Hebrew Multiword Expressions: Linguistic Properties, Lexical Representation, Morphological Processing, and Automatic Acquisition

Hassan Al-Haj

THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE MASTER DEGREE

University of Haifa
Faculty of Social Sciences
Department of Computer Science

December, 2009

Hebrew Multiword Expressions: Linguistic Properties, Lexical Representation, Morphological Processing, and Automatic Acquisition

By: Hassan Al-Haj

Supervised By: Dr. Shuly Wintner

THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE MASTER DEGREE

University of Haifa
Faculty of Social Sciences
Department of Computer Science

December, 2009

Approved by: ________________________________
Date:________
(supervisor)

Approved by: ________________________________
Date:________
(Chairman of M.A Committee)

Contents

Hebrew Multiword Expressions: Linguistic Properties, Lexical Representation, Morphological Processing, and Automatic Acquisition

Hassan Al-Haj

Abstract

Multiword Expressions (MWEs) are lexical items consisting of more than a single orthographic word. MWEs constitute a major part of any language. The number of MWEs in a speakers' lexicon (in English) is of the same order of magnitude as the number of single words. Identification of MWEs in running text is important for a variety of natural language processing applications such as information retrieval, machine translation, question-answering, word sense disambiguation, and text summarization.

In this work we investigate the properties of Hebrew MWEs, and classify these properties along three dimensions: morphological, syntactic, and semantic. Based on our linguistic investigation, we describe an architecture for lexical representation of MWEs in an existing large-scale lexicon of Hebrew. We also provide a specification of the integration of MWEs into a morphological processor of Hebrew. Then, we describe a system that extracts noun compounds from Hebrew raw text based on their idiosyncratic morphological and syntactic properties. The raw text is first morphologically analyzed and disambiguated. Then, all noun-noun constructs are extracted from the morphologically disambiguated text. For each candidate noun compound we define a set of features based on the idiosyncratic morphological and syntactic properties of noun compounds that we identified. These features are used to train a support vector machine classifier to identify the noun compounds in the list of noun-noun constructs. We show that combining linguistically-informed features with a collocation measure results in a classification accuracy of over 80%, reflecting a reduction of 36.16% in the classification error rate compared with the best collocation measure baseline classifier. The extracted nominal compounds are used to extend the exsiting Hebrew lexicon.

List of Figures

1 Introduction

1.1 Multiword Expressions

Multiword Expressions (MWEs) are lexical items consisting of more than a single orthographic word. Sag et al. (2002) define MWEs as "idiosyncratic interpretations that cross word boundaries (or spaces)". MWEs are expressions whose linguistic properties (morphological, syntactic or semantic) are not necessarily derived from the properties of the individual words that compose them.

The term MWE refers to a heterogeneous class of phenomena with diverse sets of characteristics. Semantically, the compositionality of MWEs (i.e., the degree to which the meaning of the whole expression results from combining the meanings of its individual words when they occur in isolation) is gradual. As Bannard, Baldwin, and Lascarides (2003) have noted, MWEs "do not fall cleanly into the binary classes of compositional and non-compositional expressions, but populate a continuum between the two extremes." Specifically, some MWEs have a compositional and transparent meaning (e.g., the Dutch expression *witte wijn* "white wine" (Grégoire, 2007)), while others have a non-compositional and opaque meaning (e.g., the Turkish expression *ipe sapa gelmez* (lit. "(he) does not come to rope and handle") "worthless" (Oflazer, Çetinoğlu, and Say, 2004)). MWEs can also fall between these two extremes (e.g., the English Verb-particle construction *clean up*). Syntactically, some MWEs appear in one rigid pattern, where the constituents have a fixed order, while others can undergo various syntactic variations, including permutation of the order of the constituents (e.g., the English expression *spill the beans* "reveal the secrets" can undergo different types of syntactic variations and modifications as in *The beans were spilled in the last press conference*). Moreover, MWEs can have idiosyncratic and irregular syntactic patterns (e.g., the English expression *by and large* which conjoins a preposition with an adjective).

Grammatically, MWEs may function as words (e.g., the Spanish expression *de golpe* (lit. "like a blow") "suddenly", an adverb (Dowdle, 1967)) or as phrases (e.g., the Dutch expression *bok schieten* (lit. "to shoot a male-goat") "to make a blunder", a verb phrase (Grégoire, 2007)). Morphologically, MWEs are not homogeneous, allowing some constituents to freely inflect while restricting (or even preventing) the inflections of others. As an example, consider the Turkish expression *kafayl ye-* (lit. "stand (in) respect") "to become mentally deranged." The first part of the expression, the accusative marked noun *kafayl*, is fixed, while the part starting with the verb *ye-* may be inflected and/or derived in various ways: *kafayl tedim* "I became mentally deranged", *kafayl yiyeceklerdi* "they were about to become mentally deranged", etc. (Oflazer, Çetinoğlu, and Say, 2004).

In some cases MWEs may even allow constituents to undergo non-standard morphological inflections that they would not undergo in isolation, or contain words that have no part of speech or a literal meaning. For example, the word *noizik* in Basque is an archaism of modern *noiztik* "from when", which occurs just in expressions such as *noizik behin*, *noizik noizera* and others, all meaning "once in a while", but does not occur in isolation in other contexts, and it is difficult to assign it a part of speech (Alegria et al., 2004). Orthographically, the components of MWEs can occur in the text either contiguously with intervening spaces or dispersed.

MWEs constitute a major part of any language. As Jackendoff (1997, page 156) notes, the magnitude of this phenomenon is far greater than has traditionally been realized within linguistics. He estimates that the number of MWEs in a speakers' lexicon (in English) is of the same order of magnitude as the number of single words. Sag et al. (2002) note that for a wide coverage natural language processing (NLP) systems, this is almost certainly an underestimate. In WordNet 1.7 (Fellbaum, 1998), for example, 41% of the enteries are multiwords. In an empirical study, Erman and Warren (2000) revealed that over 55% of the texts they studied were instances of what they call **prefabs** (defined somewhat more broadly than the MWEs we consider here.)[1]

The identification of MWEs is relevant for a variety of NLP applications such as information retrieval, machine translation, question-answering, word sense disambiguation, and text summarization. MWEs must be correctly processed for such NLP applications to perform accurately. Moreover, expressions with idiosyncratic features that cannot be predicted on the basis of their component words must be included in language descriptions (such as lexicons) in order to account for actual usage.

The morphological, syntactic and semantic idiosyncratic properties make MWEs a challenge for computational processing of natural languages. 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, Çetinoğlu, and Say, 2004; Alegria et al., 2004). Hebrew poses additional challenges for representing, processing and extracting MWEs due to its complex morphology and problematic orthography. We discuss some of these features and challenges below.

1.2 Hebrew morphology and orthography

Hebrew,[2] like other Semitic languages, has a rich and complex morphology. The major word formation machinery is *root* and *pattern.* The root is a sequences of three (typically) or more consonants, called *radicals.* The pattern is a combination of vowels and, possibly, consonants too, with slots into which the root consonants can be inserted. Words (lexemes) are created by interdigitating roots into patterns: the first radical is inserted into the first consonantal slot of the pattern, the second radical fills the second slot and the third fills the last slot. Consider the Hebrew root š.m.r and the patterns `_w__ and __i_h.` When the root combines with these patterns the resulting lexemes are (the noun) *šwmr* "guard", and (the noun) *šmirh* "guarding", respectively. See Shim-

[1] prefabs are as "at least two words favored by native speakers in preference to an alternative combination which could have been eqivalent had there been no conventionalization."

