TAGARAB: A Fast, Accurate Arabic Name Recognizer Using High-Precision Morphological Analysis

John Maloney and Michael Niv
SRA International Corp.
4300 Fair Lakes Court
Fairfax, VA 22033
{john_maloney,michael_niv}@sra.com

Abstract

We describe a fast, high-performance name recognizer for Arabic texts. It combines a pattern-matching engine and supporting data with a morphological analysis component. The role of the morphological analysis in accurate name recognition is discussed. We also provide evaluations of both morphological analysis and name recognition.

1 Introduction

1.1 Roadmap

Arabic named entity recognition in texts in Arabic script is, to our knowledge, a little researched topic.[1] In this paper we describe a system, TAGARAB, that uses a generic pattern-matching engine, SRA's NetOwl TurboTag^{TM}, combined with an integrated morphological analysis process, which recognizes names at a high level of accuracy.[2] We first discuss the factors involved in recognizing names in Arabic. We then present a system description, focussing on the morphological analysis and the name recognition components. We also report the results of our evaluations of each component's performance.

Finally, we discuss the contribution of the morphological analysis to the name recognition.

1.2 Background to Arabic Name Recognition

Name recognition has emerged as an NLP technology that is effective and can provide high value to several different kinds of applications, such as clustering and summarization (Aone and Maloney, 1997; Aone et al., 1997). Development of such a capability for Arabic involved meeting some new challenges.

Like other Semitic writing systems, Arabic does not exhibit differences in orthographic case. Unlike in English-language mixed-case texts, therefore, there is no obvious clue such as initial capitalized letters to indicate the presence of a name constituent. In our experience with Thai and other non-case writing systems, this seems to effectively impose a requirement that there be some understanding of the morphological nature of each token — especially part-of-speech information. Having some morphological information allows one to make distinctions between likely and unlikely name constituents, which is particularly important when deciding where a name ends and the non-name context begins.

We partially motivate this for Arabic as follows:[3]

- Closed-class items are rarely, if ever, part of an Arabic proper name: [Hasan fī], with a preposition in second position, seems an unlikely Arabic name.[4]
- Inflected forms of verbs are rarely part of proper names: [yata9allamu], "he learns," is not a permissible name constituent.[5]
- Many lexical items, of course, are not used, or highly unlikely to be used, in proper names: verbs of speaking or cognition, for example, do not appear to occur in names and they do not frequently overlap in appearance with proper name constituents: [muHammad qultu] is an unlikely name.
- Items with subject or object suffixes are rarely, if ever, part of names: [yukaatibuhu], "He will correspond with him."[6]

[1] "Named entity recognition" is meant in the sense familiar from the Message Understanding Conferences (MUC) and the Multilingual Entity Task (MET). It refers to the identification and categorizing by type in unformatted text of names of persons, entities, and locations, as well as of numerics such as percentages and times/dates. In the following, we use "name" loosely in the same sense.

[2] This work has been funded by the Department of Defense, Contract No. MDA904-97-C-3065.

[3] The Arabic source used in this work was the newspaper *Al-Hayat*.

[4] We will usually vocalize Arabic for the sake of ease of comprehension (marked with square brackets), but will present it in consonantal transliteration when appropriate (marked with italics). Unobvious symbols: t̲ = thā, H = voiceless pharyngeal fricative; 9 = voiced pharyngeal fricative; $ = alif without hamza.

[5] There are some *ism* names that contain, at least historically, imperfect verb forms, such as [yazīd], "he increases." cf. Schimmel 1988, (p. 1).

[6] Other Semitic naming traditions, e.g. Akkadian, Amorite, Hebrew, do permit "sentence names" containing a finite verb or a predication. (cf. Schimmel 1988, p. 1).

In addition, there are many positive cues, such as titles, frequent given names, and so forth, that allow a system to identify names, but some morphological characterization of non-name portions is of critical importance. We will discuss this in more detail below in Section 4.

2 System Description

Figure 1 contains the architecture of TAGARAB.

Our system has two major modules: a Morphological Tokenizer and a Name Finder. The Morphological Tokenizer has the ability, in addition to performing lexical scanning that establishes word-level units, to add morphological features to tokens. Text encoded in ISO-8859-6 is first passed through this tokenizer and then the tokenized stream is processed by the Name Finder module which identifies names and other extraction targets and annotates the text with appropriate SGML tags for each extracted item.

