A Comprehensive NLP System for Modern Standard Arabic and Modern Hebrew

Morphological analysis, lemmatization, vocalization, disambiguation and text-to-speech

Dror Kamir
Melingo Ltd.
16 Totseret Haaretz st.
Tel-Aviv, Israel
drork@melingo.com

Naama Soreq
Melingo Ltd.
16 Totseret Haaretz st.
Tel-Aviv, Israel
naamas@melingo.com

Yoni Neeman
Melingo Ltd.
16 Totseret Haaretz st.
Tel-Aviv, Israel
yonin@melingo.com

Abstract

This paper presents a comprehensive NLP system by **Melingo** that has been recently developed for Arabic, based on **Morfix**™ – an operational formerly developed highly successful comprehensive Hebrew NLP system.

The system discussed includes modules for morphological analysis, context sensitive lemmatization, vocalization, text-to-phoneme conversion, and syntactic-analysis-based prosody (intonation) model. It is employed in applications such as full text search, information retrieval, text categorization, textual data mining, online contextual dictionaries, filtering, and text-to-speech applications in the fields of telephony and accessibility and could serve as a handy accessory for non-fluent Arabic or Hebrew speakers.

Modern Hebrew and Modern Standard Arabic share some unique Semitic linguistic characteristics. Yet up to now, the two languages have been handled separately in Natural Language Processing circles, both on the academic and on the applicative levels. This paper reviews the major similarities and the minor dissimilarities between Modern Hebrew and Modern Standard Arabic from the NLP standpoint, and emphasizes the benefit of developing and maintaining a unified system for both languages.

1 Introduction

1.1 The common Semitic basis from an NLP standpoint

Modern Standard Arabic (MSA) and Modern Hebrew (MH) share the basic Semitic traits: rich morphology, based on consonantal roots (*Jiðr / Šoreš*)[1], which depends on vowel changes and in some cases consonantal insertions and deletions to create inflections and derivations.[2]

For example, in MSA: the consonantal root /ktb/ combined with the vocalic pattern CaCaCa derives the verb *kataba* 'to write'. This derivation is further inflected into forms that indicate semantic features, such as number, gender, tense etc.: *katab-tu* 'I wrote', *katab-ta* 'you (sing. masc.) wrote', *katab-ti* 'you (sing. fem.) wrote, *?a-ktubu* 'I write/will write', etc.

Similarly in MH: the consonantal root /ktv/ combined with the vocalic pattern CaCaC derives the verb *katav* 'to write', and its inflections are: *katav-ti* 'I wrote', *katav-ta* 'you (sing. masc.)

[1] A remark about the notation: Phonetic transcriptions always appear in Italics, and follow the IPA convention, except the following: ? – glottal stop, ¿ – voiced pharyngeal fricative ('Ayn), đ – velarized d, ś – velarized s. Orthographic transliterations appear in curly brackets. Bound morphemes (affixes, clitics, consonantal roots) are written between two slashes. Arabic and Hebrew linguistic terms are written in phonetic spelling beginning with a capital letter. The Arabic term comes first.

[2] For a review on the different approaches to Semitic inflections see Beesley (2001), p. 2.

wrote', *katav-t* 'you (sing. fem.) wrote', *e-xtov* 'I will write' etc.

In fact, morphological similarity extends much further than this general observation, and includes very specific similarities in terms of the NLP systems, such as usage of nominal forms to mark tenses and moods of verbs; usage of pronominal enclitics to convey direct objects, and usage of proclitics to convey some prepositions. Moreover, the inflectional patterns and clitics are quite similar in form in most cases. Both languages exhibit construct formation (*Iđa:fa / Smixut*), which is similar in its structure and in its role. The suffix marking feminine gender is also similar, and similarity goes as far as peculiarities in the numbering system, where the female gender suffix marks the masculine. Some of these phenomena will be demonstrated below.

1.2 Lemmatization of Semitic Languages

A consistent definition of **lemma** is crucial for a data retrieval system. A lemma can be said to be the equivalent to a lexical entry: the basic grammatical unit of natural language that is semantically closed. In applications such as search engines, usually it is the lemma that is sought, while additional information including tense, number, and person are dispensable.

