Improved Written Arabic Word Parsing through Orthographic, Syntactic and Semantic constraints

Nahli Ouafae **Marchi Simone**

Istituto di Linguistica Computazionale, Consiglio Nazionale delle Ricerche
Via G. Moruzzi, 1, 56124 Pisa - Italy
{firstname.lastname}@ilc.cnr.it

Abstract

English. The Arabic script omits diacritics, which are essential to fully specify inflected word forms. The extensive homography caused by diacritic omission considerably increases the number of alternative parses of any morphological analyzer that makes no use of contextual information. Many such parses are spurious and can be filtered out if *diacriticization*, i.e. the process of interpolating diacritics in written forms, takes advantage of a number of orthographic, morpho-syntactic and semantic constraints that operate in Arabic at the word level. We show that this strategy reduces parsing time and makes morphological analysis of written texts considerably more accurate.

Italiano. Le convenzioni ortografiche della lingua araba consentono l'omissione dei diacritici, introducendo così numerosi casi di omografia tra forme flesse e la conseguente proliferazione di analisi morfologiche contestualmente spurie. Un analizzatore morfologico che utilizzi i vincoli ortografici, morfo-sintattici e semantici che operano a livello lessicale, può tuttavia ridurre drasticamente il livello di ambiguità morfologica del testo scritto, producendo analisi più efficienti e accurate.

1 Introduction

Arabic is a morphologically rich language, where a lot of information on morpho-syntactic and semantic relationships among words in context is directly expressed at the word level[1]. Some prepositions, conjunctions and other particles are morphologically realized as proclitics, while all pronouns are enclitics. Orthographic, morphological and syntactic characteristics of Arabic contribute to increasing the level of ambiguity of written word forms, which is made even more complex by the unsystematic use of diacritical markers in the Arabic script[2]. In this paper we suggest that spelling rules, morpho-syntactic and semantic constraints should be jointly evaluated as early as possible in parsing an Arabic text. In particular, the analysis of spelled-out forms requires simultaneous use of morpho-syntactic and semantic information to define constraints on NLP, and "interpolate" missing vowels/diacritics (diacriticization) in Arabic written texts.

2 Morphological structure of Arabic words

2.1 Maximal and minimal words

In Arabic, written tokens correspond to either a "minimal word form" (see infra) delimited by white spaces, or a morphologically more complex token resulting from a concatenation of a minimal word form with clitics (called "maximal word form"). In (1), we offer the example of a maximal word form, consisting of the inflected form of the verb *kataba* 'write' surrounded by clitics[3].

Example 1 ***wa=ta-ktub-u=hu***
and=2MS-write$_{IPFV}$-PRS.IND=it
'and you write it'

The morphological structure of (1)[4] can be schematized as follows:
proclitics=prefix-stem-suffixes=enclitics.

By removing clitics, the remaining word form (***ta-ktub-u***) is a minimally autonomous inflected form, whose structure consists of **prefix-stem-suffixes**. Due to these levels of morphological embedding, word tokenization in Arabic must be followed by a sub-tokenization phase demarcating the boundaries between proclitics, the minimal word and enclitics.

[1] Tsarfaty et all (2013).

[2] Farghaly A., and Shaalan K. (2009).

[3] Interlinear glosses follow the standard set of parsing conventions and grammatical abbreviations explained in: "**The Leipzig Glossing Rules: Conventions for interlinear morpheme-by-morpheme glosses**" February 2008. Hyphen marks segmentable morphemes and an equal sign marks clitic boundaries, both in transliterations and in the interlinear gloss.

[4] Dichy J. (1997).

2.2 Ambiguity in tokenization

In Arabic written texts, vowels, gemination and other signs are written as diacritics added above or below consonant letters. Their marking, however, is not systematic. For instance, the word *kataba* 'he wrote' can be written in any of the following variants: *ktb*, *katb*, *katab*, *ktaba*, *katba*, etc. Furthermore, *ktb* is shared by all members of its derivational family. This means that, by vocalizing the skeleton differently, one can obtain word forms of other lexical units than the base verb: *kutub* (books), *katb* (writing), *kattaba* (dictate; make write). As a result of these powerful morphological relations, omission of diacritics in written texts causes extensive homography in Arabic. Text reading and understanding is an active process of text interpretation, based on context, grammatical knowledge and vocabulary. For example, clitics can be in grammatical combination with only some minimal forms. Hence, one can use the presence of clitics in maximal forms to cut on the level of ambiguity of their embedded minimal forms.