2.1 Morphological Tokenizer

2.1.1 Description

The Morphological Tokenizer's basic task is to identify the sequences of words, punctuation symbols, numbers, existing SGML tags, etc., that comprise the input text. For each such "token," a description of what it has found is returned as a vector of up to 32 application-definable bits (e.g., PUNCTUATION, WORD, NUMBER). The Tokenizer is a very fast program, generated using the Flex scanner-generator from a tokenizer specification.

We decided to augment the tokenizer's usual role. While it still finds numbers and punctuation tokens, it treats an Arabic word (a contiguous sequence of Arabic letters) as a collection of one or more morpheme tokens, each with its own bit-vector of properties. The properties include those listed above, as well as morphology-specific properties whose nature and linguistic motivation is discussed in the next section.

Making the morphological analysis part of tokenization has the advantage of maintaining the high speed of SRA's TurboTag. An external morphology module — with a high computation overhead — would degrade performance.

Table 1 contains the features identified by the tokenizer.

Each token in the text receives some subset of the lexical types in Table 1. For example, a string such as *šrbh*, phonetically [šaribahu], "he drank it," receives the tokenizer types ARABIC, PERFECT, and SUFFIX. The first type means that the token comprises Arabic letters, the second that it is a Perfect verb form, and the third that there is a suffix attached. Note that in this case, the string is not broken up into pieces, such as stem and suffix. It remains a single token with information being added as to what the component pieces are. The only cases where the tokenizer splits off pieces of a string are where there is an attached conjunction (*wa* or *fa*) or an attached preposition (*la*, *ba*, *ka*), or both. In these cases, in place of an original string such as orthographic *wq$l*, phonetic [waqāla], "and he said," there will be two separate tokens: [wa] with the type information ARABIC and CONJ, and [qāla] with the type information ARABIC and PERFECT.

Basic Feature	**Specification**
Token type (unique)	
ARABIC	Arabic character string
UPPER	Upper-case English
LOWER	Lower-case English
CAP	Initial-cap English
MIXED	Mixed-case English
PUNC	Punctuation marks
INT	Arabic integer
REAL	Arabic number with decimal pt.
WRITTEN	Arabic number with commas
UNKNOWN	Unknown token
White space info (optional)	
NOWS	No white space before token
NL	New line char before token
DSPACE	Double space before token
TAB	Tab before token
BL	Blank line before token
SPACE	Space before token
Morphological Feature	
Part of Speech (if P.O.S = ARABIC, unique)	
CLOSED	Closed-class items; not prep.
CONJ	*wa-* and *fa*
PREP	Prepositions
NOUN	Nouns, including Verbal Nouns
IMPERFECT	Imperfects, including all moods
ADJ	Adjectives
ADV	Adverbs
PERFECT	Perfects
PART	Participles (not used as noun or adjective)
Feature (optional)	
DEF_ART	Definite Article
SUFF	Verbal and Nominal Suffixes

Table 1: Tokenizer Features for Arabic

Some of these tokenizer types are exclusive, such as PERFECT and IMPERFECT. A token cannot be both simultaneously. Others, however, such as NOUN and DEF_ART, can both be applied to a token.

2.1.2 Implementation

We initially developed the morphological analysis module as a sequence of 31 patterns in Perl's regular expression language. This allowed us to efficiently

