Why Microsoft Arabic Spell checker is ineffective

Alexis Amid Neme

Why Microsoft Arabic Spell checker is ineffective

Alexis Amid Neme

E-mail: alexis.neme@gmail.com

Abstract

Since 1997, the MS Arabic spell checker was integrated by Coltec-Egypt in the MS-Office suite and till now many Arabic users find it worthless.

In this study, we show why the MS-spell checker fails to attract Arabic users. After spell-checking a document (10 pages - 3300 words in Arabic), the assessment procedure spots 78 false positive errors. They reveal the lexical resource flaws: an unsystematic lexical coverage of the feminine and the broken plural of nouns and adjectives, and an arbitrary coverage of verbs and nouns with prefixed or suffixed particles.

This unsystematic and arbitrary lexical coverage of the language resources pinpoints the absence of a clear definition of a lexical entry and an inadequate design of the related agglutination rules. Finally, this assessment reveals in general the failure of scientific and technological policies in big companies and in research institutions regarding Arabic.

1. Introduction

For more than two decades, English and French spell checkers are used by common and professional users on a daily basis. Since 1997, the MS Arabic spell checker was integrated by Coltec-Egypt in the MS-Office suite and till now many Arabic users find it worthless. Arabic users may choose to use the spell-checking functionality or not because of they find it untrustworthy. The users of Arabic spell checker are less confident than their counterparts for French. The main challenging issue: Arabic is a Semitic language where infixes are ubiquitously used whereas Indo-European languages use only prefixes and suffixes but not infixes.

"Arabic spell checking is an active area of research since results are not satisfactory." (Shaalan Kh. *et al*., 2003) and the state-of-the-art did not improve enough according to the author (Shaalan Kh. et al, 2012). A dozen of significant projects were initiated since 1990 in addition to, certainly, many dozens of unknown attempts. Only few Arabic spell-checkers were achieved but with poor and uneven results. Lately in May 2012, the Google Arabic spell checker has been released. It flags randomly and erroneously correct words which are fully or partially diacriticized.

One should distinguish lexicon-based projects from word-list-based projects (Shaalan Kh. *et al*., 2012). Often, Arabic word lists are made by processing a huge Arabic corpus, and then the list is controlled semi-automatically. First, word list projects are dedicated to spell checkers; while with lexicon-based project, one application may be spell checking, but others may also be information retrieval, morpho-syntactic analysis, and so on. Second, the word list should include word forms and all their possible agglutinations (prefixed and suffixed particles), but all possible agglutinations may perhaps not occur even in a huge corpus. While in a lexicon-based project, one can cover all possible agglutinated word forms by attaching to each Part-Of-Speech a grammar of agglutinations (Neme A., 2011).

As far as we know, the best commercial package on the market is still the MS-Office spell checker. The spell checkers of the GNU project for Open-Office (Ayaspell) look puzzled. We notice this disorder in the project documentation. Since developers are computer scientists, they do not have a lexicographic background and they borrow, indiscriminately, traditional concepts from traditional grammar. Developers are confused between traditional morphology and computational models; between root-and-pattern Semitic approach such as root, pattern, stem, prefixes, stem, suffixes, infixes, and concatenative techniques. They are borrowing concepts from traditional derivational morphology such as agent noun, patient noun, and instrument noun, and so on, and mixing them up with inflexional morphology. *No surprise, because all together these concepts configure a complex model; and so far, all Arabic linguistic tools without exceptions do not have clear definitions of these concepts. No matter how witty your algorithmic approach are; these borrowed linguistic concepts in the implemented model should be defined clearly and not imported blindly.*

Since MS-Office 2007 is still commonly used by Arabic users, we did not evaluate the MS-Office 2010 spell checker, and we are not sure if COLTEC is still responsible of this 2010 package. Arabic users suffered to upgrade from MS-Office 2003 to 2007. Therefore, many MS-Office users kept the version of 2007 and did not upgrade to 2010, at the individual level at least. It seems that there are some improvements in the lexical coverage of the spell checker of the version of 2010; but still many problems exist. Nevertheless, it is difficult to argue for an upgrade to 2010 version because the new spell checker has some improvements. MS-Office lost its credibility in the previous versions of the spellchecking tools. The main question to Microsoft remains: in order to have a better lexical coverage in the 2010 version, would Microsoft dare to reveal the complexity and the cost of upgrading the Arabic lexical resources?

