International Journal of Lexicography, Vol. 27 No. 2, pp. 130–170
doi:10.1093/ijl/ect036 Advance access publication 20 December 2013

LEXICAL REPRESENTATION OF MULTIWORD EXPRESSIONS IN MORPHOLOGICALLY-COMPLEX LANGUAGES

Hassan Al-Haj: Dept. of Computer Science, University of Haifa
Alon Itai: Dept. of Computer Science, Technion, IIT
Shuly Wintner: Dept. of Computer Science, University of Haifa

Abstract

In spite of the surging interest in multiword expressions (MWEs) in recent years, it is still unclear how such expressions should be stored in computational lexicons. This problem is amplified in morphologically-complex languages, where the unique properties of MWEs interact with non-trivial morphological processes. We propose an architecture for lexical representation of MWEs, augmented by a protocol for integrating MWEs into a morphological processing system. The proposal is applied to Modern Hebrew, a Semitic language with complex morphology and a problematic orthography. The result is an integrated system that can morphologically process Hebrew multiword expressions of various types. In light of the complexity of Hebrew morphology and orthography, we are confident that the proposed architecture is general enough so as to accommodate MWEs in a large number of languages.

1. Introduction

Multiword Expressions (MWEs) blur the traditionally-assumed dichotomy between the lexicon and the grammar: they are sequences of orthographic words that for various reasons must be stored in computational lexicons as a unit. The term 'MWE' refers to a heterogeneous class of phenomena with diverse sets of characteristics. Semantically, they tend to be less compositional than other phrases, but their compositionality is gradual (Bannard et al., 2003). Syntactically, MWEs may function as words or as phrases. Some MWEs exhibit a rigid pattern, in which the constituents must occur in a fixed order, while others can undergo various syntactic transformations. The components of MWEs can occur in the text either contiguously (with intervening spaces) or dispersed. Morphologically, MWEs are not homogeneous either, allowing

some constituents to freely inflect while restricting (or even preventing) the inflections of others. In some cases MWEs may allow their constituents to undergo non-standard morphological inflections that they would not undergo in isolation. MWEs may also consist of words that have no literal meaning outside the expression. A main characteristic property of MWEs is idiosyncrasy: they typically exhibit an irregular behavior, be it morphological, syntactic or semantic.

MWEs are extremely prevalent. They constitute a significant portion of the tokens in a running text, as well as the entries in the lexicon, with estimates ranging between 40% and 60%, depending on one's definition of MWEs (Jackendoff, 1997; Erman and Warren, 2000; Sag et al., 2002). However, while MWEs constitute significant portions of natural language texts, most of them belong to the long tail in terms of frequency: specific MWEs tend to occur only rarely in texts, and automatic identification of MWEs is therefore a challenge (Baldwin and Villavicencio, 2002; Graliński et al., 2010).

The prevalence of MWEs, as well as their idiosyncratic properties, make them a challenge for computational processing of natural languages (Sag et al., 2002). They are even more challenging in languages with complex morphology, because of the unique interaction of morphological and orthographical processes with the lexical specification of MWEs (Oflazer et al., 2004; Alegria et al., 2004; Savary, 2008).

We propose an architecture for lexical specification of MWEs in morphologically-complex languages, focusing on (Modern) Hebrew, a language with rich, complex morphology and problematic orthography. Motivated by a careful survey of the properties of a wide spectrum of Hebrew MWEs, we propose a solution for storing MWEs in an existing large-scale lexicon (Itai et al., 2006), as well as a protocol for integrating MWEs in an existing morphological processing system (Itai and Wintner, 2008). Morphological analyzers of morphologically rich languages are complex systems. They contain vast linguistic knowledge that is applicable to single words and to MWEs alike; redesigning such a system in order to account for MWEs would be labor intensive. The challenge is to integrate MWEs in an existing morphological analyzer with minimal changes to the analyzer. Our lexical representation focuses on the morphological and morpho-syntactic properties of MWEs, rather than on their semantics, and we do not provide a scheme for encoding the semantics of MWEs here (mainly because the existing lexicon that we extend with MWEs does not encode meanings).

The contribution of this work is manifold. The main practical outcome is an integrated system that supports morphological processing of Hebrew MWEs, thereby extending the state of the art in morphological processing of the language to account also for MWEs. More generally, while the proposed architecture is motivated by and exemplified on Hebrew, we trust that it is general enough so as to be usable for a number of other languages. Finally, this is the

first work to investigate the variety and diversity of Hebrew MWEs within a computational setup.

We begin in Section 2 with a discussion of related work. Section 3 provides an overview of Hebrew orthography, morphology, and syntax, with a focussed discussion of the properties of Hebrew MWEs in Section 4. We review the existing morphological processing system in which we implement our proposed architecture in Section 5. Section 6 then specifies the lexical representation of MWEs, including a brief discussion of the lexical acquisition of MWEs in Section 7. Section 8 details the integration of MWEs in the morphological processor and provides some implementation details and results. We conclude with suggestions for future research.

2. Related work

There has been a growing awareness in the natural language processing community of the problems that MWEs pose, both in linguistics and in computational applications (Villavicencio et al., 2005; Grégoire et al., 2007, 2008; Anastasiou et al., 2009; Laporte et al., 2010; Baldwin and Kim, 2010; Rayson et al., 2010; Kordoni et al., 2011, 2013). Recent works address the characterization and analysis of MWEs, their lexical representation, morphological and syntactic processing, and identification and extraction from data. We focus on lexical representation and morphological processing in the following survey.

The definition of MWEs is still a matter of much debate. Erman and Warren (2000) refer to prefabricated text, which consists of "at least two words favored by native speakers in preference to an alternative combination which could have been equivalent had there been no conventionalization." Sag et al. (2002) divide MWEs into two classes: lexicalized phrases, which have at least partially idiosyncratic syntax or semantics, or contain 'words' which do not occur in isolation; and institutionalized phrases, which are syntactically and semantically compositional, but occur with markedly high frequency. In this work we use a practical definition of MWEs: since our main goal is to provide a solution that would facilitate downstream computational processing, we consider as MWE any expression that exhibits some facet of idiosyncrasy, be it orthographic, morphological, syntactic, semantic, or pragmatic. Such expressions must obviously be stored in computational lexicons, if downstream processing is to treat their idiosyncrasy properly.

Various strategies for encoding MWEs in lexical resources have been employed for different languages (see Savary (2008) for an excellent survey). Some of these works focus on encoding a certain class of MWEs. For example, Baptista et al. (2004) construct an electronic dictionary of European Portuguese frozen sentences, defined as elementary sentences in which the components can inflect freely, but the main verb and at least one of its argument

noun-phrases are distributionally constrained. This work classifies the frozen expressions into formal classes, and encodes the lexical and syntactic properties of each frozen expression. Then, a finite-state transducer is built for each class, and is used to identify and tag the frozen sentences in texts. Another example is the Alvey Tools Lexicon (Carroll and Grover, 1989), which has a good coverage of (English) phrasal verbs, providing extensive information about their syntactic aspects, but which does not distinguish compositional from non-compositional entries or specify entries that can be productively formed.

Other works adopt a more general approach by proposing an architecture for lexical encoding of MWEs which allows for a unified treatment of different kinds of expressions. Villavicencio et al. (2004) present an encoding of MWEs that uniformly captures different types of expressions (e.g., nominal compounds, verb-particle constructions and idioms). They encode the properties of MWEs using a hierarchy of tables built one on top of the other. In the lowest level of this hierarchy lies a table that contains simplex entries for single words. Each of these entries encodes the orthographic, morphological, grammatical and semantic properties of a single word. Higher tables have links to lower tables through which they inherit their properties, but the tables also provide additional syntactic and lexical information such as the position of the component in the expression, whether the component is optional or not, and in case a component can be realized in different ways, all possible realizations are encoded. The tables also provide semantic information about the expression, such as the meaning of the component in the frame of the MWE (which may be different from its meaning in isolation). This work, however, does not provide an adequate solution for the vast array of morphological irregularities that MWEs tend to exhibit in morphologically-rich languages (Section 4). In contrast, we define a general representation scheme for MWEs which can account for any combination of morphological and syntactic variation exhibited by MWEs. Note, however, that unlike Villavicencio et al. (2004) we do not account for the semantics of the MWEs.

Morphological issues are considered by Oflazer et al. (2004), who describe a system for morphological processing of MWEs in Turkish, an agglutinative language with a very productive morphology. The MWE processor is composed of a number of stages, where each stage produces a morphological analysis for a certain class of MWEs, and its output is fed to the following stage. Our scheme is similar in spirit, but our solution is sensitive to the special needs exhibited by Hebrew (and related languages), in particular the complex interaction of the deficient orthography with the rich morphology.

Grégoire (2007) describes a lexicon of Dutch MWEs, defined as combinations of words that exhibit "linguistic properties not predictable from the individual components or the normal way they are combined." Expressions are lexically stored according to their equivalence class, which clusters together expressions that behave in the same way. The main classification follows

Sag et al. (2002), where each expression is categorized as either fixed, semi-flexible or flexible. Then, expressions are classified according to their pattern, which defines the syntactic properties of the expression. This work facilitates a concise lexical specification of various types of MWEs, including flexible ones, in which the order of the constituents is not fixed; but it does not account for morphological and orthographic idiosyncrasies.

In a subsequent work, Grégoire (2010) describes a lexicon of Dutch MWEs (of various types) that does account for some morphological variation. The design of the lexicon is based on the equivalence class method, whereby MWEs that share similar syntactic behavior are grouped together. This approach apparently solves many of the idiosyncrasies exhibited by Dutch MWEs, but as we show in Section 3, Hebrew poses some unique properties (e.g., the interaction between the deficient orthography and the complex morphology) that require dedicated solutions. Furthermore, the equivalence class method requires the stipulation of at least one instance of each class (granted, the effort involved is reduced due to the introduction of parameterized classes). In the present work, we describe general properties of Hebrew MWEs, focusing on the interaction of the lexicon, the orthography and the morphology. These properties drive a solution that we believe is more general, applicable to the entire inventory of MWEs in the language.

Graliński et al. (2010) compare two formalisms, Multiflex (Savary, 2009) and POLENG (Krzysztof, 2004), with respect to how they facilitate the lexical representation of nominal and adjectival MWEs in Polish. Multiflex is a graph-based generator of MWEs that can generate all the inflected forms of an expression's constituents. It can only handle MWEs whose components are contiguous, and cannot enforce constraints that hold between some component of the MWE and an external token. POLENG is a simpler formalism with a similar functionality, but it does not support an explicit indication of morphological idiosyncrasies, such as partial agreement. Nor does it allow one to describe variability in word order, except full freedom. Graliński et al. (2010) conclude that neither formalism satisfies all the needs of lexicographers incorporating MWEs into the lexicon of a morphologically-rich language.

The interactions between MWEs and morphological idiosyncrasies are discussed in detail by Savary (2008), with examples from several languages (English, French, Polish, Serbian, German, and Turkish). Savary (2008) compares as many as eleven lexical approaches to MWEs, and concludes with desiderata for lexical databases that attempt to represent MWEs in morphologically-complex languages. Our proposal here satisfies many (but not all) of these desiderata (see Section 9).

Few works address MWEs in Hebrew (or other Semitic languages) from a computational perspective. Working on Arabic, Attia (2006) proposes methods to process fixed, semi-fixed, and syntactically-flexible MWEs (adopting the classification of Sag et al. (2002)). The fixed and semi-fixed expressions are

processed by building a finite state transducer for each MWE, which is then composed with the tokenizer. The resulting transducer then complements an existing (single word) morphological transducer. Syntactically-flexible expressions are processed by the syntactic parser through the use of lexical rules. Attia (2006) does not address lexical representation or morphological idiosyncrasy.

