SETIT 2012 IEEE
Sciences of Electronics, Technologies of Information and Telecommunications
Sousse 21-24 March 2012-Tunisia

Transcription of Arabic Names into Latin

Houda SAADANE[*, **], Aurélie ROSSI[**], Christian FLUHR[**], Mathieu GUIDERE[*, ***]

** Laboratoire de linguistique et didactique des langues étrangères et maternelles (LIDILEM)*
Université Stendhal - Grenoble III - 1180, avenue centrale, F-38400 Saint Martin d'Hères
saadane_houda@doctorant.univ-grenoble.fr
Mathieu.Guidere@univ-tlse2.fr

*** Cadège Technologies & Consulting*
32 rue Brancion 75015 Paris
aurelie.rossi@geolsemantics.com
christian.fluhr@geolsemantics.com

**** Université de Toulouse 2*

Abstract: Transcription and transliteration are experiencing significant growth due to the increasingly multilingual Internet and to the exponential needs in the field of cross-lingual information retrieval. This is especially true for finding named entities (names of persons, places, companies, organizations, etc.), but these entities have a plurality of forms, spellings, and transcripts depending on languages and countries. The case of Arabic names illustrates this complex and multifaceted situation. In this article, we will briefly introduce the theoretical and practical difficulties that arise in the transcription and transliteration of Arabic names into Latin characters, as well as possible solutions and processing that can solve these difficulties.
Keywords: transcription, transliteration, Arabic, dialects, proper names.

INTRODUCTION

Transcription consist in replacing each sound or phoneme of a phonological system by a grapheme or a group of graphemes of a writing system, while transliteration consist in replacing each grapheme of a writing system by another grapheme of a group of graphemes of another writing system, regardless of pronunciation.

In the case of transcription, the objective is to reconstruct the original pronunciation using the writing system of the target language. In the case of transliteration, the objective is to represent the original grapheme with the corresponding graphemes of the target language.

To achieve these objectives, there are phonological transcription systems and graphematic standards of transliteration. But these systems and conventional standards are numerous and complex, and the more the source and the target languages differ, the more complex these systems are. Thus, for Arabic, there are several standards of transliteration, including EI (1960), ISO/R 233 (1961), UN (United Nations Group of Experts is Geographical Names, 1972), DIN 31635 (Deutsches Institut für Normung , 1982), ISO 233 (International Organization for Standardization, 1984), and the ALA-LC (America Library Association, 1997) standard. The international scientific community uses two among these standards: DIN-31635 and the standard adopted by the Encyclopedia of Islam (EI).

1. The various aspects of the subject

1.1. Linguistics aspect

But the issues raised by transcription and transliteration mostly concern the relationship between orality and scripturality when switching from one linguistic system to another, and not so much the nature of the adopted convention. Indeed, the oral and the written obey different rules: one uses sound material, the other visual material, and each material has an internal dynamics and constraints.

These constraints include the phenomena of morphological transformation which affect the words differently depending on the nature of their initial letter. Thus, if the word contains the article AL (ال), we

must distinguish between the sun and the moon letters. In the case of the sun letters, the "L" is silent and the letter that follows is doubled in pronunciation and in writing. Conversely, with the moon letters, the "L" of the article is pronounced and the letter that follows is not doubled, neither in pronunciation nor in writing1. But these rules of the Arabic language are not always respected in practice, which we can see in the examples given by [QUN 01] concerning the transliteration of names of Arabic newspapers into the Latin alphabet: النهار(An-Nahar) in Lebanon, الزمان (AZZAMANE) in London, الصباح(AL-SABAH) in Palestine, اليوم(ELYAUM) in Algeria etc...

A similar phenomenon of deviation from the phonological standards of the source language system is observed in the transliteration of the Arabic letter "tied-T" (ة), called “Marbouta” in Arabic. It is only pronounced (as ”t”) in the state of annexation. But again, some transliterations of newspaper names reflect the spelling of the Arabic word and not its actual pronunciation, for example: الثورة(AL-TAWRA) instead of (ATH-THAWRAH).

Therefore, any automated language processing should be preceded by questioning the aspects that may seem obvious at first but which deserve further analysis. It is important to analyze the following aspects prior to automate processing of the phenomenon that interests us:

– Degree of adequacy of the oral transcription (phoneme);

– Degree of adequacy of the transliteration in writing (graphemes);

– Degree of adequacy of the symbolic notation used in practice (social).

1.2. Cognitive aspect

We often forget that phonemes and graphemes can have a symbolic value that is revealed in different ways:

- The psychological reaction of users in response to the symbols used to record the oral or the written (tolerance or rejection);
- The claim of an oral tradition linked to the history or to the values of the source group (for example, the notion of honor associated to a promise in Arabic countries);
- Political pressure concerning the per-eminence of one idiom over another. In our case, it is the per-eminence of the Arabic as the language of the Koran and therefore considered as something sacred over the dialects on one hand and over foreign languages on the other hand.

This means that a language processing specialist should, prior to any processing, take into account of a number of phenomena that do not stand out directly from the NLP, but which must be taken into account in the phase of analysis and modeling. Otherwise there is a risk that the proposed solutions turn out disconnected from the linguistic reality or from social demand.

Thus, as far as the Arabic languages is concerned, a system of proper names transcription cannot be developed without taking into account the morphological characteristics and the symbolic value of the naming system [ROM 07]. For example, a great number of Arabic first names are formed by using one of the names of Allah (there are 99). The combination of these formations is not random but strictly regulated. This implies an internal coherence during transcription or transliteration, as well as the definition of a certain number of contextual rules that take into account this aspect which is sometimes symbolically important for the named person.