Revision rules are defined mainly by professional editors and revisors in the printing industry. On one hand, revision rules in Modern Arabic depend mainly on current orthographical standards. usage standards are defined by professional revisors of traditional editors and by current orthographic usage in Arab regions. Therefore, there are some regional variations in orthography but also in the editing rules of a publisher in the same Arab country. On the other hand, the normative grammar or prescriptive rules defined by Arabic language institutions are generally not respected if they do not fit the publication editor rules. For example, normative grammars advice to put the ending accusative case "-F" before the Alif such as in كتابًا "ktAbFA"[1], (book). For normative orthography, the Alif is always followed by a silent vowel, sukuun; thus, "-AF" is a mistake. Such normative rule is far from being respected. The editing industry in Lebanon uses more the form كتاباً "ktAbAF" than the normative form.

In this study, we investigate and evaluate the MS-Office 2007 Arabic spell checker using test sets. The MS-Office spell checker flags real spelling mistakes. Besides them, spell checkers also flag erroneously numerous correct words, called false positive errors, which decrease the precision of the tool. Therefore, the user's confidence in the tool decreases in proportion. We focus in this study on words flagged erroneously and on the lexical coverage of lexical resources. Generally, false positives are related to hamza spelling variations, affix agglutinations, feminine or plurals, and missing entries or orthographic variants. The test set includes 3 parts: a list of 550 verb occurrences; 10 pages of reviewed document, i.e. 3300 words without spelling errors, and finally a random list of singular/broken plural.

In Section 2, we give a brief background for spelling mistakes, especially the most common ones, and related definitions and terminology. In Section 3, we describe and analyse the result with the test sets for the MS spell checker. In Section 4, we give some concluding remarks.

[1] In this paper, examples displayed in the Latin alphabet are transliterated according to Buckwalter-Neme (BN) code, a variant (Neme, 2011, p. 6, note 4) of Tim Buckwalter's transliteration that avoids the use of special characters. In this transliteration, upper-case and lower-case letters, e.g. *E* and *e*, denote distinct, independent consonants : ء, c; آ, C; أ, O; ؤ, W; إ, I; ئ, e; ا, A; ب, b; ة, p; ت, t; ث, v; ج, j; ح, H; خ, x; د, d; ذ, J; ر, r; ز, z; س, s; ش, M; ص, S; ض, D; ط, T; ظ, Z; ع, E; غ, g; ف, f; ق, q; ك, k; ل, l; م, m; ن, n; ه, h; و, w; ى, Y; ي, y; ً, F; ٌ, N; ٍ, K; َ, a; ُ, u; ِ, i; ّ, G; ْ, o.

2. Background in spelling mistakes

For any language, spelling errors are limited to 3 types: *typographical* such as <comma> not followed by <space>; lexical *or non*-word errors; non-lexical or real-word errors. Typographical errors are simple to detect by several dozens of rules. The detection of *lexical* errors is generally based on a lookup procedure in a large dictionary of all valid words of a given language. *The non-lexical* are valid words which are invalid in the context of a sentence. The detection of non-lexical errors need complex grammars rules, and sometimes semantics which is only partially achieved even in English. So, our focus is lexical errors or non-word errors which require a large Arabic dictionary. We have demonstrated the feasibility of such a dictionary with a verbal lexicon with 15 400 entries (Neme A., 2011).

Spell checkers algorithms are based on a *lookup procedure* in a large dictionary. So-called Out-Of-Vocabulary words (OOV) may be valid words which are flagged because they are out of the lexical coverage of the lexical resources. This is because flagging a word depends on the lexical coverage of the dictionary. On the other hand, error correction proposes suggestions to correct a misspelled word and is out of the scope of this evaluation.

Spelling error types are letter substitution, letter omission, letter insertion, letter transposition and space omission. We have chosen a sample of errors which MS-Office 2007 flags by underlining them in red.