In MSA and MH a lemma is actually the common denominator of a set of forms (hundreds or thousands of forms in each set) that share the same meaning and some morphological and syntactic features. Thus, in MSA, the forms: ?*awla:d*, *walada:ni,* despite their remarkable difference in appearance, share the same lemma *WALAD* 'a boy'. This is even more noticeable in verbs, where forms like *kataba, yaktubu, kutiba, yuktabu, kita:ba* and many more are all part of the same lemma: *KATABA* 'to write'.

The rather large number of inflections and complex forms (forms that include clitics, see below 1.5) possible for each lemma results in a high total number of forms, which, in fact, is estimated to be the same for both languages: around 70 million[3]. The mapping of these forms into lemmas is inconclusive (See Dichy (2001), p. 24). Hence the question rises: what should be defined as lemma in MSA and MH.

[3] For Arabic - see Beesley (2001), p. 7 For Hebrew - our own sources.

The fact that MSA and MH morphology is root-based might promote the notion of identifying the lemma with the root. But this solution is not satisfactory: in most cases there is indeed a diachronic relation in meaning among words and forms of the same consonantal root. However, semantic shifts which occur over the years rule out this method in synchronic analysis. Moreover, some diachronic processes result in totally coincidental "sharing" of a root by two or more completely different semantic domains. For example, in MSA, the words *fajr* 'dawn' and *infija:r* 'explosion' share the same root /fjr/ (the latter might have originally been a metaphor). Similarly, in MH the verbs *pasal* 'to ban, disqualify' and *pisel* 'to sculpture' share the same root /psl/ (the former is an old loan from Aramaic).

In Morfix, as described below (2.1), a lemma is defined not as the root, but as the manifestation of this root, most commonly as the lesser marked form of a noun, adjective or verb. There is no escape from some arbitrariness in the implementation of this definition, due to the fine line between inflectional morphology and derivational morphology. However, Morfix generally follows the tradition set by dictionaries, especially bilingual dictionaries. Thus, for example, difference in part of speech entails different lemmas, even if the morphological process is partially predictable. Similarly each verb pattern (*Wazn / Binyan*) is treated as a different lemma.

Even so, the roots should not be overlooked, as they are a good basis for forming groups of lemmas; in other words, the root can often serve as a "super-lemma", joining together several lemmas, provided they all share **a semantic field**.

1.3 The Issue of Nominal Inflections of Verbs

The inconclusive selection of lemmas in MSA and MH can be demonstrated by looking into an interesting phenomenon: the nominal inflections of verbs (roughly parallel to the Latin participle, see below). Since this issue is a good example both for a characteristic of Semitic NLP and for the similarities between MSA and MH, it is worthwhile to further elaborate on it.

Both MSA and MH use the nominal inflections of verbs to convey tenses, moods and aspects. These inflections are derived directly from the verb according to strict rules, and their forms are pre-

dictable in most cases. Nonetheless, grammatically, these forms behave as nouns or adjectives. This means that they bear case marking in MSA, nominal marking for number and gender (in both languages) and they can be definite or indefinite (in both languages). Moreover, these inflections often serve as nouns or adjectives in their own right. This, in fact, causes the crucial problem for data retrieval, since the system has to determine whether the user refers to the noun/adjective or rather to the verb for which it serves as inflection.

Nominal inflections of verbs exist in non-Semitic languages as well; in most European languages participles and infinitives have nominal features. However, two Semitic traits make this phenomenon more challenging in our case – the rich morphology which creates a large set of inflections for each base form (i.e. the verb is inflected to create nominal forms and then each form is inflected again for case, gender and number). Furthermore, Semitic languages allow nominal clauses, namely verbless sentences, which increase ambiguity. For example, in English it is easy to recognize the form 'drunk' in 'he has drunk' as related to the lemma DRINK (V) (and not as an adjective). This is done by spotting the auxiliary 'has' which precedes this form. However in MH, the clause *axi šomer* could mean 'my brother is a guard' or 'my brother guards/is guarding'. The syntactical cues for the final decision are subtle and elusive. Similarly in MSA: *axi ka:tibun* could mean 'my brother is writing' or 'my brother is a writer'.

1.4 Orthography

From the viewpoint of NLP, especially commercially applicable NLP, it is important to note that the writing systems of both MSA and MH follow the same conventions, in which most vowels are not marked. Therefore, in MSA the form *yaktubu* 'he writes/will write' is written {yktb}. Similarly in MH, the form *yilmad* 'he will learn' is written {ylmd}. Both languages have a supplementary marking system for vocalization (written above, under and beside the text), but it is not used in the overwhelming majority of texts. In both languages, when vowels do appear as letters, letters of consonantal origin are used, consequently turning these letters ambiguous (between their consonantal and vocalic readings).