More recently, Al-Haj and Wintner (2010) describe a method for distinguishing between MWE and non-MWE noun-noun constructions in Hebrew, building on the morphological idiosyncrasies of MWEs. Tsvetkov and Wintner (2010, 2011, 2012) extend these results to virtually any type of linguistic construction, using a multitude of linguistic cues that help distinguish Hebrew MWEs from compositional expressions. These works, again, focus on MWE identification and extraction, and do not address their representation or morphological processing.

3. Linguistic background

3.1 Orthography

The orthography of Hebrew[1] poses several problems for computational processing (Wintner, 2004). As is well known, in the standard script most vowels are not explicit. Furthermore, many particles, including the definite article h "the", four of the most frequent prepositions (b "in", k "as", l "to" and m "from"), the coordinating conjunction w "and" and some subordinating conjunctions (such as š "that" and kš "when"), all attach to the word which immediately follows them. Viewing a text as a space- and punctuation-delimited stream of tokens, then, results in tokens that may require further analysis, depending on the task at hand.

This orthographic quirk is challenging for adequate processing of MWEs, because the rules that govern the combination of Hebrew prefix particles with the words they attach to are basically syntactic. For example, the preposition m "from" can combine with nouns but not with adverbs. The same rules govern the combination of Hebrew prefix particles with MWEs, but these combinations are constrained by the syntactic category of the whole expression, rather than its first word. For example,[2] m "from" can combine with the noun ph "mouth" when it occurs in isolation, but not with the expression ph axd ("mouth one") "unanimously"), which functions as an adverb. Thus # mph axd "from one mouth" can only be interpreted literally.

Henceforth, we use the term lexical entry (or lexical item) to refer to lexical words, i.e., items stored in the lexicon. As we are relying on an existing lexicon (Section 5.1), these are predefined, and include free morphemes but also bound morphemes, in particular the prefixes mentioned above. When MWEs are concerned, we use the term constituent to refer to lexical items that make up

MWEs. For example, the MWE blb slm ("in + heart whole") "wholeheartedly") consists of three constituents, the prefix b "in" and the free morphemes lb "heart" and slm "whole".

3.2 Morphology

Hebrew, like other Semitic languages, has a rich and complex morphology (Berman, 1978; Glinert, 1989). The major word formation machinery is root and pattern (Shimron, 2003) and is highly productive.

Nominals, namely nouns, adjectives and numerals, inflect for number (singular, plural and, in rare cases, also dual) and gender (masculine or feminine). [3] In addition, nominals have three phonologically (and orthographically) distinct forms, traditionally known as states: the absolute (citation) state; the definite state, which is indicated by the prefix h "the"; and the construct state, which is typically used in genitive (possessive) constructions. For example, xwlch "shirt" (absolute) vs. hxwlch "the + shirt" (definite) vs. xwlct "shirt-of" (construct). Furthermore, nominals (in the construct state) take pronominal suffixes, sometimes referred to as clitics, which are interpreted as possessives. These suffixes inflect for number, gender and person (e.g., xwlct + h ! xwlcth "her shirt", xwlct + nw ! xwlctnw "our shirt", etc.) As expected, these processes involve certain morphological alternations.

Verbs inflect for number, gender and person (first, second and third) and also for a combination of tense/aspect and mood, which is traditionally analyzed as having the values past, present (participle), future, imperative and infinite. For example, the verb akl "eat" has the inflected forms akl + nw ! aklnw "we ate" (first person plural past), t + akl + w ! taklw "you will eat" (second person plural masculine future), and many more.

Prepositions can combine with pronominal suffixes that are interpreted as the objects of the preposition. These inflect for number, gender, and person. For example, lid "near" yields lid + w ! lidw "near him", lid + h ! lidh "near her", lid + nw ! lidnw "near us", and many more.

3.3 Syntax

The standard constituent order of Hebrew is Subject–Verb–Object, although many other orders are possible, and some are highly frequent (Melnik, 2006). Within the noun phrase, constituents tend to occur in a fixed order (roughly, quantifier, noun, adjective, possessive, relative clause). Various elements of a noun phrase may be marked as definite (e.g., by being in the definite state); all elements of the noun phrase must agree with respect to definiteness. Another interesting syntactic phenomenon is the abundance of non-verbal predicates, with or without an explicit copula (Doron, 1983).

Hebrew has three different possessive constructions: 1. the head noun is in the construct state, indicating a genitive relation with the following noun (azrxi hmdinh "citizens-of the+ state"); 2. the two nouns are in the absolute or the definite state, and the preposition šl "of" relates them (hazrxim šl hmdinh "the+ citizens of the+ state"); and 3. the head noun is in the construct state, with a cliticized pronoun, followed by the possessive preposition and the possessor (azrxih šl hmdinh "her-citizens of the+ state"). In this case, the cliticized pronoun refers to the possessor (here, mdinh "state", which is feminine) and must agree with it in gender and number.

4. Properties of MWEs

To motivate our proposed architecture for MWE representation (Section 6), we describe in this section some of the properties of Hebrew MWEs, focusing on their idiosyncratic behavior. While some of these properties are unique to Hebrew (or to a small group of Semitic languages), many, especially the semantic properties, are observed in a large number of languages (Savary, 2008, Section 2); we begin with the latter, providing examples from Hebrew but also from English when appropriate. We then move on to describe syntactic, morphological and orthographic properties of MWEs, which are obviously more language-specific (and are demonstrated predominantly on Hebrew). As this discussion motivates and drives the lexical representation we advocate (Section 6), we provide for each property a forward reference to the section in which its effect on the representation is discussed.

4.1 Semantic properties

The most distinguishing property of MWE is their non-compositionality: their meaning cannot always be deduced from that of the constituents. Several consequences follow from this observation.

Lexical fixedness

Several MWEs are lexically fixed: replacing a constituent by a semantically (and syntactically) similar word results in an invalid or a literal expression. For example, English salt of the earth cannot be construed as salt of the soil or salt of the land. Consider also Hebrew akl at hkwby ("eat the+ hat") "eat one's hat"). Substituting the noun kwby "hat" by mcnpt "conical hat", mgbyt "brimmed hat", or qsdh "helmet" would result in a (nonsensical) literal meaning.

In some cases MWEs are more lexically flexible. For example, xTp mkwt ("snatched hits" "be hit") has several variants, e.g., xTp sTirh ("snatched a-slap") "be slapped").

Limited paraphrasing

Certain syntactic structures can be paraphrased by other constructions, generally retaining the semantics. Some MWEs resist this possibility. For example, English devil's advocate is unlikely to be paraphrased as # the advocate of the devil. Hebrew MWEs typically allow only one of the three possessive constructions discussed in Section 3.3. This is true for noun compounds, which are formed in the first manner (with construct-state nouns), and cannot occur in the other two: bit spr ("house-of book") "school"), but not #bit sl spr "house of book". As another example, the expression ymd yl dytw ("stand on his-mind") "insist") requires the possessive enclitic, and is nonsensical with the genitive preposition: #ymd yl hdyt slw "stand on the+ mind of-him".

Literal translation

MWEs tend to be translated to other languages in a non-literal, non-compositional way. Often, MWEs are translated to a single word in a foreign language (that happens to lexicalize the same concept). English by and large translates to Hebrew as a single word (byrk). Hebrew examples include itr yl kn ("more on thus") "furthermore"), qwr rwx ("coldness-of spirit") "calmness"), swlxn ybwdh ("table-of work") "desk"), bit spr ("house-of book") "school"), etc. If an expression in one language is translated to a single lexical item in another, it is an indication that it denotes a single concept, and is hence more likely to be a MWE.

Limited reference

When a head is modified by adjuncts, it is usually possible to refer back to the head, suppressing the adjuncts, in subsequent text. This possibility may be unavailable in MWEs, especially when their meaning is idiomatic. For example, English red tape does not permit reference to tape. Similarly, Hebrew awr irwq ("light green") "permission") cannot be paraphrased as # hawr "the+ light"; biqs at idh ("ask-for her-hand") "ask for one's hand in marriage") does not permit reference to id "hand".

Limited modification

Compositional expressions can have their parts modified by adjectives, adverbs, prepositional phrases, etc. These can either modify a single element of the expression (internal modification), or the entire expression (external modification). Internal modification is often restricted in MWEs, especially when their meaning is idiomatic. For example, sprwt iph ("literature pretty") "belles-lettres") cannot occur as #sprwt mawd iph "literature very pretty". Similarly, there is an important difference between the way external modification affects compositional expressions and MWEs. In the case of compositional expressions, an adjunct can modify a

part of the expression or the expression as a whole; hence prt dwdti hsmnh "cow-of my-aunt the+ fat" is ambiguous as to who is fat. In contrast, ywrk hdin hxds"editor-of the+ law the+ new" can only mean "the new lawyer", and not #"the editor of the new law".

We do not address semantics directly in this work, as our lexicon does not include any meaning representation. But it is clear from the above discussion that MWEs must be lexically specified. Moreover, our lexicon includes English translations of many of the items; since the meaning of MWEs is often non-compositional, it is clear that their translations have to be lexically specified as well.

4.2 Syntactic properties

Like other phrases, MWEs exhibit, syntactically, a great variety and some flexibility. However, many of the syntactic properties of MWEs set them apart from compositional phrases. We list some syntactic properties of MWEs below. Most of them are not specific to Hebrew; we provide examples in both Hebrew and English where possible.

Variety

The syntactic category of MWEs is highly varied, as the following examples demonstrate:

Verb

English look up; Hebrew hlk mxil al xil ("go from+ army to army") "succeed"), ymd yl dytw ("stand on his-mind") "insist"), xzr bw ("return in-him") "regret").

Noun

English roller coaster; Hebrew bit spr ("house-of book") "school"), sprwt iph ("literature pretty") "belles-lettres"), ab bit din ("father-of house-of law") "president of the court").

Adjective

English straight-faced; Hebrew isr lb ("straight-of heart") "honest"), byl sywr qwmh ("owner-of measure-of height") "honorable"), ql dyt ("light-of mind") "hasty").

Adverb

English with flying colors; Hebrew bid xzqh ("in+ hand strong") "forcefully"), bsbr pnim ipwt ("in+ expression-of face beautiful") "kindly"), xd wxlq ("sharp and+ smooth") "straightforwardly"). Note that while the constituents of the last expression are adjectives, the resultant MWE is an adverb.

Conjunction

English let alone; Hebrew kmw kn (“like thus”) “also”), ala am kn (“but if thus”) “unless”), aš̌r yl kn (“that on thus”) “therefore”), ap yl pi (“even on mouth-of”) “although”).

Preposition

English in order to; Hebrew al ybr (“to direction”) “towards”), yl awdwt (“on concerning”) “about”), yl mnt (“on portion-of”) “in order to”).

POS transformation

In many cases the part of speech of the MWE is different from that of a literal interpretation. For example, the MWE ph axd (“one mouth”) “unanimously”) is an adverb, while the POS of its literal meaning is a noun phrase.

As mentioned above, we represent the syntactic category of the MWEs in the lexicon (Section 6.1).

Open slots

Some MWEs contain open slots, which can be filled with various complements. As an example, consider the expression iš̌b yl X š̌byh (“sit on X seven (days)”) “mourn”). The open slot, indicated by X, can be filled by any noun phrase, including a pronoun, in which case it is realized as an enclitic of the preposition yl “on” : iš̌b yliw š̌byh (“sat on-him seven”) “mourn him”). Similarly, English has drive X crazy.