3. Testing MS-Office 2007

The COLTEC-Egypt package of Arabic spell-checking tools was chosen by MS-Office in 1997. For our evaluation of MS-Office, we have chosen the Arabic (Saudi Arabia) dictionary. In the rest of the study, errors flagged by MS-Office are indicated in **bold**.

a) General evaluation for detecting types of errors

Observations	Error Type	Mistake	Correct	#
	letter substitution	**شاعد**	ساعد	1
		ضدفة	صدفة	2
		هذة	هذه	4
		زراعى	زراعي	5
		زبحت	ذبحت	6
	letter omission	**محد**	محمد	8
		كتبو	كتبوا	9
MS-Office flags two errors with two agglutinated direct object pronoun suffixes		**يلزّمكمها**	**يلزّمكموها**	10
	letter insertion	**علّلل**	علّل	11
		يجنيون	يجنون	12
For letter transposition, MS-Office flags both errors (lexical coverage)	transposition	**إستتمع**	**إستمتع**	13
	space omission	**نامالولد**	نام الولد	14

Table 1: Types of spelling errors with examples. errors Flagged by the MS-Office Arabic spell checker are in **bold.**

b) Evaluating the verbal lexical coverage

We have extracted from a corpus a random list of 550 fully diacriticized agglutinated verbal occurrences. We have submitted this list to MS-Office. It appears that 2 verbs are missing in the present-perfect-3-person-singular: **يَتَنَامَى** , **يَعْتَسِف** (to grow, to shun). The Ms-Office lexical coverage of verbs is comparable to that of BAMA (Buckwalter T., 2002) which has also two missing verbs out of 8700 (cf. Neme A., 2011, section 6) . The verbal lexical coverage of BAMA is considered as medium.

Two other verbs suffixed with *–hu* **الْتَقَتْهُ**, **تَتَدَاوَلُهُ** (she met him, she exchanges it) were flagged erroneously; these forms exist in the dictionary, but without the suffix.

c) Evaluating false positives in 10 pages' documents

We selected three documents totalling 3 300 tokens, i.e. 3 300 strings between two spaces (about 10 pages), and containing popular science about three topics: pollution and fishing in Egypt, earthquakes in the world, and quality of water. It is a small sample of the NEMLAR Arabic Written Corpus (Attia M. *et al.*, 2005). This corpus was produced and annotated by RDI, Egypt, for the Nemlar Consortium. We used the documents in the fully diacriticized version.

MS-Office flagged erroneously 78 correct words out of 3300. Besides other true mistakes such as misspelled words or typographical mistakes in the document, the user must enter almost 8 decisions/page to ignore false flagged errors.

Here is a sample of correct words which are flagged (see details in the appendix). Examples are grouped by types:

- (1-3) false positives related to agglutinations, selected among 9 occurrences.
  7 occurrences of <PREPOSITION><PRONOUN> such as "bi_haA" as in (1); "muluHa**t**i_haA" as in (2) whereas "muluha**p**" is in the dictionaries. feminine nouns ending with "p" are rewritten with "t" when followed by an agglutinated possessive pronoun. The spell checker flags randomly these noun, for instance, it does not flag "OaMitati_haA" variant of "OaMitap". Similarly, "sa_naHyaA" (3) is a verb prefixed by the future particle, whereas "naHyaA" is in the dictionary, and the spell checker does not flag "sa_naktob".
- (4-5) false positives related to the feminine suffix "-t", "teh marbutah". MS-Office suggests substituting feminine adjectives with their masculine variants.
- (6) False positives related to the feminine suffix "-t" "teh marbutah". MS-Office suggests correcting a standard Arabic form into a colloquial orthography with "h" which is considered improper by professionals.
- (7-11) false positives related to *hamza-under-the-Alif*, selected among 43 occurrences. MS-Office considers many words with *hamza-under-the-Alif* as incorrect and suggests substituting it with a bare Alif. MS-Office suggests correcting a rich typographical representation by a poor one, which is unacceptable for professional proofreaders or writers. On the other hand, MS-Office suggests a rich typography with *hamza-under-the-Alif* for the two words: "انتاج وايجاد" and recommends "إنتاج وإيجاد" (production and existence).
- (12-24) false positives related to Out-Of-Vocabulary scientific terms.
- (25-31) false positives related to Out-Of-Vocabulary geographical proper nouns.