[2] To facilitate readability we use a straight-forward transliteration of Hebrew using Roman characters; the letters used, in Hebrew lexicographic order, are *abgdhwzxTiklmnsypcqršt*.

ron (2003) for a survey. After the root combines with the pattern to create a *lexeme*, morpho-phonological alterations which are sometimes non-trivial may take place.

Lexemes can be inflected in various ways. Inflectional morphology is highly productive and consists mostly of suffixes, but sometimes of prefixes or circumfixes. Nominals, namely nouns, adjectives and numerals, inflect for number (singular, plural and, in rare cases, also dual, e.g., *šwmr+im*→*šwmrim* "guards"), gender (masculine or feminine as in *šwmr+t*→*šwmrt* "a feminine guard") or for both gender and number (e.g, *šwmr+wt*→*šwmrwt* "feminine guards"). In addition, nominals have three phonologically (and orthographically) distinct forms: the definite form, which is indicated by the prefix *h*; the absolute form; and the construct form, which is typically used in genitive (possessive) constructions. For example, *xwlch* (absolute) "shirt" vs. *hxwlch* "the shirt" vs. *xwlct* (construct) "shirt of". Furthermore, nominals take pronominal suffixes which are interpreted as possessives. These 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 and aspect, which is traditionally analyzed as having the values *past, present, future, imperative* and *infinite* (e.g., the verb *akl* "eat" can occur in the form *akl+nw*→*aklnw* (first person plural masculine past) or in the form *t+akl+w*→*taklw* (second person plural masculine future), etc.

Prepositions can combine with pronominal affixes that are interpreted as the object of the preposition. These inflect for number, gender, and person (e.g., the preposition *lid* "beside" combines with a suffix as in *lid+w*→*lidw* "beside him", *lid+h*→*lidh* "beside her", and *lid+nw*→*lidnw* "beside us").

In addition to the morphological richness, the orthography of Hebrew poses several problems for computational processing. As is well known, in the standard script most vowels are not explicit. Furthermore, many particles, including 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. When a definite nominal is prefixed by one of the prepositions *b*, *k* or *l*, the definite article *h* is assimilated with the preposition and the resulting written form becomes ambiguous with respect to definiteness. Thus, a single form such as *šbth* can be read as an inflected lexeme (the verb "capture", third person singular feminine past), *š+bth* "that+field", *šbt+h* "her sitting", *š+bt+h* "that her daughter", *š+b+th* "that+in+tea", or *š+b+h+th* "that+in+the+tea". The rules that govern the combination of Hebrew prefix particles with the words they attach to are basically syntactic: They are constrained by the syntactic category of the word they attach to. For example, the preposition *m* can attach to nouns (including other syntactic categories) as in *m+bit*→*mbit* "from a house", but it cannot attach to adverbs. The same rules govern the combination of Hebrew prefix particles with the MWEs they attach to, but these combinations are constrained by the syntactic category of the whole expression. For example, the particle *m* can attach to the first word of *gn xiwt* (lit. “animals garden”) “zoo” (which is a noun) as in *m+gn xiwt*→*mgn xiwt* “from a zoo”, but cannot attach to the expression *ph axd* (lit. “mouth one”) “unanimously” (which is an adverb).

An added complexity of Hebrew orthography stems from the fact that Hebrew can be written in two ways: one in which vocalization diacritics, known as *niqqud* “diacritics”, decorate the words, and the undotted script, in which the diacritics are missing, and other characters represent some, but not all of the vowels. Most of the texts in Hebrew are of the latter kind. While the Academy for the Hebrew Language has issued guidelines for transcribing undotted texts (Gadish, 2001), they are observed only partially. Thus, the same word can be written in more than one way, sometimes even within the same document. This fact adds significantly to the degree of ambiguity.

1.3 Research goals

In this work we investigate Hebrew MWEs and their orthographic, morphological, syntactic, and semantic properties. We develop methods to represent MWEs in an existing large-scale lexicon of Hebrew (Itai and Wintner, 2008), and integrate MWEs in the existing morphological processor of Hebrew (Yona and Wintner, 2007). We describe a system that extracts nominal compounds from Hebrew raw text using their idiosyncratic morphological and syntactic properties, thereby extending the lexicon with nominal compounds.

The contribution of this work is manifold. From a theoretical point of view, this is the first systematic investigation of the properties of MWEs in Hebrew.[3] Our research provides linguists with a full picture of the diverse sets of characteristics that MWEs exhibit, enabling the classification of MWEs according to their behavior. Practically, we develop a set of tools for processing MWEs. The extended lexicon, combined with the extended morphological processors, enable much more accurate processing of Hebrew texts than the state of the art, correctly identifying several entities which were until now analyzed erroneously, if at all. The MWE acquisition module facilitates automatic expansion of the lexicon and guarantees the longevity of the developed resources, as many MWEs are constantly added to the language.

The remainder of this thesis is structured as follows: in Section 2 we discuss related work. In Section 3 we investigate the morphological, syntactic, and semantic properties of Hebrew MWEs. Section 4 describes methods for representing MWEs in the lexicon and integrating them in the morphological processor. Section 5 describes a system for automatic extraction of Hebrew nominal compounds from raw text. Finally, conclusion and future works are described in Section 6.

2 Related work

There has been a growing awareness in the NLP community of the problems that MWEs pose, both in linguistics and in NLP (Villavicencio et al., 2005).

[3] See a detailed discussion in Section 2.4.

Several projects have focused on various topics of MWEs for different languages.

2.1 Linguistic analysis of MWEs

One research topic is the linguistic analysis of MWEs. This line of work investigates MWEs from a theoretical perspective, defining appropriate linguistic descriptions for these expressions. Sag et al. (2002) investigate English MWEs and divide them into two classes: **lexicalized phrases** and **institutionalized phrases**. Lexicalized phrases have at least partially idiosyncratic syntax or semantics, or contain 'words' which do not occur in isolation. Institutionalized phrases are MWEs that are syntactically and semantically compositional, but occur with markedly high frequency. This classification and terminology were adopted in different works on other languages, including Dutch (Grégoire, 2007) and Turkish (Oflazer, Çetinoğlu, and Say, 2004). The latter work also describes a system for morphological processing of MWEs. 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 into the following stage.

2.2 Lexical and morphological processing

A related line of work is the construction of lexical resources and ontologies for MWEs. Different strategies for encoding MWEs in lexical resources have been employed for different languages, with varying degrees of success, depending on the type of MWE. Some of these works concentrate on encoding a certain class of MWEs. One example of such works is the construction of 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. Semantically, such sentences are noncompositional, and the whole expression is taken as a complex multiword lexical unit (Baptista, Correia, and Fernandes, 2004). This work classifies the frozen expressions into formal classes, and uses matrices (one per each class) to encode the lexical and syntactic properties of each frozen expression. To identify and tag the frozen sentences from a certain class in the texts, a finite-state transducer is built for each formal class. This graph describes the formal sequences of the components of the frozen sentences using variables that refer to the properties encoded in the class matrix. 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 (variation in word order, subcategorisation, etc.), 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 a 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 realised in different ways (which creates different instances of the same expression, as in *find/touch a nerve*), all possible realisations 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). When appropriate, a non-idiomatic paraphrase is kept for the idiomatic element which is used to generate a non-idiomatic paraphrase for the whole MWE.

However, this method fails to capture many syntactic and lexical variations expressed by various types of MWEs (for examples in Hebrew see Section 3). An example of such variation is morphological inflections. In many cases a word component of a MWE undergoes a (strict) subset of the full inflections that the word would undergo in isolation (we call this **partial inflection**). Villavicencio et al. (2004) provide just one of two options for each component, to inflect or not to inflect, failing to address partial inflections. One syntactic variation that this method does not account for is when MWEs take an argument between the word components. Also, it is important to note that this method does not account for changes in the orthographic and morphological properties of word components that occur as a result of changes in their position within the same expression.