Let us take as example the Arabic name "عبد الله", the first name of the King of Jordan and the King of Saudi Arabia. This name is very common in Arabic and can be transliterated in several ways depending on whether the writer wishes or not to show the original meaning of the name where "Abd" literally means "Servant of". Thus we find: {Abdallah, AbdAllah, Abd Allah, Abd-Allah, Abdullah etc. The same phenomenon can be observed for other names including one of the 99 names of Allah, such as Abdelkader "عبد القادر" or Abderrahim "عبد الرحيم".

In some cases, this symbolic aspect constitutes the heart of the naming system. Indeed, in organizations like Al-Qaïda, it has been demonstrated by [GUI 06] that the "war name" of each member – for instance, that of Abou Moussab Al-Zarqawi – was not only symbolic but also had a strategic value2.

1.3. Dialectological aspect

This symbolic aspect concerning the meaning of the name is accompanied by a phonological aspect that should be taken into account for any automated language processing and which depends on the linguistic situation of the Arab world. Indeed, the Arabic language today is characterized by a complex polyglossia [DIC 09]. Thus, there is a number of works written in literal Arabic (classical, modern, average, etc.) and a great number of dialects (different

[1]For more information, go to http://en.wikipedia.org/wiki/Sun_and_moon_letters

[2]Guidère (2006) relies on the morphological and the semantic analysis of the naming system used by the islamist organizations to deduce what the intentions of the named person are and the strategic value of the name in the context. You can see his article at the following address : http://www.c4ads.org/files/defense_concepts_I.3.pdf

kinds of Arab, regional or local) whose dominant features are noticeable to Arab-speaking people3.

From this perspective, Arabic dialectology comprises two big families of dialects: the Maghreb (Morocco, Algeria, Tunisia and Libya) and the Mashriq (Egypt, Syria and Middle East). But within these families of "geolects", there are many national dialects (natiolects) as well as regional (regiolects) or local dialects ("topolects"), spoken in a limited area (village, city).

Thus, when one wants to develop a transcriber / transliterator of Arabic names, one is confronted with the phonological features of these dialects, as the native speaker pronounces he same name differently depending on his dialect and recognizes the dialectal variation depending on the pronunciation he hears. This is even truer because in the course of history each Arabic dialect has been influenced by other languages like French, Italian and Spanish (for the Maghreb) or English (for the Mashriq). This leads to the fact that the same name or first name in Arabic can have several different pronunciations in different dialects and different transliterations depending on the specific phonological and graphematic features of the target language.

For example, the name of the Libyan leader (Gaddafi), which has a single spelling in Arabic (معمر القذافي) but several pronunciations and accents depending on the dialect, is transcribed into Latin script by over 60 different forms, including: Muammar Qaddafi, Mo'ammar Gadhafi, Muammer Kaddafi, Moammar El Kadhafi, Muammar Gadafi, Moamer El Kazzafi, Mu'ammar al-Qadhdhafi, Mu'amar Qadafi, Muammar Gheddafi, Mu'ammar Al Qathafi, Muʿammar Al-Qaḏâfî…

This multiplicity of forms causes numerous problems both in finding information about a named entity (in this case, the name of a political leader) and for interlingual enrichment of data on a particular topic (e.g. Tripoli in Libya). Indeed, the failure to list all the existing forms of the same name can be detrimental to the effectiveness of a search.

In this study, after an overview of the state of the art, we shall present the transliteration system that we propose for the conversion of the Arabic names from the Arabic into the Latin alphabet. Further we shall present the opposite case that is the conversion of names written in Latin alphabet into the Arabic alphabet. In doing so, we shall explain the system of spelling variants generation in both directions: from the Latin alphabet into the Arabic characters and from Arabic into Latin. Finally, we shall present the methodology used to validate the results and the prospects of the system development.

[3]See our research on the perception of the dialectal changes in the Arabic language (2010) based on samples taken from the media and representing polyglossia of the Arab world.

2. State of the art

Linguistic literature contains many articles dealing with the problem of transliteration, trying to assign a single transliteration to a given name. A generative model for transcription of English names written with Japanese characters (Katakana) into the Latin writing system was proposed by [KNI 97]. This approach was adapted by [STA 98] to the way of transcription into English of an English name written in Arabic. This technique has its limits when it comes to pronunciations unknown to the training dictionary. A statistical technique of transliteration and evaluation of English names into Arabic was suggested by [ABD 03]. It is based on considering the most likely form as the correct one, but this is not true in all the countries. To avoid the difficulty of pronunciation and the problem of dialects, [ALG 05] proposed a system for transliteration of vowelized Arabic names into English. This system is based on a dictionary of Arabic names where the pronunciation is adjusted by adding vowels to the listed names, with an indication of their equivalent in English spelling. But this approach has the disadvantages of the previous two: not only it does not take into account the pronunciation that is not listed in the dictionary, but it is prescriptive because it only offers one possible transliteration for a given name. We feel that the intention of the author is to promote the adoption of a transliteration standard, but this cannot be the result of an individual and isolated initiative.

In reality, the current state of research in this area does not cover the complexity of the problem of transcription and transliteration, which affects both orality and scripturality in two or more linguistic systems at the same time. In fact, transcribing a name or a first name of a source language to a target writing system is a delicate task which requires a certain number of operations requiring consideration of a set of morphological, phonological and semantic properties. These operations are necessary to ensure a correct process of transliteration, especially for security or identity verification applications, or for finding information on the Internet. However, there is now virtually no study in NLP that takes into account the relation:

- between compared phonology and interlanguage transcription;
- between compared graphematics and multilingual transliteration;
- between Arabic dialectology and Latin transliteration systems.