We account for MWEs with open slots in Section 6.8.

Limited constituent order

Major constituent order is relatively flexible in Hebrew, and objects (and adverbials) can often precede their head verbs. Most MWEs, however, are more rigid. For example, in ica mhklim (“left from+ the+ tools”) “go postal”), the complement must immediately follow the verb, as is the case with mt yl (“die on”) “love”). Still, some MWEs, especially verb phrases, exhibit some flexibility. For example, iš̌b š̌byh yl abiw “sat seven on his-father” is perfectly acceptable, in addition to the variants shown above.

Constraints on word order are discussed in Section 6.7.

Asymmetry

In compositional expressions, phrases combined by a coordinating conjunction are interchangeable. This variation is unavailable in some MWEs. For example, English more or less and first and foremost must occur in this particular order, as does Hebrew pxwt aw iwtr (“less or more”) “more or less”). Similarly, ica bs̈n wyin (“went-out in+ tooth and+ eye”) “be injured, loose”), cannot be construed as #ica byin ws̈n.

Again, such constraints can be modeled with the mechanism we propose in Section 6.7.

Limited transformation

Several syntactic transformations that apply to compositional phrases are more limited when MWEs are concerned. For example, verb phrases headed by a transitive verb can undergo passivization; in MWEs this transformation may be blocked. In Hebrew, sp̌k at lbw ("spilled his+ heart") "confess") cannot occur as # lbw nsp̌k "his+ heart was-spilled". Similarly, bnh mgdlim bawwir ("built towers in+ the+ air") "build castles in the air") cannot be realized as # mgdlim nbnw bawwir "towers were-built in+ the+ air".

Hebrew has powerful (derivational) morphological processes that account for several transformations. For example, passivization is a lexical process in Hebrew: active verbs have passive counterparts that are often morphologically related to the active forms. Thus, the verb sp̌k "spill" has a counterpart nsp̌k "be spilled". However, as noted above, the passive form cannot be used in the MWE sp̌k at lbw. When a verb in a MWE can undergo passivization, we list the active and passive variants as separate lexical entries.

Syntactic irregularity

Finally, some MWEs are constructed using syntactic patterns that would be ungrammatical for compositional phrases. Examples include English by and large; Hebrew bxwr wTwb ("young-man and+ good") "an outstanding young man"), which conjoins a noun with an adjective; am ki ("if because") "however"), which consists of two conjunctions in a sequence; ild Twb irwsl̈im ("boy good Jerusalem") "obedient"), which has the irregular pattern noun+ adjective+ proper name; and nxba al hklim ("hide toward the+ tools") "shy"), in which the preposition al "toward" is not subcategorized by the verb.

Since MWEs are lexically specified, they can violate syntactic constraints that would hold for compositional phrases.

4.3 Orthographic and morphological properties

The orthographic properties of Hebrew, described in Section 3.1, motivate two design decisions. First, it is crucial to represent particles that are realized as prefixes explicitly in the MWE lexicon; we account for this in Section 6.1. Second, since the prefixes that combine with MWEs are dependent on the syntactic type of the entire MWE, it is crucial to represent the syntactic category of MWEs in the lexicon (of course, there are further, more general reasons for encoding the POS of MWEs). This is accounted for in Section 6.1.

As far as morphology is concerned, MWE constituents may exhibit idiosyncratic morphological behavior which differs from their behavior in isolation. This is manifested in the following manners (again, examples are mostly from Hebrew, but English examples are provided when the limited morphology of English allows it).

Frozen form

Constituents can appear in one fixed (frozen) form, disallowing all inflections. This form can be the citation form, such as part in take part, or Hebrew id "hand" in the expression ain lw id bdbr ("does not have a hand in the thing") "is uninvolved"), or kptwr "button" in kptwr wprx ("a button and a flower") fantastic"). It can also be some inflected form, e.g., beans in spill the beans, or hxlwnwt "the windows" in hxlwnwt hgbwhim ("the+ high the+ windows") "upper echelon"); in this case, the singular form is unavailable: hxlwn hgbwh "the+ high the+ window" only has a literal meaning.

Frozen forms are of course trivial to represent, see Section 6.1 for the basic representation.

Partial inflection

In some cases, constituents undergo a (strict) subset of the full set of inflections that they would undergo in isolation. For example, consider the expression hlk axri lbw ("walk after his+ heart") "follow one's heart") : the noun lb "heart" takes a possessive suffix, which must agree with the verb in number, gender, and person, as in hlkw axri lbm "they followed their heart"; but the noun itself does not inflect for number, hence the invalidity of #hlkw axri lbbwtihm "they followed their hearts". As another example, consider bit xwlim ("house-of sick-people") "hospital"). The second noun, xwlim "sick-people", can take a pronominal suffix, as in xwlik "your patients"; but in the context of the MWE, this option is prohibited.

We account for partial morphological inflections in Section 6.2.

Non-standard inflection

Constituents can also undergo non-standard morphological inflections that they would not undergo in isolation. For example, consider the expression bdltiim sgwrwt ("in+ two-doors closed") "behind closed doors"). The first constituent, bdltiim "in+ two-doors", consists of the prefix b "in" followed by the dual form of dlt "door". As noted above, the dual form is unproductive in Hebrew, and in particular, it does not apply to dlt "door", except in this frozen expression.

We account for partial morphological inflections in Section 6.4.

Fossil words

The most extreme cases of idiosyncratic morphological behavior involve constituents that only occur in MWEs. English examples include run amok and without further ado. In Hebrew kmTxwwi qsṫ "a stone's throw", the word mTxwwi (the prefix k is the preposition "as"), by itself, has no literal meaning. Another example is abd yliw hklx "outdated", in which the third word, hklx, is obsolete outside this MWE. This situation is frequent in expressions borrowed from other languages, e.g., lit man dplig "without dispute" which is originally Aramaic. While it may be perfectly compositional in the source language, it is acquired as a single unit to the target language and hence its constituents do not occur in isolation.

We account for fossil words in Section 6.3.

Violated agreement

In some MWEs, constituents that generally agree in morphological features such as number, gender, person, state, or definiteness, violate the agreement constraints. For example, in yin hry ("eye the+ evil") "evil eye"), the noun yin "eye" (feminine indefinite) and the adjective hry "the+ evil" (masculine definite) should agree in number, gender, and definiteness, but agreement in both gender and definiteness is violated.

Agreement is discussed in Section 6.6.

4.4 Case study: noun compounds

To emphasize the idiosyncratic properties of Hebrew MWEs we focus in this section on one particular construction, noun compounds, and list some of its unique properties. Recall (Section 3.3) that one of the ways to express genitive relations in Hebrew involves two nouns, the first of which (the head) is in the construct state and the second (the modifier) is in the absolute or definite state. This construction, which we denote noun-noun construction (NNC) here, is highly prevalent, especially in written texts. Many instances of NNCs are MWEs; we refer to those as noun compounds.

Noun compounds are "idiosyncratic in a regular way". They exhibit various properties that distinguish them from non-MWE NNCs, but all noun compounds exhibit these properties. This is not coincidental: several MWE types, in various languages, behave in the same way. This observation will motivate our design of prefabricated templates (Section 7.1).

We list below some of the idiosyncratic properties of noun compounds. These will be demonstrated on the noun compound ywrk din ("editor-of law") "lawyer"), which will be contrasted with the non-MWE NNC ywrk yitwn ("editor-of journal") "journal editor").

Partial inflection

Compositional NNCs allow both their constituents to inflect freely; thus, ywrk yitwn (“editor-of journal”) “journal editor”) gives rise also to several inflected forms, including: ywrk hyitwn (“editor-of the+ journal”) “the editor of the journal”), ywrk yitwnim (“editor-of journals”) “an editor of some journals”), ywrk hyitwnim (“editor-of the+ journals”) “the editor of the journals”), and many others. In contrast, while the second constituent of ywrk hdin (“editor-of the+ law”) lawyer”), namely din “law”, can have a definite article prefix, it cannot inflect for number. Hence ywrk hdin (“editor-of the+ law”) “the lawyer”), is fine, but not #ywrk hdinim “the lawyers”.

Hebrew nouns can combine with pronominal suffixes to encode possessives, and this construction is available for some, but not all, noun compounds. For example, ywrk din “lawyer” yields ywrk dinw (“editor-of his-law”) “his lawyer”); but bit xwlim (“house-of sick-people”) “hospital”) does not give rise to #bit xwlik “house-of your-sick-people”, although xwlik “your-sick-people” is grammatical.

Non-standard inflection

Compositional NNCs are construed definite by adding the definite article h “the” to their second constituent; noun compounds also allow a form in which the definite article attaches to the first constituent, the construct-state noun, in violation of prescriptive constraints of the normative grammar: hywrki din (“the+ editors-of law”) “the lawyers”). Recall (Section 3.2) that Hebrew nouns can either be in the construct state, or be definite, but not both; forms like hywrki “the+ editors-of” are not prescriptively grammatical, and are not generated by the morphological generator (Section 5.2).

Limited paraphrasing

While ywrk yitwn (“editor-of journal”) “journal editor”) has an equivalent construction with an almost identical meaning, ywrk šl yitwn (“editor of journal”) “journal editor”), this construction is impossible for noun compounds: #ywrk šl din “editor of law”.

Limited modification

In compositional NNCs it is possible to modify (with an adjective) either the first or the second constituent:[4] ywrkt hyitwn hxdš (“editor-of the+ journal-m the+ new-m”) “the female editor of the new journal”) vs. ywrkt hyitwn hxdšh (“editor-f the+ journal-m the+ new-f”) “the new female editor of the journal”). Compounds only allow adjectival modification of the entire phrase: #ywrkt hdin hxdš “editor-f the+ law-m the+ new-m” vs. ywrkt hdin hxdšh (“editor-f the+ law-m the+ new-f”) “the new female lawyer”).

Limited coordination

Two compositional NNCs that share a common head can be conjoined using the coordinating conjunction w "and". This option is blocked if one of them (or both) are compounds. For example, while ywrk yitwn wirxwn "editor-of journal and+ monthly" is permissible, #ywrk din wyitwn "editor-of law and+ journal" is not.

Lexical fixedness

While it is possible to replace any of the constituents of a compositional NNC by synonyms or related terms, this is impossible in the case of noun compounds: ywrk yitwn ("editor-of journal" "journal editor") gives rise to ywrk sbwywn ("editor-of weekly" "editor of a weekly"), but ywrk din ("editor-of law" "lawyer") does not license #ywrk xwq "editor-of law", although din and xwq are synonymous.

5. The Hebrew morphological processor

We incorporate our MWE representation and processing architecture in an existing morphological processing system of Hebrew (Wintner and Yona, 2003; Itai and Wintner, 2008) that we describe in this section. The architecture of the existing system is depicted in Figure 1. It consists of two main units: the Generation Unit, which includes a Lexicon, a Generator, and a Database of inflected forms; and the Analysis Unit, consisting of a Tokenizer, a Morphological Analyzer, and an XML wrapper. Below, we briefly describe

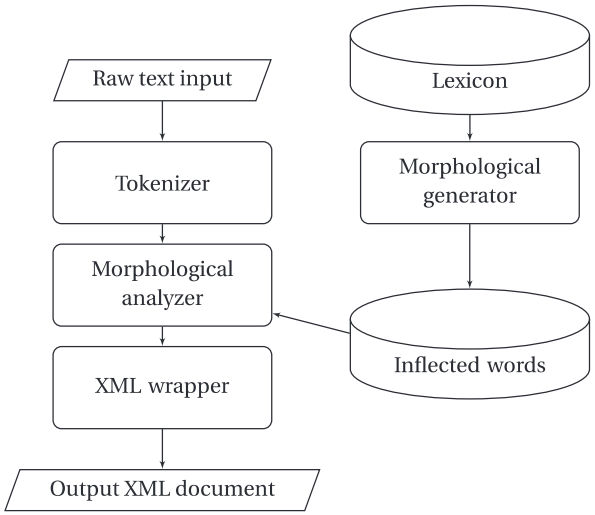

Figure 1: The architecture of the basic morphological system

each unit and explain how the modules interact to produce morphological analyses for input words (for further details, see Itai and Wintner (2008)).