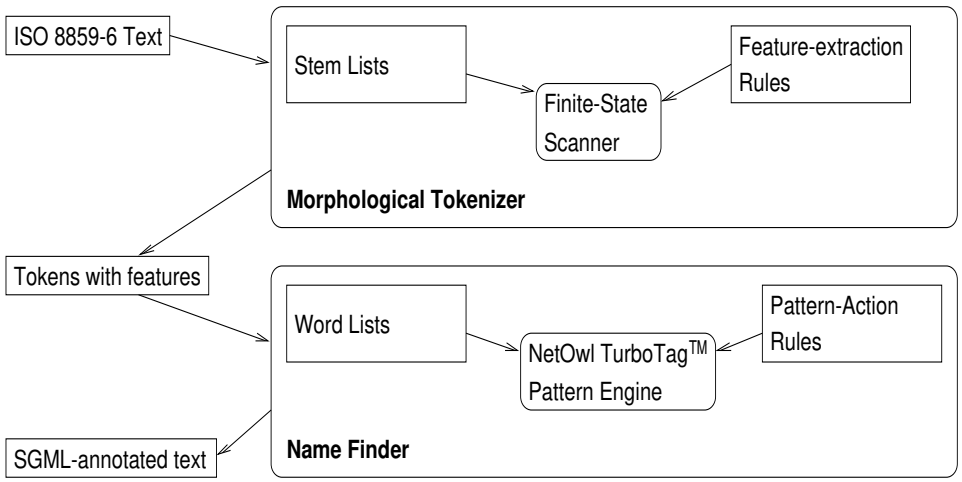

Figure 1: TAGARAB Architecture

develop and refine the patterns needed to recognize the various morphological word-shapes. When we plugged this version of the morphological analyzer into the original tokenizer, however, processing was quite slow due to the sequential nature of our morphological patterns and the backtracking nature of Perl's regular expression matcher. To compensate, we incorporated the morphological functionality directly into the Flex specification of the tokenizer. Whenever the Flex-generated scanner identifies an Arabic word, it dispatches the appropriate regular-expression to extract the separate morphemes from that word — a task that is beyond the capability of Flex.

The result is the fastest Arabic morphological analyzer we are aware of: The overall processing rate for TAGARAB is approximately 46 megabytes/hour on a Sun Ultra 1. Morphological processing by itself runs at about 190 megabytes/hour.

2.1.3 Linguistic Design

We had originally planned to develop a morphological capability that would be helpful in improving name recognition, as discussed in Section 1.2. In the following, we discuss the linguistic design of the morphological analysis.

Arabic is a highly inflected language. We believed that there are frequently enough surface cues in the shape of an Arabic word[7] to allow the assignment of the kind of morphological information described in Section 2.1.1. For example, inflected forms of derived verb stems such as [yaftatiHu], "he inaugurates," would seem to have an orthographic "shape" that is fairly unique in an Arabic text. We felt that this information could be exploited to successfully identify tokens as nouns, verbs (perfects or imperfects), etc., to a sufficiently reliable extent that the later name-recognition patterns could effectively make use of it.

The morphological analysis process consists of a series of regular expressions partially supported by lists of noun, verb, and adjective stems, as well as closed-class items. The regular expressions cover all allowable prefixes and suffixes for each stem type. Infixation phenomena, however, such as the infixed *t* of the Eighth Verbal Form,[8] are handled as variant forms in the verb stem list e.g., *'9tbr*, "he considered" and *9br*, "he crossed." No attempt is made to handle co-occurrence constraints among prefixes and suffixes, nor to assign voice. Likewise, no attempt is made to include contextual information, as is done with standard part-of-speech taggers. There is no attempt to handle ambiguity: The regular expression patterns are ordered, and the search for an analysis of a word stops at the first match. The token types are then assigned, and the form is not submitted to any other regular expressions.

Not all Arabic tokens are hit by one of these regular-expression patterns that provide morphological features. Although there is a mix of patterns supported by lexical information and patterns that operate entirely by rule (no supporting lexical data), the vast majority of matches appear to occur with the former set of patterns. In other words, the coverage of the morphological analysis is crucially dependent on the lexical data. There are 1051 noun forms, 813 verb forms, and 241 adjective forms. There is also a comprehensive list of closed-class items. The notion of "lexical item" in TAGARAB's lists is somewhat similar to the listing principle found in Landau : broken plural forms for nouns and ad-

[7] TAGARAB deals exclusively with the written form, i.e., without indication of short vowels.

[8] We use the usual terms for these forms found in Western grammars.

jectives receive an independent entry, much like different stems of verbs as mentioned previously. We make no effort to distinguish I and II forms of verbs, as these are not usually distinguished orthographically in *Al-Hayat*. In general, if there is no visible orthographic distinction in normal Arabic prose, we do not make a distinction in the lexical data.

We also entered forms both with and without *hamza*, as in *Al-Hayat* any form that may receive a *hamza* may also appear without it (even in the same text!).

Another important feature is that we entered only what seemed to us to be frequent lexical items in the lexical lists, and tried to do it in such a manner that what seemed intuitively the most likely reading of a form would be the one selected. This makes sense in the case of such a highly deterministic morphology and also given our time and resource constraints. We wanted to ensure that we got the right readings for a large number of highly frequent items, as this would be the most useful way to constrain the name-recognition patterns. Many high-frequency common nouns, verbs, adjectives, and other parts of speech in Arabic do not usually form part of names. As it turned out (See Section 4 below), this strategy worked quite well with person names but was less significant for organization names.