Section 2.3 illustrates how addition of proclitics can help morpho-syntactic disambiguation. Section 2.4 shows how semantic features of the minimum word can help constrain the number of enclitics that can be added to it.

2.3 Morpho-syntactic characteristics

Arabic clitics are important because impose morpho-syntactic restrictions on the words they are attached to. Particularly when the particle is proclitic, morphological restrictions can be of help for the morpho-syntactic analysis of a spelled-out form. Consider the example 2, where the form *ktb* is preceded by the determiner and the preposition ***li***. In this case, the form *llktb* has a single reading because, in Arabic, all prepositions require genitive case:

Example 2 ***li=l=kutub-i***
to=DET=books-GEN.DEF
'to the books'

Hence, to decrease the level of orthographic ambiguity, it is important to have a full list of clitics and the morphotactic constraints defining their compatibility with minimal words.

2.4 Verb semantics and agreement

Another peculiarity of Arabic is a complex system of N-V agreement rules. For example, when the subject refers to a rational entity (e.g. a person), its anaphoric clitic in the verb agrees with it in both number (SG, DU and PL) and gender (M and F). However, when the subject refers to an irrational entity, e.g. a non-human entity, its clitic marker in the verb is always in third person, and agrees with the noun in both number and gender only if the noun is singular or dual. If the noun is plural, the anaphoric clitic is 3SGF only. Consider the example 3 below. The verb *wahaja* requires an inanimate subject[5]. Thus, it can only select pronoun clitics in 3 SG/DU. Even if the subject is plural (3.b and 3.d), the verb is inflected in 3FSG. Furthermore, it cannot be inflected in the first and second person.

Example 3

a- **النَّارُ وَهَجَتْ**
'an=nār-u wahaj-at
DET=fire-NOM burn.PST-3SGF
'The fire burns' (cf. DET='al)

b- **النِّيرَانُ وَهَجَتْ**
'an=nīrān-u wahaj-at
DET=fires-NOM burn.PST-3SGF
'The fires burn' (cf. DET='al)

c- **العِطْرُ وَهَجَ**
'al='iṭr-u wahaj-a
DET=perfume-NOM spread.PST-3SGM
'The perfume spreads' (cf. DET='al)

d- **العُطُورُ وَهَجَتْ**
'al='uṭūr-u wahaj-at
DET= perfume -NOM burn.PST-3SGF
'The perfumes spread' (cf. DET='al)

To sum up, verbs are characterized by a conceptual structure that governs the selection and morpho-syntactic mapping of its arguments. The semantic properties of lexical units enforce constraints that can help predict their morpho-syntactic realization. Number and category of syntactic arguments are licensed by lexical restrictions imposed by the verb semantic class. These "selectional restrictions" on arguments are an essential part of the verb meaning and govern its morpho-syntactic behaviour[6]. Thanks to these restrictions, it becomes possible to successfully tackle possible ambiguities in the morpho-syntactic realization of the argument structure of a verb.

3 Word processing issues

We consider here the impact of the above-mentioned constraints on word processing in Arabic. Several software systems are available for the morphosyntactic analysis of Arabic texts.

[5] For example 'fire', which is feminine in Arabic and 'perfume', which is masculine.

[6] Jackendoff R. (2002), page 133 - 169

Buckwalter's Morphological Analyzer 1.0 (hereafter referred to as "AraMorph") is certainly one of the most popular such systems. Released in 2002, it is also offered as a Java port version, written by Pierrick Brihaye[7]. AraMorph's components are essentially two: the rule engine for morphological analysis and a repository of linguistic resources mainly composed of three lexicons: i) the dictStems lexicon, which contains 38.600 lemmas; ii) the dictPrefixes lexicon, which consists of sequences of proclitics and inflectional prefixes; iii) the dictSuffixes lexicon, which consists of sequences of inflectional suffixes and enclitics. These lexica are accompanied by three compatibility tables used for checking combinations of A (proclitics+prefixes), B (stems) and C (suffixes+enclitics). AraMorph analyzes transliterated Arabic text, and implements an algorithm for morphological analysis and for Part-of-Speech (POS) tagging that includes tokenization, word segmentation, dictionary look-up and compatibility checks. It finally produces an analytic report. In what follows, we consider some of the problems AraMorph encounters in tackling the extensive homography of Arabic written texts.[8] We then move on to our proposed solutions.

