Segmentation and tagging of vocalized Arabic texts

Djamel Eddine Kouloughli[1]

1. Preliminary remarks

In a language with rich morphology, like written Arabic, part of the morpho-syntactic information coded in a text is expressed in the form of morphological information such as inflexional morphology, case marking, etc. When an Arabic text is completely vocalized, this morphologically coded information is largely available and accessible to analysis on the level of the isolated word (in sharp contrast with strongly analytical languages in which most of the syntactic information is scattered at a more global level : word order in the sentence, presence of independent grammatical morphemes around the word, and so on. In Arabic it is consequently feasible to automatically segment (most) words into their morphological component parts without needing to take into account more global levels.

Apart from being a language with rich morphology, Arabic is also a strongly "agglutinative" language, i.e. a language where many words are formed by joining morphemes together. This is, in part, a complicating factor for the automatic segmentation (and analysis) of words. But on the other hand, considering that the agglutination process obeys a small number of linguistic constraints, the segmentation of agglutinated morphemes is not too difficult to manage. However this situation creates in some cases ambiguities in the analysis. I will come back to this point later.

For example a word like *fasayaktubuhu* may easily be analyzed thus[2] :

fasayaktubu++hu
fa++sayaktubu++hu
fa++sa++yaktubu++hu
fa++sa++yaktub+u++hu

The segmentation of proclitics (clitics that may appear before a host word) is driven by the fact that only CV morphemes or the article "*al-*" are allowed in this position, which gives hardly more than a half-dozen possible candidates.

As to the segmentation of enclitics (clitics that may appear after a host word) it is driven by the fact that only the so-called "suffix pronouns" are allo-

1. Histoire des théories linguistiques (HTL), UMR 7597, CNRS.

2. Each line shows a step in the process of segmentation. The analysis stops at the level of the word base. We will see that, for our purpose, it is not necessary to go further deep (i.e. to the level of roots and patterns).

wed in this position, and since they belong to a closed list, it is easy to scan the end of the word to see if any of these pronouns is cliticized[3].

Finally there is the question of "real" suffixes, that is the morphological markers of different grammatical morphemes such as case and mood, which, again, belong to a rather limited set : essentially the three short vowels and the –na/-ni markers of indicative mood after long wowels.

2. THE *DEKUP* SEGMENTATOR

On the basis of these very simple principles it is quite feasible to contrive a system capable of (semi)automatically segmenting entirely vocalized Arabic texts. It is along these lines that the program *Dekup* was devised. Technically this program consists in a context-free grammar resting on the very simple idea that, considering any fully vocalized Arabic word, whatever is not analyzable as a proclitic at the beginning of that word, and an enclitic or suffix at its end, belongs to the word-base. In the output of *Dekup* clitics are separated from the rest of the word by a double plus sign "++", and suffixes by a simple plus sign "+".

Of course such a program, based only on the "formal" identification of words, without any morphological or syntactic knowledge, will sometimes make mistakes and propose wrong segmentations, which will have to be corrected by the human editor. If these errors appear to be predictable, some improvements may be introduced into the system of segmentation. In the *Dekup* system, when a tentative segmentation presents a high degree of uncertainty the system adds a special sign "%" to the segmentation marks, in order to attract the editor's attention to that dubious segmentation. Admittedly, there are cases where the system does not identify a segmentation as dubious and yet it turns out to be mistaken ! In any case, all erroneous segmentations will have to be eventually spotted and corrected by a human editor, and in the "dubious but finally correct" case the caution symbol will have to be removed. Our evaluations suggest that the system generates no more than an average 5% of mistaken segmentations.

As to the inconvenience of having to supervise and possibly to correct manually the output it is counter-balanced, at least up to a point, by the very high speed of processing of texts by *Dekup* : a corpus of about 200 000 words is segmented in a matter of seconds.

One last remark: *Dekup* works on transliterated and fully vocalized text. Our system of transliteration, called TRS, was devised at a time when standard computers could not display normal Arabic text, and the specialized software proposed to perform this task was extremely expansive, a state of affairs which

3. In classical Arabic more than one pronoun may be cliticized to a word, but then in a constrained sequence : *'a'ṭā-nī-hu* but NOT *'a'ṭā-hu-nī*. This again facilitates the segmentation.

drove a number of researchers in the field of Arabic NLP to work in transliteration[4].

This said, we may have a look at the way *Dekup* works and at its output. Here is a fragment of newspaper text[5] first in Arabic :

فِي عامِ 70 ماتَ عَبْد الْناصِر. وَنَظَرْتُ إِلَى ابْنَتِي، إِلَى الْمُسْتَقْبَلِ. وَلَمْ أَمْلِكْ نَفْسِي مِنْ الْحَسْرَةِ عَلَيْهاَ. شَعَرْتُ أَنَّهاَ أَصْبَحَتْ يَتِيمَةً رَغْمَ أَنَّهاَ تَعِيشُ وَسَطَ حَنانِ الْوالِدَيْنِ؛ وَأَحْسَسْتُ أَنَّهاَ أَضْحَتْ بِلاَ أَب، رَغْمَ أَنَّني أَبُوهاَ وَماَ زِلْتُ إِلَى جانِبهاَ. لَكِنَّهاَ وَجِيلَهاَ كُلَّهُ أَصْبَحاَ بِلاَ أَب؛ فَالأَبُ الْحَقِيقِيُّ هُوَ الْزَعِيمُ وَالْقائِدُ وَمُخَطِّطُ الْمَسِيرَةِ. وَجِيلُهاَ كُلُّهُ أَصْبَحَ جِيلًا ضائِعًا بِلاَ زَعامَةٍ، وَبِلاَ هَدَفٍ يَجْتَمِعُ حَوْلَهُ، وَبِلاَ رَمْزٍ يَأْتَلِفُ عَلَيْهِ، وَبِلاَ رُوحٍ يَسْتَمِدُّ مِنْهاَ الْطُمُوحَ وَالْتَفاؤُلَ وَالإِيْمانَ.