We also decided not to enter items that are used directly by the later name-recognition patterns, such as locations and given names, as these are accessed by the patterns through that module's own word lists and therefore having a morphological reading for them is not important (see Section 2.2).

In addition to the supporting lexical data, the ordering of the regular expressions also aided in determining the analysis selected. The regular expressions are grouped functionally, and in general the ones pertaining to closed class items apply first, then nouns, adjectives, perfects, and imperfects in that order. This had the effect that items which are highly "marked" as belonging to one category (e.g., verbs with a double pronominal object) would be captured appropriately by a verb-recognizing regular expression that looks for such suffixes, but that items that are not so highly marked (e.g., a simple third-person masculine perfect form) would be biased towards a reading according to the order of the regular expressions.

We were pleasantly surprised to learn that this kind of approach — although obviously simplistic in some ways — produces a very high level of precision in analysis (i.e., the parts of speech assigned tend strongly to be correct) and surprisingly good recall (i.e., there is good coverage of the corpus). We discuss our empirical results in Section 3.

In addition, the morphological information thus produced contributed substantially to the effectiveness of the name-recognition patterns. We discuss this in Section 4.

2.2 Name Finder

The Name Finder module of TAGARAB (see Figure 1) uses as input the tokens found by the Morphological Tokenizer with the basic and morphological features attached. The pattern-matching engine is SRA's NetOwl TurboTag^{TM}. It uses data consisting of a set of Pattern-Action rules supported by Word Lists. The latter consists of items such as personal titles that are used by the patterns to recognize names. The Pattern-Action rules use contextual and structural information about names to recognize them dynamically. They also make extensive use of the feature information coming from the Morphological Tokenizer. There is minimal permanent storage of names.

The Pattern-Action rules are written in a convenient specification language. They are not compiled, but are read at run time as part of engine initialization.

3 Evaluation

3.1 Morphology

3.1.1 Preparation

To evaluate the quality of the morphological analysis, we used SRA's tagging tool, TagTool^{TM}, to manually tag a set of documents for morphological analysis and part of speech.[9]

For this test document set, we randomly selected fourteen texts from the *Al-Hayat* CD-ROM not belonging to the name recognition training or testing sets. In addition to manually tagging them, we also ran TAGARAB over these fourteen texts and used a standard MUC-style scoring program to compare the morphological output of TAGARAB with the "answers" in the hand-tagged version.

3.1.2 Tagging Rules

We hand-tagged every token in the text, except for:

- Punctuation, Digits
- Tokens within person names
- Person titles
- Locations
- Location adjectives (e.g., [kuwaytī], [lubnānī])
- Adverbs or accusative forms of nouns or adjectives marked by *fatHatayn*.

[9]Because of staffing constraints and need for knowledge of Arabic, the same person worked on both the development of the morphology component, the name patterns, and also hand-tagged the test set. To remove as far as possible any taint, we did not change the system or any supporting data once the manual tagging began.

These exceptions exist because the morphology component did not attach features to these items for the reasons given in Section 2.1.3. As a result of not hand-tagging them, the scoring program judged as spurious any morphological features found by the system for such items.

We tagged the test set contextually, again in accordance with the design of the morphological component. The most important effect was on the feature PART. We found participles which act usually as a noun, (e.g., [al-mašrū9], "the plan", [al-muwaẓẓaf], "the employee"), usually as an adjective (e.g., [al-bayt al-majhūl], "the unknown house") or seem to be freely used in either reading (e.g., [muslim], "Muslim,"). We tagged these participles contextually as nouns or adjectives. One effect of this was that the number of items tagged as participles was quite low (in effect, only when they are used predicatively).

In all, the evaluation corpus contains 3214 tokens, of which 2324 are Arabic words. 1879 of the latter received morphological features when hand-tagged.

3.1.3 Morphology Evaluation

The scores for the morphology component are given in Table 2.[10]

Since we did not have access to a morphological analyzer that produces all possible readings for forms based on a large lexicon, we do not have a picture of the total morpho-lexical ambiguities in our evaluation texts. However, despite the small lexicon we manually built, the overall recall is reasonable (73.0%), and it also holds up well for most of the major open class items: perfects (72.7%), imperfects (60.2%), and nouns (66.8%). The low recall in adjectives (28.8%) is due to the fact that we did not make many lexical entries for adjectives. Since adjectives do not come first in the Arabic noun phrase, and since we use the morphological information to constrain the name patterns, tagging the head noun in a noun phrase is what is generally necessary, not tagging the adjective.