The current goal of our research is to provide an automatic transliteration system that, for the transcription of Arabic names and first names to the

Latin alphabet, takes into account all the aspects listed above, namely the relation between phonology, graphematics and dialectology. To do this, we define a number of rules on the basis of an experimental study, which reflect the complexity of the area.

Below we explain the main steps that allowed us the development of such a system for the module of transliteration of a standard grapheme from the Arabic alphabet into Latin and vice versa.

3. Transcription of Arabic names into Latin script

A peculiarity of the Arabic language which seems interesting and challenging at the same time is the absence of short vowels in Arabic texts, which is unusual for other languages and leads to many different ways of pronouncing the name.

3.1. Methodology of construction of the transliterator

Some Arabic letters are transcribed into figures.

Table1. *Transliteration of Arabic characters into numbers*

Arabic letter	Representation as number
ء	2
ح	7
خ	7'
خ	5
ص	9
ض	9'
ط	6
ظ	6'
ع	3
ع	3'
ق	8 ou 9

This transliteration constitutes the norm in SMS messaging in Europe and the Middle East. This is very useful for the understanding of the social background of the writing person and the geographical origin of the extracted data (geolocation). The table above sums up these special numbers.

3.2. The reasons for the ambiguity in the transliteration of Arabic names

There are historical and linguistic reasons of ambiguity in transcription:

– Numerous Western civilizations that have crossed the Arab countries and the colonization that came along justify the presence of a wide range of arabization, “linguistic miscegenation”, giving rise to a dialectal Arabic varying significantly from region to region.

– To present all the scenarios of transliteration of an Arabic name in the Latin alphabet, one should base oneself on the phonetic system of the literal Arabic as well as on the majority of dialect families (Algeria, Saudi Arabia, Egypt, Morocco, Tunisia etc.), taking into account the numerous regional and local variations, because the pronunciation of certain letters and vocalization of certain words change from one province or village to another.

– Languages are characterized by different phonetic systems, and alphabets contain more or fewer letters. In this context, some letters have no equivalent in other languages, particularly Arabic letters (ع, ق, خ etc) for which each country has a different transcription.

– The lack of a common standard and of a unified Arab strategy in the field of transliteration has been an obstacle for former non-Arab writers, who had to base themselves on the pronunciation of the vernacular dialects of the speakers for the transcription of Arabic names. If we refer to the writer Lawrence of Arabia in his book of 1926, we notice that the city جدةwas transcribed as « Jeddah », 25 times, « Jidda » 6 times and only once « Jedda » in the same book. These kinds of spelling differ from the name used in Saudi Arabia, which is « Jaddah ». Lawrence of Arabia explained this saying that “Arabic names won't go into English exactly, for their consonants are not the same as ours, and their vowels, like ours, vary from district to district.” [ALS07].

– The Latin writing systems also differ from one language to another. For example, these French letters are not found in English: é, è, ô, à, ù, â, ê, ç, î. [ALB 09]. This implies a problem for transcribing Arabic names, as in general people of the Maghreb are influenced by the French literature while people from the Mashriq4 are influenced by the English literature. Let us take an example of the article (ال). The rules of its assimilation vary from one dialect to another. The people of the Maghreb transcribe it by (**El**) and the people of the Mashriq by (**Al**) with or without a gap or a hyphen between this article and name. The same problem arises for certain letters like the letter (ج) which is transcribed in (Dj) in Algeria (J) or (G) in a part of the Maghreb and Mashriq. The letter (ش) is transcribed (Ch) in the Maghreb (Sh) in Mashriq. All this results in the fact that the same Arabic name can be transcribed in different ways.

– Some letters are transcribed by numbers. This phenomenon is widely used in SMS messages (in Europe and the Middle East) and in emails.

– The existence of different systems of transliteration of Arabic can result in using the same Latin letter to transcribe different Arabic letters (the letter سand the letter صare transcribed with the same letter S, the letters **(ذ، ظ، ض)** with the letters Dh, etc.

– Some of the rules proposed for the transliteration of Arabic names (Naif Arab University for Security

[4]Mashriq is the region of Arabic-speaking countries of the Asian continent together with Egypt and Sudan.

Sciences, 1424 H) is not respected. This concerns the following rules:

- If the word contains the article (ال), we must distinguish between the « sun » and the « moon » letters.

– A part of the non-cultivated population has difficulty in using special characters for the transcription of some Arabic letters. Therefore there are certain limits as to how to use these diacritics and we establish how to pronounce them. Here are some names containing special characters: Mu`ammar, Mabruk, at Ṭulayḥah, Bū, Yaḥyá Ḥammūdah Musṭafá, Ismā`īl, Hâdî.

- An Arabic name can be as simple, like « **صالح** : Salah»**,** or compound, like « **عبد الله**: Abd Allah » that can be transcribed in several ways, with or without hyphens or gaps. The number of results found for the transcription of the name « **رحيم**: Rahim» is less important than for the name « **عبد الرحيم: Abderrahim**». We conclude that the relation between the number of results and number of transcribed names increases as increases the number of words, making the search more complex.
- We would like to remind that modern Arabic uses few short vowels, which leads to a lot of different ways of pronouncing a name. The following examples show how a name can be transcribed when it is vowelled and not: **محمد** without the vowels is transcribed by {Mohamed, Mouhammad, Muhamed, Mhamed, M'Hamed, Muhammad}. However, if the same name is vowelled, there will be fewer results, like **مُحَمَد**{Muhamad, Mouhamad, Mohamad}, or **مْحَمَّدْ** {Mehammad, Mehammade}.