It is easy to see the additional difficulty that this writing convention presents for NLP. The string {yktb} in MSA can be interpreted as *yaktubu* (future tense), *yaktuba* (subjunctive), *yaktub* (jussive), *yuktabu* (future tense passive) and even *yuktibu* 'he dictates/will dictate' a form that is considered by Morfix to be a different lemma altogether (see above 1.2). Furthermore, ambiguity can occur between totally unrelated words, as will be shown in section 1.7. A trained MSA reader can distinguish between these forms by using contextual cues (both syntactic and semantic). A similar contextual sensitivity must be programmed into the NLP system in order to meet this challenge.

Each language also has some orthographic peculiarities of its own. The most striking in MH is the multiple spelling conventions that are used simultaneously. The classical convention has been replaced in most texts with some kind of spelling system that partially indicates vowels, and thus reduces ambiguities. An NLP system has to take into account the various spelling systems and the fact that the classic convention is still occasionally used. Thus, each word often has more than one spelling. For example: the word *shi?ur* 'a lesson' can be written {š¿wr} or {šy¿wr}. The word *kiven* 'to direct' can be written {kwn} or {kywwn}, the former is the classical spelling (*Ktiv Xaser*) while the later is the standard semi-vocalized system (*Ktiv Male*), but a some non-standard spellings can also appear: {kywn}, {kwwn}.

MSA spelling is much more standardized and follows classic conventions. Nonetheless, some of these conventions may seem confusing at first sight. The *Hamza* sign, which represents the glottal stop phoneme, can be written in 5 different ways, depending on its phonological environment. Therefore, any change in vowels (very regular a phenomenon in MSA inflectional paradigms) results in a different shape of *Hamza*. This occurs even when the vowels themselves are not marked. Moreover – there is often more than one shape possible per form, without any mandatory convention. One could argue that all *Hamza* shapes should be encoded as one for our purposes. This may solve some problems, but then again it would deny us of crucial information about the vowels in the word. Since the *Hamza* changes according to vowels around it, it is a good cue for retrieving the vocalization of the word, and to reduce ambiguity.

1.5 Clitics and Complex Forms

The phenomenon which will be described in this section is related both to the morphological structure of MSA and MH, and to the orthographical conventions shared by these languages. Both languages use a diverse system of clitics[4] that are appended to the inflectional forms, creating complex forms and further complications in proper lemmatization and data retrieval.

For example, in MSA, the form: *?awla:dun* 'boys (nom.)', a part of the lemma *WALAD* 'boy', can take the genitive pronominal enclitic /-ha/ 'her' and create the complex form: *?awla:d-u-ha* 'boys-nom.-her (=her boys)'. This complex form is orthographically represented as follows: {?wladha}. Similarly in Hebrew, the form *yeladim* 'children' (of the lemma *YELED* 'child'), combined with the genitive pronominal enclitic /-ha/ 'her', yields the complex form *yelade-ha* 'children-her (=her children)'. The orthographical representation is: {yldyh}.

Enclitics usually denote genitive pronouns for nouns (as demonstrated above) and accusative pronouns for verbs. For example, in MSA, *?akaltu-**hu*** 'I ate **it**' {?klth}, or in MH *axalti-**v*** 'I ate **it**' {?kltyw}. It is easy to see how this phenomenon, especially the orthographic convention which conjoins these enclitics to the basic form, may create confusion in lemmatizing and data retrieval. However, the nature of clitics which limits their position and possible combinations helps to locate them and trace the basic form from which the complex one was created.

There are also several proclitics denoting prepositions and other particles, attached to the preceding form by orthographic convention. The most common are the conjunctions /w, f/, the prepositions /b, l, k/ and the definite article /al/ in MSA, and the conjunction /w/, the prepositions /b, k, l, m/ (often referred to as *Otiyot Baxlam*), the relative pronoun /š/ and the definite article /h/ in MH. Therefore, in MSA, the phrase: *wa-li-l-?wla:di* 'and to the boys' will have the following orthographical representation: {wll?wlad}. In MH the phrase *ve-la-yeladim* 'and to the children' will be represented orthographically as: {wlyldym}. Once again, when scanning a written text, these proclitics must be taken into account in the lemmatization process.

[4] The term "clitics" is employed here as the closest term which can describe this phenomenon without committing to any linguistic theory.