5.1 Lexicon

The MILA Lexicon of Contemporary Hebrew (Itai et al., 2006) is the broadest-coverage publicly available lexicon of Hebrew, currently consisting of over 31,000 entries. The lexicon is represented in XML[5] as a list of item elements, each with a base form which is the citation form used in conventional dictionaries.

Lexical entries have a unique ID; they list the citation form of the entry. In addition, every lexical entry belongs to a part of speech category, which is designated as a sub-element of the item. The part of speech of an entry determines its additional attributes. In addition, each lexical entry specifies features which govern the inflectional morphological behavior of the lexeme (which is used by the generator, see below). Examples of (somewhat simplified) lexical entries are depicted in Figure 2.

The precise format of the lexicon is defined by an XML schema[6] and is described in detail by Itai and Wintner (2008). For the purpose of the present paper, it suffices to realize that each item has several attributes, most crucially a unique ID (in Figure 2, attributes like `id=''8442''`) and several representations of the lemma (two variants in the Hebrew script, dotted and undotted, and one transliteration in ASCII; in the sequel we only present the latter, e.g., `transliterated=''akl''`). Then, an item has a single element that specifies its POS category (approximately twenty categories are defined). Depending on the POS, the element is specified for further attributes. For example, nouns have attributes for number and gender, but also attributes that determine whether a feminine inflection is possible, and if so, the form of the feminine suffix (in Figure 2, this is `feminine=''t''`); the form of the plural suffix (`plural=''im''`); whether a dual inflection is possible, etc. Verbs have an attribute that specifies their inflectional pattern (e.g., `inflectionPattern=''5''`);

```
<item id="8442" transliterated="akl" >
  <verb inflectionPattern="5" valence="transitive"/>
</item>
<item id="8174" transliterated="ywrk" >
  <noun deverbal="true" dual="false" feminine="t" plural="im"/>
</item>
<item id="8018" transliterated="xzq">
  <adjective feminine="h" plural="im"/>
</item>
```

Figure 2: Lexical entries of the verb akl "eat", the noun ywrk "editor" and the adjective xzq "strong" (simplified)

this is basically a pointer to a specific entry in (traditional, printed) verb paradigm tables that list all the inflections of Hebrew verbs (Zdaqa, 1974).

The schema also allows for some representation of meaning, in the form of (potentially several) sense elements that are specified for each item. Senses can include, as attributes, definitions (but this is not currently used), and as embedded items, translations (currently, only to English). For more details, see the XML schema itself; we list in Appendix A an example of a complete lexical entry.

5.2 Morphological generation

The generator is a computational implementation of the inflectional morphology of Modern Hebrew (Yona and Wintner, 2005, 2008). It creates, off-line, all the inflected forms induced by each lexical item (excluding combinations of prefix sequences with the inflected forms).[7] These forms are then stored in a database, which lists for each inflected form a pointer to its citation form (the lexicon ID of the item from which it was generated), and the complete morphological analysis, in terms of a set of feature-value pairs. We refer to the resulting database as the database of inflected forms.[8]

5.3 Morphological analysis

5.3.1 Tokenization. The tokenizer module operates on the input text (UTF-8 encoded raw data), and segments it into paragraphs, sentences and tokens. The output of the tokenizer is fed into the morphological analyzer.

5.3.2 Morphological analysis. The morphological analyzer strips possible prefixes (taken from a list of all possible prefix sequences) off each token and matches the remaining strings against the database of inflected forms. When a match is successful, the prefix and the remaining string are passed to the analyzer, which determines whether the combination of the prefix sequence and the inflected form is (syntactically) valid, in which case the analysis is fed to the XML wrapper. Note that morphological analysis is thus reduced to not much more than table lookup.

5.3.3 XML wrapper and the corpus representation schema. The XML wrapper wraps all possible analyses of each token in XML and returns an XML document corresponding to the entire input text. The XML document follows The Hebrew Corpus XML Schema, which induces a well-defined structure on the document. A morphologically analyzed corpus contains all the analyses of a word (as produced by the morphological processor), regardless of context. Each analysis consists of zero or more prefixes, a base and an optional

suffix. The base specifies the properties of the lemma of the token, including its form, part of speech and POS-dependent features (such as number, gender and nominal state in the case of nouns). Appendix B lists a fragment of a small morphologically-analyzed corpus, represented in XML.

Note that the morphological processor operates on a token-by-token basis. The tokens are acquired from the tokenizer which uses only blanks and punctuation to segment a text into tokens. In particular, the tokenizer is completely independent of the lexicon. Crucially, the lexicon includes single-word tokens only, and the morphological analyzer is completely unaware of MWEs.

6. Lexical representation of Hebrew MWEs

Motivated by the observations of Section 4, we now present an architecture for lexical representation of MWEs; our design decisions are driven by properties of Hebrew MWEs discussed and exemplified in Section 4. We only focus on the lexicon in this section; a protocol for integrating MWEs into the existing morphological processor of Hebrew (Section 5) is presented in Section 8.

Our approach is to design a representation for MWEs that is simple and consistent with the current lexicon, on one hand, and on the other hand is expressive enough to account for the properties of MWEs discussed in Section 4. We adopted the original XML schema of the MILA Lexicon (Section 5), with all its attributes and elements (retaining their values and functions), and further extended it by adding new elements and attributes, which we gradually describe in this section through multiple examples. A summary of the changes we introduced is presented in Section 6.10.

6.1 Basics

We begin with the very simplest case of MWEs, namely frozen forms. In the extended schema, each MWE is represented as an item in the lexicon, which encodes its morphological and syntactic properties. These properties serve as directives for generating all the possible forms that the MWE can appear in. A MWE lexical entry includes an element that specifies that the item is a MWE, followed by its POS. Recall that storing the syntactic category of MWEs is required for properly handling the combination of prefixes with such expressions (Section 4.3), and to model the syntactic variability of MWEs (Section 4.2).

Figure 3 depicts a fragment of the lexical entry of the MWE niw iwrq "New York". Note the element item with its attributes. Since names of towns in Hebrew are all feminine, the gender attribute of the MWE is specified as feminine.

Each of the MWE constituents has its own features and inflection directives. Each constituent is realized as an atom, and since we consider prefixes and

```
<item id="28498" transliterated="niw iwrq">
      <MWE pos="properName" type="town" gender="feminine"/>
</item>
```

Figure 3: Fragment of the lexical entry of niw iwrq "New York"

suffixes to be separate atoms, an atom represents a morpheme, rather than a word. Consequently, a MWE consists of a sequence of morphemes, either free or bound. This representation facilitates the specification of MWEs consisting of a verb and a prefix that attaches to another word that is not part of the MWE. It also facilitates referring to the inflectional properties of pronominal suffixes: recall (Section 3.2) that nouns can take suffixes that are interpreted as possessives; these inflect for number, gender, and person, and are sometimes involved in agreement constraints with other constituents of the MWE. Because they have their own agreement properties, such suffixes are listed separately, so that agreement constraints involving other constituents of the MWE can be specified.

We thus add the following new elements and attributes to the lexicon schema:

atom

Defines a constituent along with all its possible inflected forms. Each atom is specified for a unique id, and (with the exception of suffixes) a lexiconPointer that links it to the lexical entry of the constituent's citation form. The atom has the following optional sub-elements:

prefix

Specifies that the constituent is a prefix that is an inherent part of the MWE.

inflect

Restricts the possible inflections of the constituent to those specified. Each attribute restricts the inflection of a specific morphological feature. If a feature is not specified then the constituent can inflect for all the possible values of that feature. If the entire inflect element is missing then the constituent is frozen, i.e., only its base form may be part of the MWE. Note that specifying all the features of the base form is equivalent to omitting the inflect element.

suffix

Specifies that the constituent is a suffix that attaches to the previous atom. A constituent cannot take a pronominal suffix unless the following atom is specified as a suffix atom.

Figure 4 depicts a fragment of the lexical entry of the MWE mcd sňi ("from side second") "on the other hand"). It consists of three morphemes, the

preposition m "from", the noun cd "side" and the adjective sňi "second"; all three are frozen. In general, nouns may inflect for state, number and definiteness. However, in this case the second constituent cd is frozen, the values of these features being absolute, singular and indefinite, respectively. The third constituent is an adjective and hence in addition may inflect for gender, which in this case is also frozen (masculine), agreeing with the gender of the preceding noun. Since the three constituents are frozen, the values of their attributes are simply stipulated; in Section 6.6 we describe a more elaborate mechanism for specifying agreement among constituents when the values of features may vary.

6.2 Partial morphological inflections

The basic mechanism described above can support partial inflections (Section 4.3), as it facilitates the specification of a subset of the inflected forms that MWE constituents would otherwise be able to occur in.

Figure 5 depicts (a part of) the representation of the MWE iwsb raš ("sitter-of head") "chairman"). This is a noun-noun compound (Section 4.4) in which the first constituent, the noun iwsb "sitter", is a participle derived from the verb isb "sit".[9] It must be in the construct state, but is free to inflect

```
<item id="29000 transliterated="mcd šni">
  <MWE pos="adverb"/>
  <atom id="1" lexiconPointer="10418">                                    <!-- m -->
      <prefix/>
  </atom>
  <atom id="2" lexiconPointer="20473">                                   <!-- cd -->
    <inflect state="absolute" definiteness="false" number="singular"/>
  </atom>
  <atom id="3" lexiconPointer="3561">                                  <!-- šni -->
    <inflect state="absolute" definiteness="false" number="singular" gender="masculine"/>
  </atom>
</item>
```

Figure 4: The lexical entry of mcd sňi ("from side second") "on the other hand")

```
<item id="39990" transliterated="iwšb raš"
  <MWE pos="noun"/>
  <atom id="1" lexiconPointer="14020">                                <!-- iwšb -->
    <inflect tense="participle" state="construct" definiteness="false"/>
  </atom>
  <atom id="2" lexiconPointer="20910">                                 <!-- raš -->
    <inflect number="singular"/>
  </atom>
  <atom id="3" lexiconPointer="0">                          <!-- pronominal suffix -->
    <suffix/>
  </atom>
</item>
```

Figure 5: Part of the lexical entry of iwsb raš ("sitter-of head") "chairman")

for gender and for number. This is achieved simply by omitting these attributes in the inflect element, thus there are no gender and number restrictions. Similarly, the second constituent, the noun raš "head", must be singular, but can freely inflect for definiteness. This entry yields forms such as: iwsbt raš (feminine, singular, indefinite), iwsb hraš (masculine, singular, definite), iwsbi raš (masculine, plural, indefinite), etc. The third atom, a suffix, is listed in order to allow the generation of forms in which the noun raš "head" combines with a possessive suffix, e.g., ywsb rash ("sitter-of her-head") "her chairman"). It is specified as `lexiconPointer=''0''` because pronominal suffixes are not listed in the lexicon, they are handled by the generator.