What is striking in the above table is the high precision across all the categories, with the exception of adjectives and participles, the latter a very small set for the reasons set out in Section 3.1.2. Precision is consistently above 90%. We interpret this to mean that a manually built system with a moderate lexicon, having the capacity to only select one reading for a given form and not paying any attention to a word's context, is capable of a very significant amount of morphological disambiguation in Arabic.

Our results are also consistent with the results of Levinger *et al.* for the structually similar Hebrew. Levinger *et al.* discovered that non-context-based morphological analysis preferring the most likely morpho-lexical analysis (generated using a statistical algorithm) gives extremely good results.

Table 3 shows the collisions among the tags.

The most common confusions were between perfects and nouns in both directions. The system tagged the following tokens as nouns where the human tagged them as perfects: *nšr* (2x), "he/it published," *b9ṯ*, "he/it sent," *Hdṯ*, "it happened," *wṣfthm*, "we described them,"[11] *ṣdrt*, "was issued, appeared," *w$ṣl*, "he/it continued."[12]

Conversely, there were 16 cases where TAGARAB considered a token as a perfect, and the human tagged it as a noun. As with the previous case, the great majority were confusions of the perfect with a derivationally related noun or verbal noun (e.g., *qtl*, "he killed" or "killing.") Despite the small numbers of such collisions in the sample, it seems to us that this is the most difficult disambiguation task, at least at the part of speech level, since verbal nouns plus the semantic subject or object in an *iḍāfa* construction can look much like a finite verb plus subject or object. Clearly, context or a higher level of syntactic/ semantic understanding is required to differentiate the two readings.

On the other hand, the other major confusion revealed by this table, Noun/Adjective and Adjective/Noun, is one that seems easily remedied by building in some knowledge of short context into the morphology component. For example, the following three examples (and the rest resemble these) are cases where the system selects a noun reading for the adjective within the scope of a noun phrase: [lāji'ūna muslimūn], "Muslim refugees," [biṣūratin xaaṣa], "in a special form," [mu'tamaruhu al-ṣiHāfī], "his news conference." In these cases, the adjectives also have noun readings, but the local context shows clearly which reading is correct.

These results identify specific areas of morpho-lexical ambiguity bringing into focus where additional contextual cues are needed for better ambiguity resolution.

3.2 Evaluation of Name Patterns

The scores for the name recognition in TAGARAB over the training set of texts are given in Table 4. The blind scores are given in Table 5.

[10] The column headings are the standard ones from MUC: POS: possible number of points (one point for identifying a constituent boundary, another for identifying its category), ACT: actual responses given, COR: correct answers, PAR: boundary errors, INC: category labelling errors, SPU: responses given that are not in answer key, MIS: items in key missing from response, REC: recall (COR/POS), PRE: precision (COR/ACT), F-M: f-measure $((2 \cdot \text{PRE} \cdot \text{REC})/(\text{PRE} + \text{REC}))$.

[11] In this case, TAGARAB had identified the initial *w* as the conjuntion *wa* and the rest of the string as a noun, *ṣfthm*, "their property."

[12] Similar to *wṣfthm*. TAGARAB took the inital *w* as the conjunction and took the rest of the string as a noun.

Type	POS	ACT	COR	PAR	INC	SPU	MIS	REC	PRE	F-M
CLOSED	336	326	322	0	4	0	10	95.8	98.8	97.3
PERFECT	300	240	218	0	18	4	64	72.7	90.8	80.7
CONJ	286	306	284	0	0	22	2	99.3	92.8	95.9
PREP	772	696	674	0	22	0	76	87.3	96.8	91.8
ADJ	320	136	92	0	40	4	188	28.8	67.6	40.4
PARTICIPLE	12	8	6	0	2	0	4	50.0	75.0	60.0
IMPERFECT	186	124	112	0	10	2	64	60.2	90.3	72.3
ADV	32	28	26	0	2	0	4	81.2	92.9	86.7
NOUN	1514	1072	1011	1	50	10	452	66.8	94.3	78.2
TOTAL	3758	2936	2745	1	148	42	864	73.0	93.5	82.0

Table 2: Final Scores for Morphological Tokenizer (see footnote 10)

