# Stateful Augmented Sliding Window Based Arabic Pos Deep Tagging

Hatim Ibrahim

Kalmasoft, July 2014

info@kalmasoft.com

*Abstract: a new approach to Arabic PoS tagging based on augmented stateful sliding-window (SWPoST), the system assigns the part of speech to the word based on the information provided by a variable width window of words around it; this paper is the first of a series of three articles on the same topic (concept, theory, and application).*

## Introduction

Language is a sort of communication between two or more living species, its is an acquired knowledge developed for emotional satisfaction by humans and upper orders of animals; in a sense, it is considered a social skill for both; all languages have their distinct phonological system that may be accompanied by gestures which may have the same effect of giving the meaning of the message directed to the listener; some languages have optional complex or primitive scripting systems that serve as a means of remote communication "messaging" medium, this set the language as a written language.[1]

Language is described as "*natural*" if socially developed over the human history, it may also be distinguished as being "*live*" if it is evolving (being spoken does not necessarily qualify a language to be live) or "dead" such as the Hieroglyphic, Cuneiform and some other languages spoken by some minorities.

The term "natural" distinguishes human native languages from other two types of "*artificial*" languages which come in two categories:

- *Constructed Languages*: these are mastered to be used by humans mostly for scientific and academic purposes including the standardization of terminologies and concepts in an attempt to create a *universal language* of more rigid structure for more effective communication; the *Esperanto* is a good example of those; the other most notable one is *Interlingua* which is effectively used in the United Nations and has native speakers.
- *Programming Languages:* in contrast to the above, these are machine oriented set of declarative and imperative vocabulary; they do not differ much from a natural language except they do not require any phonological system; that is, they do not require a separate phonology since they simply borrow from the underlying natural language vocabulary; they are strictly controlled by their creators who may intervene to alter the grammar or syntax if necessary; computer programming languages are all members of this category.

1 Animals leave droppings and smell to warn insiders analogous to human writing.

A natural language is primarily spoken, but because of the weakness of the human voice not traveling so far, a writing system is developed for the sole purpose of delivering the message to a distance farther than that reachable by the human voice; education and heritage archiving are other secondary uses of writing.

Written languages have different writing systems (or scripts), families of languages may share a common writing system; but all these systems are consistent in that they evolve and change separately from phonology; there is no direct link between the phonology and orthography of any natural language since they belong to two completely different natural cognitive domains.

Language orthography does not stand for the meaning by default, but only when perceived by the reader as being emotionally satisfying based on cognitive social agreement; this is evident in the clear distinction between praise and insult if found in written forms. Writing systems may be divided into two main categories; alphabetical in which a chart of meaningless monographs are assigned to some set of phonemes, this category is prevalent in most of the written languages because of its flexibility; the other category makes use of primitive symbols (logographs or pictographs) that depicts an abstraction of the meaning, those ideographs evolve over time and may be associated with any form of phonemes that may indicate the meaning, pictographs are in fact the origin of most writing systems including ancient e.g. Hieroglyph and current CJK languages (Chinese, Japanese, Korean).

Of importance here is that a language script is just a parallel representation to the speech, this representation is in completely different visual cognitive domain we call “writing and reading” compared to the other audio domain “talking and listening”. To conclude the above, accurately describing a natural language's semantic content through analysis of its representing written form is unlikely straightforward and requires linguistic manipulation and deep knowledge that strongly dependent on how accurately the writing represents the language; fortunately, the evolution of different natural language processing techniques makes it feasible and easy to handle both categories of writing systems mentioned above, and even to deal with natural voice or any of its electronic representations.

**The Arabic language**

Arabic is as complex as many other natural languages, native speakers tendency to reveal rhetorical skills and delinquency more than actually required has increased its complexity by rendering both its spoken and written forms almost beyond any reasonable appropriate uses. Arabic morphology is based on a long list of meaningless radicals called *roots* that inflect systematically using morphological templates (*binyanim*) which are perceived by native speakers as being one of the most language distinguishing features -along with richness- come in handy for *morphotactics* but not as significant if implemented for natural language processes other than morphological analysis as we will see later in the following pages.

Morphologically, Arabic is excessively derivational and highly inflectional meaning it inherently has the ability to over-generate vocabulary based on basic building blocks; in fact Arabic morphological system can generate tens of thousands of words originating from the same radical (root) but not necessarily be semantically convergent; Arabic writing system uses two sets of characters, a set of 28 consonants, and another short list of diacritic vowels used for disambiguation; the system itself branched off a long lineage of ancestor systems including Aramaic, Nabataean, and Syriac and evolved through various stages until it stabilized to current modern form. In earlier times, Arabic writings were undotted free style without punctuation, the dots did not become obligatory until much later as a necessity in post-Islamic era when Arabic retained its formally accepted writing convention which was basically to preserve religious texts such as the Qur'an. Historically, Arabic remained spoken for ages in a community that did not adopt writing in most of traditional and social activities.

Syntactically, Arabic follows a Non-rigid Grammar paradigm, words can take different arrangement within the sentence with preference to (VSO) form which depends on the dialect level; Arabic has acquired at least two dialectical layers. It is also syntactically verbose since parts of a sentence compositions are not basically required in the syntax. The following examples are not deliberately mastered to support this point, the parts in red can be removed without changing the meaning in any way.

| | verbose | concise | **meaning** |
|---|---|---|---|
| 1 | إِنَّ السَّيَّارَةَ مُعَطَّلَةٌ | السَّيَّارَةُ مُعَطَّلَةٌ | The car is broken |
| 2 | نَحْنُ نَحْتاجُ إلى المَاءِ | نَحْتاجُ المَاءَ | We need water |
| 3 | وَقَدْ أَكَّدَ الرَّئِيسُ أَنَّه سَيَتِمُّ مُلاحَقَةَ المُعْتَدِينَ | وَأَكَّدَ الرَّئِيسُ مُلاحَقَةَ المُعْتَدِينَ | The President confirmed the prosecution of offenders |
| 4 | الحَيَاةُ مِنْ دُونِ ابْتِلاءٍ لا تَسْتَحِقُ العَيْشَ | الحَيَاةُ دُونَ ابْتِلاءٍ لا تَسْتَحِقُ العَيْشَ | Life is worthless without griefs |

Fortunately, this verbosity serves in many ways like in part of speech tagging in particular; the concept of our new suggested approach is to consider those overused tokens to analyze and dismantling the text in a more scientific and intuitive way.