// "Glose" – Observations	**False positive error**	#
////////// Correct agglutinated word flagged (9 occurrences)		
// " in_it" Flagged and all bi-PRO (7 occurrences)	بِهَا	1
// "salinity_its" , whereas مُلُوحَة the unagglutinated variant exists.	مُلُوحَتِهَا	2
// "we_will_live" , whereas نحيا "we_live" the unagglutinated variant exists	سَنَحْيَا	3
// "proximity" as a feminine adjective whereas the masculine exists in the dictionary.	تَقَارُبِيَّةٍ	4
// "sliding" as a feminine adjective, masculine is suggested as correction.	انْزِلَاقِيَّةٍ	5
// "alone" حده is the proposed correction variant with final "h"	حِدَةٍ	6
/////// Words beginning with hamza-under-the-Alif: false positives (43 occ.).		
// "monopole"	إِحْتِكَارَ	7
// "undertaking"	إِتِّخَاذِ	8
// "tremblings"	إِهْتِزَازَاتِ	9
// "the_sliding"	الإِنْجِرَافِ	10
// "in _the_isolation"	بِالإِنْفِصَالِ	11
////////// Missing vocabulary – Flagged errors - Out-Of-Vocabulary		
// "far" as a feminine adjective.	تَبَاعُدِيَّةٍ	12
// "in_the_deep" as a feminine adjective.	الْقَاعِيَّةِ	13
// "authentic" as a feminine adjective	حَقِيقِيَّةٍ	14
// "the_tectonic" as a feminine adjective,	التَّكْتُونِيَّةِ	15
// "the_nitrogenised"	النَّيْتْرُوجِينِيَّةِ	16
// "the_magma", this word is listed in the Larousse Arabic dictionary	الصَّهِيرِ	17
// "The_tsunami"	التِّسُونَامِي	18
// "the_cracks" : Broken plural is missing whereas the singular exists.	الصُّدُوعِ	19
// "the_hunting_field" missing; the broken plural is proposed as a correction	الْمَصِيدَ	20
// "and_the_magnesium"	وَالْمَغْنِيسْيُومْ	21
// "the_chlore"	الْكُلُورْ	22
// "the chlorated" as a masculine adjective	الْمُكَلْوَرْ	23
// "and_the_shrimp(s)" and 5 occurrences of common Mediterranean fish	وَالروبْيَانْ	24
////////// Missing Proper Nouns – Geographical or related		
// "Sumatra"	سُومَطْرَةْ	25
// "Himalaya", existing undiacriticized form variant هملايا	هِمَالَايَاْ	26
// "the_Andes"	الْأَنْدِيزْ	27
// "Aegean_sea", existing other orthographic variants إيجة , إيجيه	إِيجَهْ	28
// "Sri Lanka".	سِرِيلَانْكَاْ	29
// "Americans", variant missing	أَمِيرِكَانْ	30
// "the_Aztecs"	الْأَزْتِيكْ	31

Table 3: A sample of false positive errors flagged by MS-Office 2007 in 10 pages documents test set.

d) Evaluating lexical coverage of plurals: on a sample list of broken plurals

In this section, we investigate the lexical coverage of related pairs of singular/broken plural forms. We have submitted to MS-Office a list of 850 pairs of singular and broken plural. Many broken plural nouns (or adjectives) are missing in the dictionary, even when the corresponding singular is present. Some singulars are missing also, even though their BP is present in the dictionary. The table below lists a sample of missing BPs and missing singular word forms.

Gloss	**Broken plural missing**	singular
hamstring	**عَراقِيب**	عُرْقُوب
incision	**حُزُوز**	حزّ
crew	**طَواقِم**	طَاقِم
dredge	**مَناكِيش**	مِنْكَاش
rat	**جَراذِين**	جُرَذ
elbow	**أكْواع**	كُوْع
crack	**فُلُوق**	فَلَق
ring	**أشْراج**	شَرَج
eye	**نَواظِر**	نَاظِر
sieve	**غَرابِيل**	غِرْبَال
little	**خَناصِر**	خِنْصِر
phoneme	**صَواتِم**	صَوْتَم
robust	**شُطُوب**	شَطَب
letter	**مَكاتِيب**	مَكْتُوْب
lieutenant	**فُرَقاء**	فَرِيْق
monarch	**عَواهِل**	عَاهِل
balance	**قَساطِيس**	قُسْطَاس
tomb	**ضَرائِح**	ضَرِيْح
cover	**مَلاحِف**	مِلْحَف
portion	**شُطُور**	شَطَر
queue	**أرْتال**	رَتْل
anklet	**خُجُول**	حَجْل
anklet	**أحْجال**	حَجْل
bell	**جَلاجِل**	جُلْجُل
escape	**مَهارِب**	مَهْرَب
	Broken plural	**Singular form missing**
group	فَصائِل	**فُصَيْل**
tear	مَدامِع	**مَدْمَع**
crack	فُلُوع	**فُلْع**

