRFs are so an modelization of the conditional distribution $P(Y/X)$ with the objective of predicting a sequence: $Y^* = \arg\max_Y P(Y/X)$, ignoring the correlations between the observation variables and not caring about whether they are independent or not, where, in our case of O2P transcription:

- X is the grapheme sequence of a word;
- Y is a candidate pronunciation (one possible phonetic transcription);
- F_k is the k-th potential function expressed in terms of feature functions f_k;
- α_k is a weight of the feature;
- $Z(X)$ is a normalization quantity in order to get a probability at the end equal to 1, given by:

$$Z(X) = \sum_X \exp\left\{\sum_{k=1}^{m} \alpha_k F_k(Y, X)\right\}. \qquad (2)$$

The graph is separated into cliques, each of which constitutes two consecutive phonemes and the entire grapheme sequence. Thus $F_k(X, Y)$ can be expressed in terms of features f_k of cliques, given by:

$$F_k(Y, X) = \sum_{j=1}^{m-1} \{f_k(Y_j, Y_{j-1}, X, j)\}. \qquad (3)$$

That is to say:

- training set: input and target sequence pairs $\{(X_j, Y_j)\}$;
- the j-th input sequence of vectors:

$$\mathbf{X}_j = [\mathbf{X}_1, ..., \mathbf{X}_m];$$

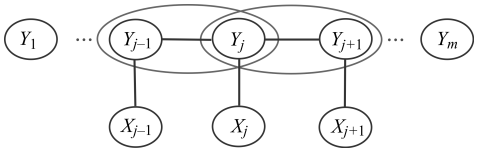

Fig. 5. The chain structured case of CRFs for sequences (first-order linear chain model), as a non-oriented graph.

- the j-th target sequence of labels (i.e., the sequence of the phonemes retained in the transcription of the sequence of the corresponding graphemes):

$$\mathbf{Y}_j = [\mathbf{Y}_1, ..., \mathbf{Y}_m];$$

- and m is the sequence (orthographic word) length.

We will use *features* that are binary functions that look for presence of graphemes and phonemes at various positions in the clique. The *features* are real-valued functions. It is through them that all the domain knowledge (conversion process) is integrated in the model. These *features* take as parameters the values of the random variables of the clique on which they apply (Y_j) and the entire observation X. Therefore, the value taken by a random variable can depend on all the observation X. For example, in the case of a prediction of pronunciation of a sequence, the choice of the label (the phoneme) associated to the last element of the sequence may be related to the value of the first element of this sequence (Jousse *et al.*, 2006).

To these *features* are associated weights α_k. These weights are the model parameters. They allow to attach more or less importance to certain *features* or even to indicate that the phenomenon characterized by a *feature* must not happen (if the weight is negative). CRF is therefore defined by an independence graph G and a set of *features* f_k to which weights α_k are associated.

The first problem associated to CRF is the annotation problem, which consists in finding the most likely annotation according to Eq. (1) associated to an observation. The second problem is that of *inference* or CRF *training*, which is to estimate the parameters $\{\alpha_k\}$ that maximize the likelihood of the model with respect to an annotated observations sample. These parameters can be learned using a conventional method of maximizing of the log-likelihood (Illina *et al.*, 2012). The optimum parameters cannot be calculated analytically, approximate methods – such as *gradient-descent* – are alternately used. Wallach (2002) showed that the most effective method in this context is the *limited-memory* BFGS (Broyden-Fletcher-Goldfarb-Shanno) algorithm (L-BFGS). The interest and efficiency of CRFs come from the fact that they take into account the dependencies between labels connected to each other in the graph. When looking for the best y, that is, the best sequence of labels associated to a complete data x, they generally behave better than a series of classifications of isolated units. But this consideration has a price: the training phase of a CRF can be long. Once this phase has been completed, annotating a new sequence x of n input words then amounts to finding the y which maximizes $p(y/x)$. The O2P Transcription problem has been so reduced to a supervised classification problem with structured output.