6.3 Fossil words

Recall that fossil words (Section 4.3) include constituents that never occur in isolation; such constituents are not listed in the lexicon. Therefore, for each of these constituents we create a new item in the lexicon, with all the standard attributes that an item is specified for, but with a special designation, fossil, for the POS, so that the generator does not generate it (or any inflections thereof) in isolation.

Figure 6 depicts the lexical entry of the fossil word kmTxwwi, followed by (part of) the lexical entry of the MWE kmTxwwi qšt "stone's throw".

6.4 Irregular inflections

Components of MWEs may undergo non-standard morphological processes (Section 4.3). Consequently, MWEs may have to specify forms that are not listed in the database of inflected forms (Section 5.2). We use the same approach as in the case of fossil words (Section 6.3) to solve this problem: we add the special forms to the lexicon, and refer to them explicitly in the MWE.

One case of irregular inflections, however, is very common: this is the case of noun-noun compounds, where the definite article can combine with a construct-state noun (Section 4.4). We devised an ad-hoc mechanism to address

```
<item id="27000" transliterated="kmTxwwi">
      <fossil/>
</item>
<item  id="23999 transliterated="kmTxwwi qšt">
  <MWE pos="adverb"/>
  <atom id="1" lexiconPointer="27000"/>                    <!-- kmTxwwi -->
  <atom id="2" lexiconPointer="3507">                          <!-- qšt -->
    <inflect definiteness="false" state="absolute" number="singular"/>
  </atom>
</item>
```

Figure 6: Part of the lexical entry of kmTxwwi qšt and its fossil constituent kmTxwwi

this idiosyncrasy: we add the attribute hprefix to the MWE. This serves as an indication to the generator to allow the colloquial forms in which the definite article h "the" is attached as a prefix to the entire MWE. See Figure 7.

6.5 Retrieving morphological features

Often the MWE inherits some of its morphological features from those of the constituents. Figure 8 depicts a fragment of the lexical entry of the MWE kwx adm ("power-of man") "manpower"). Note that the second constituent inflects for definiteness, and the entire MWE inherits the definiteness feature from this constituent. In the XML schema, this is modeled by having the value of such an attribute equal the ID of the constituent from which it inherits the value (e.g., `definiteness=''2''` in Figure 8).

Figure 9 revisits the lexical entry of the MWE iwsb raš ("sitter-of head") "chairman"). Note the possible inflected forms of both constituents. Observe that the number and the gender of the MWE are inherited from the first element, whereas the definiteness is inherited from the second element, as is typical for noun-noun compounds in Hebrew. See also Section 7.1.

6.6 Agreement among constituents

Some MWEs require agreement among the features of some of their constituents (Section 4.3). For example, milh nrdpt ("word chased") "synonym") can inflect for number and for definiteness, and both constituents must agree in these features. Conversely, some MWEs involve constituents that violate agreement constraints that would have been enforced in compositional phrases. We use a similar mechanism to the one introduced in Section 6.5 to model agreement (and disagreement) among constituents. Instead of explicitly specifying the value of some attribute, an atom can list the ID of some other atom; this implies that the value of the attribute is taken from the other atom.

Figure 10 depicts the lexical entry of milh nrdpt ("word chased") "synonym"). Observe that the values of the attributes number and definiteness in the

```
<item id="28579" transliterated="ywrk din" hprefix="true">
  <MWE pos="noun" definiteness="2" state="2" number="1" gender="1"/>
  <atom id="1" lexiconPointer="8174">
   <inflect state="construct"/>                                    <!-- ywrk -->
  </atom>
  <atom id="2" lexiconPointer="5208">                               <!-- din -->
    <inflect number="singular"/>
  </atom>
  <atom id="3" lexiconPointer="0">                        <!-- pronominal suffix -->
    <suffix/>
  </atom>
</item>
```

Figure 7: The lexical entry of ywrk din "lawyer"

```
<item id="29100" transliterated="kwx adm">
  <MWE pos="noun" gender="masculine" number="singular" definiteness="2"/>
  <atom id="1" lexiconPointer="4192">                                    <!-- kwx -->
    <inflect state="construct" number="singular" definiteness="false"/>
  </atom>
  <atom id="2" lexiconPointer="11357">                                   <!-- adm -->
    <inflect state="absolute" number="singular"/>
  </atom>
</item>
```

Figure 8: The lexical entry of kwx adm (“power-of man”) “manpower”)

```
<item id="39990" transliterated="iwšb raš"
  <MWE pos="noun" state="2" definiteness="2" number="1" gender="1"/>
  <atom id="1" lexiconPointer="14020">                                <!-- iwšb -->
    <inflect tense="participle" state="construct" definiteness="false"/>
  </atom>
  <atom id="2" lexiconPointer="20910">                                 <!-- raš -->
    <inflect number="singular"/>
  </atom>
  <atom id="3" lexiconPointer="0">                          <!-- pronominal suffix -->
    <suffix/>
  </atom>
</item>
```

Figure 9: The lexical entry of iwšb raš (“sitter-of head”) “chairman”)

```
<item id="39991" transliterated="milh nrdpt"
  <MWE pos="noun" state="absolute" definiteness="1" number="1" gender="feminine"/>
  <atom id="1" lexiconPointer="3265">                                  <!-- milh -->
    <inflect state="absolute"/>
  </atom>
  <atom id="2" lexiconPointer="10097">                                <!-- nrdpt -->
    <inflect tense="participle" type="adjective" state="absolute"
             definiteness="1" number="1" gender="feminine"/>
  </atom>
</item>
```

Figure 10: The lexical entry of milh nrdpt (“word chased”) “synonym”)

second atom, the adjective nrdpt “chased”, are not specified explicitly, but rather “point to” the attributes of the first atom, the noun milh “word”. In addition, the values of the number and definiteness attributes of the MWE are inherited from the first atom, as in Section 6.5: the two mechanisms can be used in parallel in the same lexical entry. Note that since the gender of the first constituent is (inherently) feminine, the gender of the second atom and the entire MWE are fixed as feminine.

Sometimes the morphological features of a suffix constituent must also agree with, or control, the features of other constituents. This is the case with ymd yl dytw (“stand on his-mind”) “insist”) (Figure 11), in which the pronominal suffix of dyt “mind” must agree with the verb ymd “stand”. We model such cases in exactly the same way, as suffixes are represented as standard constituents. The person, number and gender attributes of the first atom, corresponding to the verb ymd “stand”, all “point” to the fourth atom, namely

```
<item id="40015" transliterated="ymd yl dytw"
  <MWE pos="verb" person="4" number="4" gender="4" tense="1"/>
  <atom id="1" lexiconPointer="19724">                              <!-- ymd -->
    <inflect person="4" number="4" gender="4"/>
  </atom>
  <atom id="2" lexiconPointer="17434"/>                              <!-- yl -->
  <atom id="3" lexiconPointer="8300"/>                              <!-- dyt -->
  <atom id="4" lexiconPointer="0">
    <suffix/>                                                    <!-- suffix -->
  </atom>
 </item>
```

Figure 11: The lexical entry of ymd yl dytw (“stood on his mind”) “insisted”)

pronominal suffix attached to the noun dyt “mind”. The attributes of the suffix are not specified since their values are unrestricted and can vary freely.

6.7 Accounting for syntactic flexibility

MWEs may allow variation in the order of their constituents (Section 4.2); we add a set of attributes and elements in order to account for this variability. The default is that all the constituents must appear consecutively in the order determined by the atoms. If other orders are possible, all the allowed permutations are prescribed within perm items. Consider the MWE ysh lilwt kimim (“made nights like-days”) “work intensively”) (Figure 12). Since it may also appear as ysh imim klilwt “make days like-nights”, with an identical meaning, we list also the order “1 4 3 2” (the preposition k is a prefix and is prepended to the following atom).

Specifically, we add the following elements and attributes to the schema:

perms
: defines a set of possible orders for the constituents of a MWE. It consists of a list of permutation sub-elements.

perm
: defines a particular order of the constituents of a MWE. It has two attributes, an ID and an order attribute that defines the surface order of the atoms specified earlier for this MWE.

6.8 Open slots

The mechanism described above facilitates proper treatment of MWEs with open slots (Section 4.2), such as akl at X bli mlx (“eat X without salt”) “defeat”). We allow each permutation to include, in addition to atom IDs, also the values “+ ”, indicating an open slot of one or more words, and “*”, an open slot of zero or more words.[10] In Figure 13, the last five permutations

```
<item id="39991" transliterated="yšh imim klilwt">
  <MWE pos="verb" tense="1" person="1" number="1" gender="1"/>
  <atom id="1" lexiconPointer="376">                                    <!-- yšh -->
    <inflect/>
  </atom>
  <atom id="2" lexiconPointer="9475">                                   <!-- iwm -->
    <inflect state="absolute" definiteness="false" number="plural"/>
  </atom>
  <atom id="3" lexiconPointer="20001">                                    <!-- k -->
    <prefix/>
  </atom>
  <atom id="4" lexiconPointer="8024">                                  <!-- lilh -->
    <inflect state="absolute" definiteness="false" number="plural"/>
  </atom>
  <perms>
    <perm id="1" order="1 2 3 4"/>
    <perm id="2" order="1 4 3 2"/>
  </perms>
</item>
```

Figure 12: The lexical entry of yšh imim klilwt (“made days like-nights”) “work intensively”)

```
<item id="23986" transliterated="akl at + bli mlx">
  <MWE pos="verb" person="1" number="1" gender="1" tense="1"/>
  <atom id="1" lexiconPointer="8442">
   <inflect/>                                                           <!-- akl -->
  </atom>
  <atom id="2" lexiconPointer="3382"/>                                   <!-- at -->
  <atom id="3" lexiconPointer="0">                            <!-- pronominal suffix -->
    <suffix/>
  </atom>
  <atom id="4" lexiconPointer="21542"/>                                  <!-- bli -->
  <atom id="5" lexiconPointer="608"/>                                    <!-- mlx -->
  <perms>
    <perm id="1" order="1 2 3 4 5"/>                          <!-- akl awtw bli mlx -->
    <perm id="2" order="2 3 1 4 5"/>                          <!-- awtw akl bli mlx -->
    <perm id="3" order="2 3 4 5 1"/>                          <!-- awtw bli mlx akl -->
    <perm id="4" order="4 5 1 2 3"/>                          <!-- bli mlx akl awtw -->
    <perm id="5" order="4 5 2 3 1"/>                          <!-- bli mlx awtw akl -->
    <perm id="6" order="1 2 + 4 5"/>                          <!-- akl at + bli mlx -->
    <perm id="7" order="2 + 1 4 5"/>                          <!-- at + akl bli mlx -->
    <perm id="8" order="2 + 4 5 1"/>                          <!-- at + bli mlx akl -->
    <perm id="9" order="4 5 1 2 +"/>                          <!-- bli mlx akl at + -->
    <perm id="10" order="4 5 2 + 1"/>                         <!-- bli mlx at + akl -->
  </perms>
</item>
```

Figure 13: The lexical entry of akl at X bli mlx “defeat”

replace the suffix by one or more words. The effect is that in the first five forms the accusative marker “at” has a pronominal suffix, whereas in the last five it appears in isolation followed by one or more words.