and then transcribed in TRS :

```
fiy &aami 70 maata &abd elnaaSir. wanaZartu 'ilaY ebnatiy,
'ilaY elmustaqbali. walam 'amlik nafsiy min elHasra#i &alayhaa.
$a&artu 'an*ahaa 'aSbaHat yatiyma#aN ragma 'an*ahaa ta&iy$u wasaTa
Hanaani elwaalidayni; wa'aHsastu 'an*ahaa 'aDHat bilaa 'abiN,
ragma 'an*aniy 'abuwhaa wamaa ziltu 'ilaY jaanibihaa. laakin*ahaa
wajiylahaa kul*ahu 'aSbaHaa bilaa 'abiN; fael'abu elHaqiyqiy*u
huwa elza&iymu waelqaaIidu wamuxaT*iTu elmasiyra#i. wajiyluhaa
kul*uhu 'aSbaHa jiylaN DaaIi&aN bilaa za&aama#iN, wabilaa hadafiN
yajtami&u Hawlahu, wabilaa ramziN ya'talifu &alayhi, wabilaa
ruwHiN yastamid*u minhaa elTumuwHa waeltafaaUula wael'iymaana.
```

And here is the out-put of the segmentation of this text by *Dekup*:

4. Our system resembles in many respects the one designed independently by T. Buckwalter (cf. Habash, Soudi and Buckwalter, 2007). This likeness is probably due to the fact that the two systems had the same basic objectives, among which the ease of reading is crucial. There are however some marked differences between them, notably that TRS has no symbol for *sukūn* which linguistically represents nothing and which can be generated, whenever needed, by a very simple context analysis. For more information on the requirements of good transcription systems for Arabic and the TRS system, cf. Kouloughli, 2004-2009.

5. This fragment is from a completely vocalized (and transcribed) corpus of newspaper extracts produced by Adnan Aljuburiy in the mid-eighties. I reeived a copy of this text (in transcription) by courtesy of Everhard Ditters, Nijmegen University. I am responsible for the conversion in TRS and hence to Arabic script.

fiy &aam+i 70 maat+a &abd el++naaSir. wa++%naZart+u 'ilaY ebn+at++iy,
'ilaY el++mustaqbal+i. wa++%lam 'amlik nafs++iy min el++Hasra#+i &alay++haa.
$a&art+u 'an*a++haa 'aSbaH+at yatiyma#+aN ragm+a 'an*a++haa ta&iy$+u wa++%saT+a
Hanaan+i el++waalid+ayni; wa++%'aHsast+u 'an*a++haa 'aDH+at bi++laa 'ab+iN,
ragm+a 'an*a++niy 'abuw++haa wa++%maa zilt+u 'ilaY jaanib+i++haa. laakina++haa
wa++%jiyl+a++haa kul*+a++hu 'aSbaH+aa bi++laa 'ab+iN; fa++el++'ab+u el++Haqiyqiy*+u
huwa el++za&iym+u wa++el++qaaIid+u wa++%muxaT*iT+u el++masiyra#+i. wa++%jiyl+u++haa
kul*+u++hu 'aSbaH+a jiyl+aN DaaIi&+aN bi++laa za&aama#+iN, wa++%bi++laa hadaf+iN
yajtami&+u Hawla++hu, wa++%bi++laa ramz+iN ya'talif+u &alay++hi, wa++%bi++laa
ruwH+iN yastamid*+u min++haa el++TumuwH+a wa++el++%tafaaUul+a wa++el++'iymaan+a.

As we said earlier, suffixes are separated from the base-words by "+", and clitics from the base-word or suffixes by "++". Remember too that dubious cases are marked by "%". Let's analyse those dubious cases, highlighted below in green if, after control, they turned out to be good guesses, and in yellow if not.

fiy &aam+i 70 maat+a &abd el++naaSir. wa++%naZart+u 'ilaY ebn+at++iy,
'ilaY el++mustaqbal+i. wa++%lam 'amlik nafs++iy min el++Hasra#+i &alay++haa.
$a&art+u 'an*a++haa 'aSbaH+at yatiyma#+aN ragm+a 'an*a++haa ta&iy$+u wa++%saT+a
Hanaan+i el++waalid+ayni; wa++%'aHsast+u 'an*a++haa 'aDH+at bi++laa 'ab+iN,
ragm+a 'an*a++niy 'abuw++haa wa++%maa zilt+u 'ilaY jaanib+i++haa. laakina++haa
wa++%jiyl+a++haa kul*+a++hu 'aSbaH+aa bi++laa 'ab+iN; fa++el++'ab+u el++Haqiyqiy*+u
huwa el++za&iym+u wa++el++qaaIid+u wa++%muxaT*iT+u el++masiyra#+i. wa++%jiyl+u++haa
kul*+u++hu 'aSbaH+a jiyl+aN DaaIi&+aN bi++laa za&aama#+iN, wa++%bi++laa hadaf+iN
yajtami&+u Hawla++hu, wa++%bi++laa ramz+iN ya'talif+u &alay++hi, wa++%bi++laa
ruwH+iN yastamid*+u min++haa el++TumuwH+a wa++el++%tafaaUul+a wa++el++'iymaan+a.

We first notice that all cases of dubious segmentation involve "wa" in initial position, except the last which marks the boudary between the definite article "el" and the following word. The questions one might ask is why are things that way ? The answer is that the program *Dekup* working without dictionary does not know whether an initial "wa" is the cliticized coordination particle or if it is an integral part of the word. So it does insert a clitic boundary "++" but indicates (with a "%") that it is in doubt about the validity of this decision.