5.2. Data for training and test

We will now study this technique that currently claim to provide the best results in this field, and exploit them to the standard Arabic language. We will discuss our work developed with this method and expose the results obtained. We use as training and test corpus, a phonetic dictionary of about 35 000 words that we have developed by ourselves assisted by phoneticians and linguists from the University of Tlemcen, we have called MCAW-Dict (Dictionary of the ***M**ost **C**ommonly used **A**rabic **W**ords*) (Cherifi, 2020). The MCAW-Dict is an open-source machine-readable of the *Most Commonly used Arabic Words* pronunciation dictionary. This dictionary contains over 35 000 words in modern standard Arabic "MSA" and their pronunciation using the 34 phonemes in Speech Assessment Methods Phonetic Alphabet (SAMPA) notation (Tables 5 and 6). MCAW-Dict is being maintained and expanded. Table 7 gives an overview of MCAW-Dict.

Table 7. An overview of MCAW-Dict.

Diacritized word	SAMPA transcription	Sens
أَبَّ	?abba	desire
آبْ	?a:b	August
بَتَرَ	batara	amputate
بُرْكَانْ	burka:n	volcano
تَأْنِيثْ	ta?ni:T	feminization
تَجْرِيدِيٌّ	taʒri:dijjun	abstract
حَجَزَ	xaʒaza	retain
حَدٌّ	xaddun	border
زِينَهْ	zi:nah	embellishment
سَاهَمَ	sa:hama	participate
سَائِلٌ	sa:?ilun	asking
ضَاقَ	d'a:qa	be_narrow
ضَائِعٌ	d'a:?i?'un	lost
طُفُولِيَّهْ	t'ufu:lijjah	infancy
طَفِئَ	t'afi?a	die_out
نَفْخٌ	nafXun	inflation
.....	.....	.....
يَئِسَ	ja?isa	despair

5.3. Tests and results

As an example, Fig. 6 represents a simple finite-state model designed to distinguish between the two phonetic transcriptions [?'anbar] and [?'ambar] for the same orthographic word عنبر (we have not shown the vowels in this example, because their transcriptions present no problems).

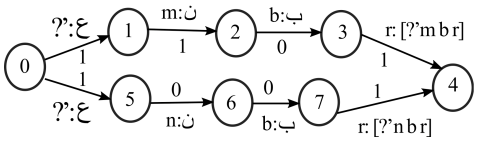

Fig. 6. We place observation-label pairs o:l on transitions.

To set up the experiment we need two things. First, we need a tool for aligning the training examples. Second, we require a tool to perform training on the aligned training data. We use the Giza++ toolkit (Casacuberta, Vidal, 2007) to get "1-1" alignments. Giza++ treats the set of words as a source language and the set of pronunciations (i.e. their corresponding phonetic transcriptions) as a target language. Then learning the mapping between these two languages is modelled as a statistical translation problem. To do the actual CRF training and testing, we use the CRF++ toolkit. This tool is developed by the NTT Communication Science Laboratories in Japan (Kudo, 2005). This open source tool, written in C++, uses the limited memory BFGS algorithm (Apostolopoulou *et al.*, 2009) for training CRF. This speeds up execution while making sure the memory requirements do not escalate to unmanageable proportions. In addition, this allows us to specify fairly large number of feature functions. It also gives us an option to use L1 or L2 regularization[1]. This toolkit allows the user to specify feature templates in advance. A macro $\%x$ [row, column] is used to specify the location of a token in the input file corresponding to the current token. It expands these macros using the training data to generate the appropriate binary indicator functions. There are two types of feature functions, unigram and bigram. The unigram feature involves only the current output token, while bigram features, if specified, contain a combination of previous and current output token. Consider the following example:

A:ah B:b A:ae C:k K:null

The symbol *null* represents the null output label (*No phoneme selected*). With this as reference the template "T" is "$\%x$ [0,0]" would expand to generate functions of the following form:

func1=if(output=k and feature="T:C")
return 1 else return 0
func2=if(output=b and feature="T:B")
return 1 else return 0 ...

Thus, there would be a *feature* function for each combination of grapheme token and label phone. We incorporate information about the grapheme context by using an n-gram of input tokens. So a macro of the form $\%x$ [1,0]/$\%x$ [0,0] would consider both the current and previous grapheme token along with the current label phone.