Our approach is to design a general representation scheme for MWEs which can account for any combination of syntactic and lexical variation exhibited by the many various types of MWEs that we invistigated. In our method (Section 4), each MWE has its own entry in the lexicon, written in XML, where all its properties are encoded. These include morphological properties such as inflections (whether full, partial, or non-standard) that the word components can undergo, and the morphological features of the whole expression. They also include orthographic properties such as the orthography of the word components, including all their different realizations, and the part of speech of both the components and the expression as a whole. Each entry also encodes syntactic properties such as the permutations (all possible orders) that the components could appear in, including the morphology and the orthography of the constituents in each of these permutations. Note, however, that unlike Villavicencio et al. (2004) we do not account for the semantics of the MWEs.

2.3 Automatic identification and acquisition

The last research topic that we address here is the identification and extraction of MWEs from written corpora. Many works in different languages have focused on the acquisition of various types of MWEs. Some of the classical and earliest approaches for automatically extracting MWEs concentrated on their collocational behavior (Church and Hanks (1989b); Church and Mercer (1993); Dunning (1993); etc.) A great deal of work has been done on the extraction of collocations in the last decade and a half employing various association measures. These measures include *T-score*, *pointwise mutual information* (PMI), *log likelihood*, χ^2, etc. (see Pecina (2005) for a survey). Pecina (2008) compare 55 different association measures in ranking German Adj-N and PP-Verb collocation candidates. This work shows that combining different collocation measures using standard statistical-classification methods (such as Linear Logistic Regression and Neural Networks) gives a significant improvement over using a single collocation measure. Other works show that using linguistic information with collocation measures can improve the performance. Ramisch et al. (2008) evaluate a number of association measures on the task of identification of English Verb-Particle Constructions and German adjective-noun pairs. They show that adding linguistic information to the association measure by counting the number of times the expression appears in different linguistic patterns gives significant improvement in performance over using pure frequency.

Hashimoto and Kawahara (2008) perform a large-scale study of token classification into idiomatic versus literal for all types of Japanese MWEs. They annotate a web based corpus for training data. They identify 101 idiom types in the training data, and experiment with 90 idiom types for which they had more than 50 examples. They define two types of features and use support vector machines to classify idiomatic and literal expressions. The first type of features have been commonly used in word sense disambiguation (WSD). These include POS, lemma information, token and n-gram features. The second type of features have been designed for Japanese idiom identification and include mostly inflectional features such as voice, negativity, modality, in addition to adjacency and adnominal features. Hashimoto and Kawahara (2008) report an overall improvement of 16.27% in classification accuracy (over the baseline) using all the features.

Other works concentrated on the **lexical fixedness** property of (certain types of) MWEs in order to extract them from texts. An expression is considered lexically fixed if the replacement of any of its constituents by a semantically (and syntactically) similar word generally results in an invalid or literal expression. One example of such work is the extraction of Dutch verb+noun idiomatic combinations (VNICs) done by Van de Cruys and Villada Moirón (2007). The main idea in this work is to try to substitute the noun within a verb+noun composition (VNC) with other nouns that have similar or close meaning (taken from clusters of semantically related nouns which are automatically extracted from the corpus using distributional similarity measures). For each of these nouns the preference of the verb for it is measured using a number of statistical measures inspired by Resnik (1993). The VNC candidate is considered idiomatic only if the verb significantly prefers the original noun over the other nouns in the cluster.

Some works use the **syntactic fixedness** of MWEs in order to distinguish them from superficially similar literal combinations. Syntactically fixed expressions are expressions that prohibit (or restrict) syntactic variation. One example of such work is the identification of (English) VNICs in corpora (Bannard, 2007). Three kinds of (non-morphological) variation that VNCs can undergo are identified: addition or deletion of a determiner; internal modification such as adding an adjective to the noun; and passivization. Each kind of these variations is associated with a single component word (either the noun or the verb). Bannard (2007) estimates the extent to which the probability that the whole VNC undergoes a variation deviates from the variation probability one would expect for the component associated with it (using a measure called the **conditional pointwise mutual information** (CPMI)). The total syntactic flexibility for a VNC (denoted by **SynVar**) is taken as the sum of the CPMIs of all the kinds of variations. The VNC candidates are first ranked according to their SynVar (the less flexible at the top). Then, this ranking is evaluated using a list of idioms taken from published dictionaries, by observing how many of the gold standard items are found in each top n (for different values of n), and calculating the accuracy score. Another work uses both the syntactic and the lexical fixedness of VNICs in order to distinguish them from non-idiomatic ones, and eventually to extract them from corpora (Fazly and Stevenson, 2006).

In this work we describe a system that extracts nominal compounds (defined in Section 3.5) from raw Hebrew text. The text is first morphologically analyzed and disambiguated. Then, all noun compounds are extracted from the morphologically disambiguated text. We exploit the rich morphology of the Hebrew language, and define a set of idiosyncratic morphological properties for nominal compounds. This set of idiosyncratic morphological properties, combined with a set of idiosyncratic syntactic properties, are fed to a support vector machine classifier which is then used to identify the nominal compounds in the list of noun compounds.

2.4 MWEs in Semitic languages

Little work has been done on MWEs in Hebrew and other semitic languages compared to amount of work done on Indo-European languages. In this section we discuss some of the works done on MWEs in Hebrew and semitic languages, which are related to our work. One linguistic Hebrew MWE work was done by Berman and ravid (1986). In their work, they investigate the "dictionay degree" of noun compounds in Hebrew, which measures the extent of their closeness to a single word from a grammatical point of view, and from the manner they are grasped by the language speaker. A group of 120 Hebrew speakers were asked to assign a dictionary degree (from 1 to 5) to a list of 30 noun compounds. An analysis of the questionnaire results revealed that language speaker share a common dictionary, where the highest degree of agreement was achieved on the ends of the dictionary degree spectrum. Another conclusion is that both the pargamtic uses of the noun compound and the semantic relation between its constituents define the dictionary degree of the compound.

Another more recent linguistic Hebrew MWE work was done by Paltiely and Ephrat (2006). In their work they investigate the semantic, lexical, and syntactic properties of idioms in Hebrew. Two categories of idiom properties are identified. The first category is **lexical fixedness**, which includes properties that stem from the non-compositional semantics of idioms (e.g., "lexical fixedness of the constituents"). The second category is **syntactic fixedness** and includes syntactic properties that characterise idioms (e.g., "Is there a change in the order of the constituents"). 19 Hebrew "accepted idioms", i.e., idioms that appear in two different Hebrew lexicons, three dictionaries, and at least four out of six idiom resources are chosen. For each idiom, they investigate which properties hold for it. From amongst these properties six are chosen that hold for at least 85% of the given idioms. Based on the choosen properties they formulate an operative definition for Hebrew idioms, and provide a flowchart for the identification of idioms.

Attia (2005) proposes methods to process fixed, semi-fixed, and syntactically-flexible *Arabic* MWEs (he adopts the classification of MWEs and the terminology 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 resultant MWE transducer is used to complement the existing (single word) morphological transducer. Processing of syntactically-flexible expressions is done by the syntactic parser through the use of lexical rules accommodated in LFG. This work neither investigates the properties of Arabic MWEs, nor addresses the issue of representing MWEs in the lexicon.

Fabri (2007) provides an overview of the different types of compounds (14 in total) in present-day Maltese, focusing on one type of compounds consisting of an adjective followed by a noun (**A+N**). He also provides morphological, syntactic, and semantic properties of this group which distinguishes them from other non-compound constructions. This work is purely descriptive and non-computational.

