Blog-based corpora: some issues

Cristina Solimando[1]

0. Introduction

During last events of the "Arab Spring", Social Networks played an important role in gathering demonstrators and creating a virtual communicative space of protesters that involved Arab speakers from all over the world. These events brought attention to the new forms of online communication in the Arab world and probably increased their use through the Arab countries. Beside these web spaces which have been placed in limelight by the media, Internet is swamped with conversations inside groups of friends such as Facebook, Twitter and several other blogs that are dedicated to various topics as wellness and beauty, love affairs and marriage, religion, technologies, etc. This study does not address the sociological role of these communicative means, but is most focused on the language employed by the users in their correspondence. Most of these conversations consist in more or less personal exchange of opinions about very different topics and the type of communication is at a midway distance between a speech and a written text.

In the last sixty years scholars have discussed about the presence and the identification of different linguistic levels and varieties inside Arabic language: perhaps, for the first time we have access to a huge quantity of texts that have the peculiarity of being a sort of conversation, a quasi-oral texts. In order to investigate the linguistic peculiarities in these texts it is necessary to collect as many texts as possible: corpus linguistics and the tools of computational linguistics permit nowadays to work on a large number of data. The realization of a corpus is based on the identification of a group of texts as belonging to a textual genre[2]. Miller and Shepherd[3], in an article published in *Into the Blogsphere,* reckon that "when a type of discourse acquires a common name within a given context or community, that's good sign that it's functioning as a genre", an assumption which is supported by the definition of blogs' writers as "blogger". Obviously, the quantity of texts available on the Web is not measurable and the dimension of a corpus that is based on the Internet collection of texts depends on our goal: if we were to investigate on the use or presence of a specific morphological elements, even a 100 000 words corpus, that is very small among the corpuses that are usually employed in corpus linguistics, can be valued as representative. We will

1. Roma Tre Università.

2. Poncini, 2007.
3. Miller & Shepherd, 2004.

explore different aspects of this topic with the aim to give some examples of the strategies that we can apply to the linguistic analysis of a blogs-based corpus.

In paragraph (1) we will provide a brief overview on the Arabic linguistic varieties; in (2) we will dwell upon the linguistic and typological peculiarities of blog and we will focus on the differentiation among blog types and suggesting a classification in two main categories; in paragraph (3) through the employment of some tools of interrogation of the texts, we will illustrate some examples of examination of specific linguistic aspects recoverable in a blog corpus in order to demonstrate the advantages of these tools; new perspectives of research in this field will be suggested in the conclusion (4).

1. An overview on Arabic varieties

"Arabic" is a label that we use to refer to different realizations of the same language, a language that boasts an ancient history and a large number of speakers. Classical Arabic (CA), that is the language of the Quran and of texts belonging to the classical literature and to the traditional Islamic tradition, represents the prestigious variety chosen by all Arabic countries as official language and is used mostly in written texts. It is a strongly institutionalized and grammatically ruled language in its syntactic structure and its lexicon is largely impervious to foreign loanwords. Contacts native speakers have with CA are relatively occasional and confined to specific contexts such as educational environments and religious texts. Beside the classical form, Modern Standard Arabic (MSA), that differs from CA mainly its vocabulary and stylistic structure, is the Arabic of newspapers, modern literature and all those texts, even oral texts, that require a formal shape. Many scholars have focused on the phenomenon that Ferguson[4] defined as diglossia. The identification of two functional varieties of the Arabic language, an higher level for formal contexts of communication and a lower level for familial and friendly conversations, opened the way to a complex discussion about the Arabic language varieties. In fact, the sociolinguistic reality is more diversified than this representation would imply, and a deeper analysis[5] of the various, oral and written, texts would demonstrate that it is not possible to base a plausible linguistic representation of Arabic reality upon this rigid separation. The variety of the language used to communicate by two educated speakers belonging to two different areas within the Arabophone domain is a mixed variety that oscillates between Modern Standard Arabic and dialect; moreover, even a large presence of dialectal interference does not make a text necessarily a dialectal text. The status of formal or informal communication is based more on the text itself than on the extra-linguistic context: in the same context, such as in religious preaches[6], quotations from Quran and *Ḥadīṯ* are strictly in Classical Arabic, while the rest of the

4. Ferguson, 1959, p. 325-340.
5. Badawi, 1973.
6 Al-Šarqāwī, 2007.

text is a mixture between Classical and vernacular. It is what happens in almost all interviews broadcast by Arabic channels, where the interviewee introduces some vernacular elements in the discourse or switches to the vernacular in some passages. These oral texts cannot be considered as samples of "dialect", but rather as a colloquial realization of the Arabic language with a different degree of dialectal interference.