We consider experiments with grapheme context windows of size 2 and 3 on MCAW-Dict and report Phone Error Rates (PER) using one best scoring and L2 regularization. We found out that L2 regularization works better than L1 to prevent over fitting. Indeed, the regularization by L1 norm tries to minimize the sum of the absolute differences between real values and predicted values of the model parameters (such as weights and feature functions). Linear, it offers the possibility for the model to easily set a weight to 0 and can therefore, among other things, facilitate the selection of characteristics by forcing a sparse representation. The regularization by L2 norm tries to minimize the sum of the squares of the differences between real (training) values and predicted values. This term is, among other things, faster to calculate than the L1 term. Exponential, it rather promotes a diffuse representation and, therefore, generally performs better than the L1 (Sînziana, Iria, 2011). The results are shown in the Table 8.

Table 8. CRF result for MCAW-Dict (PER: Phoneme Error Rate).

CRF window	PER [%]
CRF (+1, –1)	15.0
CRF (+2, –2)	13.0
CRF (+3, –3)	11.0
....	....

We observe (from Table 8) that as we capture longer context (m – length of the sequence), we get an improvement in the performance. This is in line with our intuition as longer context allows learning the mapping between grapheme clusters and phonemes, because in this case, all the phonetic and phonological phenomena of the Arabic language will be covered. However longer context also means more feature functions. We encountered memory limitations for contexts greater than three, when we used CRF++ on MCAW-Dict. For standard Arabic, a 3-gram context is more than sufficient, because it will thus cover all the phonetic phenomena that may arise between the phonemes of a word, or between the borders of words in a sentence.

[1]For more details on the different types of regularizations used in machine learning models (like L1 and L2), the reader can refer to the following document: http://ai.stanford.edu/~ang/papers/icml04-l1l2.pdf.

6. Conclusions

Our work describes the efforts to exploit, in order to perform Orthographic-To-Phonetic Transcription for standard Arabic, an advanced data-driven probabilistic approach, i.e. Conditional Random Fields (CRF). The current results are quite satisfactory on the dictionary adopted for test and learning. Even if these results do not surpass the best scores of the baseline existing systems but point the way towards future innovations. On the other hand, the CRF can be exploited in O2P Conversion by formulating this task as a sequence-labelling problem. In addition, the CRF are discriminative classifiers that can integrate complex *features* functions. However, modeling by CRF requires alignment between letters and phonemes. In addition, traditional CRF linear chains usually employ in output only bi-grams information for practical reasons, and this is not sufficient for our task. As we can see, the system accuracy shows steady improvement as we incorporate longer histories.

Our future research will focus on the possibility of combining CRF and other technics in tandem to achieve a hybrid system. Such a system takes advantage of both individual approaches. An interesting perspective is to associate discriminative models (based on CRF) with generative models (based on HMM). The syllabic features may be incorporated in the CRF as additional functionalities, which will allow a significant improvement in the score of the hybrid system. It is the discriminative power of CRFs that makes the difference compared to a model based simply on HMMs.

Acknowledgment

The authors would like to thank Mrs. Amina Hadjiat, professor and researcher at Mathematics Department of University of Tlemcen, Algeria, for her comments on earlier versions of this article.