As far as we know, no work has yet been done on computational treatment of Hebrew MWEs.

3 Linguistic properties of Hebrew MWEs

In this Section we thoroughly investigate the properties of Hebrew MWEs, and classify these properties along three dimensions: morphological, syntactic, and semantic. Below, we define the properties in each of these categories, describe the values these properties can have, and provide examples of Hebrew MWEs for each case. To the best of our knowledge, this is the first exhaustive investigation of MWEs in Hebrew (or any other Semitic language). Later on in Section 3.5 we focus on the linguistic properties of noun-noun construct MWEs.

3.1 Morphological properties:

MWE constituents may exhibit idiosyncratic morphological behavior which differs from the their behavior in isolation. This is manifested in the following manners:

Frozen form: Constituents can appear in one fixed (frozen) form. This form can be their citation (canonical) form like the word *id* in *ain lw id bdbr* (lit. “does not have a hand in the thing”) “is uninvolved”, or the word *kptwr* in the expression *kptwr wprx* (lit. “a button and a flower”) “fantastic”. It could also be a frozen inflected form, like the word *hxlwnwt* (the plural and definite form of *xlwn* (lit. window) in *hxlwnwt hgbwhim* (lit. “The high windows”) “upper echelon”.

Partial inflection: In some cases, constituents undergo a (strict) subset of the full inflections that they would undergo in isolation. For example, the verb *npl* (lit. fall) in the expression *npl ‘l hraš* (lit. “(he) fell on his head”) “(he) lost his mind” can inflect for number, gender, person, and all tenses except imperative. So, this expression can appear in the forms *nplnw ‘l hraš* “(we) lost our mind”, or as *nplh ‘l hraš* “(she) lost her mind”, but not in the imperative form *pwl ‘l hraš* (lit.“fall on your head”). Another example of partial inflection is the expression *hlk axri lbw* (lit. “walk after his heart”) “to follow one’s heart”: the third constituent *lb* “heart” can take possessive suffixes as in *hlkw axri lbm* “they followed their heart”, but it does not inflect for number. So, it does not appear in the form **hlkw axri lbbwtihm* “they followed their hearts”.

Non-standard inflection: Constituents can also undergo non-standard morphological inflections that they would not undergo in isolation. For example, consider the word *iwšb* (lit. sitting) in the expression *iwšb raš* (lit. “sitting head”) “chairman”. This expression has a (colloquial) definite form, *hiwšbi raš* “the chairmen”, whereby the first word *hiwšbi* is the definite construct-state (plural form) of the word *iwšb*, which is a ungrammatical outside of MWEs. Another example is the expression *bdltiim sgwrwt* (lit. “in close two doors”) “behind closed doors”. The first constituent in this expression, *bdltiim*, consists of the prefix *b* “in” followed by the dual form of *dlt* “door”, which is a form that this word does not appear in outside the MWE.

Hapax legomena: Other examples of idiosyncratic morphological behavior are constituents that have no other usage or literal meaning outside the expression they appear in. For example, the expression *kmTxwwi kšt* "a stone's throw". The first word *mTxwwi* (the prefix *k* is the preposition "as"), by itself, has no literal meaning in modern Hebrew. Another example is *abd 'liw qlx* "outdated". The third word, *qlx*, has no literal meaning or part of speech outside the MWE. This situation can also occur in expressions borrowed from other languages, like the expression *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.

Violated agreement: There are MWEs in which constituents are supposed to agree on morphological properties such as number, gender, person, status, or definiteness, but they do not. For example, in the expression *'in hr'* (lit. "eye the evil") "evil eye", the noun *'in* and the adjective *hr'* should agree on number, gender, and definiteness. However, in this case they do not agree on both gender and definiteness. While *'in* is feminine and absolute, the word *hr'* is masculine and definite.

It is important to note that different constituents of the same MWE can exhibit different kinds of morphological idiosyncrasy. For example, in the expression *npl 'l hraš*, the first constituent *npl* undergoes partial inflect2ion , while the other constituents are fixed. Another example is the expression *iwšb raš*, the first constituent *iwšb* undergoes both partial inflection and non-standard inflection, and the second constituent *raš* undergoes partial inflection.

3.2 Syntactic properties

We discuss below some syntactic properties of MWEs, confronting them with compositional phrases:

Variety: MWEs can belong to various part of speech categories (as a whole):

- **verb phrase**: *hlk mxil al xil* (lit. "go from army to army") "be very successful", *hxziq awtw qcr* (lit. "held him short") "keep on a short leash", *'md lw 'l hraš* (lit. "stand on someone's head") "bothered someone".
- **noun**: *bit spr* (lit. "house of book") "school", *sprwt iph* (lit. "beautiful literature") "belles-lettres", *ab bit din* (lit. "father of house of law") "President of the Court".
- **adjective**: *išr lb* (lit. "straight heart") "honest", *b'l š'wr qwmh* (lit. "owns a measure of height") "honorable", *kl d't* (lit. "light knowledge") "hasty".
- **adverb**: *bid xzqh* (lit. "in hand strong") "forcefully", *xd wxlq* (lit. "sharp and smooth") "straightforwardly", *bsbr pnim ipwt* (lit. "in expression face beautiful") "kindly".

- **conjunction**: *kmw kn* (lit. “like so”) “also”, *ala am kn* (lit. “but if yes”) “unless” , *ašr ’l qn* (lit. “that on thus”) “therefore”, *ap ‘l pi š*— (lit. “none on of — ”) “even though”, *itr ’l kn* (lit. “more on thus”) “moreover”, *am kn* (lit. “if thus”) “therefore”.
- **preposition**: *al ’br* (lit. “to direction”) “towards”, *’l awdwt* (lit. “on concerning”) “about”, *’l bsis* (lit. “on base”) “based on”, *’l drk* (lit. “on way”) “by way of”.

Compositionality: Some MWEs contain *open slots*, which can be filled with complements of certain parts of speech. For example, consider the expression *akl at — bli mlx* (lit. “eat — without salt”) “defeated — easily”. The open slot in this expression must be filled by a noun phrase, as in *akl at iribiw hxzqim bli mlx* (lit. “eat his powerful oponents without salt”) “defeated his powerful oponents easily”. Another example is the expression *prq at — mnšqw* (lit. “disassemble — from his weapon”) “disarm —”, in which the open slot can be filled with a noun phrases as in *bT-mn prq at hgwqr mnšqw* (lit. “batman disassembled the joker from his weapon”) “batman disarmed the joker”. These open slots are constrained with selectional restrictions imposed by other parts of the expression. One such restriction is the meaning of another part (or other parts) of the expression. While it is the literal meaning in the case of compositional expressions, it is the idiomatic meaning in the case of MWEs. For example, the open slot in the expression *akl at — bli mlx* cannot be filled by a type of food as in **akl at hcips bli mlx* (lit. “eat the chips without salt”) “defeated the chips easily”.

Constituent order: The order of the constituents in most MWEs tends to be fixed, but still in some cases a limited kind of change in order could occur. We chose a number of syntactic structures that, when used compositionally, permit a change in the order of constituents while preserving the meaning of the whole expression. We collected examples of MWEs that follow each of these syntactic structures, and checked if these MWEs also allow the same change. Below are the patterns we chose, including examples of some of the MWEs that we investigated.