In most cases (in this example !) *Dekup*'s guesses are correct, but in one case (highlighted in yellow) the doubt turns out to be justified since the correct analysis is "wasaT+a". Notice that in three cases, namely in line 7 "wa++el+qaaIidu", and in line 10 "wa++el+tafaaUula" and "wa++el+'iymaan+a", *Dekup* did not insert the "%" mark after "wa". This is because, in this configuration, the solution is sure to be the one proposed.

Notice moreover that in the case of "wa++el+tafaaUula" *Dekup* did insert a "%", but it is between the definite article "el" and the following word. Again this is because lacking lexical knowledge, the system cannot be sure whether the sequence of segments "elt..." consists of the article plus a noun beginning with "t" or belongs to a single word as in "eltafata" (he turned).

3. THE *ETIKET* TAGGER

Now that we have an idea of the way *Dekup* performs text segmentation, let's turn to the program *Etiket* designed to use *Dekup*'s output to perform parts of speech tagging on it. Here again, considering that the whole process of analysis is conducted on vocalized texts, the output of segmentation gives a considerable number of formal clues allowing in many cases correct guesses of the POS tags to be associated with each segmented word.

To illustrate this, let us consider the following four paragraphs, extracted from the Arabic translation[6] of a well known fairy tale, "the tale of the sleeping beauty" or "qiSSa# al'amiyra# elnaaIima#".

kaan+a fiy qadiym+i el++zamaan+i malik+uN wa++malika#+uN ya&iy$+aa+ni fiy qaSr+i+himaa el++jamiyl+i &iy$a#+a hanaa'a#+iN wa++sa&aada#+iN. laakin*+a $ayI+aN waaHid+aN kaan+a yuHzin+u++humaa, wa++huwa 'an*a++hu lam yakun lahum+aa walad+uN.

wa++%kam e$tahay+aa 'an yakuwn+a la++humaa wa++lad+uN! wa++maa mar*+a yawm+uN 'il*aa rad*ad+aa fiy++hi haa@ihi el++jumla#+a: "'aah+iN yaa layta+naa nurzaq+u wa++lad+aN!".

fa++fiy 'aHad+i el++'ay*aam+i, baynamaa kaan+at el++malika#+u tastajim*+u, ra'+at Difda&a#+aN taxruj+u min el++maa'+I wa++tukal*im+u++haa qaaIila#+aN: "laa taHzan+iy, &am*+aa qaliyl+iN turzaq+iyna Tifla#+aN!"

fa++%riH+at el++malika#+u fa++%raH+aN &aZiym+aN, wa++%@ahab+at musri&a#+aN 'ilaY zawj+i++haa el++mal+i++ki, fa++%raw+at la++hu el++xabar+a.

The text segmented by *Dekup* and manually corrected looks like the following:

kaan+a fiy qadiym+i el++zamaan+i malik+uN wa++malika#+uN ya&iy$+aa+ni fiy qaSr+i+himaa el++jamiyl+i &iy$a#+a hanaa'a#+iN wa++sa&aada#+iN. laakin*+a $ayI+aN waaHid+aN kaan+a yuHzin+u++humaa, wa++huwa 'an*a++hu lam yakun la++humaa walad+uN.

wa++kam e$tahay+aa 'an yakuwn+a la++humaa walad+uN! wa++maa mar*+a yawm+uN 'il*aa rad*ad+aa fiy++hi haa@ihi el++jumla#+a: "'aah+iN yaa layta++naa nurzaq+u walad+aN!".

fa++fiy 'aHad+i el++'ay*aam+i, baynamaa kaan+at el++malika#+u tastajim*+u, ra'+at Difda&a#+aN taxruj+u min el++maa'+I wa++tukal*im+u++haa qaaIila#+aN: "laa taHzan+iy, &an+maa qaliyl+iN turzaq+iyna Tifla#+aN!"

fariH+at el++malika#+u faraH+aN &aZiym+aN, wa++@ahab+at musri&a#+aN 'ilaY zawj+i++haa el++malik+i, fa++raw+at la++hu el++xabar+a.

6. It was translated by *Rūz Ğurayyib* and published by *Maktabat Lubnān*, Beyrouth, 1981.

The POS tagging process works on the (corrected) output of *Dekup* and makes use of a list of about 100 tags established according to the following principles:

The tag names always begin with an upper-case letter indicating the main part of speech of the word based on the traditional three parts of speech distinction of traditional Arabic grammar, namely Noun, Verb and Particle. So, in principle any word has a tag beginning with one of the Acronym letter N, V or P. However two other classes are recognized, namely that of pronouns with the two letter symbol PR (to differentiate it from Particles), and that of Relatives tagged as R.

Following that major classification, the POS tag name generally consists of a lower-case letter specifiying it within the major class to which it belongs. For example, in the class of verbs, Va means Verb-perfect (in French "accompli"), Vi means Verb-imperfect (French "inaccompli"). Similarly, in the class of Nouns Nn means Noun-nominative, Na means Noun-accusative etc. Next in the tag name come indications specifying further the word analyzed. Most of them are self-explanatory and are easily remembered once you have perused the list of tags and looked at some instances of tagged text. For example Va3fp means "Verb-perfect-3 person-feminine-plural", and Nai means "Noun-accusative-indefinite.

Some words are tagged as simple acronyms of the word concerned. It is generally the case for "special" words such as the verbe kaana (to be) which is tagged as K (followed by its specifying small case tags of course), so that kunt+u is tagged as Ka1s, and yakuwn+uw+na as Kiim3mp for "kaana-imperfect-indicative-third person-masculine-plural[7]. Other words tagged this way are verb *laysa* (not to be) tagged as L, the conjunctive particles *wa-* (and) and *fa-* (then/so) tagged as (lower case !) w- and f- respectively, and the word *maa* (what, which, that which) tagged as "maa" because it is impossible to identify its actual status without a contextual interpretation. Following is the list of most of the tags used in *Etiket*[8].