1.6 Syntax

The syntactic structure of MSA and MH is very similar. In fact, the list of major syntactic rules is almost identical, though the actual application of these rules may differ between the languages.

A good demonstration of that is the **agreement rule**. Both languages demand a strict noun-adjective-verb agreement. The agreement includes features such as number, gender, definiteness and in MSA also case marking (in noun-adjective agreement). The MH agreement rule is more straightforward than the MSA one. For example: ***ha**-yelad**im** **ha**-gdol**im** halx**u*** 'the-child-pl. the-big-pl. go-past-pl. (=The big children went). Note that all elements in the sentence are marked as plural, and the noun and the adjective also agree in definiteness.

The case of MSA is slightly different. MSA has incomplete agreement in verb-subject sentences, which are the vast majority. In this case the agreement of the verb will only be in gender but not in number, e.g. *ðahaba l-?awla:du* 'go-past-masc.-sing. boy-pl. (=The boys went)'. MSA also distinguishes between human plural forms and non-human plural forms, i.e. if the plural form does not have a human referent, the verb or the adjective will be marked as feminine rather than plural, e.g. *ðahab**at** el-kila:bu l-kabi:r**at**u* 'go-past-fem.-sing. the-dog-masc.-pl. the-big-fem.-sing. (=The big dogs went)'.

The example of the agreement rule demonstrates both the similarities and the differences between MSA and MH. Furthermore, it demonstrates how minor are the differences as far as our purposes go. As long as the agreement rule is taken into account, its actual implementation has hardly any consequences in the level of the system. This example also demonstrates a very useful cue to reduce ambiguity among forms. This cue is probablyused intuitively by trained readers of MSA and MH, and encoding it into the Morfix NLP system turns out quite useful.

1.7 Ambiguity

Perhaps the major challenge for NLP analysis in MSA and MH is overcoming the ambiguity of

forms. In this respect, Morfix has to imitate the rather sophisticated reading of a trained MSA or MH speaker, who continuously disambiguates word tokens while reading.

The reason for ambiguity can be depicted in three main factors:

i. The large amount of morphological forms, which are sometimes homographic.

   For example, both in MSA and MH the verbial inflection of the imperfect for the singular is the same for 2nd person masculine and 3rd person feminine: MSA – *taktubu*, MH – *tixtov*.

ii. The possibility of creating complex forms by conjoining clitics, which raises the possibility of coincidental identity.

   For example, in MSA: *ka-ma:l* 'as money', *kama:l* 'perfection, Kamal (proper name)' → **{kmal}.** Similarly in MH: *ha-naxa* 'the-resting-fem.', *hanaxa* 'an assumption, a discount' → **{hnẖh}**.

iii. The orthographical conventions, such as the lack of vowel marking and various spelling alternatives.

   For example, in MSA: *muda:fi¿* 'defender', *mada:fi¿* 'cannons' → **{mdaf¿}**, and in MH *baneha* 'her sons' *bniya* 'building' → **{bnyh}**.

In many cases ambiguity is the result of the combination of two factors or even all three. This makes ambiguity rate rather high, and its resolution such a major component of NLP mechanism.

Disambiguation is based on syntactical structures and semantic cues that can be retrieved from the text, which might resemble the way a human reader copes with these problems. It is the objective of NLP systems dealing with MSA an MH to formalize these cues.

2 A Description of the Morfix Architecture and its Application

2.1 Architecture

On one hand, as can be expected in the light of the similarities described above, a single NLP system is applicable for both MSA and MH, including code infrastructure, database structures, and methodology. On the other hand, in adapting a previously existing MH system to MSA some minor adaptations are nonetheless needed.

Morfix is comprised of two lexical databases: a lemma database and an idiom/collocation information database, and two rule databases: a morphological rule database and a syntactical rule database.

The lemma database contains all crucial information about each lemma, including lexical features such as part of speech, gender, number, meaning, root, verb pattern (*Wazn* / *Binyan)* etc. Most of these features are common to MH and MSA, and have the same morphological implications. All inflectional forms of a lemma are generated by applying algorithms that process these features. These algorithms make use of the morphological rule database. These rules generate forms by superimposing verb patterns and morphophonemic principles. Exceptions are allowed, i.e. the lexicographer may edit a specific form. The exception mechanism is much less used in MSA than in MH, due to the higher consistency of MSA inflections (but see below 2.2 for the treatment of the MSA "Broken Plural" in Morfix). By the conclusion of this inflection procedure, the entire 70 million forms inventory is accessible.