In other linguistic domains, colloquials are recognized as full-fledged variety of the language whereas colloquial Arabic still lacks a defined linguistic status: indeed, despite the appearance of a fair number of novels written in *ʿāmmiyya* during the last decade in Egypt[7], colloquial Arabic written form is anything but fixed and standardized Nowadays, the widespread use of social networks is creating the opportunity for million users to sharing a still unstable use of the colloquial and the dialectal in their written version.

2. The specificity of blogs

We are accustomed to associating written texts with Modern Standard Arabic or Classical Arabic and oral texts with vernacular varieties: to this categorization belongs also the differentiation between formal and informal texts. The formal or informal feature of a text is often considered strictly linked to its planning, so a written text should be from this point of view a planned text since it needs a reflexion upon the language used. But the variability and flexibility of the language permit us to analyze the different stylistic nuances inside the texts, thus we can differentiate among the various situations where formal or informal language is required, whether in written or spoken texts. Moreover, the rigid separation between written-planned texts and oral-unplanned texts is contradicted by the evident mixed nature of blogs texts, where users move between the two categories.

Arabic speakers, mainly educated young users of Internet, move in a multilanguage environment[8]: they resort very often to a western language[9] (English and French) whether because of easiness to communicate in those languages due to the historical contacts with Anglophone and Francophone countries or because of their aim to spread their message as far as possible. Beside English and French, chats and blogs are in the majority of cases written in a variety of Arabic which is very close to the colloquial: the language used is also very diversified according, for example, to the topic (religious blogs are often closer to the classical variety), the geographical location and origin of the blogger, the level of familiarity between the interlocutors. Consequently, the level of influence of the spoken language can vary: the immediacy and the spontaneity in transmitting a message, the personal involvement in the topic which is discussed, the degree of familiari-

7. Rosenbaum, 2010.
8. Peel, 2004.
9. Benrabah, 2009, p. 254.

ty felt inside the Web community push the blogger to use a more familiar style, closer to the language of everyday life. In fact, it is very difficult to categorize the nuances in the different levels of use of Arabic language in this communicative space.

Some websites have a regional nature and the majority of bloggers who participate in the chats speaks the same dialect, so the texts' language is strongly dialectally connoted. But an interesting perspective of analysis can be focused on other forums that are more "panarab" and that, consequently, host Arabs who speak different dialects. The question that evidently arises is what variety they use or until what level they introduce dialect elements in their writing.

An element which should be analyzed is the standardization of the orthography: it is well known that the habit of writing in dialect in not private texts is relatively recent and still not consolidated. The exam of the writing of one or more authors is probably not sufficient for an exhaustive analysis of this issue, so the building of a corpus in order to answer some of these interrogatives seems to be an obligatory operation.

A transversal typology consists in blogs written in latin characters: the initial absence of Arabic keyboards for computers and mobiles pushed the users to adapt their language to available characters, that are the latin alphabets and the numerical symbols. The use of some numbers is now pretty standardized as it happened for 7, 2 and 3 that represent respectively the letters *ḥā'*, *hamza* and *'ayn* as in *Baddi ru7 halla2* – "I want to go no" – or *3n jadd?* – "really?". When the Arabic keyboards became available not all writers changed their habits and now the transcribed texts are very widespread. It is a fact that all this material deserves to be investigated, but for the moment we postpone this analysis and we focused on the target of building a corpus based on Arabic characters' texts since the transcribed texts pose some problems of annotation and textual interrogation.