3.1 Problems and solutions

Case 1

In processing the written form *yaktub*, Aramorph produces the different parses listed in Table 1.[9]

	Analyses	Lemma
1	***ya-ktub***	*kataba* 'write'
2	**yu-ktab*	
3	**yu-ktib*	*ʾaktaba* 'dictate'
4	**yu-ktab*	

Table 1 – Aramorph's analyses for "*yaktub"*

Note that the AraMorph engine simply ignores the vowels present in the original spelling, and proposes a number of alternative parses, some of which are simply incompatible with the input form *yaktub*. This is the result of AraMorph's normalization strategy of written texts. To tackle lack of consistency in the Arabic spelling of diacritics, AraMorph gets rid of all diacritics marked in the original text, and parsed undiacriticized forms only. Buckwalter justifies this approach by claiming that writing without diacritics "is a common feature" of Arabic scripts. However, the approach generates spurious output analyses, based on a drastically underspecified spelling.[10] We suggest that diacritics marked in the original text should never be dispensed with, but rather used to filter out the set of candidate parses provided by AraMorph. For this reason, we designed a component assessing the compatibility of the vowel structure of AraMorph multiple parses with the original spelling in the text, to discard all candidates that are not compatible with the original spelling. Another noticeable aspect of Table 1 is that all parses simply ignore omission of the word final vowel in *yaktub*, a vowel used in the Arabic verb system to convey features of time and mood, as shown in example 4 below. This is due to AraMorph's suffix dictionary (dictSuffixes) lacking this information.

Example 4 ***ya-ktub-u***
IPFV.3-read-IND
ya-ktub-a
IPFV.3-read -SBJV
ya-ktub-Ø
IPFV.3-read -JUSS

To improve resulting parses, we augmented AraMorph's prefix and suffix dictionaries with missing information. Furthermore, it was necessary to update compatibility tables.

Case 2

Table 2 shows the analyses output by Aramorph upon processing the spelled-out form *whajt*.

solutions	Analyses	Lemma
1	**wa=hij-tu*	*hāja* 'be agitated'
2	**wa=hij-ta*	
3	**wa=hij-ti*	
4	**wa=hajj-ato*	*hajja* 'burn'
5	***wa=hajj-ato***	*hajjā* 'spell'
6	***wa=haj-ato***	*hajā* 'satirize'
7	**wahaj-tu*	*wahaja* 'burn; spread'
8	**wahaj-ta*	
9	**wahaj-ti*	
10	***wahaj-ato***	

Table 2 – Aramorph's analyses by "*whajt"*

Note that in this case, word segmentation differs depending on the output lemma. In solutions 1-6, each spelled-out form is an inflected form of the verbs *hāja/hajja/hajjā/hajā*, preceded by the clitic conjunction "wa=" (and). Solutions 7-10 are inflected forms of the verb *wahaja*. As in Case 1 parses 1, 2 and 3 may be filtered out if we take into account diacritics in the original spelling.

[7] AraMorph is downloadable from the LDC site at: http://www.nongnu.org/aramorph

[8] Hajder S. R. (2011).

[9] Wrong analyses are marked with an asterisk (*).

[10] Farghaly A., and Shaalan K. (2009).