References

1. Abu-Salim I.M. (1988), Consonant assimilation in Arabic: An auto-segmental perspective, *Lingua*, **74**(1): 45–66, doi: 10.1016/0024-3841(88)90048-4.
2. AbuZeina D., Al-Khatib W., Elshafei M., Al-Muhtaseb H. (2012), Within-word pronunciation variation modeling for Arabic ASRs: a direct data-driven approach, *International Journal of Speech Technology*, **15**(2): 65–75, doi: 10.1007/s10772-011-9122-4.
3. Ahmed M.E. (1991), Toward an Arabic text-to-speech system, *The Arabian Journal for Science and Engineering*, **16**(4): 565–583.
4. Al-Daradkah B., Al-Diri B. (2015), Automatic grapheme-to-phoneme conversion of Arabic text, [in:] *2015 Science and Information Conference (SAI)*, pp. 468–473, doi: 10.1109/SAI.2015.7237184.
5. Alduais A.M.S. (2013), Quranic phonology and generative phonology: formulating generative phonological rules to non-syllabic Nuun's Rules, *International Journal of Linguistics*, **5**(5): 33–61, doi: 10.5296/ijl.v5i1.2436.
6. Al-Ghamdi M., Al-Muhtasib H., Elshafei M. (2004), Phonetic rules in Arabic script, *Journal of King Saud University – Computer and Information Sciences*, **16**: 85–115, doi: 10.1016/S1319-1578(04)80010-7.
7. Al-Ghamdi M., Elshafei M., Al-Muhtaseb H. (2009), Arabic broadcast news transcription system, *International Journal of Speech Technology*, **10**(4): 183–195, doi: 10.1007/s10772-009-9026-8.
8. Apostolopoulou M.S., Sotiropoulos D.G., Livieris I.E, Pintelas P. (2009), A memoryless BFGS neural network training algorithm, [in:] *Proceeding of the 7th IEEE International Conference on Industrial Informatics* (INDIN), pp. 216–221, doi: 10.1109/INDIN.2009.5195806.
9. Bagshaw P.C. (1998), Phonemic transcription by analogy in text-to-speech synthesis: novel word pronunciation and lexicon compression, *Computer Speech and Language*, **12**(2): 119-142, doi: 10.1006/csla.1998.0042
10. Biadsy F., Habash N., Hirschberg J. (2009), Improving the Arabic pronunciation dictionary for phone and word recognition with linguistically-based pronunciation rules, [in:] *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the ACL*, Boulder, Colorado, pp. 397–405.
11. Casacuberta F., Vidal E. (2007), *Systems and tools for machine translation. GIZA++: Training of statistical translation models*, Universitat Politécnica de Valéncia, Spain, https://www.prhlt.upv.es/~evidal/students/master/sht/transp/giza2p.pdf.
12. Cherifi E.H. (2020), *MCAW-Dict, Phonetic Dictionary of the Most Commonly used Arabic Words with SIMPA Transcription*, https://drive.google.com/file/d/1h_dPwAXKone7nGIKgelMt8mIzGYFF7d2/view?usp=sharing.
13. Cherifi E.H., Guerti M. (2017), Phonetisaurus-based letter-to-sound transcription for standard Arabic, [in:] *The 5th International Conference on Electrical Engineering (ICEE-B 2017)*, pp. 45–48, October 29th to 31st, 2017, Boumerdes, Algeria, doi: 10.1109/ICEE-B.2017.8192073.
14. El-Imam Y.A.(1989), An unrestricted vocabulary Arabic speech synthesis system, *IEEE Transactions on Acoustics, Speech and Signal Processing*, **37**(12): 1829–1845, doi: 10.1109/29.45531.
15. El-Imam Y.A. (2004), Phonetization of Arabic: rules and algorithms, *Computer Speech and Language*, **18**: 339–373, doi: 10.1016/S0885-2308(03)00035-4.
16. Elshafei M., Al-Ghamdi M., Al-Muhtaseb H., Al-Najjar A. (2008), Generation of Arabic phonetic dictionaries for speech recognition, [in:] *Proceedings of the International Conference on Innovations in Information Technology IIT2008*, pp. 59-63. doi: 10.1109/INNOVATIONS.2008.4781716.

17. Ferrat K., Guerti M. (2017), An experimental study of the gemination in Arabic language, *Archives of Acoustics*, **42**(4): 571–578, doi: 10.1515/aoa-2017-0061.

18. Habash N., Rambow O., Roth R. (2009), Mada+ tokan: a toolkit for Arabic tokenization, diacritization, morphological disambiguation, pos tagging, stemming and lemmatization, [in:] *Proceedings of the 2nd International Conference on Arabic Language Resources and Tools (MEDAR)*, Cairo, Egypt, pp. 102–109.

19. Illina I., Fohr D., Jouvet D. (2012), Pronunciation generation for proper names using Conditional Random Fields [in French: Génération des prononciations de noms propres à l'aide des Champs Aléatoires Conditionnels], *Actes de la Conférence Conjointe JEP-TALN-RECITAL 2012*, Vol. **1**, pp. 641–648.