7. Notice that the second i is identified unequivocally as the mood indicator because of its position, after the « aspect » marker.

8. The tag names are in French but should not prove too difficult to understand. A full justification of the tags chosen would require a long development, unnecessary for the purpose of this presentation.

POS tags in *Etiket*

Ka3fs = kaana accompli 3e personne féminin singulier
Ka3mp = kaana accompli 3e personne masculin pluriel
Ka3ms = kaana accompli 3e personne masculin singulier
Kii2ms = kaana inaccompli indicatif
Kii3fs = kaana inaccompli indicatif
Kii3ms = kaana inaccompli indicatif 3e personne masculin singulier
Kij3ms = kaana inaccompli jussif 3e personne masculin singulier
L1s = laysa 1re personne singulier
L3ms = laysa 3e personne masculin singulier
N0- = Nom- à cas morphologique indéterminable (eg: maqhaY)
N0d = Nom défini à cas morphologique indéterminable (eg: elmaqhaY)
NDfs = Nom déicitique féminin singulier (haa@ihi)
NDms = Nom déicitique masculin singulier (haa@aa)
NUxy = Nom Numéral (variable ou non, défini ou non)
Na0 = Nom accusatif nu (eg : kitaaba)
Nad = Nom accusatif défini (eg: elkitaaba)
Nai = Nom accusatif indéfini (eg: kitaabaN)
Ng0 = Nom génitif nu
Ngd = Nom génitif défini
Ngi = Nom génitif indéfini
Ni = Nom interrogatif (eg : 'ayna, kayfa)
Nn0 = Nom nominatif nu
Nnd = Nom nominatif défini
Nni = Nom nominatif indéfini
No0 = Nom "oblique" (a/g) nu (eg: mu&al*imiy...)
Nod = Nom "oblique" défini (eg "elmuslimiyna" où a et g sont confondus)
Noi = Nom "oblique" indéfini (eg "muslimiyna" où a et g sont confondus)
Nxd = Nom "croisé" défini (comme "elcaaniy" où n et g sont confondus)
Nxi = Nom "croisé" indéfini (comme "caaniy" où n et g sont confondus)
PR1p = Pronom 1re personne pluriel
PR1s = Pronom 1re personne singulier
PR2fs = Pronom 2e personne féminin singulier
PR2ms = Pronom 2e personne masculin singulier
PR3d = Pronom 3e personne duel
PR3fp = Pronom 3e personne féminin pluriel
PR3fs = Pronom 3e personne féminin singulier
PR3mp = Pronom 3e personne masculin pluriel
PR3ms = Pronom 3e personne masculin singulier
Pa- = Particule- accusative
Pc = Particule conjonctive
Pe = Particule énonciative
Pe- = Particule- énonciative
Pf = Particule phrastique
Pf- = Particule- phrastique
Pi = Particule interrogative
Pj = Particule interjective ('aahiN)
Pjf = Particule interjective féminin ('ay*atuhaa)
Pm- = ? Particule- modale (layta)
Pn = Particule négative
Pp = Particule prépositionnelle
Pp- = Particule- prépositionnelle
Px = Particule exceptive (eg : 'il*aa)
R = Relatif
Rfs = Relatif féminin singulier (eg : el*atiy)
Rfp = Relatif féminin pluriel (eg: allawaatiy)
Rmp = Relatif masculin pluriel (eg : el*a@iyna)
Rms = Relatif masculin singulier (eg: el*a@iy)
Va3d = Verbe accompli 3e personne duel
Va3d- = Verbe- accompli 3e personne duel
Va3fs = Verbe accompli 3e personne féminin singulier
Va3fs- = Verbe- accompli 3e personne féminin singulier
Va3mp = Verbe accompli 3e personne masculin pluriel

Va3ms = Verbe accompli 3e personne masculin singulier
Va3ms- = Verbe- accompli 3e personne masculin singulier
Vi03fp = Verbe inaccompli 3e personne féminin pluriel
Vii1p = Verbe inaccompli indicatif 1re personne pluriel
Vii1s = Verbe inaccompli indicatif 1re personne singulier
Vii2fs = Verbe inaccompli indicatif 2e personne féminin singulier
Vii3d = Verbe inaccompli indicatif 3e personne duel
Vii3fs = Verbe inaccompli indicatif 3e personne féminin singulier
Vii3fs- = Verbe- inaccompli indicatif 3e personne féminin singulier
Vii3ms = Verbe inaccompli indicatif 3e personne masculin singulier
Vii3ms- = Verbe- inaccompli indicatif 3e personne masculin singulier
Vij2fs = Verbe inaccompli jussif 2e personne féminin singulier
Vij3fs = Verbe inaccompli jussif 3e personne féminin pluriel
Vij3fs- = Verbe- inaccompli jussif 3e personne féminin pluriel
Vij3ms = Verbe inaccompli jussif 3e personne masculin singulier
Vis1s = Verbe inaccompli subjonctif 1re personne singulier
Vis1s- = Verbe- inaccompli subjonctif 1re personne singulier
Vis3d = Verbe inaccompli subjonctif 3e personne duel
Vis3fs = Verbe inaccompli subjonctif 3e personne féminin pluriel
Vis3mp = Verbe inaccompli subjonctif 3e personne masculin pluriel
hVis3ms = Verbe inaccompli subjonctif 3e personne masculin singulier
Vmfs- = Verbe- impératif féminin pluriel
f- = conjonction "fa"
maa = particule "maa"
w- = conjonction "wa"

The *Etiket* POS tagger consists of an expandable data base containing as many segmented and tagged words as possible. Initially it consisted essentially of very frequent words (identified by their frequency in texts) and manually tagged, together with a few simple rules allowing it to assign POS tags to words whose identification was particularly straightforward.