Beyond these cases, AraMorph outputs further unlikely candidate parses. For example, Buckwalter includes obsolete lexical items[11]. In fact, the fourth proposed analysis is derived from the verb *hajja* that is not used in Arabic[12]. Focusing now on the last four solutions (7-10), they correspond to different inflected forms of the verb *wahaja* depending on what word final vowels are interpolated in the original spelling:

Solution 7 * ***wahaj=tu***
* burn$_{.PST}$=I
*'I burn'

Solution 8 * ***wahaj=ta***
* burn$_{.PST}$=You$_{.M}$
*'You burn'

Solution 9 * ***wahaj=ti***
* burn$_{.PST}$=You$_{.F}$
*'You burn'

Solution 10 ***wahaj-at***
burn$_{.PST}$-she
'She burn'

The inflectional suffixes -tu, -ta, -ti and –at respectively convey 1S, 2SM, 2SF and 3SF. However, we know that the verb *wahaja* requires an inanimate subject. Therefore it cannot be inflected for 1S, 2SM and 2SF. To capture this restriction and cut down on parse overgeneration, one has to enforce further restrictions in compatibility tables, e.g. the verb's ability to accept nominative and accusative pronouns, and to select a rational subject. We then augmented verb entries with subcategorization information such as case assignment and the restriction on rational subjects. At the same time, it was necessary to update compatibility tables. Table 3 shows how many entries are contained in AraMorph's original dictionaries (Original), and how many entries form the current improved version of the same dictionaries (Plus). Note that the number of stems is smaller in Plus than in Original, due to removal of obsolete entries and a number of foreign names that are unlikely to be found in Arabic texts[13]. Table 4 shows compatibility rules for tables AB, AC and BC in both Original and Plus.

AraMorph	entries		
	Prefixes	dictStems	dictSuffixes
Original	299	38600	618
Plus	335	35475	876

Table 3 - Entries in AraMorph's dictionaries

AraMorph	Compatibility		
	Table AB	Table AC	Table BC
Original	1648	598	1285
Plus	2698	1295	2161

Table 4 - Entries in compatibility tables

Finally, Table 5 shows how many parses of the same text[14] are output by AraMorph (Original) and AraMorph Plus. Figures are higher in the former case, in spite of the parser's failure to recognize 656 word tokens, due to lexical gaps in the stem dictionary. In addition, AraMorph Original presents a number of spurious parses. In Plus, on the other hand, restrictions on word grammatical behavior help improve results, and the number of proposed parses significantly decreases, despite Plus more extensive coverage (0 "Not found" parses).

Aramorph	Arabic forms	parses	Not found
Original	9502	21544	656
Plus		20847	0

Table 5 - Arabic text parsing by Original and Plus AraMorph

In addition, original AraMorph presents severely underspecified parses especially concerning morphosyntactic features. By augmenting information in clitics dictionaries and updating compatibility tables, AraMorph Plus provides more thorough morphosyntactic features[15].

4 Conclusion and future research

Automatic text processing requires annotation of different levels of linguistic analysis: morphological, syntactic, semantic and pragmatic. For some languages, like English, it makes sense to analyze those levels in a serial way, by taking the output of an early level of analysis as the input of the ensuing level. Purpose of this article is to demonstrate that specific characteristics of Arabic appear to recommend a different approach. Inflectional, derivational and non-concatenative characteristics of Arabic morphology require interdependence and interaction between different levels of analysis for segmentation of spelled-out forms and their analysis to be adequate. This suggests that Arabic processing may require substantial revision of traditional NLP architectures. For improvement and future work, we plan to complete and refine language resources for Arabic. As a further step, we consider including other contextual factors, such as knowledge about the immediate syntactic context of a word token, as restrictions on diacriticization.

[11] Attia M., Tounsi, L., and Van Genabith J. (2010)
[12] Lisān al-arab. Volume 2, page 170.
[13] Lancioni et al. (2013).
[14] Badawī A. (1966).
[15] Nahli O. (2013).

Reference

Alansary S., Nagi M., and Adly N. (2009). *Towards analysing the international corpus of Arabic (ICA)*. In International conference on language engineering. Progress of Morphological Stage, Egypt. Pp. 241–245.

Alkuhlani S. and Habash N. (2012). *Identifying Broken Plurals, Irregular Gender, and Rationality in Arabic Text*. In Proceeding EACL '12 Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics. Pages 675-685.

Attia M., Tounsi, L., and Van Genabith J. (2010) *Automatic Lexical Resource Acquisition for Constructing an LMF-Compatible Lexicon of Modern Standard Arabic*. Technical Report. The NCLT Seminar Series, DCU, Dublin, Ireland.