Arabic semantics can also be labeled verbose; a considerable amount of tokens do not contribute to the semantic structure of the sentence; this is due to the tendency to use excess of words, a common practice that prevail even with shortest and straightforward expressions that carry simple meanings, the following examples reflect this point:

| | Natural form | Basic form | Gloss |
|---|---|---|---|
| 1 | غَايَةُ الأَدَبِ أَنْ يَسْتَحِي المَرْءُ مِنْ نَفْسِهِ | الحَيَاءُ مِنْ الأَدَبِ | Modesty is a virtue |
| 2 | لا نُقَاتِلْ لِنَغْلِبْ، لَكِنَّنَا نُقَاتِلُ لِنُحَرِّرْ ، وَنَحْنُ لا نُقَاتِلْ لِلتَوَسُّعِ ،وَلَكِنَّنَا نُقَاتِلُ لِنَحْيَا | لا نُقَاتِلْ لِنَغْلِبْ وَنَتَوَسَّعُ بَلْ لِنُحَرِّرْ وَنَحْيَا | We fight not to conquer and annex but to liberate and live |

Having this said, Arabic is by nature highly ambiguous, in the sense that semantic information does not appear directly from the text, in other words, a text doesn't accurately represents the language's semantic content to the extent the text represents usage. The wordiness of language is beyond control since its early formation, only the context and the etymology of the vocabulary itself that can help in disambiguation as we will soon see in a simple example sentence. Generally, all languages are vague if the measure is how succinct or accurate is the semantic content that lies in the text, it depends on the author's culture and of course readers' good perception of the language thus, processing natural languages is subject to how systematic and consistent is the language semantics.

**Applications of natural language processing**

Being natural, Arabic doesn't follow specific shear logic in any aspect, although some scholars like Noam Chomsky presents theories and provisions that make natural languages look as if deliberately logically masterminded while in fact they are not, the whole matter is due to the activity of instinctive social practice happens to come in repetitive or regular patterns appears to be consistent or so linguists perceive it.

Technically, natural language processing is practiced nowadays through development of computer software applications that adapt to language principles; for this purpose we have to isolate different domains that specify language including but not necessarily limited to the two most intuitively known aspects (phonology and morphology), another visual domain does exist i.e. gestures e.g. nodding which are beyond the scope of this paper and cannot be referenced here[(2)], this paper is discussing NLP from the orthographical aspect, though a natural language cannot dispense with any two of the domains mentioned so far.

2 Writing is meant for readers, other practices include non-verbal communication e.g. facial expressions and gestures, sign language for deaf, and tactile representation as in Braille.

**Current status of Arabic language processing techniques**

Principally, language morphological processing is the analysis/synthesis of text; research community has developed an interest in applying latest NLP techniques to Arabic, this revealed a misunderstanding of the mechanism of language and computer capabilities considering the work accomplished so far as explained below:

1. most applications for Arabic language processing are directed to visual agents, this results in two issues:
    - some applications are designed to output solutions that mimic Arabic Syntactic Analysis (إعْرَاب) which is just a context-sensitive declarative analysis in nature, not to mention that it is basically a native speaker prospective.
    - output of these applications are mostly to be examined by experienced native speakers, not an automated application software, this is inadequate and add to the problem by putting the major load on human rather than machines.[3]
2. lots of applications are deeply involved in morphological analysis which in itself is not the final goal; stemmers, tokenizers and lemmatizers are good tools but has little role in disambiguating the language.
3. many resort to having multiple potential solutions to every word input, see model (2); mostly not required from the context; this is realistic and logical but impractical since multiple solutions on word level result in thousands of combination of possible solutions without feasible means to limit the number within reasonable range.
4. many applications directed to solve problems that are already solved, e.g. re-vocalizing Qur'anic verses, this contradicts the initial concept of using computers, the vocalized text -in itself- is not actually preferred in practice.

**Ambiguity of Arabic sentence**

We may list here plenty of models that clearly reflect the ambiguity Arabic, we have one simple example we use to test applications of different complexities including machine translation systems. This is to give an insight of how versatile and dramatic a change in meaning can occur to an supposedly simple sentence; in the following example we simulate the intuitive way that an Arabic sentence must be interpreted by an automated parser, same the way as human mind works in different contextual circumstances, the result is shown in Table 1 on the next page, parts in red denote personal names which are potential solution among any Arabic ordinary text.

3 The ultimate goal is to have final output adapted for human, perhaps in an intuitive written or audible form e.g. voice.

| | Structure | Example | Glossary |
|---|---|---|---|
| 1 | VSO | شَرِبَ أَحْمَدُ الحَلِيبَ | *Ahmad* has drank the milk |
| 2 | | شَرِبَ أَحْمَدُ الحَلِيب | It was Ahmad who drank the milk |
| 3 | | شَرِبَ أَحْمَدَ الحَلِيبِ | Someone has drank the most praised milk |
| 4 | VS | شُرِبَ أَحْمَدُ الحَلِيبِ | Ahmad the halib has been drunk (adj: halib) |
| 5 | | شُرِبَ /أَحْمَدُ الحَلِيبُ/ | *Ahmad Alhalib* has been drunk |
| 6 | | شُرِبَ /أَحْمَدَ/ الحَلِيبُ | the most praised milk has been drunk |
| 7 | | شُرِبَ أَحْمَدُ الحَلِيبُ | |
| 8 | VSO | شَرَّبَ أَحْمَدُ الحَلِيبَ | Ahmad has forced someone to drink milk |
| 9 | | شَرَّبَ /أَحْمَدُ الحَلِيبُ/ | *Ahmad Alhalib* has forced someone to drink something |
| 10 | | شَرَّبَ /أَحْمَدَ الحَلِيبَ/ | Someone has forced *Ahmad Alhalib* to drink |
| 11 | | شَرَّبَ أَحْمَدُ الحَلِيبَ | |
| 12 | | شُرِّبَ أَحْمَدُ الحَلِيبَ | Ahmad is forced to drink the milk |
| 13 | | شُرِّبَ /أَحْمَدُ الحَلِيبُ/ | Someone called *Ahmad Alhalib* is forced to drink |
| 14 | | شُرِّبَ /أَحْمَدَ الحَلِيبَ/ | |
| 15 | | شُرِّبَ أَحْمَدُ الحَلِيبَ | |
| 16 | | شُرْبُ أَحْمَدَ الحَلِيبَ | The way *Ahmad* drinks |
| 17 | | شُرْبُ /أَحْمَدُ الحَلِيبُ/ | The way *Ahmad Alhalib* drinks |
| 18 | | شُرْبُ /أَحْمَدَ/ الحَلِيبِ | The way the praised milk is drunk |
| 19 | | شُرْبُ أَحْمَدُ الحَلِيبَ | Drinking the most praised milk |
| 20 | VSO | شَرْبٌ أَحْمَدَ الحَلِيبَ | Sharb praised the milk |
| 21 | | شَرْبٌ أَحْمَدُ الحَلِيبِ | Sharb is the most praised kind of milk |

*Table 1: model analysis of simple sentence*

The table above summarizes the potential of Arabic language to defy most theories of an optimistic NLP specialist but we have to make it clear here that we started from an early stage in the analysis dealing with tokens in a context-free basis; we also have taken into account the fact that (شرب) can produce about 30,000[4] in addition to the possibility of it representing masculine or feminine personal name, a basic feature in Arabic; this may not be the case with other approaches but, basically, PoS taggers should at least assume the twenty solutions listed in the table for any other sentence of similar syntactical structure.