- The order of verb complements is relatively flexible in Hebrew. Thus, a sentence like *hwa ica mbito al h’bwdh* “he left home for work”, which follows the pattern subject+verb+object, can be rephrased as *mbitw hwa ich al h’bwdh* “from home he left for work”. However, MWEs that follow the same pattern tend not to undergo his change in order. For example, in *hwa ica mhqlim axri hpgišh* (lit. “he left from the tools after the meeting”) “he got very angry after the meeting”, does not appear in the form **mhqlim hwa ica axri hpgišh* “from the tools he left after the meeting”. However, some cases of MWEs undergo a change of order of the verb complements. For example, in the expression *akl at — bli mlx* (lit. “eat — without salt”) “defeated — easily”, the space — can be filled by a noun phrase complement, as in *nakl at kwlm bli mlx* (lit. “(we) will eat everyone without salt”) "(we) will defeat everyone easily". This expression can also appear in the form *at kwlm nakl bli mlx* (lit. "everyone (we) will eat without salt") "everyone (we) will defeat easily".

- The order of the adverbials is also flexible in Hebrew. In particular, adverbials can appear after the subject, as in *hwa Tmn at hawcr lid hxwp* "he buried the treasure beside the beach", or before the subject, *lid hxwp hwa Tmn at hawcr* "beside the beach he buried the treasure". However, the order of adverbials in MWEs tends to be fixed. For example, the expression *Tmn (at) rašw bxwl* (lit. "buried his head in the sand") "to bury one's head in sand" does not appear in the form *bxwl hwa Tmn at rašw* "in sand he buried his head". Similarly, the expression *hxziq at hraš m'l hmim* (lit. "(he) held his head above the water") "keep one's head above water" does not appear in the form **m'l hmim hxziq at hraš* "above the water he held his head".
- In compositional expressions, phrases combined by a coordinating conjunction are interchangeable. For example, the sentence *prxim cwmxim bhrim wbšdwt* "flowers grow in mountains and in fields" can be rephrased as *prxim cwmxim bšdwt wbhrim* "flowers grow in fields and in mountains". However, this variation is not possible in MWEs such as *ica bšn w'in* (lit. "went out in a tooth and an eye") "be injured, loose", which does not appear as **ica b'in wšn* "went out in an eye and a tooth". Another example is the expression *pxwt aw iwtr* (lit. "less or more") "more or less", consisting of two adjectives combined by a conjunction, which does not appear in the form **iwtr aw pxwt* "more or less".
- Verb phrases headed by a transitive verb can undergo passivation, which results in a change in the order of the verb's arguments. For example, *haišh špkh at hmim* "the woman spilled the water" can undergo passivation as in *hmim nšpkw 'l idi haišh* "The water was spilled by the woman". In MWEs this transformation may be blocked. For example, the expression *špk at lbw* (lit. "(he) spilled his heart") "confessed one's true thoughts and feelings" does not appear in the passive form **lbw nšpk* "his heart was spilled". Another example is the expression *bnh mgdlim bawwir* (lit. "(he) built towers in the air") "build castles in the air", which cannot be realised as **mgdlim nbnw bawwir* "towers were built in the air".

Syntactic variants: Certain syntactic structures in Hebrew have a syntactic variant which allows paraphrasing the expression. Below we give examples of such syntactic structures.

Noun-Noun Constructs: A Noun-Noun Construct (henceforth NNC[4]) is a pair of consecutive nouns where the first noun (the head) appears in construct state, and the second noun (the modifier) is typically in

[4] Called *smixut* in Hebrew.

the absolute state. Such an expressions can be paraphrased by introducing a preposition between its two parts. For example, *bit hild* (lit. "house the boy") "the boy's house" can be paraphrased as *hbit šl hild* "the house of the boy". This variation may be unavailable for MWE constructs, especially when their meaning is idiomatic . For example, the expression *bit hspr* (lit. "house the book") "school" cannot be paraphrases as **hbit šl hspr* "the house of the book", and the expression *raš krwb* (lit. "head cabbage") "idiot" cannot be paraphrase as **raš mkrwb* "a head from cabbage" (see further details in Section 3.5).

possessiveness: Hebrew has (at least) two possessive constructions. One uses pronominal suffixes as in *bit+w→bitw* "his house", and the other uses the possessive pronoun *šlw* "his" (which can inflect for person, gender, and number) as in *hbit šlw* (lit. "the house his") "his house". So, a compositional expression such as *haiš ica mbitw* "the man left his home" can be paraphrased as *haiš ica mhbit šlw* (lit. "the man left home his") "the man left his home". However, MWEs tend not to undergo this kind of paraphrasing. For example, the expression *ica md'tw* (lit. "(he) went out of hi2s mind") "gone crazy" does not appear in the form **ica mhd't šlw* (lit. "went out of mind his") "went out of his mind"; the expression *biqš at idh* (lit. "asked for her hand") "proposed(marriage)" does not appear in the form **biqš at hid šlh* (lit. "asked for hand her") "asked for her hand".

Reference: In the case of compositional expressions, it is possible to refer to certain parts of the expression in subsequent sentences, in various ways. Below we mention two such phenomena variations and show that MWEs tend not to allow them.

Category transformation: Here, the referring word is a category transformation of the part (of the expression) to which it refers. Consider the compositional expression *hildh hiph* (lit. "the girl the beautiful") "the beautiful girl". In subsequent sentences, one can refer to the "beauty" of the girl as in *hiwpi šlh mrhiv* (lit. "the beauty of her is ravishing") "her beauty is ravishing". The noun *iwpi* "beauty" is a category transformation (adjective to noun) of the adjective *iph* "beautiful". Category transformation cannot be used to refer to constituents of MWEs. For example, in the expression *hxlwnwt hgbwhim* (lit. "the windows the high") "upper echelon", it is not possible to refer to the adjective *gbwhim* "high" using the previous category transformation as in **hgwbh šl hxlwnnwt* "the hight of the windows". Another example of a category transformation is verb to noun. Consider the compositional expression *akl at htpwx* "(he) ate the apple". One can then refer to the action in this sentence as in *akilt htpwx grmh lw kabi bTn* (lit. "the eating of the apple caused him ache stomach") "the eating of the apple caused him stomach aches". Such transformations may be blocked in MWEs. For example, in the expression *akl at iribw bli mlx* (lit. "ate his opponent without salt") "defeated his opponent easily", it is not possible to refer to the action literally, as in **akilt hirib* "the eating of the opponent".

Deletion: When a head is modified by adjuncts, it is usually possible to refer back to the head, suppressing the adjuncts, in subsequent text. Consider the expression *mnw’ xzq* (lit. "engine powerful") "powerful engine". One can, in subsequent sentences, refer to the head *mnw’* "engine" without the adjective as in ***hmnw*** *mtwcrt arcwt hbrit* "the engine is made in the U.S". This possibility may be unavailable in MWEs, especially when their meaning is idiomatic. For example, the expression *awr irwq* (lit. "light green") "authorization". We cannot then use deletion in order to refer to the first constituent *awr* "light".

In some cases of MWEs (especially the semantically compositional ones), deletion can be used to refer to constituents of the expressions. For example, consider the expression *mkwnt qph* (lit. "machine coffee") "coffee machine". One possible sentence that uses deletion to refer to the first constituent *mkwnh* "machine" is *qninw* ***mkwnt qph*** *xdšh, abl* ***hmkwnh*** *htklklh axri xwdš* (lit. "we bought a machine coffee new, but the machine broke down a month later") "we bought a new coffee machine, but the machine broke down a month later".

Syntactic irregularity Some MWEs are syntactically irregular. This is manifested in a number of ways, such as irregular syntactic patterns, and the unusaul use of a certain part of speech. Below are some examples.