2.1. Which blogs?

It is sufficient to go through the forum chats available on the Web to realize that the nature of the online conversations is very diversified: we can roughly distinguish between the actuality blogs that consist in the commentary inside websites dedicated to the news and the personal blogs that show a more private nature, sometimes very close to a diary. Other typologies as thematic, politics and directory blogs can be considered as belonging to these two main categories.

Our demonstration of the analytical strategies that we can adopt to work on a blog-based corpus will start from this differentiation of the two categories, so we will build two different corpuses: one corpus that collects all the commentaries of the website [www.shabayek.com] (until may 2011) that contains political and commercial news and another corpus based on the collection of personal blogs taken from websites dedicated to topics such as beauty, marriage, car and technologies that are very visited and host a lot of discussions (we quoted their references in the bibliography). The building of two different corpuses is due to

the hypothesis that the personal blogs are much less linguistically ruled and, consequently, richer of dialectal interference phenomena because of their communicative peculiarities, such as the virtual or real friendship among writers, the use of a shared language, etc. On the other side, the actuality blogs, in particular the comments to the news, should show a stronger relation with the standard Arabic: the topic is not personal, they write the comment just after having read a text in standard Arabic, so they keep the reference to the article remaining unconsciously on the same linguistic level (use of lexicon and morpho-syntactical structures once read) and they could use the standard Arabic in order to address other Arabic speakers independently from their current geographical location. All these observations make us expect a slight number of dialectal interferences' phenomena in these blogs. In order to demonstrate this hypothesis we can: 1) build the two different corpuses; 2) identify some dialectal peculiarities which can or cannot be present in the corpus' texts; 3) verify through specific skills of linguistic interrogation their presence in the corpuses. The results obtained can suggest a perspective of linguistic analysis and of comparison between the two blogs categories. We will not take into consideration the lexicon used or the style of the texts, but we will focus on some grammatical phenomena such as the presence of the preverbs *b-* for the "imperfect" and *ḥa-/ha-* for the future.

3. Skills of linguistic interrogation of the corpus: Python and regular expressions

The actuality and personal blogs corpuses reached respectively about one million and 350 thousand words. The question of the representativity of a corpus arises every time we quote the quantity of words present in a corpus: there is no univocal answer to this question, it depends on the goal of the exploration of the corpus. If we wanted to get data about the richness of the vocabulary used in the two corpuses the dimensions of 1 million and 353,000 words are not so representative for this target, but in this case we want to verify the presence and the frequency of some precise morphological elements so we consider our corpuses representative.

To explore the two corpora we resorted to Regular Expressions which, to our ends, are essentially a tiny and highly specialized programming language embedded inside Python[10] and made available through the "RE" module. Regular Expression patterns are compiled into a series of bytecodes, which are executed by a matching engine written in C. This little language permits to specify the set of possible strings that we want to match in a text[11]. On this level, we will focus on the morphological aspects of texts and will give no consideration to syntactic analysis, nor to meaning: a "word" will be considered as just a sequence of cha-

10. See Bird, Klein & Loper, 2009.
11. See El-Kourdi, 2004, p. 19.

racters divided by spaces, so the article and the particles that are attached to the nouns or verbs are considered as a constituent of the string of the word.

Since what we are looking for is the presence and frequency of preverbs such as *b-* and *ḥa/ha*, some morphological constraints allow us to suppose some elements of the interrogation: the preverbs are regularly attached to the *muḍāri'*, so they are necessarily followed by the verbal prefixes *'-, t-, y-, n-*. Since the writing of a regular expression poses some bidirectional problems if we use Arabic characters, we defined a simple system that permits us to use Latin characters, through the use of variables[12]:

```
wb = ur'[^ي-ء]' #word boundary
non_alif = ur'[^ ا]' #non alif
ba = ur'[ب]' # ba-
tny = ur'[تني]' # tā', yā' or nūn
ya = ur'[ي]' # yā'
cons = ur'[ي-ء]' # any consonant of the alphabet
alm_3 = ur'{3,}' # at least 3 consonants
tn = ur'[تن]'
alif = ur'[ا]'
Ha = ur'[ح]'
non_ya= ur'[^ي]'
ha = ur'[ه]'
```