**Why using taggers?**

Two of the most important applications in modern time are Information Extraction and Information Retrieval that are heavily rely on text analysis which in turn depends on PoS tagger to obtain a tagged text. Most of recent works on Arabic language show tendency toward such respected areas depend on processing large amounts of text and its prospecting specific features such as patterns and clues that can be used to analyze the language; part of speech tagging comes early in the hierarchy of natural language processing applications, actually the second stage i.e. syntactic analysis as shown in Figure1, other functions of taggers include the following:

1. Word sense disambiguation.
2. Suggesting missing information in texts by investigating the missing parameters, such as the missing diacritics, in the case of Arabic.

The two above are vital processes in preparing tagged corpora used in statistical analysis which prove to be promising in this domain, untagged corpus is useful only in the statistical morphological analysis; tagged corpus is not a goal in itself, but is used as input for further processes.[5]

4 Including all affixes, please see Kalmasoft web page www.kalmasoft.com/KMAPS/mrlconj.htm
5 ???

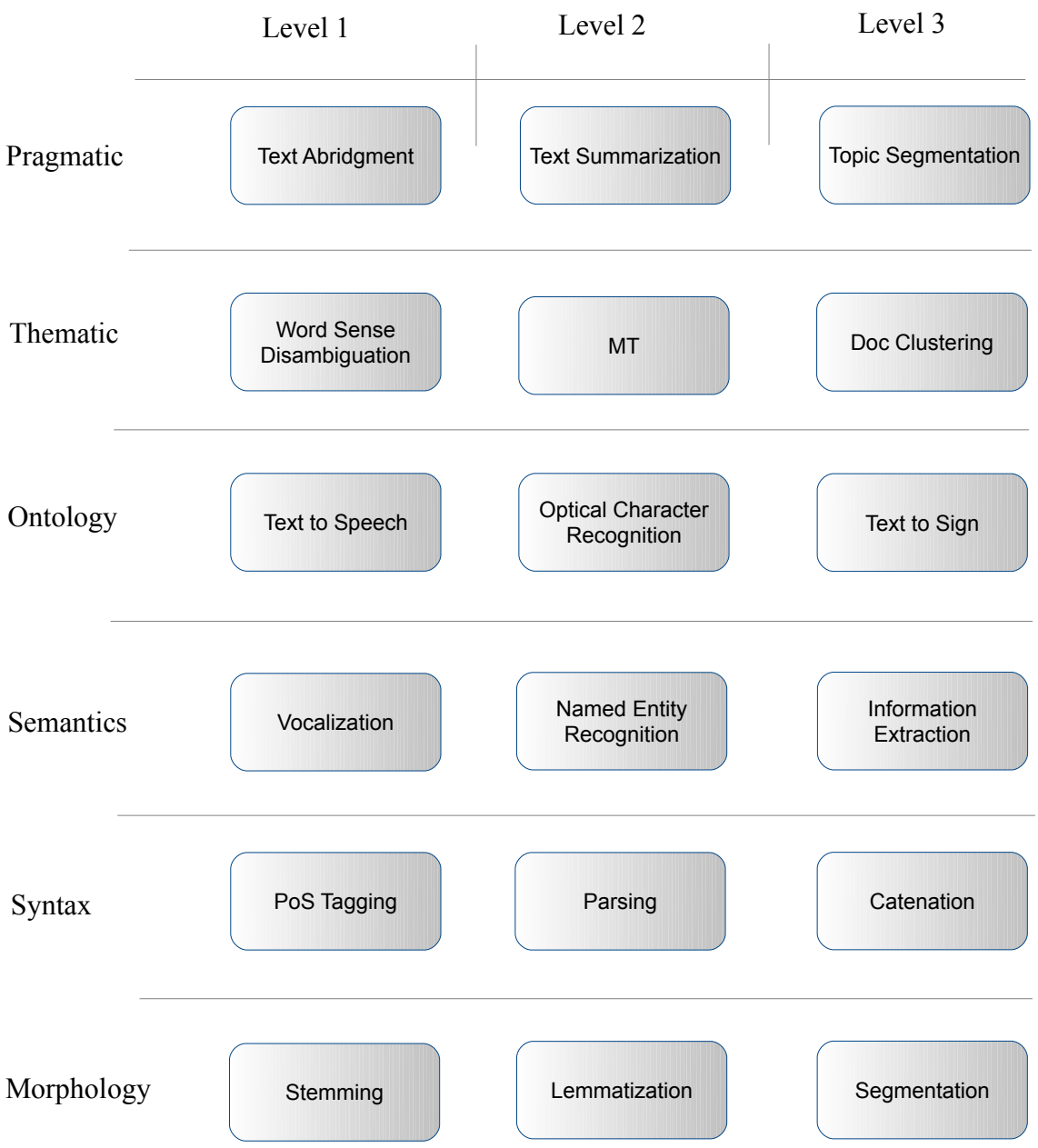


*Figure 1: hierarchy of the applications of natural language processing*

**Basics of taggers**

Part of speech taggers deals with texts; depending on the nature of language, for languages that have unambiguous sentence-ending markers (Chinese, Japanese) it is straight forward to identify the boundaries between sentences but as for Arabic the only available input in most cases is individual word (token) that is separated from the nearest neighbors by space or any other non-visual character; so Arabic used to be PoS tagged on a context-free fashion but taggers may also work on full sentence as the concept introduced in this paper. Generally, there are two approaches to tag text:

- A statistical technique (Example Based) based on Information Theory, successfully implemented to Arabic since it does not depend on the nature of language; application is initially given a training input (Training Sample), the output is highly text domain dependent and also taggers using this technique fail to overcome the OOV (Out of Vocabulary) gaps or to give smart solution for our simple example early presented; the reason behind is that people tend to use a small proportion of the total vocabulary of the language, the most remain unrepresented in daily usage. It goes without saying that the statistical technique requires huge samples of text which in itself requires manual pretagging; this obvious weakness manifests itself as the sample preparation process is undertaken by humans.
- Analytical technique (Rule Based) and is more difficult technique used in natural language processing, the applications using this technique require a profound knowledge of the underlying language.
- A hybrid technique which is a combination of the two above.

**Difficulties of Arabic language processing**

So many NLP applications with almost identical results only reveal one thing, similar methodologies; this partially caused by the misunderstanding of some aspects of the language we explain in the following pages; the most prominent issues of these can be attributed to the language itself; listed below are the reasons why Arabic NLP applications suffer such limitation:

**1. Multiple dialectical layers**

Some languages are described as being diglossiac that is, having two dialect layers one for elite and another for commoners; a cultural rather than social distinction since a native speaker may switch to any layer in any situation or even make a "blend" of all dialects a case that is not exceptional; Arabic-language is, in particular, *Triglossiac;* it comprises a classical highly disciplined dialect used for literary works such as the pre-Islamic poetry; and a simpler dialect (*Fus'ha*) a.k.a "Modern Standard Arabic" used in contemporary media in addition to less rigid colloquial version that varies depending on several social and geographical factors. The problem with this matter that overlap always happens between the three layers specially in public affairs and when quoting from historical cultural resources, colloquial Egyptian Arabic, for example, is used officially in the media.