For example, in the first line of the text given as an illustration above, words such as "*qadiym+i*", "*el++zamaan+I*", "*malik+uN*", and "*wa++malika#+uN*" can easily be recognized, on a purely formal basis, as nouns with specific case markers and determiners. As for the three other words in the line not immediately identifiable, namely "*kaan+a*", "*fiy*" and "*ya&iy$+aa+ni*", the first two belong to the class of very frequent words originally tagged in the data base, and the last can be guessed by the program on the basis of its form. But then the guess could be mistaken because the form of this word is compatible with two POS assignments: either as a Vii3d (i.e. Verb, imperfect, indicative, 3rd person, dual) or as a Nnid (i.e. Noun, nominative, indefinite, dual) on the analogy of "kitaab+aa+ni" (two books, nom.). In the present case only the first analysis is correct. But whatever the decision of the program may be[9], it will have to be validated by a human editor.

The general scheme is that, in the course of time, and as we go along with processing more texts, and correcting more errors, Etiket will become more and more knowledgeable in its field and commit less and less mistakes.

As to the practical aspects of POS assignment to a given text, things work as follows: Prior to the process of POS tagging, the segmented text is fed word by word to a processing data base in order to allow *Etiket* to examine each segmented word sequentially and take a decision concerning its tagging. Initially the POS fields of the words are empty.

Then *Etiket* is run on the text's data base, record by record, looking up each time in its own data base to verify whether the processed word is already there. If so *Etiket* copies the POS tag of its reference data base to the empty POS field of the word being processed. If not the guessing process is started and a decision is taken concerning the new word.

In some cases the POS field is left empty because no acceptable decision could be taken by Etiket concerning the word under examination. At the end of the process of tag assignment, a competent human editor revises its results and corrects them wherever necessary.

Finally, the words constituting the newly analyzed text are fed to *Etiket*'s data base (avoiding doubloons, of course) so as to increase the program's knowledge. In this way it is hoped that *Etiket* will become progressively more competent in correctly assigning POS tags to the words of the texts it processes.

This is an example of (part of) the output of *Etiket*'s analysis of our tale :

9. This decision will actually depend on the order of the rules of POS assignment in *Etiket* : the first rule dealing with the formal configuration in question will apply *IF the word has not already been stored with its correct tagging in the data base.*

kaan+a	Ka3ms
fiy	Pp
qadiym+i	Ng_
el++zamaan+i	Ngd
malik+uN	Nni
wa++malika#+uN	w_Nni
ya&iy$+aani	Vii3d
fiy	Pp
qaSr+i++himaa	Ng_PR3d
el++jamiyl+i	Ngd
&iy$a#+a	Na_
hanaa'a#+iN	Ngi
wa++sa&aada#+iN	w_Ngi
laakin*a	Pf
$ayI+aN	Nai
waaHid+aN	Nai
kaan+a	Ka3ms
yuHzin+u+humaa	Vii_PR3d
wa++huwa	w_PR3ms
'an*a++hu	Pf_PR3ms
lam	Pn
yakun	Kij3ms
la+humaa	Pp_PR3d
walad+uN	Nni

Note that agglutinated morpheme POS tags are linked to the main word's tag by an underscore (which may be replaced, as below, with hyphens for the sake of readability).

At the final stage, the data base containing our tale looks something like the following[10] :