Resp\Keys	NOUN	PREP	PERFECT	CLOSED	ADV	PART	IMPERFECT	ADJ	TOTAL
NOUN		2	7		1	1	4	17	32
PREP									0
PERFECT	16	9		1			1	1	28
CLOSED									0
CONJ	6		1						7
ADV				1					1
PART								2	2
IMPERFECT			1						1
ADJ	3								3
TOTAL	25	11	9	2	1	1	5	20	74

Table 3: Morphological Tag Collisions (Rows=System Responses; Columns=Keys)

Pattern performance followed our experience with other languages, except for the recognition of time expressions. Usually, these scores run in the mid- to high-nineties on a test corpus, but the rich variety of time and date formulas hindered scoring very high here. The scores for Arabic are consistent with scores for other languages (Thai, Chinese, Japanese) where there is no orthographic case information.

4 Contribution of Morphology to Name Recognition

As described in Section 2.2, name recognition is performed by a set of Pattern-Action rules ("patterns") supported by data in the form of word lists, and in the case of TAGARAB, reliance on morphological information.

To investigate the role of the latter in supporting name recognition, we performed an experiment where we turned off the morphological analyzer portion of the tokenizer to see what the impact on the patterns would be. "Turning off" the morphology simply means that we substituted a tokenizer that does not cleave off clitics, and only generates the token types under Basic Feature in Table 1 and none of those under Morphological Feature. This had the effect that name-finding patterns which had previously accessed the morphological features provided by the tokenizer could no longer do so. The patterns themselves were not changed in any way. The results are contained in Table 6. It should be compared with Table 5.

NUMBER and TIME do not show much change with morphological token information removed. This is not surprising, since the patterns identifying these elements rely on word lists of month names, written-out numerals, etc. We handled the different inflected forms of such items as words for time units ([9ām], "year," etc.) by simply listing all possibilites (singular, dual, plural, masculine, feminine) as separate entries. This is a reasonable strategy given the relatively small number of inflected forms for these items. In addition, relying on "static" word lists rather than "dynamically" generated morphological information coming from the tokenizer reduces the risk of error.

PERSON names are affected the most by not having the morphological features available: Recall drops 12.2 points and Precision 24.1. By contrast, ENTITY names are less affected: Recall loses 7.4, and Precision 7.2. This large difference in the drop in Precision (24.1 points for Person vs. 7.2 for Entity) is due to the difference in pattern-writing "styles" for the two name types. For PERSON names in Arabic, there are few structural or contextual clues: we had

Type	POS	ACT	COR	PAR	INC	SPU	MIS	REC	PRE	F-M
NUMBER	314	314	307	5	0	2	2	97.8	97.8	97.8
ENTITY	2144	2070	1740	58	44	228	302	81.2	84.1	82.6
TIME	1612	1574	1467	41	2	64	102	91.0	93.2	92.1
LOCATION	2640	2680	2534	30	10	106	66	96.0	94.6	95.3
PERSON	1926	1946	1778	36	4	128	108	92.3	91.4	91.8
TOTAL	8636	8584	7826	170	60	528	580	90.6	91.2	**90.9**

Table 4: Name Recognition Scores for Arabic: Training Set

Type	POS	ACT	COR	PAR	INC	SPU	MIS	REC	PRE	F-M
NUMBER	264	262	256	4	0	2	4	97.0	97.7	97.3
ENTITY	2024	1880	1557	77	84	162	306	76.9	82.8	79.8
TIME	1596	1416	1288	60	4	64	244	80.7	91.0	85.5
LOCATION	3468	3128	2957	61	20	90	430	85.3	94.5	89.7
PERSON	2466	2180	1879	157	16	128	414	76.2	86.2	80.9
TOTAL	9818	8866	7937	359	124	446	1398	80.8	89.5	**85.0**

Table 5: Name Recognition Scores for Arabic: Blind Set

available a list of titles and given names which served as the major indicators of the presence of a personal name. For the remainder of the name, consisting of an uncertain number of tokens extending out beyond the given name, there was not much to distinguish where the name ended and the non-name context began, except the morphological information provided by the tokenizer.