So, if we want to find how many times it is used in the blog the preverb *b-*, we look for a precise sequence of characters; the regular expression that we used is relatively small and compact:

shabayek_search (ur'(' + wb + ba + tny + cons + alm_3 + ur')')

that means that we are looking for a word starting with *b-*, followed by the prefixes *t-, n-* or *y-* and another sequence of at least three consonants. The matches found (2212 occurrences) are displayed in a file .text and highlighted in the text through the use of asterisks. Some examples:

Text no. 2 - par. no. 5
ربنا يوفقك ويوفق كل معتصم .. مش عارف انا مش مقبل مشروع خمسات ليه .
بهرب لما ألاقي بوست ليك * بيتكلم* عليه

Text no. 2 - par. no. 26
و بعدين شبايك اذا بدك اتعامل مع شركة ميديا بوكس الي * بتوفر* اعلانات
-اعلانات لصحاب المدونات و المواقع – * بتجبلك*

Text no. 3 - par. no. 23
و بعد 20 سنة عدت الى لبنان و بعد ان اكتسب خبرة في التجارة و الاستثمار و -
كبر الاطفال قررت بدأ مشروع العمر اعادة ترميم المباني القديمة في لبنان تحديدا في
مدينتي طرابلس و بما ان العقارات الشئ الوحيد السليم في بلدنا قررت ان اشتري المباني
القديمة و المتهدمة و اعيد ترميمها و ادخال الخدمات اليها من ماء وكهرباء و مصاعد – بما

12. The "ur" prefix in the examples below stands for "unicode raw" strings.

ان الخدمات رائعة لدينا في لبنان و الكهرباء لا تنقطع ابدا لدينا – ههه ثم اكيد * بيعها* لعلي اجعل من مدينتي داون تاون لبنان رقم 2.

Text no. 3 - par. no. 26
لكن الذكاااء هو بان تتقبل الواقع فلن نتحول بين ليلة وضحاها من افغانستان الى فنلندا و النرويج بل علينا ان نتمهل رويدا رويدا و خطوة خطوة قد نصبح دبي ثم ريو دي جينيرو ثم ربما روما و باريس .. بس بعد 100000 سنة *بتكون* صارت عظامنا مكاحل

Text no. 0 - par. no. 46
وبخصوص المنتدى التدوينه دى رووووووعه فيها فعلا نصايح فادتنى كتيييييير على الرغم انى مش من معجبيه قووى ولكن انا بشوف ان اى شئ اذا اتعمل بشكل محترف (مميز _ ليه فكره _ *بيطرح* حاجه حلوه _.....) اكيد هيكون ليه نصيب من النجاح

Text no. 2 - par. no. 3
انا لما كنت برد علي الناس كنت مش بعرف اكتب قررت أتوقف عن الرد لقيت العدد بيقل
المهم أنا كنت صراحه بتغاظ منك في الأول ، بس دلوقتي مش بتغاظ ليه لأن عارف عذرك ـ بس لما حضرتك بترد علي واحد ومش بترد علي 10 ـ طبيعي إن الواحد *بيعتبر* رأيه مفدكش او مش مهم ورأي الشخص دا مهم
ثانيا حضرتك كان عندك غلطه وانا احرجت اقولها ليك إنت احيانا بترد علي شخص بعينه .

Text no. 2 - par. no. 36
السلام عليكم ورحمة الله تعالى وبركاته
أشكرك أخ شبايك أنك تفهمت تعليقات وآرائنا وتقبلتها بروح رياضي إلى حد شجعك لكتابة هذه التدوينة التي يشبه سرد حساب عما يجري *بيننا* وبينك [...]

The matches identified in these five passages satisfies only partially our research: cases as *بيعها* (verbe *māḍī* + suffix pronoun) and *بيننا* (preposition+suffix pronoun) are false positives and their high frequency is a problem avoidable in two ways: we can decide to exclude directly the word from our research or to refine the regular expression as follows :

shabayek_search (ur'(' + wb + ba + tny + cons + alm_3 + non_alif + wb + ur')')

In this way we excluded the sequences ending in *alif* : the operation reduced the matches to 1265. The same method has been applied to the personal blogs corpus with the result of 76 occurrences (they were 105 before the exclusion of the *alif* at the end of the sequence).