Attia M. (2008). *Handling Arabic Morphological and Syntactic Ambiguity within the LFG Framework with a View to Machine Translation*. Ph.D. Thesis. The University of Manchester, Manchester, UK. Pages 35-39.

Attia M. (2002*). Implications of the Agreement Features in Machine Translation*. Phd Thesis. Faculty of Languages and Translation, Al-Azhar University, Cairo, Egypt.

Badawī A. (1966). *'aflūṭīn 'inda-l-'Arab*, Dār al-Nahḍat al-'arabiyya, Cairo.

Bahou Y., Belguith Hadrich L., Aloulou C., and Ben Hamadou A. (2006). *Adaptation et implémentation des grammaires HPSG pour l'analyse de textes arabes non voyellés* In Actes du 15e congrès francophone AFRIF-AFIA Reconnaissance des Formes et Intelligence Artificielle (RFIA'06).

Boudlal A., Lakhouaja A., Mazroui, A., Meziane A., Ould Abdallahi Ould Bebah, M., and Shoul M. (2011). *Alkhalil MorphoSys: A Morphosyntactic analysis system for non-vocalized Arabic*, Seventh International Computing Conference in Arabic (ICCA 2011). Riyadh.

Buckwalter T. (2004***).*** *Issues in Arabic orthography and morphology analysis.* COLING 2004, in Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages, edited by Ali Farghaly and Karine Megerdoomian, Association for Computational Linguistics, Stroudsburg PA, USA. Pages 31-34.

Dichy J. (1997). *Pour une lexicomatique de l'arabe: l'unité lexicale simple et l'inventaire fini des spécificateurs du domaine du mot*. Meta: journal des traducteurs / Meta: Translators' Journal, vol. 42, n° 2, pages 291-306.

Farghaly A., and Shaalan K. (2009). *Arabic Natural Language Processing: Challenges and Solutions*. Journal ACM Transactions on Asian Language Information Processing (TALIP), Volume 8 Issue 4, December; New York, USA.

Hajder S. R. (2011). *Adapting Standard Open-Source Resources To Tagging A Morphologically Rich Language: A Case Study With Arabic*. Proceedings of the Student Research Workshop associated with RANLP 2011, Hissar, Bulgaria. pages 127–132.

Jackendoff R. (2002). *Foundations of language, Brain, Meaning, Grammar, Evolution.* Published in the United States by Oxford University Press Inc., New York.

Kenneth R. B. (1998). *Arabic morphology using only finite-state operations*. In Proceeding Semitic '98 Proceedings of the Workshop on Computational Approaches to Semitic Languages. Pages 50-57.

Lancioni, G., Pepe, I., Silighini, A., Pettinari, V., Cicola, I., Benassi, L., & Campanelli, M. *Arabic Meaning Extraction through Lexical Resources: A General-Purpose Data Mining Model for Arabic Texts*. IMMM 2013 "The Third International Conference on Advances in Information Mining and Management". Copyright (c) IARIA, 2013. ISBN: 978-1-61208-311-7

Lisān al-arab, edited by Ḥaydar A. and ʿibrāhīm A. Dār al-kutub al-ʿilmiyyah, Beirut, Lebanon.

Manning Christopher D., and Schuetze H. (1999) *Foundations of Statistical Natural Language Processing*. The MIT Press Cambridge, Massachusetts, London, England.

Nahli O. (2013). *Computational contributions for Arabic language processing Part 1. The automatic morphologic analysis of Arabic texts.* In Studia graeco-arabica vol.3, Published by ERC Greek into Arabic Philosophical Concepts and Linguistic Bridges European Research Council Advanced Grant 249431, C. D'Ancona (a cura di), Pacini Editore, Pisa. Pages 195-206. ISSN 2239-012X.

Tsarfaty R., Seddah D., Kubler S., and Nivre J. (2013). *Parsing Morphologically Rich Languages: Introduction to the Special Issue*. Computational Linguistics, Vol. 39, No. 1: 15–22.

Zemirli Z., and Elhadj, Y.O.M. (2012). *Morphar+: an Arabic morphosyntactic analyzer*. In Proceedings of ICACCI. 2012, International Conference on Advances in Computing, Communications and Informatics, CHENNAI, India. ACM New York, NY, USA ©2012. Pages 816-823.