The variety of possible ways of transliteration depending on dialects, countries, education or mode of communication did not facilitate the establishment of rules of transliteration.

3.3. Proposed solutions

Our system consists in automating the process of transliteration of Arabic names in real time. It is based on finite state machines and on the family of Arabic dialects and the literary Arabic.

These hypotheses of transcription can be simplified for search on the Internet because we know that, for example, search engines remove diacritics. This will greatly reduce the number of requests to perform.

At the first stage of our work, we tried to test our transliteration hypotheses using contextual rules and without using a search engine or dictionaries. This module can generate all possible Latin forms from an Arabic proper name. Latin variants cover most Latin languages, including French and English.

```
Function Main Algorithm ( NAME : Words) : String
  Result : String list
  For each input NAME execute
    If (NAME) is non-vowelled Then
      Apply the contextual rules of transliteration.
      Processing of the name into Latin script
      Result ← Weighted list of Latin names
    Else
      Delete short vowels
      Apply the contextual rules of transliteration
      Processing of the name into Latin script.
      Result ← Weighted list of Latin names
    End If
  End For
  Return Result
End
```

Algorithm1. *Transliteration of Arabic names using the Latin alphabet*

3.4. Prototype running

Input: a proper name, either simple or compound, written in Arabic.

Output: a sorted list of Arabic names written in Latin, resulting from the transliteration of the proper name of the input.

The functioning of our approach to transliteration of Arabic names written in Arabic into Latin script is described in Figure1.

a. **Removal of the vowels:** We have chosen to delete all the short vowels from the input name, as the objective of our approach is not only to find the best solution but all possible ways of spelling of a proper name. We remind that the pronunciation of a name can vary from one region to another within the same country.

b. **Transliteration:** When all the short vowels are removed, first the system converts the Arabic letters into the Latin letters according to the of transliteration standards. Then, Latin vowels are added according to the rules of transliteration.

First we look at the form of the input name: is it an exceptional case, is the name preceded by an article or not?

Depending on the form, first we apply the rules to transcribe the part which does not constitute the name, and then we apply the rules for the transcription of names depending on the number of consonants in the name, and in a determined priority order. This phase is broken down the following way:

b.1. Preprocessing: this stage consist in removing the gaps before and after the name if the name is simple. Otherwise, this step allows putting each proper name in one line in order to remove the gaps.

b.2. Processing of exceptional cases: at this stage, proper names that do not follow the general rules of transliteration are processed in a special way. When

the input name is mentioned on the list of exceptions, it is transcribed directly. Among the names on this list are: (Ibn) ابن، (Abd) عبد, (Taha) طه etc.

b.3. Analysis of the article: this stage consists in processing of specific cases such as the articles used with proper names, such as ال(Al) and آل(Aal) that have their own rules of transliteration.

b.4. Transliteration of proper names: after processing the parts of the exceptional names, this stage consists in transcribing the consonants. It also allows the insertion of vowels between the consonants following the rules of transliteration, which constitute the core of our system of transliteration and dependent on well-defined contexts. Everything depends on the number of consonants and long vowels. After the transliteration of all the parts of the name, other rules are applied to concatenate them together. Finally, all the results of the transliteration of the input name are displayed.

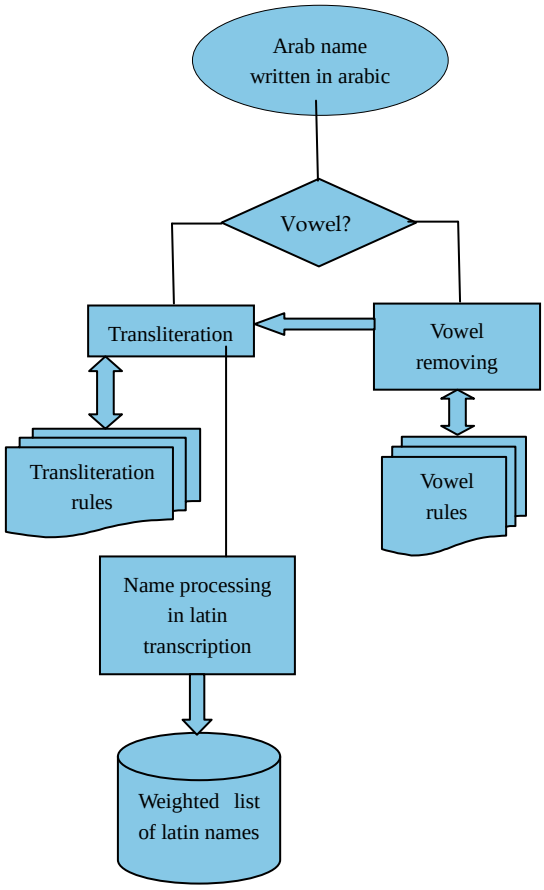

Figure1. *Running organization*

Example: if the word is composed by عبد(Abd)+ ال (Al)+ Name رحيم(Rahim) as the example below :

- We translate عبد
- We translate { ال}
- We translate the name
- We can concatenate the equivalent of the word عبدand { ال} connecting them with a hyphen or a gap. Then we put the equivalent of the word.
- We can also concatenate the two names together.