As quoted before, the use of computational tools, fundamental to work on very large corpuses, can provide useful information about the process of orthographical standardization: in dialectology studies, for instance, it is mentioned the alternative use of the preverb *ḥ-* or *h-* to mark the future. The analysis of a big corpus provides us useful data about the actual use of the preverb: we'll apply the same operation through the following regular expression:

shabayek_search (ur'(' + wb + Ha + tny + cons + alm_3 + ur')')
shabayek_search (ur'(' + wb + ha + tny + cons + alm_3 + wb +ur')')

The results obtained in the two different corpuses are: in the actuality blogs corpus *ha-* recurs 282 times while *ḥa-* 81 times; in personal blogs *ha-* recurs 10 times and *ḥa-* 18 times.

4. Corpus and computational tools: Issues and new perspectives of research

In paragraph (4) we suggested the hypothesis that personal blogs could be richer of dialectal phenomena of interference than the actuality blogs: the intimate nature of the conversation and the topics treated could lead us to suggest a stronger involvement of the blogger and, consequently, a minor control of the linguistic style. The analysis of these corpuses contradicted this hypothesis and demonstrated a very strong presence of dialectal intrusion in the text even into the comments to politics articles. The new tendency of Arabs to write in dialect is an aspect worth of interest and that deserved to be analyzed: Social Networks and Internet in general represent the privileged testing ground.

The huge quantity of linguistic material available on the Web is unmanageable without the contribution that the new computer applications gave to the analysis of natural language. The few examples quoted herein show some linguistic phenomena that blogs' texts share with the spoken language. This is the starting point of a more complex operation which consists of the identification of the most meaningful linguistic features of this new means of communication: its nature obliges millions of bloggers or simple visitors of forum spaces to an unaware homogenization of their "writing" which is necessary if they want to convey a message. Corpus linguistics and the available computational tools help us in investigating this unique typology of texts. The quantification of the data can help us to define general issues or to delineate linguistic features that could be themselves object of analysis, as the case of the use of preverbs.

References

AL-ŠARQĀWĪ M., 2007, "Tasā'ulāt fī al-radd 'alā al-izdiwāğiyya al-luġawiyya", *Language*, n° 6, G. Mansour and M. Doss eds, p. 117-140.

BADAWI E., 1973, *Mustawayāt al-'arabiyya al-mu'āsira fī Misr,* Cairo, Dār al-Ma'ārif.

BENRABAH M., 2009, *Devenir langue dominante mondiale. Un défi pour l'arabe*, Genève, Droz.

BIRD S., KLEIN E. and LOPER E., 2009, *Natural Language Processing with Python*, Sebastopol, O'Reilly.

EL-KOURDI M., 2004, *Arabic Word Root Extraction and Automatic Categorization of Arabic Web Documents*, Rabat, Al Akhawayn University in Ifrane.

FERGUSON C., 1959, "Diglossia", *Word,* n° 15, p. 325-340.

MILLER C. R. and SHEPHERD D., 2004, "Blogging as social action: a genre analysis of the weblog", *Into the Blogsphere: Rethoric, Community and Culture of Weblogs*, Minneapolis, University of Minnesota Libraries.

PEEL R., 2004, "L'Internet et l'utilisation des langues : une étude de cas dans les Émirats arabes unis", *International Journal on Multicultural Societies (IJMS),* vol. 6, n° 1, p. 159-172.

PONCINI G., 2007, "Corporate podcasts and blogs: exploring the voices of emerging genres", *Multimodality in Corporate Communication Web Genres And Discursive Identity*, Milan, Franco Angeli, p. 147-166.

ROSENBAUM G., 2010, "I want to write in the colloquial: an exemple of the language of contemporary egyptian prose", *Folia Orientalia,* vol. 47, p. 71-97.