## 2. Influence of other languages

This is not an issue specific to Arabic, languages may affect each other lexically by borrowing words and terminology, or syntactical by adopting the grammar as in the case of Lebanese and Syrian colloquial Arabic which adopted many Aramaic grammatical structures, Moroccan colloquial Arabic retains cognates from *Tamazight*; the same can be said for Egyptian colloquial Arabic which clearly adopted some Coptic grammatical rules and considerable vocabulary from old Egyptian. Romance languages such as French for example possess most of Latin grammatical features even more than other Romance languages while Italian is closest to Latin in terms of vocabulary. Most importantly is the literal translation that occurs due to incapability of Arabic to transmit scientific and technical terminologies as well as neologisms, most of loanwords kept intact but slightly adapted to fit in Arabic morphological structure, for example (تَلْفَـزَة) /təl'fəzə/ which is the Arabic word for "TV broadcasting", most of technological terms underwent similar treatment. In Arabic Gulf (e.g. Emirates, Oman, Bahrain, and Qatar) where is overlap between languages Arabic, Farsi, Urdu, Hindi, and English resulted in a creole emerged as a result of of social-economic interaction between heterogeneous groups from South Asia and Asia Minor.

## 3. Regional varieties

Arabic vernaculars have never been standardized or even officially discussed as Arabic extends in wide different regions each with its own culture and local preferences; in contrast to Chinese which spreads with minimal differences in Taiwan, Hong Kong, and main land China; also English has major differences between United Kingdom and the United States but minimal in Australia, Canada, India and South Africa. There are as many differences as the number of Arabic speaking states, unfortunately this exists in the most important dialect layer, the Modern Standard Arabic which is officially adopted by official media and represents a significant portion of the targeted Arabic text available. The table below shows part of MSA varieties across sample countries.

| Example | KSA | UAE | Egypt | Syria | Iraq | Tunis | Morocco | Algeria |
|---|---|---|---|---|---|---|---|---|
| Street | شارع | شارع | شارع | شارع | شارع | نَهْج | نَهْج | زَنْقة |
| Examination | إمتحان | إمتحان | إمتحان | إختبار | إختبار | مُسابَقة | مُسابَقة | مُسابَقة |
| Club | نادي | نادي | نادي | نادي | نادي | جَمْعِية | جَمْعِية | جَمْعِية |
| Procedures | إجراءات | إجراءات | إجراءات | إجراءات | إجراءات | مَسْطرة | مَسْطرة | مَسْطرة |
| Enrollment | تَسْجيل | تَسْجيل | إنخراط | تَسْجيل | تَسْجيل | إلتحاق | إلتحاق | إلتِحاق |
| Cellphone | جَوال | هاتف مُتَحرك | مَحْمُول | جوال | موبايل | هاتِف | أغنجي | بورتابل |
| Driving | قِيادة | سِياقة | قِيادة | سَوْق | سَوْق | سِياقة | سِياقة | سِياقة |

*Table 2: Arabic regional varieties*

Clearly, there is no de facto MSA that can be admitted; moreover, vernaculars are of cultural nature and most caused by urban interaction with western civilization inflows as evident an can be noted in modern terminologies such as (cellphone, driving) in the table above; the most notable case is of "Google" which has the orthographic variants ( گوگل، گوگل، جوجل، غوغول، غوغل، قوقل). Regional varieties also come as natural tendency to prefer specific morphological paradigm which can be referred to as (*Lexical Bias*)[6] for example, with C being the radical letters, it is common to use of the word form muCaC~aC (مُحَطَّم) "crushed" but not maCCuwC (مَحْطُوم) which is legitimate yet not found in any written text, it has the same meaning except that the former is slightly energetic while the later is more conforming to the common structural usage; morphological synthesis based on morphological templates may not be practical since the majority of vocabulary generated may not fit within the contemporary language paradigm; if applied blindly, the output will come up with words that look very uncommon to native speakers e.g. "مَرْغُوم، كاسُور، جَابِن، مِخْواف، كَواذِب، زُبَنَاء، رَابُوط، ضَوَارِب، شَبِيع" though legitimate and can actually be used.

## 4. Fuzzy boundaries between word classes

Some languages distinguish the classes of words with different types of affixes; Romance languages for example use capitalization to distinguish proper nouns and uses Latin meaningful affixes that distinctively determine the word class (e.g. tion, ity, ed, ese, ing) that is why stemmers and lemmatizers work fine with such agglutinative languages; Japanese uses three different sets of characters, one of which (Katakana) fully customized to represent foreign vocabulary, e.g. US President Barack Obama is written (バラク·オバマ). In contrast, Arabic anthroponyms (personal names) do not share such distinction because they fit many other parts of speech in the running text since they are ordinary dictionary words that may be come naturally in the lexicon; the same name is used for various purposes as being verb or adjective; this dynamic nature of Arabic vocabulary resulting in multiple faulty solutions given by NLP applications on both syntactic and semantic levels though a few markers and prefixes do exist to tag proper nouns such as (Abu, Ibn, Al, and Nisba).[7]

## 5. Semantic inconsistency

Arabic morphological system is inflectional based on roots, this feature is shared by the rest of the Semitic family languages (Aramaic and Hebrew) and Afro-Semitic (Tighrinia and Amharic); each Tri-consonantal root can hypothetically derive some 30,000[8], root would provide a great service to language if all those derivatives conform semantically to some unique set; unfortunately, this applies only to a limited number of roots, the principle of *Semantic Consistency* is not present in Arabic.

6 Lexical Bias is a new factor suggested by Kalmasoft to measure the deviation/closeness of a word to specific preferred morphological paradigm.

7 Nisba "relation" and Nasab "lineage" are now defaulted to geographical Nisba e.g. "Saddam Al-Tikriti" or Clan Nisba "Saiid Al-Hajiri"

8 With all possible inflections, please refer www.kalmasoft.com/KMAPS/mrlconj.htm.

Semantic consistency differs from synonymy which is defined with respect to certain senses of words e.g. (دَرَسَ التِّلْمِيذُ الوَاجِبَ ، دَرَسَ الحِمَارُ الشَّعِيرَ) "*the student studied his homework*; *the donkey thrashed the barley"*, both homographs are represented by the perfective (دَرَسَ) derived from the same root yet have very different meanings (study, thrash) see figure 1below for extra examples. The issue of *Semantic Dispersion*[9] can only be mitigated by generating all possible root inflections and possibly match them to their relative meanings a process that results in the so called *Full Form Dictionary* which referred to previously, this lexicon is by far useful for statistical purposes but not as part of interactive software e.g. spell checkers or text processing tool since a huge amount of vocabulary is engaged.