kaana fiy qadiymi elzamaani malikuN wamalika#uN ya&iy$aani fiy qaSrihimaa eljamiyli &	Ka*ms Pp Ng· Ngd Nni w-Nni Vii*d Pp Ng·-PR*d Ngd Na· Ngi w-Ngi
laakin*a $aylaN waaHidaN kaana yuHzinuhumaa	Pf Nai Nai Ka*ms Vii*ms-PR*d
wahuwa 'an*ahu lam yakun lahumaa waladuN	w-PR*ms Pf-PR*ms Pn Kij*ms Pp-PR*d Nni
wakam e$tahayaa 'an yakuwna lahumaa waladuN!	w-Pi Va*d Pf Kis*ms Pp-PR*d Nni !
wamaa mar*a yawmuN 'il*aa rad*adaa fiyhi haa@ihi eljumla#a:	w-Pn Va*ms Nni Px Va*d Pp-PR*ms NDfs Nad :
""aahiN yaa laytanaa nurzaqu waladaN!"	Pj Pj Pm-PR\p Vpii\p Nai !
fafiy 'aHadi el'ay*aami, baynamaa kaanat elmalika#u tastajim*u, ra'at Difda&a#aN taxruju	f-Pp Ng· Ngd, CS Ka*fs Nnd Vii*fs, Va*fs Nai Vii*fs Pp Ngd w-Vii*fs-PR*fs Nai :
"laa taHzaniy	Pn Vij*fs
&am*aa qaliyliN turzaqiyna Tifla#aN!"	Pp-maa Ngi Vpii*fs Nai !
fariHat elmalika#u faraHaN &aZiymaN	Va*fs Nnd Nai Nai
wa@ahabat musri&a#aN 'ilaY zawjihaa elmaliki	w-Va*fs Nai Pp Ng·-PR*fs Ngd
farawat lahu elxabara	f-Va*fs Pp-PR*ms Nad
waba&da $uhuwriN qaliyla#iN taHaq*aqa qawlu elDifda&a#i,	w-Na· Ngi Ngi Va*ms Nn· Ngd
fawaladat elmalika#u Tifla#aN mala'at qalbahaa waqalba zawjihaa faraHaN	f-Va*fs Nnd Nai Va*fs Na·-PR*fs w-Na· Ng·-PR*fs Nai
kaanat elTifla#u jamiyla#aN jid*aN	Ka*fs Nnd Nai Nai
maa ra'aahaa 'aHaduN min elzaaliriyna 'il*aa Saraxa:	Pn Va*ms-PR*fs Nni Pp Nod Px Va*ms :
""aahiN maa 'ajmalahaa !"	Pj maa Va*ms-PR*fs !
'am*aa waaliduhaa elmaliku, fali$id*a#i 'i&jaabihi biTiflatihi 'amara bi'an tuqaama lahaa	Pe Nn·-PR*fs Nnd, f-Pp-Ng· Ng·-PR*ms Pp-Ng·-PR*ms Va*ms Pp-Pf Vpis*fs P
yud&aY 'ilayhaa jamiy&u 'aSdiqaalihi	Vpii*ms Pp-PR*fs Nn· Ng·-PR*ms
wama&ahum elmuluwku waelmalikaatu wael'umraa'u wael'amiyraatu min jamiy&i elbuld	w-Pp-PR*mp Nnd w-Nnd w-Nnd w-Nnd Pp Ng· Ngd Ngd
qaala elmaliku:	Va*ms Nnd :
""uriydu 'an 'ad&uwa ka@aalika jin*iy*aati elmamlaka#i 'ilaY HuDuwri Hafla#i el&imaadi, f	Vii\s Pf Vis\s Pp-NDms No· Ngd Pp Ng· Ng· Ngd, f-Vis\s-PR*fp No· Ngd Vii*fs
kaana fiy elmamlaka#i calaaca &a$ra#a jin*iy*a#aN	Ka*ms Pp Ngd NU·· NU·· Nai
waaHida#uN minhun*a &ajuwzuN ta&iy$u waHiyda#aN fiy baytihaa	Nni Pp-PR*fp Nni Vii*fs Nai Pp Ng·-PR*fs
falaa taraY 'aHadaN walaa yaraahaa 'aHaduN	f-Pn Vii*fs Nai w-Pn Vii*ms-PR*fs Nni
walam*aa kaana elmaliku laysa &indahu siwaY 'icnay &a$ara SaHnaN @ahabiy*aN faqad	w-Pf Ka*ms Nnd L*ms Na·-PR*ms Px NUo· NU·· Nai Nai f-Pf Va*ms NUo· NU·
walam yad&u eljin*iy*a#a el&ajuwza	w-Pn Vij*ms Nad Nad
ba&damaa entahat Hafla#u el&imaadi, eqtarabat eljin*iy*aatu min elTifla#i liyuqad*imna l	Na·-maa Va*fs Nn· Ngi, Va*fs Nnd Pp Ngd Pf-Vi·*fp Pp-PR*fs N··-PR*fp Nad
faqaalat el'uwlaY:	f-Va*fs N·d :
"sayakuwnu wajhuki jamiylaN jid*aN"	Pf-Kii*ms Nn·-PR*fs Nai Nai
waqaalat elcaaniya#u:	w-Va*fs Nnd :
"satakuwnu 'afkaaruki jamiyla#aN"	Pf-Kii*fs Nn·-PR*fs Nai
waqaalat elcaalica#u:	w-Va*fs Nnd :
"hadiy*a#iy laki hiya elluTfu waelmaHab*a#u"	N··-PR\s Pp-PR*fs PR*fs Nnd w-Nnd
waqaalat elraabi&a#u:	w-Va*fs Nnd :
"sayakuwnu raqSuki ra$iyqaN karaqSi eljin*iy*a#i"	Pf-Kii*ms Nn·-PR*fs Nai Pp-Ng· Ngd
waqaalat elxaamisa#u:	w-Va*fs Nnd :

10. Some of the data are only partially displayed in this image, but of course they are complete in the database.

4. A PRACTICAL APPLICATION

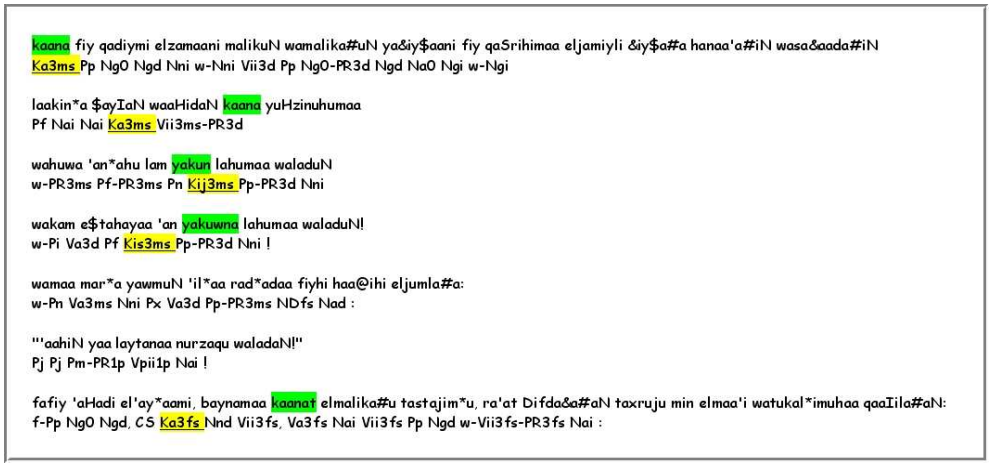
The whole idea behind the *Dekup* and *Etiket* programs was, initially, pedagogical. The aim was to provide teachers (and more indirectly learners) of Arabic with a convenient means of searching texts processed along the lines illustrated above for given linguistic forms whether words, phrases or whole sentences. To attain such an objective one further step must be taken once a text has been processed by *Dekup* and *Etiket* : sentences and word tags must be aligned horizontally so as to make a sort of "interlinear" text with actual words on the first level and POS tags at the second one. This is what the final thing looks like :

```
kaana fiy qadiymi elzamaani malikuN wamalika#uN ya&iy$aani fiy qaSrihimaa eljamiyli &iy$a#a hanaa'a#iN wasa&aada#iN
Ka3ms Pp Ng0 Ngd Nni w-Nni Vii3d Pp Ng0-PR3d Ngd Na0 Ngi w-Ngi

laakin*a $ayIaN waaHidaN kaana yuHzinuhumaa
Pf Nai Nai Ka3ms Vii3ms-PR3d

wahuwa 'an*ahu lam yakun lahumaa waladuN
w-PR3ms Pf-PR3ms Pn Kij3ms Pp-PR3d Nni

wakam e$tahayaa 'an yakuwna lahumaa waladuN!
w-Pi Va3d Pf Kis3ms Pp-PR3d Nni !

wamaa mar*a yawmuN 'il*aa rad*adaa fiyhi haa@ihi eljumla#a:
w-Pn Va3ms Nni Px Va3d Pp-PR3ms NDfs Nad :

"'aahiN yaa laytanaa nurzaqu waladaN!"
Pj Pj Pm-PRlp Vpiilp Nai !
```