6.9 Proper names

The same person’s name may appear in several different ways, middle names may be omitted, nicknames may be used. When processing documents it is

```
<item id="28605" transliterated="hnri wiliam pwrd">
  <MWE pos="properName" type="person" number="singular" gender="masculine"/>
  <atom id="1" lexiconPointer="7356"/>                    <!-- Henry -->
  <atom id="2" lexiconPointer="2266"/>                  <!-- William -->
  <atom id="3" lexiconPointer="222"/>                       <!-- W. -->
  <atom id="3" lexiconPointer="8544"/>                    <!-- Ford -->
  <perms>
    <perm id="1" order="1 2 4"/>              <!-- Henry William Ford -->
    <perm id="2" order="1 3 4"/>                   <!-- Henry W. Ford -->
    <perm id="3" order="1 4"/>                        <!-- Henry Ford -->
    <perm id="4" order="3"/>                             <!-- Ford -->
  </perms>
</item>
```

Figure 14: The lexical entry of hnri wiliam pwrd "Henry William Ford"

important that all realizations refer to the same entity. The permutation mechanism may be used to connect them all. Consider Henry William Ford. He may be referred to as Henry Ford, Henry W. Ford or just as Ford. Figure 14 shows how all such realizations can refer to a single lexicon item.

6.10 An XML schema for representating MWEs

We listed above several changes that we have introduced into the existing XML schema of the MILA lexicon in order to support MWEs. We now summarize the resulting lexical representation of MWEs. As mentioned above, this is an extension of the XML schema that was used for the MILA lexicon before the introduction of MWEs (Section 5.1).

Each MWE is represented as an item, with exactly the same attributes as other lexical items (i.e., an ID, three representations of the lemma, including the transliteration that we use here, etc.) Items have one MWE element, followed by several atom elements, and then a perms element.

The MWE element has several attributes. Most important is pos, which specifies the part of speech category of the item (Figure 3). We reuse all the twenty categories that simple lexical items are classified by, but other categories can of course be easily added (we do add the category fossil for fossil words, see Figure 6).

As in the case of standard lexical items, the POS determines the subsequent attributes of MWE elements. However, the values of those attributes is more elaborate. It can either be a standard value, e.g., `gender=''feminine''` (Figure 3); or a number, e.g., `gender=''1''` (Figure 9). The number refers to (i.e., is the ID of) one of the atom elements that follow the MWE element. This means that the value of the feature for the entire MWE is inherited from one of its atom constituents.

Atom elements correspond to the constituents of the MWE. They have two attributes, an ID (which can be used for specifying feature-value sharing, as mentioned above) and a lexiconPointer referring to the lexical entry of the

constituent. Atom elements can have at most one of the following optional sub-elements: prefix, indicating that the constituent is a prefix; suffix, indicating that the constituent is realized as a suffix on the previous atom; or inflect.

Inflect elements restrict the possible inflected forms of the atoms they are embedded in. If no such element is specified, the atom is frozen and cannot inflect (e.g., the first atom of Figure 6). Otherwise, the presence of some attribute under inflect constrains the value of that attribute. The possible attributes are determined by the POS of the atom in the same way that the main attributes of atom are determined. Furthermore, the value of an attribute can be a number, just like in the case of atoms, with the same meaning: if an attribute has a numeric value i, its value is taken from the atom whose ID is i. This can be used to force agreement among constituents of the MWE (Figure 10).

Last, the perms element can have several perm sub-elements, each with two attributes: an ID and an order attribute whose value is some permutation of (a subset of) the IDs of atom elements in the MWE. Such permutations determine the surface order of the MWE constituents (Figure 12). The special values + and * can be used instead of numbers, to indicate open slots that must be filled by at least one or at least zero words, respectively (Figure 13).

7. Adding MWEs to the lexicon

We listed in the previous section several examples that demonstrate both the idiosyncrasy and the variety of Hebrew MWEs. Lexical acquisition of MWEs, that is, the process of adding MWE entries to the lexicon, along with their unique idiosyncratic properties, is therefore a complex process. We focus on manual acquisition in this section, and suggest two mechanisms that ease the work of the lexicographer: using prefabricated templates, and a graphical user interface (GUI) we developed that guarantees the consistency of the underlying lexical database and speeds up the lexicographic work. While we do employ some automatic acquisition techniques for harvesting MWE candidates from corpora (mentioned briefly in Section 8.4), we insist that a lexicographer manually approve each and every change to the lexicon, thereby ensuring its continuous high quality.

7.1 Prefabricated templates

We exemplify the use of prefabricated templates by considering the special case of noun compounds, discussed in Section 4.4. We claimed that such MWEs, like many other types of MWEs in various languages, are "irregular in a regular way". We describe here the representation of noun compounds only, but the same mechanism can be applied to other types of frequent MWEs, e.g., verb-prepositions.

Noun compounds consist of two nouns, the first in the construct state and the second in the absolute state. All the morphological features of the construction, except definiteness, are inherited from the first constituent. Thus the MWE bit spr (“house-of book”) “school”) is singular, while bti spr (“houses-of book”) “schools”) is plural. Definiteness is inherited from the second constituent, as in bit hspr (“house-of the+ book”) “the school”). However, the non-standard form hbit spr (“the+ house-of book”) “the school”) also exists (see Section 6.4).

Since noun compounds are rather regular, their morphological features and inflection directives can be derived from a small number of parameters; we refer to this set of parameters as the template of noun compounds. The value of each parameter is set to a default value (see Table 1) but a lexicographer can change the values for specific compounds that behave in irregular ways. The default values reflect the observations that noun compounds typically exhibit both internal (prescriptive) and external (colloquial) definiteness processes, as in bit spr “school”; that they do not inflect for gender; that the two constituents must be consecutive; etc. The templates are associated with all noun compounds in the lexicon; they are expanded to the full lexical representation of MWEs using straight-forward rules. Thus, it is sufficient to specify that iwsb raš (“sitter-of head”) “chairman”) is an instance of the noun compound template, in order to obtain the lexical representation listed in Figure 9. This facilitated the rapid acquisition of 1214 NNC’s.

Other common templates include frozen expressions (1798 expressions), prepositions (13), and noun-adjective (125).

7.2 The flexible GUI tool

We hoped to cover all MWE by a small number of templates. However, as we examined more expressions, the number of templates grew as well as the number of parameters that needed to be specified. The situation became unmanageable when we tried to cover verbs and MWEs containing more than

Table 1: Features and their default values

Feature	Default value
definiteness	internal and external
inflect for gender	false
inflect for number	true
suffix	none
consecutive	true
spelling	standard

two constituents. Thus we were led to create a web-based graphic interface (GUI) to enable us to express all possible idiosyncrasies. The GUI closely follows the XML descriptions of the lexicon. The main page of each MWE contains the POS of the MWE, the lexicon IDs of its constituents and the template that defines the various orders. Each constituent leads to a separate page where the agreement and inflectional restrictions of that constituent are specified. Here, too, we use default values.

Finally, we are attempting to acquire the parameters automatically from a large corpus. This will work for common MWEs, but owing to the long tail of rare MWEs it is not clear whether we will have enough data to soundly deduce the behavior of all MWEs.

8. Morphological processing of MWEs

In the existing Hebrew morphological system (Section 5), a complete set of the inflectional morphological rules is applied to the lexicon, generating all the possible inflected forms off-line; these forms are stored in a database, and at analysis time, the main task is database lookup. The program that generates the inflected morphological forms embodies vast linguistic knowledge which is applicable to MWEs and to single words alike. However, the analyzer operates on a token by token basis, while MWEs involve the interaction of several tokens.

In designing the incorporation of MWEs in the morphological processor, we therefore decided not to interfere with the generator and analyzer, and instead to add a post-processing layer. The solution that we describe below first applies the existing morphological analyzer to full sentences; then, the post-processor identifies MWEs in the analyzed output using information derived from the MWE lexicon. We now detail all the changes that were introduced to the lexicon and morphological tools in order to support MWEs. The architecture of the upgraded morphological system is graphically depicted in Figure 15.

8.1 The MWE Lexicon

Our architecture uses a database, the MWE lexicon, that reflects all the information associated with each MWE, as demonstrated in Figures 3-14. For each MWE, the MWE lexicon contains the following information: 1. The ID of the expression, 2. The POS of the expression, 3. The lexicon ID of each constituent, 4. Agreement and other restrictions on the inflections, 5. Specification of the possible surface orders of the MWE constituents. This lexicon is now stored as part of the MILA lexicon of Hebrew, and is a natural extension of the previous lexicon that did not include MWEs. It is efficiently stored as an SQL database (although we depict it in its XML view here.) The MWE lexicon is used both by the generator (Section 8.2) and by the post-processor (Section 8.3).

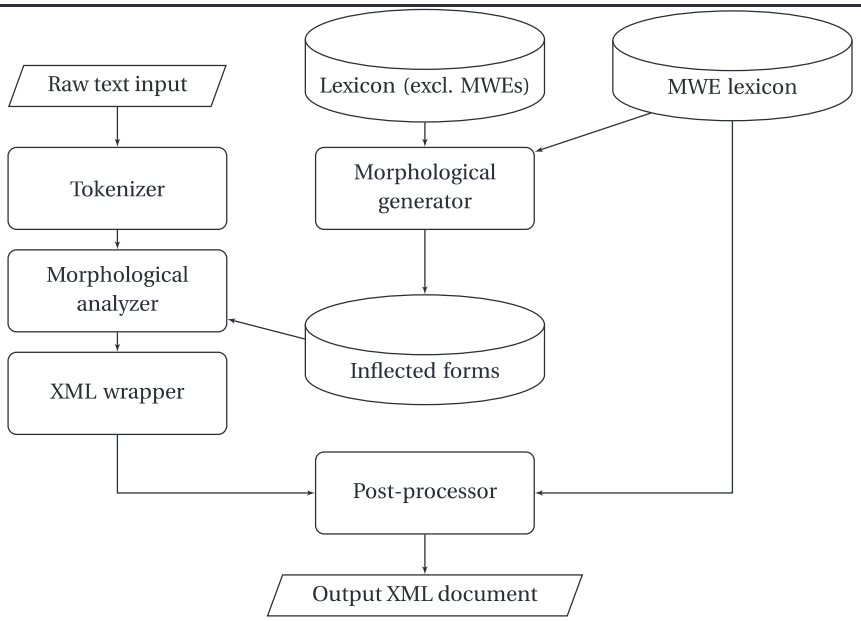

Figure 15: The architecture of the extended morphological system

8.2 Morphological generation

Recall that the generator (Section 5.2) produces all the inflected forms of each lexical entry. We would like to apply the same approach also to MWEs, and generate from each lexical entry of an MWE all its possible realizations, accounting for morphological inflections and order variations. However, it is impractical to associate this potentially large number of realizations with each constituent of the MWE.

The reason is that some words (e.g., frequent prepositions and some common nouns such as id "hand") participate in a large number of MWEs, causing such words to obtain a large number of analyses, one for each of the MWEs that contain the word as a constituent. For example, a large number of Hebrew MWEs begin in bit "house-of": bit spr ("house-of book") "school"), bit knst ("house-of gathering") "synagogue"), bit xwlim ("house-of sick-people") "hospital"), and more.

To overcome this problem, we choose for each MWE an anchor word. The anchor must be a free morpheme, not a prefix or a suffix, and it must appear in all the surface forms of the MWE. Under these constraints, the anchor can be chosen freely from among the constituents of the MWE, and it is automatically chosen to be the least frequent permissible constituent: of all the constituents that are free morphemes and that are not optional, we choose the one that is a constituent of the least number of MWEs. For example, consider again MWEs that begin in bit "house-of". If for all these expressions bit had been chosen as

```
<token id="2" surface="hdin">
  <analysis id="1">
    <base  lexiconPointer="5208" transliterated="din">
      <noun state="absolute" definiteness="true" gender="masculine"
            number="singular"/>
    </base>
  </analysis>
  <analysis id="2"/>
    <base lexiconPointer="28579" transliterated="din">
      <MWE lexiconPointer="28579" atom="2" definiteness="true"/>
    </base>
  </analysis>
</token>
```

Figure 16: The augmented morphological analysis of din "law"

the anchor, each occurrence of bit (even not in the context of a MWE) would have to receive additional analyses as anchors of these MWEs. Since the words spr, knst and xwlim are less frequent, they are automatically chosen as anchors rather than bit and no additional analyses are generated for bit.