Table 3: Singular-broken plural pairs test set. False positives are in bold

e) Conclusion for MS-Office 2007 - Coltec

1. The verbal lexical coverage of MS-Office is medium or below 9000 entries. Verbs in a common Arabic dictionary are at least 15 000 entries.
2. Some familiar scientific, geographical proper nouns are missing, as well as some of their orthographic variants.
3. The correct rich typography of the initial “Alif hamza under” is flagged; and a “bare Alif without hamza” is suggested as a correction which is unacceptable for editors and professionals.
4. The lexical coverage of agglutination of prefixes and suffixes of verbs is *unsystematic.*
5. The lexical coverage of feminine and singular/broken plurals is *arbitrary*.

This arbitrary and unsystematic lexical coverage pinpoints two main flaws in the lexical resource: the absence of a clear *definition of a lexical entry* and the *inadequate design of related agglutination rules*. The lexical resource cannot be fixed through superficial adjustments and needs a new design.

4. Conclusions and Arabic spell checker prospects

Why is the MS-Arabic spell checker for Microsoft so disappointing? We see two main reasons: the first is linked to the company's general strategy; and the second is linked to research funding.

It appears that Microsoft added the Arabic spellchecking tools to its basic products in order to improve product sales, not product quality. At least till 2007, Microsoft did not develop research on the spell checker through their companies, but they bought it from specialized companies. They are not interested in the quality of research behind. Their goal is to add an Arabic spell checker to their product in order to sell it to the world with the spell checker regardless of its completeness or usefulness. Thus, the value of the company's shares increases, shareholders do not care about the program effectiveness, but merely its existence. But, the Arabic user will relinquish the ineffective tool.

For 15 years, the introduction of the Arabic spell checker in the MS-Office Suite has not improved noticeably, nor its use increased or decreased; several copies have been issued by MS-Office later, but users did not notice any difference in the spell checker. It seems that Microsoft NLP Group does not have the expertise to control the quality of Arabic language resources nor to improve its accuracy. Since it is assumed that the Arabic spell checker is commercially safe, this reason has limited the possibility of funding research and to develop a reliable Arabic spell checkers by Microsoft or by any research institution.

The second reason is related to the efficiency of specialists in linguistics and related language technologies: the Arabic language expertise is limited and constrained by the previous results of research mainly focused on techniques adapted to Indo-European languages. For example, the most prominent research centres and mechanization of the Arabic language are in America and Europe. Egypt and Algeria have recently introduced a number of centres in the Arab countries; but they are likely to reproduce the same problems and loopholes as in Western centres.

Research is still active to produce a powerful Arabic spell checker in order to conquer the trust of Arabic users, students, writers and revisors. The experience of automating Arabic language is not similar to Indo-European languages, and thus the foundations of algorithms should be adjusted to the Semitic nature of the Arabic language and its implications in language technologies. Few researchers are working on the basis of “Semitic" approaches and modern lexicography in order to create practical tools in Arabic Natural Language Processing.

References

Neme, Alexis, Laporte Éric (2013). Pattern-and-root inflectional morphology: the Arabic broken plural. *Language Sciences*. http://dx.doi.org/10.1016/j.langsci.2013.06.002

Neme, Alexis (2011). A lexicon of Arabic verbs constructed on the basis of Semitic taxonomy and using finite-state transducers. In *Proceedings of the International Workshop on Lexical Resources* (WoLeR) at ESSLLI.

http://alpage.inria.fr/~sagot/woler2011/WoLeR2011/Program_files/WoLeR%202011%20-%20Neme.pdf.

Buckwalter Arabic Morphological Analyzer Version 1.0. (2002). LDC Catalog No.: LDC2002349.

Attia., M., Yaseen., M., Choukri., K. (2005). Specifications of the Arabic Written Corpus produced within the NEMLAR project, www.NEMLAR.org.

Shaalan, K., Allam A., Gomah A. (2003). Towards automatic spell checking for Arabic. Conference on Language Engineering.

Shaalan, Khaled, Samih, Younes, Attia, Mohammed, Pecina, Pavel, & van Genabith, Josef (2012). Arabic Word Generation and Modelling for Spell Checking. Language Resources and Evaluation (LREC). Istanbul, Turkey. Pages: 719-725.