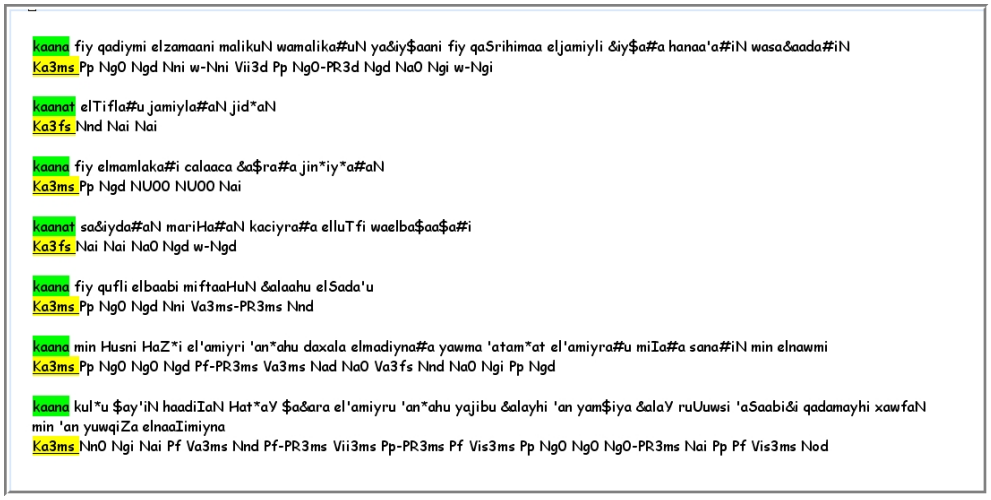
At this stage, it becomes possible to query the text for given configurations. The most convenient tool to performs such queries is, of course, Regular Expressions (henceforth Regex-es).

Suppose we want to query the database for all instances of verb *kaana* in the perfect or imperfect form, whatever its person and whatever its position in the sentence. The Regex to perform such a query would be something like this : "K[ai].?[123]..\W". Performed on the database this Regex would recognize the following sentences as answering the query[11] :

11. The yellow highlighting is added by the software used to perform the queries, the green one has been added manually to set off the words concerned. The image is, of course, just a small bit of what is retrieved. In fact there are many more occurrences of forms satisfiying the query.



It is possible to limit the query to only those instances of verb *kaana* where it occurs at initial sentence position, a typical textual configuration used for introducing new information in narratives. The relevant Regex would then be: "^K[ai].?[123]..\W", and the result would be (after collecting together the results):



Likewise, we could want to search the text for verbs in the subjunctive mood. The Regex query would be: "Vis[123].{1,3}\W" and the result would look like the following :

The number and type of queries which can be performed in this way are virtually infinite. Their result could be used to conduct corpus based text explorations centered on a specific grammatical or textual question.

5. Pending problems

The general scheme for the *Dekup* and *Etiket* programs was devised more than ten years ago, and has been used by students on a variety of texts to feed its database. This rather large experimentation has revealed both the interesting potential of this approach to Arabic text processing and its drawbacks.

The fact that it functions on trancribed text is often seen as an inconvenience by teachers of Arabic if only because their require the user to learn a system of trancription basically alien to the subject matter being taught. Correcting this problem does not seem to be an unattainable goal considering the present capabilities of Arabic word processing software and its availability to all. The main difficulty remains, in this perspective, the unreliability of integrally vocalized Arabic text when the vowelling is performed directly in Arabic script[12]. One option could be to perform the vowelling (and the text segmentation and tagging too) in transcription and then to revert to Arabic script at the stage of actual exploitation of the data. This should not prove an insuperable task.

More serious objections could be raised to the very idea of tagging individual words out of their context for, even in the case of integrally vocalized texts, many ambiguities remain if a word is processed in isolation from its context. Consider for exemple a word like *ḏahaba* : in isolation it could be construed as the verb “to go” and tagged Va3ms or as the word *ḏahab* “gold” in the accusative and

12. For more explanations on this point, cfr. Kouloughli, 2010.

be tagged Noa[13]. Admittedly it would be possible in such cases to insert the two alternative tags and let the human editor select the appropriate one. Yet it is difficult to imagine a real context of use where one would hesitate between the two interpretations. Consequently, tagging individual words does not seem a good idea both from the linguistic and computational points of view. The more recent developments in corpus linguistics further strengthen this criticism in the light of all the research done on collocations and formulaic language. This type of research suggests that identifying whole chunks of words which go together could make it possible to tag them (and perhaps even to vocalize them !) at one go.