We modify the generator such that when applied to the anchor, it generates not only all the inflected forms of that word, but also an additional analysis, as a constituent of the MWE that this word anchors. This additional analysis is associated with the ID of the MWE. For example, consider ywrk din "lawyer", whose anchor is din "law". When the word din (or any of its inflected forms, e.g., hdin "the + law") occurs in a text, it is assigned the analyses listed in Figure 16. The first analysis is standard; the second analysis identifies the word as a potential constituent of the MWE ywrk hdin "the lawyer", without considering its immediate context. The attribute `atom=''2''` indicates that if this word is indeed a constituent of the MWE, it is its second constituent.

No further change is introduced to the generator for other constituents of the MWE, except one minor issue: we must take care of inflections that can only appear as part of a MWE. For example, consider the MWE bit spr ("house-of book") "school"), which allows the colloquial form hbit spr ("the + house-of book") "the school"), where the definite article h is prepended to a noun in the construct state (see Section 7.1). This inflection does not occur in isolation. Since bit is not the anchor, the single word generator does not generate the form hbit. To enable the analysis of hbit spr, the MWE generator adds hbit to the database of inflected words. This is done based on the specification `hprefix=true` in the lexical entry of bit spr ("house-of book") "school"); see Section 6.4, in particular Figure 7.

8.3 Post-processing

The post processor works on a sentence-by-sentence basis, and has access to the output produced by the analyzer for an entire sentence. It checks the

analyses of the tokens to find anchors of MWEs (e.g., the second analysis of hdin in Figure 16). For each such anchor, the post-processor retrieves the entry of the corresponding MWE from the MWE lexicon (Section 8.1). This record contains the IDs of the remaining constituents, thereby enabling the post-processor to search for them in the sentence and verify that they satisfy the agreement and order requirements of the MWE (these are also stored in the MWE lexicon). Thus only one database search is needed for each anchor. For example, since spr "book" has an analysis as an anchor of bit spr "school", the post processor will access the database and find out that in order for spr to be part of this MWE, the lemma of the previous word must be bit. Note that had the frequent word bit been chosen as an anchor, each of its occurrences would have had additional analyses as anchors of MWEs. Since in each context, bit may participate in at most one MWE, much time and space would have been wasted to generate and check these superfluous analyses.

As an additional example, refer back to the lexical entry of the MWE ywrk din "lawyer" (Figure 7). Figure 17 depicts the (partial) output of the analyzer for the expression ywrkwt hdin "the (female) lawyers". In this example the first token, ywrkwt, has 4 analyses. The second token is hdin "the+ law"; it has two analyses, one as a single word and one as the anchor of the MWE, as in Figure 16.

From analysis 2 of hdin the post-processor realizes that it should search for MWE 28579 in the MWE lexicon. Finding the relevant record (Figure 7), it concludes that in order to construct this MWE the previous word has to have `lexiconPointer=8174` and `state=''construct''`. These requirements match those of the second analysis of ywrkwt (Figure 17). So the post-processor produces an analysis of these two tokens as a MWE. The MWE lexicon record of ywrk din "lawyer" also indicates that the number (plural) and gender (feminine) attributes should be inherited from the first constituent, while the definiteness (true) is inherited from the second one. Figure 18 depicts the result of post-processing as applied to the token hdin "the+ law" in the context of the entire expression (the analyses of the first token, ywrkwt "editors", are not modified, and are the same as in Figure 17).

At the end of the sentence the post-processor erases the analyses of tokens as anchors of MWE that were not part of a completed MWE. Thus only the analyses of single words and analyses of complete MWE remain. For example, is the token hdin "the+ law" had occurred without a preceding (inflected form of) ywrk "editor", analysis 2 of Figure 16 would be deleted.

8.4 Implementation

The modifications discussed in Section 6 were implemented as part of the MILA tools (Itai and Wintner, 2008), and are currently part of the lexicon and the morphological processor. The current MWE lexicon includes a total of

```
<token id="1" surface="ywrkwt">
  <analysis id="1">
    <base  lexiconPointer="8174" transliterated="ywrk">
      <noun state="absolute" definiteness="false" gender="feminine"
            number="plural"/>
    </base>
  </analysis>
  <analysis id="2">
    <base  lexiconPointer="8174" transliterated="ywrk">
      <noun state="construct" definiteness="false" gender="feminine"
            number="plural"/>
    </base>
  </analysis>
  . . .
 </token>
<token id="2" surface="hdin">
  <analysis id="1">
    <base  lexiconPointer="5208" transliterated="din">
      <noun state="absolute" definiteness="true" gender="masculine"
            number="singular"/>
    </base>
  </analysis>
  <analysis id="2"/>
    <base lexiconPointer="28579" transliterated="din">
      <MWE lexiconPointer="28579" atom="2" definiteness="true"/>
    </base>
  </analysis>
</token>
```

Figure 17: A partial analysis of ywrkwt hdin "the (female) lawyers" before post-processing

```
<token id="2" surface="hdin">
  <analysis id="1">
    <base  lexiconPointer="5208" transliterated="din">
      <noun state="absolute" definiteness="true" gender="masculine"
            number="singular"/>
    </base>
  </analysis>
  <analysis id="2">
    <base lexiconPointer="28579" transliterated="ywrk din">
      <MWE pos="noun" definiteness="true" number="plural"
           gender="feminine"/>
    </base>
  </analysis>
</token>
```

Figure 18: A partial analysis of ywrkwt hdin "the (female) lawyers" after post-processing

3270 MWEs; Table 2 shows their distribution by POS. We are actively adding new items, using both manual and automatic methods for identifying MWEs in corpora (Al-Haj and Wintner, 2010; Tsvetkov and Wintner, 2010, 2011, 2012, Forthcoming).

Table 2: Distribution of MWE in the lexicon by part of speech

POS	Noun	Adj	Prep	Adv	Interject	PropName	Other	Total
Count	1950	105	23	248	38	1215	139	3718

9 Conclusions

We proposed an architecture for lexical representation of MWEs in morphologically-complex languages, accompanied by a specification of the integration of MWEs into morphological processing. The architecture is used for representing Hebrew MWEs in an existing lexicon, along with their morphological and syntactic properties. Existing morphological processing tools for Hebrew have been extended, along the lines delineated here, to accommodate MWEs. While the proposal specifically addresses the special needs of Hebrew, it is in principle appropriate for a large array of morphologically interesting languages, which in most cases exhibit only a subset of the complexity that Hebrew poses.

Savary (2008, Section 4) lists several properties of MWEs that a lexical representation approach should account for. Many of them are indeed supported by our solution. These include facilities for representing exocentric MWEs (expressions with no head that determines their inflectional properties); irregular agreement; defective paradigms; insertions and omissions; order change; abbreviations; and non-contiguous MWEs. Furthermore, our solution addresses computational issues including inflectional analysis and generation and basic facilities for automated MWE lexicon creation.

Of course, not all possible MWEs can be represented using our scheme. For example, while we address the representation of MWEs with open slots (Section 6.8), we do not provide a way to constrain the syntactic structure of potential fillers of the open slot. We leave this to future work. More intricate interactions of MWEs with productive syntactic structures are also not addressed.

Our system identifies MWEs before syntactic analysis. This has several pros as well as some cons. On one hand, a high quality syntactic parser is not always available and is not necessary to recognize contiguous MWE. Identifying MWEs before syntactic analysis will help reduce morphological ambiguity. Also, a syntactic parser will not identity irregular syntactic constructions such as those described at the end of Section 3.3. On the other hand, a syntactic parser can address the problem described in the previous paragraph by restricting non-contiguous MWEs to syntactically legal constructions.

Our architecture is designed with Hebrew in mind; while Hebrew is challenging, other languages will exhibit other phenomena for which our solution will

not necessarily be optimal. As an example of such a phenomenon, note that our account of agreement among constituents (Section 6.6) is orthogonal to the specification of constituent order (Section 6.7). This is reasonable for most languages, but not for Arabic, in which the gender of the verb, in some cases, depends on whether the subject precedes or follows the verb. Our architecture will have to be extended in order to permit such specifications.

Having said that, we find our architecture adequate for Hebrew. In the future, we plan to focus on automatic acquisition of Hebrew MWEs from corpora, in order to improve the process of populating the lexicon. Several preliminary works in this direction have already been published (Al-Haj and Wintner, 2010; Tsvetkov and Wintner, 2010, 2011, 2012, Forthcoming).

Acknowledgements

This research was supported by THE ISRAEL SCIENCE FOUNDATION (grants No. 137/06, 1269/07). We are grateful to the anonymous IJL reviewers for very constructive comments that greatly improved this paper. All remaining errors are of course our own.

Notes

1 To facilitate readability we use a straight-forward transliteration of Hebrew using Roman characters; the letters used, in Hebrew lexicographic order, are abgdhwzxTiklmnsypcqrśt.

2 Hebrew examples are accompanied by word-by-word English glosses and, in the case of idiomatic expressions, also by the non-literal translation. Ungrammatical expressions are preceded by '*'; expressions whose meaning is literal rather than the expected idiomatic meaning are preceded by '#'. In glosses, we use '-' for features that are not manifested as independent morphemes and '+' for bound morphemes.

3 Nouns only inflect for gender when they denote entities with natural gender (Ordan and Wintner, 2005).

4 The glosses in this example indicate the gender of nouns and adjectives to better explain the internal structure of the NNCs.

5 The linguistic databases are stored in MySQL, but are automatically converted to Extensible Markup Language (XML, Connolly (1997)) according to schemas (van der Vlist, 2002) that enforce structure and are also used for documentation and validation purposes. We use the XML view throughout this paper.

6 The schema is available from http://mila.cs.technion.ac.il/resources_standards.html.

7 This organization is justified by Wintner (2007, 2008).

8 The format of entries in this database is designed to be machine-readable rather than human-readable. It is the XML wrapper that converts it to a more legible XML format.

9 Participles have properties of both verbs (hence `tense = ''participle''`) and nouns (hence `state = ''construct''`).

10 It is often possible to constrain the filler of such slots to a particular construction, e.g., a noun phrase. We currently do not support such constraints in the specification.

References

Al-Haj, H. and S. Wintner. 2010. 'Identifying Multi-word Expressions by Leveraging Morphological and Syntactic Idiosyncrasy'. In Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010). Beijing, China: Coling 2010 Organizing Committee, 10–18.

Alegria, I. n., O. Ansa, X. Artola, N. Ezeiza, K. Gojenola and R. Urizar. 2004. 'Representation and Treatment of Multiword Expressions in Basque'. In T. Tanaka, A. Villavicencio, F. Bond and A. Korhonen (eds), Second ACL Workshop on Multiword Expressions: Integrating Processing. Barcelona, Spain: Association for Computational Linguistics, 48–55.

Anastasiou, D., C. Hashimoto, P. Nakov and S. N. Kim (eds) 2009. Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications. Singapore: Association for Computational Linguistics.