Semantic dispersion is a persistent problem inherent in the language, it has nothing to do with *Semantic Shift* which is associated with the evolution of the language, newly shifted Arabic terms like (عَرَبَة، حَافِلَة، سَيَّارَة) "*car, bus, automobile*" should not be taken for granted to have their corresponding meanings regardless of the context, they are well "known" today but may have different connotations if found within literary texts as old as in pre-Islamic poetry, a (حَافِلَة) /ħɑːfilə/ "bus" was then never been a vehicle run on four or six wheels but merely a female camel tamed for luggage transportation. Even worse, semantic shift affects even Arabic's most common vocabulary such as derivatives of root (ضرب) e.g. (ضَرْب، ضَرِيبَة، مَضَارِب، ضِرَاب، ضُرُوب، إضْراب) (*beating|multiplication, tax, rackets, fight, types, strike*) all are used today but are rather of very different meanings i.e. semantically dispersed. The problems simply lies in that assuming which is archaic/modern meaning comes with caveats since old meanings are still effectively in use specially in literature; below are few more examples:

| Arabic | IPA | Original meaning | Current meaning |
|---|---|---|---|
| عَرَبَة | /ʕərəbə/ | floating lumber drifted in a river current | car |
| حَافِلَة | /ħɑːfilə/ | female camel tamed to transport luggage | bus |
| سَيَّارَة | /səjˈjɑːrə/ | pedestrian | automobile |
| قُنْبُلَة | /qumbulə/ | troop | bomb/grenade |
| مُسَدَّس | /musədˈdəs/ | hexagonal | hand gun (revolver) |
| قَامُوس | /qɑːmuːs/ | deep bottom sea | dictionary |

*Table 3: Arabic semantic shift*

9 Semantic Dispersion is a new factor of value in the range [0:1] suggested by Kalmasoft to measure semantic divergence of set of words of same radical.

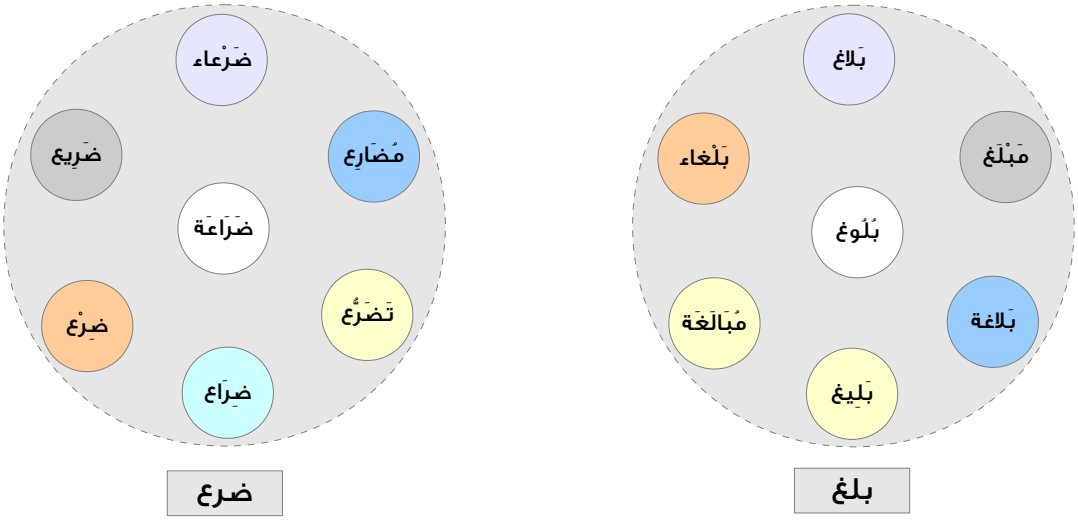


*Figure 1: semantic dispersion*

Left sphere: *busty, beseech, breast, puss, parallel, parallel, continuous, etc* . Right sphere: *eloquent, adolescence, amount, exaggeration, report, eloquence*

## 6. Metaphor excess

A key concept in Arabic language is the excessive usage of metaphors unmistakably evident to any linguist; the problem with metaphors that they do not add information as much as they do to improve the rhetorical characteristics or psychological effect on the receiver; good example sentences (رُزِقَ فُلان بِفـارِس – أزاحَـت الشَّـرِكَة السِّـتار) (*literally: A man be-gifted a knight*, *The company unveiled the curtain*) neither is even semantically close to *"gave birth to a child"* or *"started marketing a new commercial product"*, but rather rhetorical metaphors that depend entirely on the cognitive expertise of the listener and his cultural background that supports a good interpretation of the real meaning behind them; this obfuscation is beyond the control of a computer or any application software works solely on the basis of direct reasoning based solely on the written text. Arabic relies heavily on descriptive vocabulary, analyzing a relatively simple adjective like (أبْيَض) "white" will not only yield the color "white" but also the shape and size of something i.e. "egg shaped, oval", these attributes originated solely from the noun "egg", this is just to show how the language works.

## 7. Lack of diacritics

The second minor set of Arabic characters is usually not used in practice, they must be considered as real characters from the point of view any Arabic NLP application, since each is simply a grapheme and governed by spelling rules that do not differ substantially from other characters only in limited use. Diacritics come in two sets, three short vowels (*Fat'ha*, *Damma*, *Kasra*) roughly representing (a,u,i) and (*Shadda*, *Sukun*, *Madda*, *Fat'hatan*, *Damatan*, *Kasratan*), the ('*Alif Wasla* and Dagger '*Alif*) fall archaic but always present in heavily decorated text in calligraphy; Arabic shares this feature with modern Hebrew (Israeli) which abandoned the the *niqqud* (נִקּוּד).

The lack of diacritics in electronic documents is due to the difficulty of using typing techniques to add them to the document; technically, this can be traced back since the advent of typewriters when electronic keyboards started to be in Latin alphabet, Arabic letters added later with extra combination keys assigned for each single diacritic that most typists are unwilling to discover; moreover, diacritics are undesirable in text books and media since they are notorious for slowing down the reading speed to one third (triple time) compared ordinary unvocalized text.[10]

Taggers actually should not add these diacritics but be aware whether they are important from the context and how to deal with multiple solutions given in Table 1.

## 8. Non-standard punctuation

Less common in Arabic are the modern punctuation marks, since not until the 20th century (roughly around 1919) does punctuation became evidently present; punctuation is used today only in small-scale and in informal manner that not governed by convention nor themed style; it is up to the authors' educational backgrounds and editing skills to add them as necessary. Punctuation should not count in the tagging process but can be of a secondary importance even though it is not present in most of historical texts; basic punctuations marks such as question or exclamation marks or even numbering are never encountered in historical literature, at some point tagger may use punctuation on the assumption of the logical and pragmatic integrity of the text.