The information for the lemma and collocation databases is gathered by two techniques. In the first phase words are extracted from several dictionaries[5], while the second phase involves analyzing text corpora, mainly through Internet sources, using the dictionary based lexicon. Any unanalyzed word (usually new loan words, neologisms and new conventions of usage), as well as collocations found in the corpora, are the basis for enriching the lexicon. The information for the morphological and syntactical databases is retrieved both from conventional grammar textbooks[6] and from additional linguistic analysis of the corpora.

By contrast, derivational morphology is by and large not algorithmic or rule derived. That is, nouns, adjectives and verbs of different patterns that share the same root are each entered as separate lemmas. As mentioned above (1.2), there is a fine line between inflectional morphology and derivational morphological. For example, the decision whether to create a new lemma for a nominal

[5] For MSA: Wehr (1973), Al-Munjid (1992), Ayalon and Shinar (1947) and others; for MH: Even Shoshan (1991), Alcalay (1990) and others.

[6] For MSA: Wright (1896), Holes (1995); for MH: Glinert (1989).

inflection of verb is left to the lexicographer. Criteria are usually morphological, since semantic criteria are often too vague. For example, the fact that the form *ka:tib* has two possible plural form: *ka:tibuna* 'writing masc. pl.' and *kutta:b* 'writers' indicates that the form should have a lemma of its own, on top of being associated with the verb lemma.

While the lemma in Morfix is defined as an **inflectional** lemma, **derivational** morphology is also accounted for in the database in a mechanism called **word families,** namely the root-based lemma grouping described above (1.2), whose members also share a semantic field. For example, *infija:r* 'explosion' and *mufajjira:t* 'explosives' would be members of the same family, whereas *fajr* 'dawn' would not belong to this family.

The idiom/collocation database stores information about co-occurrence of words. Idioms are lexicalized word combinations (e.g. in MSA *bunya taħtia* 'infrastructure', or in MH *bet sefer* 'a school'), while collocations are combinations of words that do not have specific meanings when combined, yet often appear together in texts (e.g. in MSA *waqqa¿a* {wq¿} *¿ala l-ittifa:q* 'to sign the agreement' as opposed to *waqa¿a* {wq¿} *fi tta:ri:*x 'occured on the date' or in MH *hamtana* {hmtnh} *ba-tor* 'to wait on line' as opposed to *kabalat hamatana* {hmtnh}'accepting the gift').

Finally, the syntactical rule database is comprised of rules such as agreement rules and construct formation rules (*Iđa:fa / Smixut*). Some rules are not absolute, but rather reflect statistical information about distribution of syntactical structures in the language. These rules play a major role in the context analysis module.

Each morphological analysis has a vocalization pattern (*Taški:l / Nikud*). When analyzing word tokens in context, Morfix produces a best bet for the vocalized text.

Finally, for text-to-speech purposes, a string of phonemes is created, based on the vocalization patterns. Stress markings are added per word, and a prosody pattern is applied, based on syntactical analysis at the clause level. Prosody patterns are expressed as duration and pitch values per phoneme.

2.2 Adaptation of the technology to Arabic

Most of the elements of Morfix are common to MSA and MH. However, some features had to be specifically supplemented for MSA database. For example, MH plural markers are few and are usually suffixes. MSA on the other hand, often uses "**Broken Plural**" (a plural formed by changing the vocalic pattern of the singular, as opposed to affixation marking, e.g. *ka:tib* (sing.) → *k:atibu:na* (pl.) 'writing'; *ka:tib* (sing.) → *kutta:b* (pl.) 'writer'), which is only partially predictable, and therefore must be included in the lemma records. Coding this feature did not require major change in the database, since the MH database had optional coding for exceptional plural forms.

By contrast, a field in MH lemma records redundant in MSA is **stress location**, which, as apposed to MH, is always predictable in MSA given the phonemic structure of the form.

Case inflection in MSA (*?i¿ra:b*) is entirely predictable, hence depicted by rules in the morphological rule database. However, a field for case had to be created in the database especially for MSA, as case does not occur in MH.

Dual inflection exists in MH, though usually unproductive. This means that the "number" category throughout the Morfix database could have one of three values: singular, dual, or plural, so that MSA handling, again, demanded no general change, but only a more widespread application of an existing option in the Hebrew Morfix.