- The expression *bxwr wTwb* (lit. "a young man and good") "an outstanding young man", which conjoins a noun with an adjective.
- The expression *’šh xwšbim* (lit. "do thinking") "to hold on and think", which consists of two finite verbs in sequence.
- The expression *ild Twb irwšlim* (lit. "boy good jerusalem") "obedient", which has the irregular pattern noun+adjective+proper name.
- The expression *nxba al hklim* (lit. "hiding to the tools") "shy". The preposition *to* is not subcategorized by the verb .

Modification: Compositional expressions can have their parts modified by modifiers such as adjectives, adverbs, prepositional phrases, etc. These can either modify a single element of the expression (*internal modification*), or the entire expression (*external modification*). An example of a compositional expression that undergoes internal modification is the expression *iwm iph* (lit. "day beautiful") "a beautiful day". This expression can be (internally) modified as in *iwm mawd iph* (lit. "day very beautiful") "a very beautiful day", or as in *iwm n’im wiph* (lit. "day lovely and beautiful") "a lovely and beautiful day". Internal modification may be highly restricted in MWEs, especially when their meaning is idiomatic. For example, the expression *sprwt iph* (lit. “a beautiful literature”) “belles-lettres” does not appear in the form **sprwt mawd iph* (lit. “a literature very beautiful”) “a very beautiful literature”, or in the form **sprwt n’imh wiph* (lit. “a literature lovely and beautiful”) “a lovely and beautiful literature”. It is important to note that some MWEs do allow internal modification. For example, the expression *’bwdh šxwrh* (lit. “a black work”) “a physical unskilled labor” can be internally modified (by an adjectival phrase) as in *’bwdh hrbh iwtr prwzait wšxwrh* (lit. “a much more boring black work”) “a much more boring physical labor”.

3.3 Semantic properties

Semantic compositionality: The semantic compositionality of a given MWE is defined as the degree to which the meaning of the whole expression results from combining the meanings of its individual words when they occur in isolation. We found that the semantic compositionality of Hebrew MWEs lies along a continuous spectrum which ranges from highly idiomatic to completely transparent. Below we provide examples of MWEs arranged according to the degree of their semantic compositionality (its place on the spectrum) from low to high.

Low degree: *ap ’l pi kn* (lit. “even on the mouth thus”) “nevertheless”, *lwbš at hmknšim* (lit. “(he) wears the pants”) “he is the dominant spouse”, *ica dwpn* (lit. “leave side”) “be exceptional”, *lxm xwq* (lit. “bread law”) “routine”.

Higher Degree: *cxq ‘d dm‘wt* (lit. “laughed till tears”) “laughed hysterically”, *gan xiwt* (lit. “animal garden”) “zoo”, *’l qch hlšwn* (lit. “on tip of the tongue”) “on the tip of your tongue”, *hrim raš* (lit. “(he) raised his head”) “to feel proud”.

Highest Degree: *mkwnt qph* (lit. “machine coffee”) “coffee machine”, *bdwar xwzr* (lit. “by returning mail”) “by return mail”, *’wbd zr* (lit. “worker foreign”) “foreign worker”, *slaT irqwt* (lit. “salad vegetable”) “vegetable salad”, *mxšb ’l* (lit. “computer super”) “supercomputer”.

Lexical fixedness: Recall that an expression is considered lexically fixed if replacing any of its constituents by a semantically (and syntactically) similar word generally results in an invalid or a literal expression. We checked the lexical fixedness of MWEs by substituting their constituents with semantically related words, and examining whether the resultant expression is a MWE. We found that MWEs, in most of the cases, tend to be lexically fixed (non-productive), but some cases, especially the semantically compositional MWEs, exhibited some lexical flexibility. Consider the following examples:

- The expression *hxlwnwt hgbwhim* (lit. “high windows”) “upper echelon”. Substituting the word *xlwnwt* (lit.“windows”) with each of the following semantically related words: *2ašnbim* (lit.“hatches”), *ptxim* (lit."openings"), and *dltwt* (lit."doors") results in a literal expression. Also, substituting the second word *gbwhim* (lit."high") with the words *'liwnim* (lit."upper"), *mwrmim* (lit."elevated"), or *nmwkim* (lit."lower") results in a literal expression.

- The expression *akl at hkwb'* (lit. "eat the hat") "eat one's hat". Substituting the verb *akl* (lit."eat") with the synonyms *Trp* (lit."to devour") or *bl'* (lit."to engulf") results in a literal expression. Also, substituting the third word *kwb'* (lit."hat") with the words *mcnpt* (lit."conical hat"), *mgb't* (lit."brimmed hat"), or *qsdh* (lit."helmet") results in a literal expression.
- In other (fewer) cases MWEs are more lexically flexible. For example, the expression *mkwnt qph* (lit. "machine coffee") "coffee machine". Substituting the noun *qph* (lit."coffee") with the semantically related nouns *asprsw* (lit."espresso"), or *qpwcinw* (lit."capucino") result in another MWE. Also, substituting the noun *mkwnt* (lit."machine") with the semantically related word *mkšir* (lit."machine or instrument") results in a plausible MWE.

Translation equivalents Some MWEs translate, as a whole, to a single word in some other language. Examples include *bit spr* (lit. "house book") "school", *išr lb* (lit. "straight heart") "honest", *ap 'l pi kn* (lit. "even on thus") "nevertheless", *abd 'liw qlx* (lit. "lost on him") "outdated", *bid xzqh* (lit. "in hand strong") "forcefully". Still, some semantically non-compositional MWEs, if translated literally (and adjusted syntactically) result in an MWE in the target language. For example, *Tmn (at) rašw bxwl* (lit. "buried his head in sand") "to bury one's head in sand", *mlx harc* (lit. "the salt of the earth") "the salt of the earth", *šbr lw at hlb* (lit. "broke his heart") "break someone's heart", *hwšiT id* (lit. "to lend (someone) a hand") "to lend (someone) a hand", *ph gdwl* (lit. "mouth big") "big mouth". Many of those are borrowed expressions.

3.4 Hebrew MWEs: A constructive definition

The characterizing properties of multi-word expressions listed above can serve as indications to when an expression is less compositional. Our main motivation in this work is to support the computational processing of Hebrew expressions, and since our goal is to extend the available lexicon and morphological processors of Hebrew, we define MWEs in this work in light of whether or not they require special representation. Specifically, an expression is considered an MWE if it exhibits idiosyncratic lexical properties (for example, hapax legomena); morphological properties (for example, partial or irregular inflection patterns); syntactic properties; or semantic properties (for example, it does not translate compositionally). Such expressions must be listed in the lexicon in order for computational processors to handle them correctly, and this is our guideline in the present work.

3.5 NNC MWEs in Hebrew

Recall from Section 3.2 that an NNC is a pair of consecutive nouns where the first noun (the *head*) appears in construct state, and the second noun (the *modifier*) is in the absolute state[5]. In this section we give a list of idiosyncratic morphological, semantic, and syntactic properties of NNC MWEs in Hebrew (henceforth noun compounds). For each of these idiosyncratic properties we provide examples of noun compounds that exhibit these properties and non-MWE NNCs that do not. In Section 6, these idiosyncratic properties will be used to distinguish noun compounds from non-MWE NNCs and to extract noun compounds from text. In this work we chose to focus on the extraction of noun compounds for a number of reasons. One, noun compunds are prevalent in Hebrew texts. Second, we have all the tools needed to detect their idiosyncratic properties. Third, the approach we develop in this work, to extract noun compounds, can be straightforwardly expanded to identify other types of Hebrew MWEs, such as Adj-N and N-Adj expressions.