This objection is certainly the most serious one to the approach presented here, and it is almost certain that it points to the right direction for further developments. At any rate the *Dekup/Etiket* system is only a step in a long road now trodden by many competent researchers[14].

13. By the way "No" means « a noun without article, definite or indefinite".
14. Cf. the bibliography for some useful references, many general and a few specific to Arabic.

References

ALANSARY S., NAGI M. and ADLY N., 2008, "Building an International Corpus of Arabic (ICA): progress of compilation stage", Alexandria [http://www.bibalex.gov.eg/isis/UploadedFiles/Publications/Building%20an%20Intl%20corpus%20of%20arabic.pdf]

BEAL J. C., CORRIGAN K. P. and Moisl H. L., 2007, *Creating and Digitizing Language Corpora (1. Synchronic Databases / 2. Diachronic Databases)*, Basingstoke, Palgrave MaCMillan.

BEN OTHMANE-ZRIBI C., TORJMEN A. and BEN AHMED M., 2006, "A multi agent system for pos tagging vocalized Arabic texts", RIADI Laboratory, University of La Manouba, Tunis.

DANIELSSON P., 2007, "What constitutes a unit of analysis in language?", *Linguistik online*, n° 31 [http://www.linguistik-online.com/31_07/danielsson.html]

DUKES K., ATWELL E. and SHARAF A. M., 2010, *Syntactic Annotation Guidelines for the Quranic Arabic Dependency Treebank*, Valletta (Malta), Language Resources and Evaluation Conference (LREC).

DUKES K. and HABASH N., 2010, "Morphological annotation of quranic Arabic", Valletta (Malta), Language Resources and Evaluation Conference (LREC).

DUKES K. and BUCKWALTER T., 2010, "A dependency treebank of the Quran using traditional Arabic grammar", Cairo, 7th International Conference on Informatics and Systems.

ELLIS N. C., 2002b, "Reflections on frequency effects in language processing", *Studies in Second Language Acquisition*, n° 24, p. 297–339.

FACCHINETTI R. ed., 2007, *Corpus Linguistics 25 Years on*, Amsterdam, Rodopi.

FĀYID W. K., 2003, *Buḥūṯ fī l-ʻarabiyya l-muʻāṣira*, Cairo, ʻĀlam al-Kutub.

FITZPATRICK E. ed., 2007, *Corpus Linguistics Beyond the Word: Corpus Research from Phrase to Discourse*, Amsterdam, Rodopi.

FRIEDL J., 2006, *Mastering Regular Expressions*, Sebastopol (Ca), O'Reilly.

GOYVAERTS J. and LEVITHAN S., 2009, *Regular Expressions Cookbook*, Sebastopol (Ca), O'Reilly.

HABASH N., SOUDI A. and BUCKWALTER T., 2007, "On Arabic transliteration", *Arabic Computational Morphology*, Dordrecht, Springer.

KORHONEN A., 2006, "Current trends and future challenges in computational linguistic research on multiword expressions", *Collocations and Idioms 2006: Linguistic, Computational, and Psycholinguistic Perspectives*, Bonn, Alexander Von Humboldt Foundation.

KOULOUGHLI D. E., 1999, "Sur l'analyse syllabique automatique de l'arabe", *Langues et Littératures du Monde Arabe*, n° 1, p. 29-42.

— 2004-2009, "Initiation pratique à la constitution et à l'exploitation de corpus électroniques en langue arabe", *Langues et Littératures du Monde Arabe*, ENS Lyon [http://icar.univ-lyon2.fr/llma/numeros.htm].

— 2010, "Traitement automatique de la métrique arabe : réalisations et perspectives", *Bulletin d'Études Orientales*, vol. LIX, p. 17-31.

LÜDELING A. and KYTÖ M. eds, 2008-2009, *Corpus Linguistics: an International Handbook*, vol. I and II, Berlin, De Gruyter.

MAAMOURI M., BIES A., BUCKWALTER T. and MEKKI W., 2004, "The Penn Arabic Treebank: building a large-scale annotated Arabic corpus", NEMLAR Conference on Arabic Language Resources and Tools (2004), Philadelphia, University of Pennsylvania.

O'KEEFFE A., MCCARTHY M. and CARTER R., 2007, *From Corpus to Classroom: Language Use and Language Teaching*, Cambridge, Cambridge University Press.

O'KEEFFE A., MCCARTHY M., 2010, *The Routledge Handbook of Corpus Linguistics*, London, Routledge.

RYAN R. R., RAMBOW O., HABASH N., DIAB M. and RUDIN C., 2008, "Arabic morphological tagging, diacritization, and lemmatization using lexeme models and feature ranking", *Proceedings of the Conference of American Association for Computational Linguistics (ACL08)*, Columbus (Ohio).

SMRZ O. and HAJIC J., 2006, "The other Arabic treebank: Prague dependencies and functions", *Arabic Computational Linguistics: Current Implementations*, Stanford, CSLI Publications.

VAN KEULEN P. S. F. and VAN PEURSEN W. Th., 2006, *Corpus Linguistics and Textual History. A Computer-Assisted Interdisciplinary Approach to the Peshitta*, Assen, Van Gorkum.

WIKIPEDIA (n. d.), "Romanization of Arabic" [http://en.wikipedia.org/wiki/Romanization_of_Arabic].

WILSON A., ARCHER D. and RAYSON P., 2006, *Corpus Linguistics around the World*, Amsterdam, Rodopi.

WYNNE M. ed., 2005, *Developing Linguistic Corpora: a Guide to Good Practice*, Oxford, Oxbow Books.