Table2. *Transliteration of the name* عبد الرحيم

1) Abd Al-Rahim	13) Abd ar-Rahim
2) Abd Al Rahim	14) Abd ar Rahim
3) Abd al-Rahim	15) Abd ar Rahim
4) Abd al Rahim	16) Abdal Rahim
5) Abd El-Rahim	17) Abdarrahim
6) Abd El Rahim	18) *Abdel Rahim*
7) Abd el-Rahim	19) Abderrahim
8) Abd el Rahim	20) Abdar-Rahim
9) Abd Ar-Rahim	21) Abdul Rahim
10) Abd Ar Rahim	22) 'Abd Arrahīm
11) *`Abd Ar-Rahîm*	23) 3abd *ara7im*
12) *`Abd Arrahîm*	*24)* …

c. **Processing of the name in the Latin script**: This stage consists in processing the output name in the Latin script. For example, a capital letter is added at the beginning of the proper name, since Arabic names transcribed in Latin begin with a capital. This notion of the capital letter is preserved for the use in databases. However, this notion is useless in the case of search engines such as Yahoo, Google, etc. that automatically replace capital letters with small ones. As for special characters, they are also deleted in the case of a search engine.

d. **Weighted list of Latin names**: This stage consists in assigning a weight to the rules used to generate the list, so that the output results are presented from the most likely to the least likely, or vice versa. To accomplish this weighting, we use various search engines, marking the number of occurrences for each generated form of the proper name. For example, for the Arabic name جمال, the system generates three distinct transliterations attested in the texts (Djamel, Jamel, Gamel), and the calculation of frequencies gives the following results:

Table3. *Results obtained by Google for transliterations of the name* جمال

Latin transcription	Number of pages found by Google
Djamel	4170000
Jamel	5390000
Gamel	490000

From the weighting point of view, this example shows that the Arabic letter (ج) is transcribed, in terms of frequency, mainly by (J), followed by (Dj) and finally by (G).

This procedure was applied to all the forms of transliteration of Arabic characters. It allowed us to establish a list of equivalents weighted on the level of

graphemes, which is used to display the output results from the most to the least likely.

4. Transcription of Arabic names written in the Latin script into Arabic

After the transliteration of Arabic names into Latin, we are interested in the transliteration of Arabic names written in Latin into the Arabic script. This interest is motivated mainly by two practical reasons. Firstly, it is the need for the Arab speakers to keep the same keyboard when searching for names of Arab people. Secondly, it is the need to facilitate the search of interlanguage information and data collection in multiple languages using Arabic names written in Latin characters.

4.1. Problems in the transcription from Latin to Arabic

We have encountered the following problems :

– In reality, various transliterations of the names do not respect the proposed standard rules of transliteration of Arabic names into Latin. This also influences the transliteration into Arabic script of Arabic names written in Latin. For instance, problems can arise when no distinction is made between the sun and the moon letters concerning the article (ال). To understand it better, let us take a few examples: the name (Azzamane) is transcribed by (**الزمان**). However, the name (Azziz) is transcribed by (**عزيز**). If we take the name (Annahar), it is transcribed by (**النهار**), unlike (Annwar) which is transcribed by (**أنوار**), etc. These examples show that this problem occurs whenever we have a name that begins with the letter (A) followed by a sun letter. Transcription is even more difficult if the the solar (L) is not followed by a gap or a hyphen to distinguish the letters **(أ)** or (**ع**), as it is shown in the previous examples.

– When the letter (a) appears at the end of the word, it can be transcribed by several letters that are **{ة, ى, ا, اء}**. For example, Houda is transcribed by **هدى**; Karima (instead of Karimah) is transcribed by **كريمة** and Wafa (instead of Wafaa) that is transcribed by **وفاء**.

– Transliteration of Arabic names written in Latin script is affected by the phenomenon of « pluriglossia ». We observe that transliteration of Arabic names written in Latin into Arabic script is also characterized by the fact that some Latin letters can be transcribed by different letters in Arabic. Here are some examples:

– The Arabic letter (ث) in Arabic script transcribes one of the following Latin letters: (Th) in Tunisia and Iraq; (Th) or (T) in Algeria, depending on the region; (T) in Morocco; (T) or (S) in the Middle East, depending on the country.

– The Arabic letter (ج) in Arabic script transcribes one of the following Latin letters: (J) in Morocco and Tunisia; (Dj) in Algeria and in some parts of the Middle East; (g) in Egypt. Note that the Egyptian transliteration is a source of ambiguity as it can be confused with the transcription of the dialectal sound (ڤ), which is also transliterated by (g), as well as with the transliteration of the Arabic letter (ق), pronounced the same way as (g) in some dialects.

– The grapheme «aa» is transliterated into Arabic by several letters depending on the country: { ة, ى, ا, اء}. (See the example of the name «Wafaa».

Example1: Ambiguity in Arabic transliteration of Latin consonants

Table4. A*mbiguities in latin transliteration of Arabic vowels*

Nom en écriture latine	Transcription into Arabic
Mu**h**amad	مهمد or مححمد
Houda	حدى or هدى
Hind	حند or هند
Ni**h**ad	نحاد or نهاد
Sa**h**ar	سهر or سحر

Example2: Ambiguities in Arabic transliteration of Latin vowels

Table5. *Results returned for the name Hamid*

حامد	هامد
حميد	هميد
حاميد	هاميد
حمد	همد

Besides the phenomenon of ambiguity, these examples illustrate the lack of a unified strategy for transliteration of Arabic names into Arabic script. Indeed, transliterated forms depend on the customs in each region or an Arab country. They reflect the complex interaction within the general system of the Arabic language, between the literary and the dialectal Arabic on one hand, and between Standard Arabic and foreign languages on the other, French and English in particular.