Islamic names involving elements such as [ibn], [bin], [abu], etc., were exploited as much as possible, but names with these elements were surprisingly rare in the *Al-Hayat* articles.[13]

In effect, the patterns recognizing person names relied heavily on the presence of any morphological feature to rule out a given token as being part of the personal name. Without this information, there was, for example, a huge increase in the number of spurious names (names tagged by TAGARAB but with no equivalent in the human-tagged keys; under the SPU column in the score reports). One typical example is the vocalized string [mudīr al-9amaliyāt alsiyāsīya], "Director of Political Operations," which the system took as a name, but which is actually an appositive to a preceding actual person name. The word [mudīr] is on the title list signalling the onset of a name and there is nothing to constrain the next two tokens from being consumed. If morphological type information were available signalling that [al-9amaliyāt] is not a likely name constituent, then the pattern would not have succeeded.

There was also increase in boundary errors (under the PAR column in the score reports). These were those cases where the name pattern "didn't know where to stop," as in [al-sayyid muHammad albaGdādī ka'annahu], where the last element would have received the feature CLOSED from the Morphological Tokenizer since it is a subordinating conjunction with a person suffix. A further side-effect of such a boundary error, as well as the spurious names, is that an element such as [ka'annahu], once it is considered part of a name by the system, will be used to recognize variant short forms of that name. This has the result that any [ka'annahu] in the article is subject to being classified as a person name.

By contrast with person names, the patterns generating entity names were less affected by the lack of morphological information since they were able to exploit more specific name structure. For example, a pattern for a specific class of entity names might include both the initial word [majlis], "council," and have as final boundary marker an element such as a location adjective. The latter is a separate word list independent of the morphological component of the system. In effect, there are more specific structural indicators of entity names than there are of person names, so the patterns are written differently, with less reliance on morphological information.

The final category of names, LOCATION, shows a large increase in missing names (under the column MIS in the score reports). The reason for this might not be obvious, since location names do not receive any morphological features (see 2.1.3). However, the morphological capability within TAGARAB identifies clitic items such as the conjunctions [wa] and [fa], as well as the prepositions [ba], [la], and [ka], as described above in Section 2.1.1. If there is no mor-

[13] We do not have direct statistics on different name types in the source (e.g., Islamic, Russian). However, we do have statistics on the relative yields of the patterns, which are organized themselves in terms of name types.

Type	POS	ACT	COR	PAR	INC	SPU	MIS	REC	PRE	F-M
NUMBER	264	262	255	5	0	2	4	96.6	97.3	97.0
ENTITY	2024	1862	1407	105	148	202	364	69.5	75.6	72.4
TIME	1596	1392	1255	67	6	64	268	78.6	90.2	84.0
LOCATION	3468	2802	2536	82	36	148	814	73.1	90.5	80.9
PERSON	2466	2542	1579	289	8	666	590	64.0	62.1	63.1
TOTAL	9818	8860	7032	548	198	1082	2040	71.6	79.4	**75.3**

Table 6: Name Recognition Scores for Arabic Without Morphological Features: Blind Set

phological capability, then these clitics are not recognized anywhere. This has the effect that location names with a clitic attached will not be recognized, since the system no longer tokenizes the location as a separate item.

5 Summary

We have described a system for recognizing names, dates, times, and numerics in Arabic-language text through a combination of high-precision morphological analysis and a subsequent component that recognizes the named entities. Although highly deterministic and not taking account of context, the morphological analysis component removes a great deal of morpho-lexical ambiguity and has the side-effect of demonstrating that the true difficulties in Arabic morphological ambiguity might be limited to specific contexts.

In addition, we have shown that named entity recognition in Arabic can be performed at levels consistent with other non-case languages despite great differences in structure between them. Finally, we have shown that morphological information is crucially important to effective Arabic name recognition.

References

Chinatsu Aone and John Maloney: "Reuse of a Proper Noun Recognition System in Commercial and Operational NLP Applications," in *Proceedings of ACL'97 Workshop on From Research to Commercial Applications: Making NLP Technology Work in Practice*, Madrid, Spain, 1997.

Chinatsu Aone, Mary Ellen Okurowski, James Gorlinsky, and Bjornar Larsen: "A Scalable Summarization System Using Robust NLP," in *Proceedings of ACL'97 Workshop on Intelligent Scalable Text Summarization*, Madrid, Spain, 1997.

Jacob Landau. *A Word Count of Modern Arabic Prose.* American Council of Learned Societies, New York, 1959.

Moshe Levinger et al. "Learning Morpho-Lexical Probabilities from an Untagged Corpus with an Application to Hebrew." *Computational Linguistics*, Vol. 21/3, 1995.

Annemarie Schimmel, *Islamic Names.* Edinburgh University Press. Edinburgh, 1989.