3.5.1 Morphological idiosyncrasy

1. **Non-standard inflection:** The head of a noun compound can occur in a construct definite form, violating standard morphological rules. For example, the word *hiwšbi* (lit. "the sitting") which is the head of the noun compound *hiwšbi raš* (lit. "sitting head") "chairmen". This irregualr inflection does not occur in compositional NNCs. For example, the compositional NNC *dlt awTw* (lit. "door car") "a car's door" cannot occur in the form **hdlt awTw* (lit. "the door car").

2. **Partial inflection:** Often, the second element of a compound (which is an ordinary noun) is limited in its inflection. Specifically, such nouns are frequently limited to either singular or plural form (but not both). For example, consider *'wrq din* (lit. "editor law") "lawyer". In this expression the modifier *din* has the plural form *dinim* but it does not appear in this form within the expression. Another example is the noun compound *bit bwbwt* (lit. "house dolls") "dollhouse". In this expression the modifier *bwbwt* has the singular form *bwbh* but does not appear in this form within the expression. In the case of compositional NNCs, the modifier *can* inflect for number. For example, the modifier *xlwn* (lit. "window") in the compositional NNC *xlwn hbnin* (lit. "window the building") "the building's window" can appear also in its plural form as in *xlwnwt hbninim* (lit. "windows the buildings") "the buildings' windows".

[5] A more general defintion of NNCs includes expressions composed of a noun followed by a noun phrase. Typically, the noun phrase consists of a single noun, in which case it must be in the absolute state. But it can (recursively) be a noun-noun construct, in which case its first noun will be in the construct state.

3.5.2 Syntactic idiosyncrasy

1. **Limited syntactic variants:** Compositional NNCs can be paraphrased by a construction that uses the genitive preposition *šl* "of" instead of the construct state. For example, *bit hild* (lit. "house the boy") "the boy's house" can be paraphrased as *hbit šl hild* "the house of the boy". Another example is the compositional NNC *slT krwb* (lit. "salad cabbage") "cabbage salad" which can be paraphrased as *slT mkrwb* (lit. "salad made of cabbage") "salad made of cabbage" by introducing the preposition *m* "from". These syntactic variants are restricted in the case of noun compounds. For example, the noun compound *qli rqb* (lit. "tool car") "vehicle" cannot be paraphrased as **qli šl rkb* (lit. "tool of car"). Another example is the noun compound *raš krwb* (lit. "head cabbage") "idiot" which cannot be paraphrase as **raš mkrwb* (lit. "a head from cabbage").

2. **Constituents cannot be modified:** In compositional NNCs it is possible to modify (by an adjective) either the first or the second constituent. For example, consider the compositional NNC *'wrkt 'itwn* (lit. "editor newspaper") "newspaper editor(female)". It is possible to modify the first constituent *'wrkt* as in *'wrkt 'itwn xdSh* (lit. "editor newspaper new") "new newspaper editor" or the second constituent *'itwn* as in *'wrkt 'itwn xdSh* (lit. "editor newspaper new") "[new newspaper] editor". This is not true for noun compounds. For example, consider the noun compound *'wrkt din* (lit. "editor law") "lawyer(female)". While it is possible to modify the whole expression *'wrkt din wtiqh* (lit. "editor law seasoned") "seasoned law editor", it is not possible to modify the second constituent as in **'wrkt din wtiq* (lit. "editor law seasoned").

3. **Noun compounds cannot be conjoined:** Two compositional NNCs that have a common head can be conjoined using the coordinating conjunction *w* "and" as in the following example: The NNCs *mp'l plsTik* (lit. "factory plastic") "plastic factory", *mp'l mtqt* (lit. "factory metal") "metal factory" can be coordinated yielding *mp'l plsTik wmtqt* (lit. "factory plastic and metal") "plastic and metal factory". However, in case of noun compounds this is not possible. For example, the noun compounds *gn pirwt* (lit. "garden fruits") "fruit garden" , *gn xiwt* (lit. "garden animals") "zoo" cannot be coordinated as in **gn pirwt wxiwt* (lit. "a fruit and animal garden").

3.5.3 Semantic idiosyncrasy

1. **Semantic compositionality:** Recall from Section 3.3 that the semantic compositionality of Hebrew MWEs lies along a continuous spectrum which ranges from highly idiomatic to completely transparent. We found that noun compounds fall in the middle of the spectrum while compositional NNCs are at the extreme of the spectrum (transparent). Examples of noun compounds and compositional NNCs are given above.

2. **lexical fixedness:** Recall from Section 3.3 that an expression is considered lexically fixed if replacing any of its constituents by a semantically (and syntactically) similar word generally results in an invalid or a literal expression. We found that noun compounds tend to be lexically fixed while compositional compounds are not. For example substituting semantically related words as a head or modifier of the noun compounds *bit spr* (lit. "house book") "school" as in **bnin spr* (lit. "building book"), **bit xwbrt* (lit. "house booklet") results in ungrammatical expressions. An example of the lexical "flexibility" of compositional NNCs is *bit hild* "house of the boy" where other NNCs could be obtained by substituting semantically related words instead for the head or the modifier as in *bit hgbr* "house of the man" , *dirt hild* "apartment of the boy".

3. **Translation equivalents** Many of the noun copmpounds translate to a single word in English, while compositional NNCs translate to more than one. Examples of noun compounds include *bit mšpT* (lit. "house of trial") "court", *'wrq din* (lit. "editor of law") "lawyer", *šwmr raš* (lit. "gaurd head") "bodygaurd", *lwx zmnim* (lit. "board time") "schedule". Examples of compositional NNCs include *dlt hbit* (lit. "door of the house") "the house's door", *swp šnh* (lit. "end of year") "end of year", *xbr w'dh* (lit. "member committee") "committee member", *twšbi h'ir* (lit. "residents of the city") "the city residents".

4 Lexical representation and morphological processing of Hebrew MWEs

Based on our linguistic investigation of the properties of Hebrew MWEs in Section 3, we describe here an architecture for lexical representation of MWEs, accompanied by a specification of the integration of MWEs into a morphological processor of Hebrew. The developed system can represent MWEs in the lexicon along with their morphological and syntactic properties. It can identify MWEs in texts, morphologically analyze them, and provide their analysis using XML. The system is implemented as an extension of two existing components of the current morphological system, a Hebrew lexicon and a morphological processor (Itai and Wintner, 2008). We start this section by describing the existing morphological system and its components. Then, we describe the changes we incorporated into the existing components, the new components that we added, and the reasons behind our design decisions.

4.1 The overall architecture of the current morphological system

The architecture of the morphological system is depicted in Figure 1. It is composed of two major units: *The Generation Unit* (drawn inside the dashed box in Figure 1) consists of three modules: a *Lexicon*, a *Generator*, and a *Database of inflected forms*. *The Analysis Unit* is composed of a *Tokenizer*, a *Morphological Processor*, and an *XML wrapper*. Below, we briefly describe each unit and explain how the modules interact to produce morphological analyses for the words appearing in the input text (for further details see Itai and Wintner (2008)).

4.1.1 The Generation Unit

Lexicon: The *Haifa Lexicon of Contemporary Hebrew* is the broadest-coverage publicly available lexicon of Hebrew, currently consisting of over 25,000 entries. The lexicon is represented in XML[6] as a list of *item* elements, each with a base form which is the citation form used in conventional dictionaries. For nouns and adjectives it is the absolute singular masculine, whereas for verbs it is the third person singular masculine, past tense.

Lexicon items are specified for the following attributes:[7] a unique *id*, three representations of the lexical entry (undotted, dotted and transliterated) and *script*, which encodes deviations from the standard script as well as register (see a discussion of Hebrew orthography in Section 1.2). In

[6] The linguistic databases are represented in 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.