The number of inflectional forms of a verb entry is larger in MSA than it is in MH, most notably due to the additional mood paradigms (*Al-Muđa:re¿ Al-Majzu:m and Al-Muđa:re¿ Al-Manśu:b*). This, however, is of no major consequence to Morfix, apart from the fact that another field had to be added to the morphological analysis structure, namely "mood".

The higher number of inflections per verb, along with the generality of the dual inflection, would have resulted in a larger overall number of tokens in MSA, had it not been for the *Ktiv Male* orthographical system in MH that results in a 25% increment to the overall number of MH tokens (see also above 1.4).

The phenomenon of incomplete agreement (see also above 1.6) does not require an actual change in the code of Morfix, since the term AGREEMENT (e.g. between noun and adjective)

has an external definition, independent for each language. Syntactical rules in the system refer to the term AGREEMENT, hence, rules that make use of the term AGREEMENT will apply, in many cases, to both languages. In general, while some of the syntactical rules in the system are similar in both languages, other rules are defined specifically for each of the two languages. All rules for both languages are specified using the same mechanism.

In the MH database there are supplementary placeholders for the semi-vocalized spelling alternatives, which are often redundant for MSA, though they do become useful especially in recent loan words.

In MSA the verb predicate usually precedes its subject (VSO), while in MH the subject tends to appear first (SVO), though in both languages word order is not fixed. This difference is handled in the contextual analysis for disambiguation purposes.

MSA is used in various countries, each having its own linguistic idiosyncrasies. This entails lexical differences and a few phonetic variations, as well as some minor writing convention differences. This is handled by the MSA lemma database by assigning an additional field, where the relevant areas are specified.

2.3 Software modules

- Morphological analyzer:

This is the basic building block of our system. It analyzes an input string, and returns an array of records containing detailed information regarding each analysis: the lemma, part of speech, clitic details, as well as gender, number; person, tense, mood, case, clitics and the like.

- Lemmatizer

This is a version of the morphological analyzer, the difference being that its output is lemmas, not full morphological descriptions. This means that when several morphological analyses share a single lemma, these analyses are united into a single answer record, each includes just the lemma and its part of speech.

For example, the string {waldy} has several morphological analyses (dual construct form: 'the two parents of', dual form with genitive pronominal enclitic: 'my two parents', or singular form with genitive pronominal enclitic: 'my father'); however, the lemmatizer produces just one lemma for all the above analyses: ***wa:lid* 'a parent'**.

- Context analyzer

The input for the context analyzer is a text buffer. It returns a set of morphological analysis record arrays, an array for each token found in the buffer. In the records there is one extra field as compared to the basic morphological analyzer: the score field, which reflects the effect of the context analysis. The answer arrays are sorted according to the declining order of the score.

- Vocalizer

Given a word and a morphological analysis record as input, this module outputs the input word with its vocalization.

- Text to phoneme

Given a vocalized word, and a morphological analysis record as input, this module produces its phonemic representation, including stress marking.

- Text to speech

A module on top of the text-to-phoneme module, whose inputs are a text buffer and a morphological analysis per word. The text to phoneme module is called upon to produce the phonemic representation of the buffer. Then a prosody function is called upon to assign duration values and pitch contours to each phoneme, and the output of this function is sent to a diphone based synthesis engine.

2.4 Results and performance

The Hebrew version of Morfix has achieved the following results:

Morfix generates exceptionally accurate lemmatization. When indexing for full text search, the matching rate of the lemma receiving the highest score to the correct lemma stands at above 98%. In typical Internet texts, between 1% and 2% of words remain unanalyzed (by and large, these are proper names not included in lexicon; in search engine application, these undergo a morphological soundex algorithm designed to enable the retrieval of proper names with prepositional proclitics).

Performance depends on hardware and system environments. On a typical (as of date of publication) Intel III 800 MHz CPU, with 256 MB RAM

running Windows 2000, Morfix analyzes c. 10,000 words per second.

In text-to-speech (TTS) applications, the degree of words read correctly (fully correct phonetic transcription and stress location) is also 98%. This number is no different than the number for lemmatization, but is derived differently: on one hand, sometimes an error in lemmatization does not yield an error in phonetization (in case of homonymic tokens); on the other hand, TTS has to deal with phonetization of proper names not in the lexicon, which it carries out according to algorithms. The Hebrew TTS system is successfully implemented in systems for reading e-mail and Internet texts. Performance of the TTS system is around 20% slower than lemmatization, due to extra processing that computes the phonetic transcription given the morphological analysis.