## 9. Linguistic inconsistencies

Being natural, Arabic language is undoubtedly has its own pitfalls of retaining integrity in all aspects, it is not a language that people officially agreed on but a one that naturally evolved and influenced by many other factors; anomalies do exist in Arabic in forms of instances that not following the presumed orthographical and syntactical rules; those are rather many and versatile ranging from Broken Plurals, extra non-voiced characters (e.g. final masculine plural *Waw*, *Alif Wasla*, final *Waw* in some personal names such as in (عَمْـرْو) mistakenly pronounced /ʕəmru/; plenty of homographs that span a wide range of text categories such as ACCaC (أَحْمَـد) "Ahmad"; strange morphological behavior of vowels, and so many exceptions as in diptotes to name a few.

Phonemically, half of Arabic letters (called *Sun Letters*) render the definite article (*al*) unvoiced if come following it, some diacritics add voice to word endings (*Tanwiin*), many nouns are phonemically modal and cannot be altered in any way regardless of their syntactical role. There never been agreement on how to represents special sounds like (CH, G, V, P), also dealing with dialectical phonemic differences is a complicated issue manifests itself in the way different groups pronounce letters like (ث، ج، ذ، ز، ص، ض، ظ، غ، ق، ك) count for more than one third of the language's alphabet.

10 A simple test run.

## 10. Common mistakes

Highly present in main stream literature and media spelling mistakes mostly affect the morphological analysis phase and thus have a direct impact on the PoS tagging; mostly introduced by new generations not showing much care for the language disciplines; language control institutions are not doing applausible efforts to protect it and the whole matters is left to community who are not obliged to abide by systematic old fashioned language.

### Religious considerations

Native speakers of Arabic have long claimed that Arabic is far more than a language or just a liturgical language of over 1 billion Muslims, rather, the language chosen by God to speak to mankind, this influences how Arabs perceive the world and express reality. This, in turn, has a profound impact on all linguistic aspects considering Arabic as (*sacred*) by regarding the language as the container of mostly important and as a highly honored liturgical representative text as the Qur'an; Consequentially, terms like (*language of Islam*), (*language of the Qur'an*) or even (*language of the dwellers of Paradise*) are so common between native speakers that can at best be regarded as profound speculations rather than serious academic judgment.

The issue is primarily conceptual and harm lies not only in the misjudgment of primacy and highly standardized medium of yet an arcane and powerful religion but in the assumption that it excels the other languages; such superiority assumption imposes a legendary rhetoric values that simply do not exist, the way of thinking of many linguists is deviated by the illusion that functional solutions come implied within the language itself in forms of covert formulas that decipher its complexity ready for practical implementation.

Many believe that Arabic's intrinsic features retained its authenticity over the time (though this sets Arabic as a dead language by definition); interestingly, this even led to some recent claims that strongly declaring Arabic as (*the origin of all languages*), technically speaking, the language is at best considered "an almost live" natural language.[11]

### What is the function of taggers?

Word sense disambiguation is the most important function of a PoS tagger; based on features and clues found in sentence the tagger may or may not output a true solution, this is partly true, the word (tea) for example does not provide enough information whether it is tea leaves, tea drink, or a tea gathering unless preceded (or followed) by some other clues to clarify the meaning; in contrary to this, chances are that a word may be self-evident but additional words render it ambiguous which is always the case when using metaphors.

11 Language is said to be "live" if it continues to evolve over time; having native speakers does not necessarily qualify it being "live" since many dead languages still have active speakers e.g. Latin in Vatican, Coptic in Coptic Church, Navajo to name a few; Arabic, accordingly, is almost live.

**What is the Deep PoS Tagging?**

The need for Deep PoS Tagger is to prove the fact that current works to analyze text in a context free basis is impractical; in the case of the Arabic in particular, a tagger should benefit from rule-based techniques that can take advantage language's wordiness; determining the major categories of speech (nouns, verbs, tenses, etc) will not help much in information extraction.

Analyzing our early example (Ahmad drank the milk) in Table 1 is not as easy as assumed but things are much different if we consider adding (عِنْدَما) /ʕindəmə/ "when" that functions as a conditional, if comes leading the whole sentence is then indicative active sentence; or subjunctive otherwise i.e. if the conditional is positioned at the end of the sentence. That is not all, in the first case, the token (شرب) is a verb by default, in the latter the word preceding the conditional is a noun by default; this situation is much better than the confusing analysis we have had before for the same example.

| | | | |
|---|---|---|---|
| 1 | عِنْدَما ‹شرب أحمد الحليب› | When ‹Ahmad has drunk the milk› | Indicative active sentence |
| 2 | ‹شرب أحمد الحليب› عِنْدَما | ‹Ahmad has drunk the milk› when ... | Subjunctive active sentence |

*Table 4: Applying function words*

By far and as long as the morphology involved, we may resort to this system to disambiguate sentences; more to this is the mixed model below which is extracted from a real-world example written in a basic pseudo-language:

(▯▯▯) أحد (▯▯▯) في أحد الـــ(▯▯▯) عليه، و(▯▯) لذلك (▯▯▯)،وفي الـــ(▯▯▯)، لكنه (▯▯▯▯▯▯)؛ لقد (▯▯▯▯)، لأنه (▯▯) أنـــي (▯▯▯) عليـــه، و(▯▯▯▯) أن

الـــ(▯▯▯) ليس له (▯▯) !وأثناء (▯▯▯▯▯) عن (▯▯▯) ثم (▯▯▯)؛ وأنا في (▯▯▯▯) إلى هنا، لم (▯▯) به الـــ(▯▯▯▯) على الـــ(▯▯▯▯)، وفجأة (▯▯▯) في الـــ(▯) من ال(

▯) مع (▯▯) حتى (▯▯▯)، ثم (▯▯▯) مني، فـــ(▯▯) إليكم. (▯▯▯)؛ إن هذا الـــ(▯▯▯)، فكيف (▯▯▯) دون أن (▯▯▯) أحد، وكيف (▯▯▯) بدون (▯▯▯) من (▯▯▯)؟!

*Model 1: model paragraph of pseudo language*

Though no one knows the meaning of words between parentheses, however, native speakers can easily realize that the paragraph is about a short story and soon become aware that (▯▯▯▯) and (▯▯▯) in the first line should be verbs or verb phrases and that (▯▯▯) is definitely a noun, and so on. It should be noted that content words in the paragraph were replaced with the funny meaningless pictographs between parenthesis leaving only function words which count for about 50% of the whole paragraph, this will finally help much in identifying the linguistic functions of all sentences in the paragraph.