Attia, M. A. 2006. 'Accommodating Multiword Expressions in an Arabic LFG Grammar'. In T. Salakoski, F. Ginter, S. Pyysalo and T. Pahikkala (eds), FinTAL. Berlin, Heidelberg and New York: Springer, volume 4139 of Lecture Notes in Computer Science, 87–98.

Baldwin, T. and S. N. Kim. 2010. 'Multiword Expressions'. In N. Indurkhya and F. J. Damerau (eds), Handbook of Natural Language Processing. Second edition Boca Raton, FL: CRC Press, Taylor and Francis Group. ISBN 978-1420085921.

Baldwin, T. and A. Villavicencio. 2002. 'Extracting the unextractable: a case study on verb-particles'. In COLING-02: proceeding of the 6th conference on Natural language learning. Morristown, NJ, USA: Association for Computational Linguistics, 1–7.

Bannard, C., T. Baldwin and A. Lascarides. 2003. 'A Statistical Approach to the Semantics of Verb-Particles'. In D. M. Francis Bond, and A. Villavicencio (eds), Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment. Association for Computational Linguistics, 65–72.

Baptista, J., A. Correia and G. Fernandes. 2004. 'Frozen Sentences of Portuguese: Formal Descriptions for NLP'. In T. Tanaka, A. Villavicencio, F. Bond and A. Korhonen (eds), Second ACL Workshop on Multiword Expressions: Integrating Processing. Barcelona, Spain: Association for Computational Linguistics, 72–79.

Berman, R. A. 1978. Modern Hebrew Structure. Tel Aviv: University Publishing Projects.

Carroll, J. and C. Grover. 1989. 'The derivation of a large computational lexicon for English from LDOCE'. In Computational lexicography for natural language processing. White Plains, NY, USA: Longman Publishing Group, 117–133.

Connolly, D. 1997. XML: Principles, Tools, and Techniques. O'Reilly.

Doron, E. 1983. Verbless Predicates in Hebrew. Ph.D. thesis, University of Texas at Austin.

Erman, B. and B. Warren. 2000. 'The idiom principle and the open choice principle'. Text, 20.1: 29–62.

Glinert, L. 1989. The Grammar of Modern Hebrew. Cambridge: Cambridge University Press.

Graliński, F., A. Savary, M. Czerepowicka and F. Makowiecki. 2010. 'Computational Lexicography of Multi-Word Units: How Efficient Can It Be?' In Proceedings of the

Workshop on Multiword Expressions: from Theory to Applications (MWE 2010). Association for Computational Linguistics, 1–9.

Grégoire, N. 2007. 'Design and Implementation of a Lexicon of Dutch Multiword Expressions'. In Proceedings of the Workshop on A Broader Perspective on Multiword Expressions. Prague, Czech Republic: Association for Computational Linguistics, 17–24.

Grégoire, N. 2010. 'DuELME: a Dutch electronic lexicon of multiword expressions'. Language Resources and Evaluation, 44.1-2: 23–39.

Grégoire, N., S. Evert and S. N. Kim (eds) 2007. Proceedings of the Workshop on A Broader Perspective on Multiword Expressions. Association for Computational Linguistics.

Grégoire, N., S. Evert and B. Krenn (eds) 2008. Proceedings of the LREC Workshop: Towards a Shared Task for Multiword Expressions (MWE 2008). European Language Resources Association.

Itai, A. and S. Wintner. 2008. 'Language resources for Hebrew'. Language Resources and Evaluation, 42.1: 75–98.

Itai, A., S. Wintner and S. Yona. 2006. 'A Computational Lexicon of Contemporary Hebrew'. In Proceedings of The fifth international conference on Language Resources and Evaluation (LREC-2006). Genoa, Italy.

Jackendoff, R. 1997. The Architecture of the Language Faculty. Cambridge, USA: MIT Press.

Kordoni, V., C. Ramisch and A. Villavicencio (eds) 2011. Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World. Portland, Oregon, USA: Association for Computational Linguistics.

Kordoni, V., C. Ramisch and A. Villavicencio (eds) 2013. Proceedings of the 9th Workshop on Multiword Expressions. Atlanta, Georgia, USA: Association for Computational Linguistics.

Krzysztof, J. 2004. 'Applying Oxford-PWN English-Polish Dictionary to Machine Translation'. In Proceedings of the 9th European Association for Machine Translation Workshop on Broadening horizons of machine translation and its applications.

Laporte, E., P. Nakov, C. Ramisch and A. Villavicencio (eds) 2010. Proceedings of the Workshop on Multiword Expressions: from Theory to Applications (MWE 2010). Beijing, China: Association for Computational Linguistics.

Melnik, N. 2006. 'A Constructional Approach to Verb-Initial Constructions In Modern Hebrew'. Cognitive Linguistics, 17.2: 153–198.

Oflazer, K., O. Çetinoğlu and B. Say. 2004. 'Integrating Morphology with Multi-word Expression Processing in Turkish'. In T. Tanaka, A. Villavicencio, F. Bond and A. Korhonen (eds), Second ACL Workshop on Multiword Expressions: Integrating Processing. Barcelona, Spain: Association for Computational Linguistics, 64–71.

Ordan, N. and S. Wintner. 2005. 'Representing natural gender in multilingual lexical databases'. International Journal of Lexicography, 18.3: 357–370.

Rayson, P., S. Piao, S. Sharoff, S. Evert and B. n. Moirón. 2010. 'Multiword expressions: hard going or plain sailing?' Language Resources and Evaluation, 44: 1–5.

Sag, I., T. Baldwin, F. Bond, A. Copestake and D. Flickinger. 2002. 'Multiword Expressions: A Pain in the Neck for NLP'. In Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2002). Mexico City, Mexico, 1–15.

Savary, A. 2008. 'Computational Inflection of Multi-Word Units, a contrastive study of lexical approaches'. Linguistic Issues in Language Technology, 1: 1–53.

Savary, A. 2009. 'Multiflex: A Multilingual Finite-State Tool for Multi-Word Units'. In S. Maneth (ed.), CIAA. Springer. volume 5642 of Lecture Notes in Computer Science, 237–240.

Shimron, J. (ed.) 2003. Language Processing and Acquisition in Languages of Semitic, Root-Based, Morphology. Number 28 in Language Acquisition and Language Disorders. John Benjamins.

Tsvetkov, Y. and S. Wintner. 2010. 'Extraction of Multi-word Expressions from Small Parallel Corpora'. In Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010). 1256–1264.

Tsvetkov, Y. and S. Wintner. 2011. 'Identification of Multi-word Expressions by Combining Multiple Linguistic Information Sources'. In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. Edinburgh, Scotland, UK: Association for Computational Linguistics, 836–845.

Tsvetkov, Y. and S. Wintner. 2012. 'Extraction of multi-word expressions from small parallel corpora'. Natural Language Engineering, 18.4: 549–573.

Tsvetkov, Y. and S. Wintner. Forthcoming. 'Identification of Multi-word Expressions by Combining Multiple Linguistic Information Sources'. Computational Linguistics.

van der Vlist, E. 2002. XML Schema. O'Reilly.

Villavicencio, A., F. Bond, A. Korhonen and D. McCarthy. 2005. 'Introduction to the special issue on multiword expressions: Having a crack at a hard nut'. Computer Speech & Language, 19.4: 365–377.

Villavicencio, A., A. Copestake, B. Waldron and F. Lambeau. 2004. 'Lexical Encoding of MWEs'. In T. Tanaka, A. Villavicencio, F. Bond and A. Korhonen (eds), Second ACL Workshop on Multiword Expressions: Integrating Processing. Barcelona, Spain: Association for Computational Linguistics, 80–87.

Wintner, S. 2004. 'Hebrew Computational Linguistics: Past and Future'. Artificial Intelligence Review, 21.2: 113–138.

Wintner, S. 2007. 'Finite-state Technology as a Programming Environment'. In A. Gelbukh (ed.), Proceedings of the Conference on Computational Linguistics and Intelligent Text Processing (CICLing-2007). Berlin and Heidelberg: Springer. volume 4394 of Lecture Notes in Computer Science, 97–106.

Wintner, S. 2008. 'Strengths and weaknesses of finite-state technology: A case study in morphological grammar development'. Natural Language Engineering, 14.4: 457–469.

Wintner, S. and S. Yona. 2003. 'Resources for processing Hebrew'. In Proceedings of the MT-Summit IX workshop on Machine Translation for Semitic Languages New Orleans, 53–60.

Yona, S. and S. Wintner. 2005. 'A Finite-State Morphological Grammar of Hebrew'. In Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages. Ann Arbor, Michigan: Association for Computational Linguistics, 9–16.

Yona, S. and S. Wintner. 2008. 'A Finite-State Morphological Grammar of Hebrew'. Natural Language Engineering, 14.2: 173–190.

Zdaqa, Y. 1974. Luxot HaPoal (The Verb Tables). Jerusalem: Kiryath Sepher. In Hebrew.

A. The MILA lexicon

Figure 19 lists the complete lexical entry of the noun htnsˇawt "arrogance/soaring". Being a deverbal noun, it is specified for attributed of both nouns and verbs. It also has two distinct senses, each with two English translations.

```
<item dotted="הִתְנַשְּׂאוּת" id="I7" register="formal"
spelling="standard" transliterated="htneawt" undotted="התנשאות">
  <noun acronym="false" definiteness="true" deverbal="true"
  direction="false" dottedPlural="" dual="false"
  feminine="unspecified" foreign="false" gender="feminine"
  inflectConstructP="true" inflectConstructS="true"
  inflectPossessiveP="true" inflectPossessiveS="true"
  inflectionBase="" inflectionPattern="" ipSource=""
  lexicalLink="2427" number="singular" pattern="" plural="wt"
  root="" />
  <sense definition="" id="3905" weight="1">
    <english id="5623" te="raising" weight="1" />
    <english id="5624" te="soaring" weight="1" />
  </sense>
  <sense definition="" id="3906" weight="1">
    <english id="5625" te="arrogance" weight="1" />
    <english id="5626" te="haughtiness" weight="1" />
  </sense>
</item>
```

Figure 19: The complete lexical entry of the noun htnsˇawt "arrogance"

B. XML representation of morphologically-analyzed texts

Figure 20 depicts a fragment of a small morphologically-analyzed corpus, represented in XML.

```
<corpus comment="versions info: lexicon :30/11/2010;  morphologicalAnalyzer:1.6
         (30/11/2010); corpus schema 16/06/2009; lexicon schema 16/06/2009" >
<article id="1">
  <paragraph id="1">
    <sentence id="1">
      <token id="1" surface="šbi">
        <analysis id="1">
          <base lexiconPointer="1541" transliterated="šbh">
            <verb gender="feminine" number="singular" person="2"
                  tense="imperative"/>
          </base>
        </analysis>
        <analysis id="2">
          <base lexiconPointer="1636" transliterated="išb">
            <verb gender="feminine" number="singular" person="2"
                  tense="imperative"/>
          </base>
        </analysis>
...
        <analysis id="5">
          <base lexiconPointer="7863" transliterated="šbi">
            <noun definiteness="false" gender="masculine"
                  number="singular"
                  state="absolute"/>
          </base>
        </analysis>
        <analysis id="6">
          <prefix function="relativizer/subordinatingConjunction" id="1"
                  surface="š"/>
          <base lexiconPointer="26553" transliterated="b">
            <preposition />
          </base>
          <suffix function="pronominal" gender="masculine and feminine"
                  number="singular" person="1"/>
        </analysis>
      </token>
    </sentence>
  </paragraph>
</article>
</corpus>
```

Figure 20: XML representation of morphologically-analyzed text