20. Jousse F., Gilleron R., Tellier I., Tommasi M. (2006), Conditional random fields for XML trees [in:] *Proceedings of the International Workshop on Mining and Learning with Graphs*, ECML/PKDD 2006, pp. 141–148.

21. Kudo T. (2005), *CRF++: Yet another CRF toolkit. User's manual and implementation*, https://aithub.com/UCDenver-ccp/crfpp (retrieved September 20, 2020).

22. Lafferty J., McCallum A., Pereira F. (2001), Conditional Random Fields: probabilistic models for segmenting and labeling sequence data, [in:] *Proceedings of the International Conference on Machine Learning ICML'01*, pp. 282–289.

23. Luk R.W.P., Damper R.I. (1996), Stochastic phonographic transduction for English, *Computer Speech and Language*, **10**(2): 133–153, doi: 10.1006/csla.1996.0009.

24. McCallum A., Li W. (2003), Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons, [in:] *Proceedings of the Seventh Conference on Natural Language Learning at {HLT}-{NAACL}2003*, pp. 188–191, https://www.aclweb.org/anthology/W03-0430.

25. Polyakova T., Bonafonte A. (2005), Main issues in grapheme-to-phonetic transcription for TTS, *Procesamiento Del Lenguaje Natural*, **2005**(35): 29–34, https://www.redalyc.org/articulo.oa?id=5157/515751735004.

26. Priva U.C. (2012), *Sign and signal deriving linguistic generalizations from information utility,* Phd Thesis, Stanford University.

27. Ramsay A., Alsharhan I., Ahmed H. (2014), Generation of a phonetic transcription for modern standard Arabic: A knowledge-based model, *Computer Speech and Language*, **28**(4): 959–978, doi: 10.1016/j.csl.2014.02.005.

28. Roach P. (1987), *English Phonetics and Phonology*, 3rd ed., Longman: Cambridge UP.

29. Sejnowsky T., Rosenberg C.R. (1987), Parallel networks that learn to pronounce English text, *Complex System*, **1**(1): 145–168.

30. Selim H., Anbar T. (1987), A phonetic transcription system of Arabic text, [in:] *ICASSP'87. IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 1446–1449, doi: 10.1109/ICASSP.1987.1169472.

31. Sha F., Pereira F. (2003), Shallow parsing with conditional random fields, [in:] *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 213–220, doi: 10.3115/1073445.1073473.

32. Sindran F., Mualla F., Haderlein T., Daqrouq K., Nöth E. (2016), Rule-based standard Arabic Phonetization at phoneme, allophone, and syllable level, *International Journal of Computational Linguistics (IJCL)*, **7**(2): 23–37.

33. Sînziana M., Iria J. (2011), L1 vs. L2 regularization in text classification when learning from labeled features, [in:] *Proceedings of the 2011 10th International Conference on Machine Learning and Applications*, Vol. 1, pp. 168–171, doi: 10.1109/ICMLA.2011.85.

34. Toutanova K., Klein D., Manning C.D., Singer Y.Y. (2003), Feature-rich part-of-speech tagging with a cyclic dependency network, [in:] *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 252–259, https://www.aclweb.org/anthology/N03-1033.

35. Tsuruoka Y., Tsujii J., Ananiadou S. (2009), Fast full parsing by linear-chain conditional random fields, [in:] *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2009)*, pp. 790–798, https://www.aclweb.org/anthology/E09-1090.

36. Van Coile B. (1991), Inductive learning of pronunciation rules with the Depes system, [in:] *Proceedings of ICASSP 91: The IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 745–748, doi: 10.1109/ICASSP.1991.150448.

37. Wallach H. (2002), *Efficient training of conditional random fields,* Master's Thesis, University of Edinburgh.

38. Wells J.C. (2002), *SAMPA for Arabic*, OrienTel Project, http://www.phon.ucl.ac.uk/home/sampa/arabic.htm.

39. Yvon F. (1996), *Grapheme-to-phoneme conversion using multiple unbounded overlapping chunks*, [in:] *Proceedings of the Conference on New Methods in Natural Language Processing, NeMLaP'96*, pp. 218–228, Ankara, Turkey.