4.2. The operation of the module transliteration from Latin to Arabic

The transliteration module processes Arabic names written in Latin characters on the input and produces a list of Arabic names transliterated into Arabic script on the output.

The forms generated in Arabic are spelling variations that are not all certified in practice, as some Latin characters can be transliterated by different Arabic characters. But comparing the results to the texts of the corpus can solve this overgeneration and drastically reduce the number of forms in order to retain only the forms relevant for document search.

We have developed two strategies to resolve the ambiguity associated with the multiplicity of forms generated for the same name.

4.2.1. Strategy1: *Disambiguation by the*

consonantal root

In this strategy, the disambiguation is based on the relative position of the constituent letters of the name, by using the root of the word. Indeed, the order of consonants in Arabic roots makes it possible to distinguish two names with the same letters and therefore to generate probable equivalents of each letter of the name depending on its position in the root.

To establish such equivalence, first we remove all short and long vowels (ا, ي, و), if there are any, in order to obtain the consonantal root of the word. Then we create a list of names with respect to the position of the letter the exact transliteration of which we want to carry out. If the letter is considered first in the sequence of consonants, the root is stored in the first list. However, if the letter in question is not in the first position, the root is stored in the second list.

Example1:

Input: Hatim (Arabic name transliterated in English; French variant: Hatem).

The transliteration of the letter H is a source of ambiguity in this case because this name can be transliterated into Arabic script by هاتمor by حاتم.

To resolve this ambiguity, the system first deletes the Arabic letter (ا) to generate the consonantal root of the name (حتم).

Then the system searches in the first list for each root having two (حت) or three (حتم) consonants.

Answer: This root exists in the list.

Output: Accept only the transliteration (حاتم) and reject the transliteration (هاتم).

Example2:

Input: Wahab (Arabic name transliterated in French, variant: Waheb).

The transliteration of the letter H is a source of ambiguity in this case because this name can be transliterated into Arabic script by وهابor by وحاب.

To resolve this ambiguity, the system first deletes the Arabic letters (و) and (ا) in order to generate the consonantal root name (حب). As the name does not start with the letter (H), the system should look for the root (حب) in the second list.

Answer: This root is not in the list.

Output: Accept only the transliteration (وهاب) and reject the transliteration (وحاب).

Limits of the strategy: this strategy of disambiguation by the consonantal root has its limits. In fact, each list has examples of exceptions because for each root, one can find names whose initial letters can be transliterated by two different Arabic letters, while still presenting a relevant result (the generated name perfectly exists but does not match the original).

For example, the name «Houda» may have two relevant Arabic transliterations: (هدى) and (حودة). For such cases,a complementary strategy of disambiguation had to be developed.

4.2.2. Strategy2: *Disambiguation by the phonological context*

This solution is based on the identification of the phonological context of the letters to transliterate.

Inferred rules make it possible to distinguish between two graphemes for the same original letter.

Let us take the aforementioned example of the letter H which can be transliterated into Arabic by the letters (ه) or (ح). The rules deduced from an analysis of the phonological context include the following:

– If the phonological sequence is the letter {h, s}, the phoneme is transliterated into Arabic by the grapheme (ح). For instance: Houssin : حسين, Houssam : حسام

– If the phonological sequence is {h... i / ee / y /... C}, the phoneme (h) is transliterated into Arabic by the grapheme (ح), for example: Hamid : حميد, Habiba : حبيبة

– If the phonological sequence is {S / M ... h}, the phoneme (h) is transliterated into Arabic by the grapheme (ح), like: Masbah : مصباح, Samah : سماح

Limits of the strategy: despite its strengths, this strategy of disambiguation through the phonological context has its limits. Indeed, the large number of rules necessary for the disambiguation makes the system somewhat cumbersome and slow in the processing phase, especially on large volumes of data. Thus, if the application of contextual rules does improve the functioning of the system, it does not eliminate all the forms of ambiguity. This is mainly due to the fact that there is considerable variation in the transliteration of Arabic names even between languages using the same Latin alphabet.

5. Transcription of Arabic names written in Latin script into Latin

As the transliteration system is interlingual, it also transliterates Arabic names written in Latin to other Latin alphabets. Below you will find an overview of the general functioning of the latin → latin transliteration module.

Input: Arabic proper name written in Latin

Output: sorted list of Arabic names written in Latin, resulting from the double transliteration of the input name.

General Description: This module can generate all possible Latin spelling variants of an Arabic proper name written in Latin. For example, this lets you make queries on the Internet concerning the Arabic proper name « Oussama Ben Laden », and to find not only the documents containing « Oussama Ben Laden » but also those containing other forms of the same

name, such as: Ussama Bin Laden, Osama Ben Laden, Usama bin Ladin, etc.

As for the processing, it is a double transliteration, for the input proper name is first transliterated into Arabic and then, the generated Arabic forms are transliterated into Latin characters. Latin variants cover as many languages as possible, but are currently based on the forms that are most frequently observed in English, French, Spanish and Italian.

This processing is modular as it is possible to generate only those variants that are useful for searching on the Internet (without capital letters or diacritics) or variants useful for querying a database (therefore, with capital letters and diacritics, but without special characters in the transliteration of names). Finally, one can generate an exhaustive list of variants including capital letters, diacritics and special characters in the directory of the possible transliterations.

6. Validation of results

At this stage of the development of the system, the process of the validation of the hypotheses and of the results was primarily carried out using Internet search engines.