By contrast, the example below shows part of typical tagger output, a context-free tagging for the single word that appears in the very first line showing how following this technique results in large amount of solutions intended for human eye though it can directly be input to subsequent applications for further processing since it has the minimal organized structure but of course the output is not perfect nor practical.

```
INPUT STRING: ذكرت
LOOK-UP WORD: *krt
 Comment:
 INDEX: P1W8
 SOLUTION 1: (*akarotu) [*akar-u_1] *akar/PV+tu/PVSUFF_SUBJ:1S
 (GLOSS): mention/cite/remember + I [verb]
 SOLUTION 2: (*akarota) [*akar-u_1] *akar/PV+ta/PVSUFF_SUBJ:2MS
 (GLOSS): mention/cite/remember + you [masc.sg.] [verb]
 SOLUTION 3: (*akaroti) [*akar-u_1] *akar/PV+ti/PVSUFF_SUBJ:2FS
 (GLOSS): mention/cite/remember + you [fem.sg.] [verb]
* SOLUTION 4: (*akarat) [*akar-u_1] *akar/PV+at/PVSUFF_SUBJ:3FS
 (GLOSS): mention/cite/remember + it/they/she [verb]
 SOLUTION 5: (*ak~arotu) [*ak~ar_1] *ak~ar/PV+tu/PVSUFF_SUBJ:1S
 (GLOSS): remind + I [verb]
 SOLUTION 6: (*ak~arota) [*ak~ar_1] *ak~ar/PV+ta/PVSUFF_SUBJ:2MS
 (GLOSS): remind + you [masc.sg.] [verb]
 SOLUTION 7: (*ak~aroti) [*ak~ar_1] *ak~ar/PV+ti/PVSUFF_SUBJ:2FS
 (GLOSS): remind + you [fem.sg.] [verb]
 SOLUTION 8: (*ak~arat) [*ak~ar_1] *ak~ar/PV+at/PVSUFF_SUBJ:3FS
 (GLOSS): remind + it/they/she [verb]
 SOLUTION 9: (*krt) [DEFAULT] *krt/NOUN_PROP
 (GLOSS): NOT_IN_LEXICON
```

*Model 2: typical tagger output*

Kalmasoft pursues a leading advanced work in PoS tagger that works beyond morphological analysis and based on functional logic based on semantic analysis to extract more information about the text by combining two of the most important stages, namely the morphological and syntactic analysis to keep multiple solutions as minimal as possible and to output more intuitive solutions based on a context-sensitive technique.

Deep PoS Tagger segments the text into semantically linked paragraphs (catenas) and then analyze each using First Order Logic relying not on ordinary sentence boundaries but on keywords and clues that other techniques neglected as being redundant or Stop Words, some can be found in references [6] and [7].

Stop Words are usually discarded when processing English for the reason of being highly repetitive tokens but in the Arabic such tokens are potential keywords that rather be considered as Function Words that come to serve as text deciphering tools.

| | 1 | | | | | 2 | | | 3 | | 4 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| عِنْدَما | Morphology | | | | | Syntax | | | Semantics | | Ontology |
| | class | pos | trans | phras | | S1 | S2 | S3 | temp | redun | category |
| 1 | | pre | | | | > | | | when | | |
| 2 | | app | | | | | –> | | | if | |
| 3 | | | | | | | <–> | | | | |
| 4 | | | | | | | | <#> | | | |

*Table 5: Deep Tagger implementation chart*

Using these function words is not straightforward, each should be carefully diagnosed profiled in a thorough map that lists its complete morphological, syntactic, and semantic functions together with a set exceptions each may have; abve is partial map of one of those function words.

The majority of such vocabulary (function words) will need only a simple effort to classify and implement in the tagger; most of these function words have very specific syntactical functions that seldom change even in local Arabic dialects e.g. the definite article (*al*) prepositions which precede only nouns, some may be quite versatile as the conjunction (*waw*) "and" that can be used with all word classes and comes in different semantic flavors shown in the example below:

| | | |
|---|---|---|
| <noun+noun> | حَضَرَ <الرَّئِيسُ وَالوُزَراءُ> | The <President and ministers> arrived |
| <verb+verb> | ثُمَّ <قامَ وَخَطَبَ> فِي الحُضُورِ | he <stood and addressed> the audience |
| <noun+verb> | وَتَحَدَّثَ <الرَّئِيسُ وَصَفَّقَ> الوُزَراءُ كَثِيراً | the <President> talked and ministers <applauded> much |
| <verb+noun> | ثُمَّ <غَادَرَ وَالوُزَراءُ> إِلى مَأْدُبَةِ غَداءٍ رَسْمِيَّةٍ | he <left with ministers> to an official lunch |

*Table 6: conjunction "Waw" as function particle*

Put simply, Kalmasoft Deep Tagger parses sentences in a context sensitive analysis on the basis of Lexical Functional Grammar (LFG), the tagger is best suited to unattended corpus preparation, the output may be directly fed into Statistical Analysis or Corpus Based Analysis application.

The tagger handles text as an infinite series of paragraphs (open ended linked catenas) in contrast to other morphological analyzer that handles text on a single isolated token basis, making use of the fact that people try not to maintain a "clean vocabulary" with no redundancy or contribute more to the structure of a succinct or plain Arabic, which seem to be in favor to the concept of context-sensitive technique regarding the abundance of 50% or more promising special extra text components; the illustration below reflects the mechanism underlying the deep tagger, please note that BPMN is used in a specific meaning.

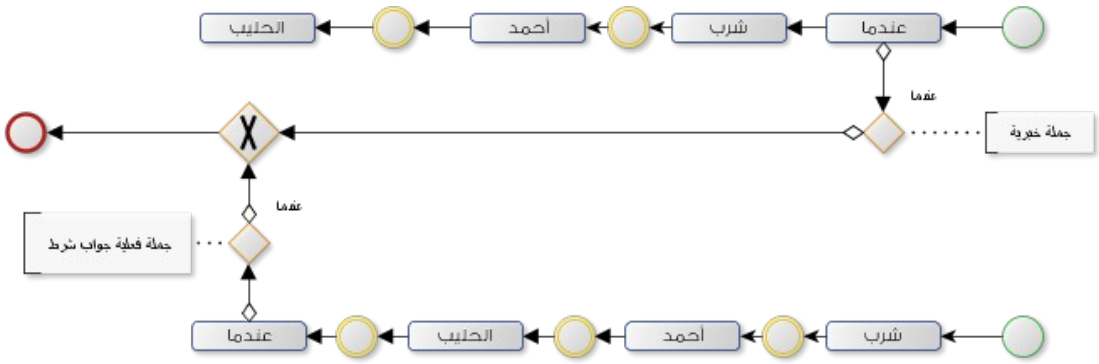


*Illustration 1: context sensitive sentence analysis*

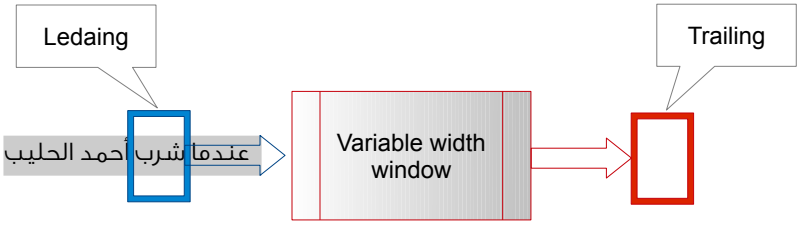


*Figure 4: augmented stateful sliding window*

In the example above, the underlying sentence is (Ahmad drank the milk), the selected function word is <عِنْدَما> that marks the whole sentence grammatical class and identify its sub-component according the preliminary Arabic grammatical rules, hence excluding four false positive solutions that assumes the token (شـرب) being infinitive; and also effectively cancels the possibility of the whole sentence being an Arabic full name[12]. Ideally, all the solutions shown in Table 1 are to be considered potentially equal except those with yellow background (16, 17, 18, 19).