The final equivalent numbers for Arabic are still not available as of date of publication. Nonetheless, because the system is similar, and MSA is quite close to MH in terms of total number of inflections and in degree of ambiguity, it is expected to reach similar results.

2.5 Applications

Various modules of the system are used by various applications. Main application beneficiaries include **full text search, categorization and textual data mining** (where context sensitive morphological analysis and lemmatization are crucial for Semitic languages), **screen readers and e-mail-to-voice converters** in telephony usage (especially the text-to-speech module), **automatic vocalizers** for schools and book publishers (especially the vocalization module), and **online dictionaries** (especially context sensitive lemmatization, to enable the retrieval of the correct entry when clicking on a word in context).

A special thought was given in order to assist the non-fluent speaker of MSA and MH. Besides the fact that all applications trace the basic forms of words, sparing the process usually done by the speaker himself, additional assistance is given, such as transliteration into Latin script.

3 Conclusion

When designing the adaptation of the MH system to MSA, the similarity between the languages on the structural level became even more apparent than was anticipated.

It became clear to us that unified studies and applications for both languages can benefit both commercial and theoretical academic fields, and we hope that this report can be a starting point for further incorporating NLP works in MH and MSA, namely, works that deal with the Semitic language phenomena, rather than with a specific language from this linguistic family. This work can be implemented in other NLP systems, mainly of other Semitic languages such as colloquial dialects of Arabic (e.g. Egyptian or Syrian which are more and more used in writing) and Maltese, but also languages that share some of the Semitic traits, mainly rich and complex morphology, or that use alternative writing conventions. This work lays the infrastructure for further adaptation, though creating special databases for each languages remains to be done.

References

Alcalay R. (1990). The Complete Hebrew-English Dictionary, Massada and Yediot Aharonot Pub., Tel-Aviv.

Al-Munjid fi l-Lugha wa-l-I'lam (1992), Dar El-Mashriq Pub., Beirut.

Ayalon D. and P. Shinar (1947). *Arabic-Hebrew Dictionary for the Modern Arabic Language*, Hebrew University, Jerusalem.

Beesley, K.R. (2001). "Finite-State Morphological Analysis and Generation of Arabic at XeroxResearch: Status and Plans in 2001", in *Arabic Language Processing: Status and Prospects - 39th Annual Meeting of the Association for Computational Linguistics*, pp. 1-8.

Dichy, J. (2001), "On Lemmatization of the Arabic Entries of Multilingual Lexical Databases", in *Arabic Language Processing: Status and Prospects - 39th Annual Meeting of the Association for Computational Linguistics*, pp. 23-30.

Even-Shoshan, A. (1992). *Ha-Milon He-Ħadash*, Kiryat Sefer, Jerusalem.

Glinert, L. (1989). *The Grammar of Modern Hebrew*, Cambridge University Press, Cambridge.

Holes, C. (1995). *Modern Arabic: Structures, Functions and Varieties*, Longman Linguistics Library, London.

Wehr, H. (1976). *A Dictionary of Modern Written Arabic*, Ithaca, NY.

Wright, W. (1896). *A Grammar of the Arabic Language*, Cambridge: University Press.

ARLEX, for Modern Standard Arabic (MSA) that explicitly lists ambiguity at the lexical and syntactic levels for each token. Arabic orthography is known for being underspecified for short vowels and other markers such as letter doubling and glottal stops, known as diacritics. This leads to further ambiguity in orthography with real impact on natural language processing (NLP) applications, not to mention readability and human language processing. Modern Standard Arabic (MSA), or Modern Written Arabic (shortened to MWA), is a term used mostly by Western linguists to refer to the variety of standardized, literary Arabic that developed in the Arab world in the late 19th and early 20th centuries. It is the language used in academia, print and mass media, law and legislation, though it is generally not spoken as a mother tongue, similar to Classical Latin or the literary register of French. MSA is a pluricentric standard language taught throughout Modern Standard Arabic and the spoken dialects are so vastly different in terms of grammar, vocabulary and pronunciation that a person who is totally 'fluent' in MSA may not have any idea what a person's saying in a local dialect. I've witnessed this personally many times here in the Middle East with advanced students of MSA who can't hold a simple conversation with an average Joe on the street. Sure, you'll be understood by many people (though not all!) when you speak but don't expect to understand the reply. What that means is that someone who spends all th