Thanks to the developed strategies, the search engine can find all the relevant documents, and that no matter what the language of the original document and the form of the proper name are. Thanks to the developed disambiguation procedures, the system can also confirm whether or not the person in question is the person searched for.

The process of finding the existence of the transliteration by a search engine (e.g. Google) can be described as follows:

```
Function Rech-Google (NAMES: Words): Integer
  Result: Integer
  Connecting to the remote machine www.google.fr, in
  socket S.
  Sending a query (sequence of words) through
  socket S.
  Reading the webpage returned by S in the buffer B.
    If (the string "No result found" exists in B) Then
      Result ← 0;
    Else
      Result ← N: the number of Web pages sent back
      by Google in buffer B
    End If
  Return Result (web pages)
End
```

Algorithm2. *Search of transliterations on google*

This verification concerns the correlation between the transliterated word (query) and the documents retrieved for that name. Transliteration is considered relevant if the result of the queries for the transliterated form is not null, and if the search engine can find at least one answer for each transliterated form concerning the same person.

Let us examine the example of a query on the name of theAlgerian president.

Input name: بوتفليقة

Output: transliterated forms marked in bold in the documents retrieved in a dozen different languages.

1. KING SADDRESS AT ARAB SUMMIT IN ALGIERS-By:IMRA
algiers, march 23 (petra-jordan news agency)--his majesty king abdullah ii said that the roadmap peace plan is the only available means to settle the palestinian ... thanks and appreciation for his excellency president abdul aziz **botafliqah** and to the algerian people for their kind hospitality ... israpost.com/Community/articles/show.php?articleID=5361>Cached

2. The Angry Arab News Service " As'ad
the angry arab news service. a source on politics, war, the middle east, arabic poetry, and art by as'ad abukhalil **...** posted by as'ad at 6:52 am 04/10/09. **butufliqa** wins. The **algerian** president wins re-election with 99.99% of the **...** angryarab.net/author/falastin>Cached

3. كونا: Arab League congratulates Algeria"s Boutfalika...
cairo, april 11 (kuna) -- the arab league congratulated saturday president abdelaziz boutaflika for his win in the **algerian** presidential elections, hoping that he would continue the development process in the north african nation. arab league kuna.net.kw/NewsAgenciesPublicSite/...&Language=en>Cached

4. Times of Oman : it comes in implementation of the directives of his majesty sultan qaboos bin said and president abdulaziz **boutafliqah** aimed at cementing bilateral relations. later at a press conference, alawi said that the two sides signed a number of agreements and mous**....**
timesofoman.com/innercat.asp?detail=33983>Cached

5. Abdelaziz Bouteflika - Wikipedia, the free encyclopedia
bouteflika lived and studied in **algeria** until he joined the front de libération nationale **...** on boumédienne's unexpected death in 1978, **bouteflika** was seen as one of the two main **...** en.wikipedia.org/wiki/Abdelaziz_Bouteflika>Cached

6. Maliki : If Sultan Hsahim is not executed I will resign
kurdish aspect covers issues related to kurds and kurdistan within the larger context of middle eastern concerns. the website offers readers a **...** he revealed to

them that in the opec meeting the **algerian** prime minister abd-al-aziz **botafliqa** had asked him: are you from an **iranian** origin? **...** kurdishaspect.com/doc020208AWENE.html>Cached

7. ...النهار الجديد- ثورة في عالم الإعلام - لأول مرة ..أسرار عن
لأول مرة ..أسرار عن بوتفليقة الرجل. **...** i love you **boutafli8a**. 0. أضف تعليقك. اسمك: أضف 66 - 1 | عرض: 66المجموع: . تعليقاتك: اقرأ أيضا فى: الوطني. وزارة التربية الوطنية تلغي التسجيل في بكالوريا النظام القديم. المتهم الرئيسي في مقتل سارة يواجه تهم الخلوة**...**
ennaharonline.com/ar/national/29309.html>Cached

8. Butaflika fires Benflis, brings back Oyahya, due to Algerian ...
butaflika fires benflis, brings back oyahya, due to algerian presidential elections, algeria, politics. arabicnews.com - your source for daily news about the arabic world. **...** **algerian** news agency quoted benflis as saying after a meeting with **butfalika** that he did not take part in taking the decision of **...**

9. YouTube - viva l'algerie , algeria is back HMD algeria mon amour
algeria is back no matter what **algeria** is still standing ya rab hamdolah , maghrab united ya **botaflika** , les marocain khawatna toujour m3a jazair youtube.com/watch?v=IAF1AOmnQlM>Cached

10. The Angry Arab News Service " As'ad
the angry arab news service. a source on politics, war, the middle east, arabic poetry, and art by as'ad abukhalil **...** posted by as'ad at 6:52 am 04/10/09. **butufliqa** wins. the **algerian** president wins re-election with 99.99% of the **...** angryarab.net/author/falastin>Cached

11. Saylac | Somalia News and Information
botofliqa ayaa lagu soo warramayaa inuu si weyn ugala dooday siyuu masfen ciidamada itoobiyaanka ah ee soo **...** mr: musfin waxa uu intaasi ku daray inuu guddoonsiiyay madaxweynaha dalka **algeria** c/caziiz **botofliqa** farriin qoraal ah oo uu uga siday ra'isul wasaaraha dowladda itoobiya melas zenawi in **...** saylac.com/news/warJan1907.htm>Cached