The ideal tagger specifications is thus can be summarized assuming error free input:

- Must be able to predict the the grammatical role of the following based on the current function word.
- It should ideally give one solution for every single sentence.

Illustration 2 on the next page shows the functional structural components of the overall application.

12 A potential solution we dropped for brevity.

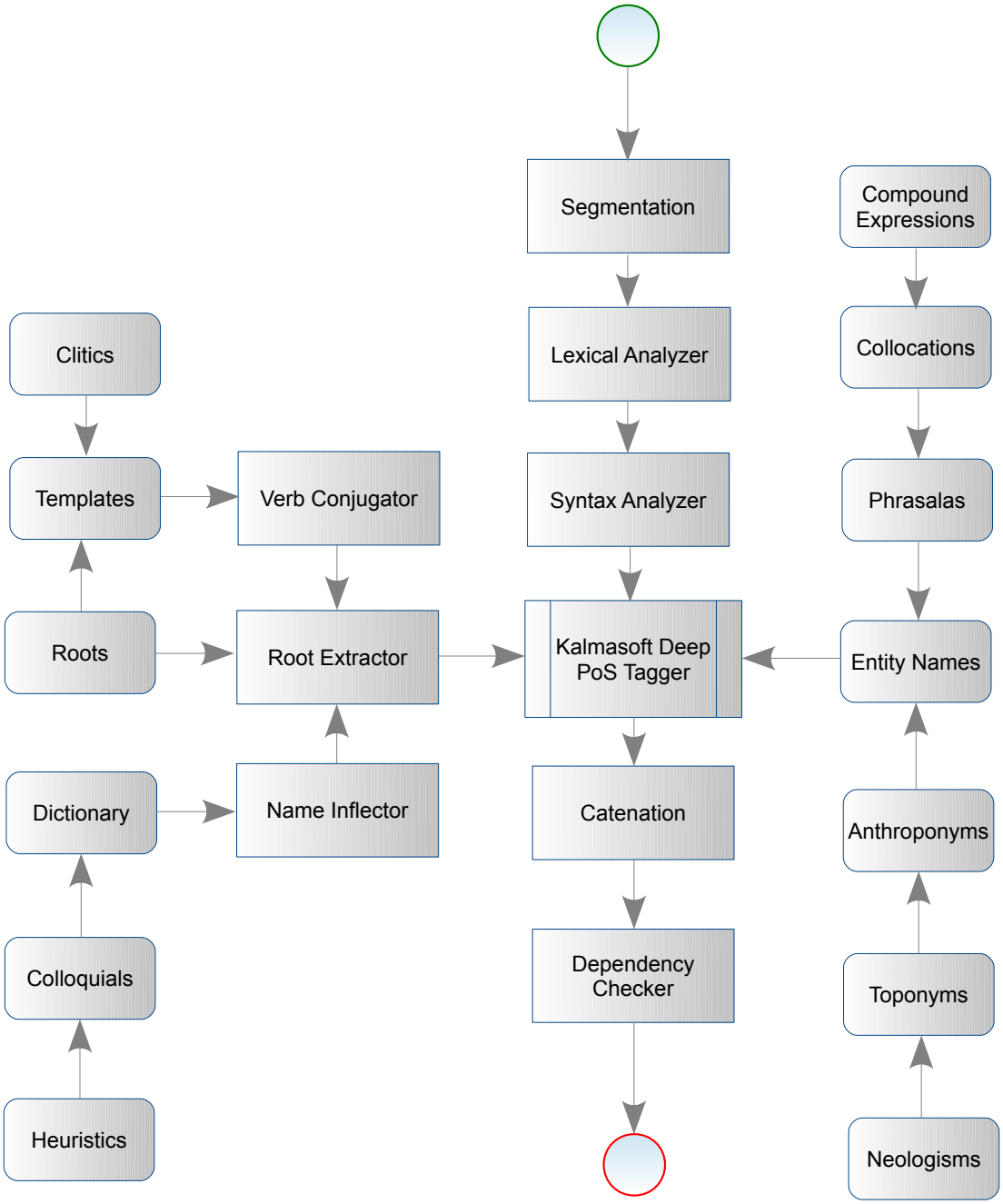


*Illustration 2: hierarchy of deep part of speech tagger*

Optimizing the solution is observably a bit more complicated and involves some ontology based on de jure knowledge (heuristics) that the computer lacks, this may finally give accurate results over other applications. The following facts can be reformulated and applied to determine and fine tune solution:

- milk is not a often "praised".
- (*Alhalib*) "the milk" is not used as a personal name except in rare extreme cases.
- Active voice may predominate over passive voice in text with no historical clues.
- Coercion (e.g. forcing someone to drink) may not be the default context of a sentence.

Applying the common heuristics above on facts such as that contained in the example, the number of possible solutions is limited to four which is the minimum for similar cases since we cannot identify the implications of any phrase from by simply analyzing the text as explained earlier in this paper.

### Output of Deep PoS Tagger

Kalmasoft Deep Tagger s designed for large text corpora, its output is formatted in such a way to be directed to applications for further automatic processing rather than human verification.

### Subsequent work

Finally, a tagger is not meant for itself but as an essential stage in the construction of Arabic language processing software and it should act separately but in a more comprehensive large system such as Kalmasoft Multitasking Arabic Processing System (MAPS).

### References

1. Kalmasoft Deep Tagger www.kalmasoft.com/KMAPS/msltag.htm
2. Interlingua Based Machine Translation www.kalmasoft.com/papers/PE-11-MT01.zip
3. Diglossia http://en.wikipedia.org/wiki/Diglossia
4. Stop_words http://en.wikipedia.org/wiki/Stop_words
5. Function_word http://en.wikipedia.org/wiki/Function_word
6. Stop_words http://arabicstopwords.sourceforge.net/
7. Function_words http://quizlet.com/10176800/arabic-grammar-function-words-flash-cards/
8. Effects of Stop Words Elimination for Arabic Information Retrieval, A Comparative Study http://www.ijcis.info/Vol4N3/Vol4N3PP119-133FS.pdf
9. Stop-Words Removal Algorithm for Arabic Language http://www.cs.wayne.edu/~eyad/sw_algo_arabic_2004.pdf
10. http://en.wikipedia.org/wiki/Niqqud