12. Boutifliqa the only candidate for Algerian elections
earlier, **boutafliqa** opposed the participation of foreign observers in the elections in order to investigate the "honesty" of any voting process **...** in an interview with **french** television, **boutafliqa** said: "i am a committed nationalist, **...** arabicnews.com/ansub/Daily/Day/990415/1999041510.html>Cached

13. Answers.com - Algeria Questions including "Do you need a visa ...
algeria questions including "do you need a visa to go to algeria" and "approximately how far apart are the capital of us and algeria" **...** abd al-aziz **botafliqa** عبد العزيز بوتفليقةabdelaziz bouteflika is the president of **algeria**, having taken the... wiki.answers.com/Q/FAQ/2837>Cached

14. ...النهار الجديد- ثورة في عالم الإعلام - نقل ابنة يزيد زرهوني
نقل ابنة يزيد زرهوني للعلاج بالخارج في حالة خطيرة **...** et la bonne santé et je dit aussi mabrouk à tous les algeriens et les algeriennes avec cette victoir ce qui est .**boutoflika** .rabi yoslah halo **...**

15. 葡萄牙每周信息(2006年9月15-22日) 15日席尔瓦总统分别接受中国、斯洛文尼亚、爱沙尼亚、澳大利亚、德国常驻葡大使及利比亚、卢旺达、赞比亚、黑山、菲律宾、马其顿等非常驻葡大使递交的国书。 阿尔及利亚总统boutoflika致信葡总统席尔瓦,肯定两国拥有的共同战略视角,愿加强两国政治协调,推动双方相互合作更加深入。 16日 社会党和社民党赞成明年1月全国就堕胎合法化进行全民公决。 葡海关和特别税总局透露,在两天的缉毒行动中,共逮捕3名毒贩并缴获34公斤海洛因。
财长表示,为使2007年财赤降至占gdp的3.7%,明年将继续紧缩经济,削减财政开支5%,达23.9亿欧元,占gdp的1.5 ...
pt.china-embassy.org/chn/ptyshx/t273560.htm>Cached

16. maliweb.net :: De la rébellion au terrorisme : Ibrahim... il existe bien une jonction entre le bandit ibrahim bahanga et le réseau al qaïda par le biais du groupe salafiste pour la **...** bahanga avait exécuté le chef des salafistes sur ordre de **boutouflika**. **..**

7. Conclusion

In this study, we explained the framework and the methodology that we used to develop a system of transliteration of Arabic names written in Latin script into Arabic script and vice versa. The system generates all the possible forms of spelling for a name using regular expressions and contextual rules.

It is possible to improve the system of transliteration from Latin to Arabic by refining the obtained results via calculating the frequency of the forms found in each considered language. But this requires the extension of the language coverage of the system, including analysis of non-Arabic names.

In the immediate future, we plan to focus our research on geolocalized transliteration in order to answer the question of how the different transliterations can provide information on the origin and / or the profile of the person who is using them (English or French, coming from the Maghreb or the Mashriq, north or south, etc.) This focus is explained by both the urgent industrial demand and the interest that this research represents.

The next step is to extend the system of transliteration on to names of non-Arab origin.

ACKNOWLEDGEMENT

The author acknowledges the support of the French National Agency of Research ANR: Agence Nationale de la Recherche, referenced by ANR-09-CSOSG-08-01.

REFERENCES

[ABD 03] N. AbdulJaleel, L. Larkey. “Statistical Transliteration for English-Arabic Cross Language Information Retrieval”, *Proceedings of the Twelfth ACM International Conference on Information and Knowledge Management*, New Orleans, LA, 3-8 Novembre 2003, New York, p. 139-146.

[ALB 09] I. Al-Balawi, A. Al-Banyan. “Différentes façons d'écrire les noms arabes en latin : formes et raisons”, Research Report of King Saud. (In Arabic)

[ALG 05] M. Alghamdi. “Algorithms for Romanizing Arabic names”, Journal of King Saud University: Computer Sciences and Information. , Riyadh, 2005, n° 17, 2005, p. 1-27.

[ALS 07] A. Alsalman, M. Alghamdi, K. Alhuqayl , S. Alsubai. “Romanization System for Arabic Names”, The First International Symposium on Computer and Arabic Language (ISCAL – 07), Riyadh, Novembre 2007, p. 214-227.

[DIC 09] J. Dichy. “La polyglossie de l'arabe illustrée par deux corpus”, In M. Bozdemir et L.-J. Calvet (eds), Politiques linguistiques en Méditerranée, Paris: Honoré Champion, 85–102.

[GUI 06] M. Guidère. “Al-Qaeda's Noms de Guerre : How Should We Decode Terrorists' Names?”, In Defense concepts, CADS Press, Vol. 1, Edition 3, Fall 2006, 6-16.

[KNI 97] K. Knight, J. Graehl. “Machine transliteration”, Journal version Computational linguistics, 24(4), 1997, p.599-612.

[QUN 01] H. Qunaiir “Romanizing Arabic names”, Journal Ar-Riyadh, Riyadh, 2001, article n° 12314. (In Arabic)

[ROM 07] A. Roman. “la création lexicale en arabe : étude diachronique et synchronique des sons et des formes de la langue arabe”, Lyon, Presses universitaires de Lyon, 2001, éd. revue et augmentée 2007.

[STA 98] B. Stalls, K. Knight, “Translating names and technical terms in Arabic text”, Proceedings of the COLING/ACL Workshop on Computational Approaches to Semitic Languages, 1998, Montreal, Quebec.

[WAT 01] For journal Al-Watan, Kingdom of Saudi Arabia, n° 407, 2001. (In Arabic)