[7] This is just a list of the attributes relevant to our discussion. For a full list of the attributes see the *The Hebrew Lexicon XML Schema* at http://www.mila.cs.technion.ac.il/english/resources/standards/hebrewlexicon

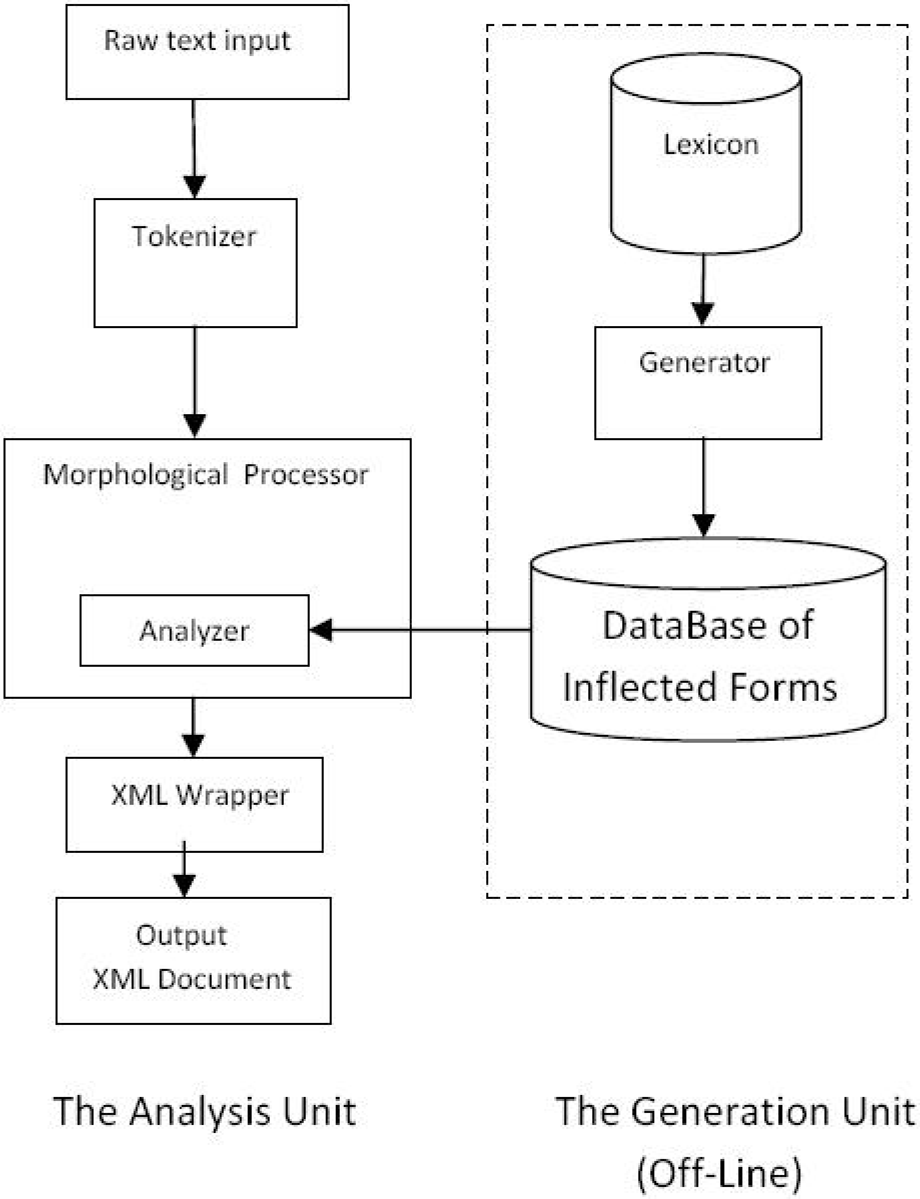


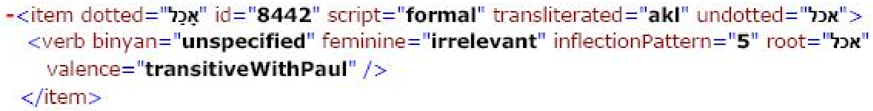
Figure 1: The architecture of the current morphological system

addition, every lexicon item belongs to a *part of speech* category,[8] which is designated as a *sub-element* of the item. The part of speech of an entry determines its additional attributes. For *nominals*, i.e., nouns, adjectives

[8] The POS categories are: noun, verb, auxiliary verb, proper name, adjective, adverb, preposition, conjunction, pronoun, numeral, interjection, quantifier, modal, prefix, interrogative, negation, existential, foreign, participle, existential, impersonal.

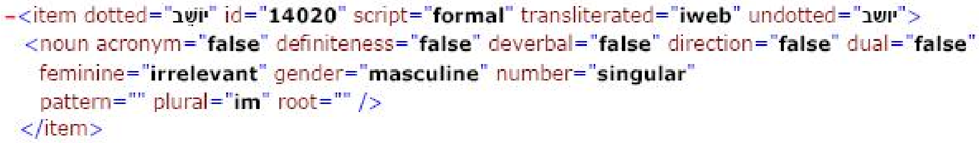
and numerals, these include number and gender. Verbs are specified for root and *inflection pattern* (IP). We also list the type of proper names (person, location, organization or date).

In addition, each lexicon item specifies features which govern the inflectional morphological behavior of the lexeme (which is used by the generator, see below). For example, nouns specify whether they inflect for gender and if so, what the feminine suffix is (either *im* or *wt*). In addition, the lexicon utilizes a special mechanism for lexical specification of idiosyncrasies: *add, replace* and *remove* directives are used to control the generation of irregular inflected forms of the lexeme. The *add* directive can be used to add a special form, *remove* removes a form which would have been generated by default, and *replace* substitutes an irregular form for the default one. Examples of lexicon entries are depicted in Figures[9] 2, 3, and 4.



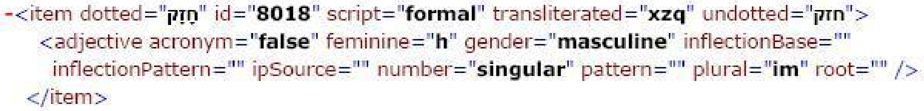
```
-<item dotted="אָכַל" id="8442" script="formal" transliterated="akl" undotted="אכל">
  <verb binyan="unspecified" feminine="irrelevant" inflectionPattern="5" root="אכל"
    valence="transitiveWithPaul" />
 </item>
```

Figure 2: Lexicon entry of *akl*



```
-<item dotted="יוֹשֵׁב" id="14020" script="formal" transliterated="iweb" undotted="יושב">
  <noun acronym="false" definiteness="false" deverbal="false" direction="false" dual="false"
   feminine="irrelevant" gender="masculine" number="singular"
   pattern="" plural="im" root="" />
 </item>
```

Figure 3: Lexicon entry of *iweb*



```
-<item dotted="חָזָק" id="8018" script="formal" transliterated="xzq" undotted="חזק">
    <adjective acronym="false" feminine="h" gender="masculine" inflectionBase=""
     inflectionPattern="" ipSource="" number="singular" pattern="" plural="im" root="" />
  </item>
```

Figure 4: Lexicon entry of *xzq*

The Generator and the Database of inflected forms: The generator goes over the items in the lexicon, and for each item it generates, *off-line*, all the inflected forms induced by the item (excluding combinations of prefix sequences with the inflected forms). Then, the inflected words are stored in a database. For each inflected form the database stores a transliteration of the word, a pointer to the citation form of the word

[9] In the figures a different transliteration is used, which replaces š with *e*, and ' with *y*

(the lexicon ID of the item from which it was generated), and its complete morphological analysis.

4.1.2 The Analysis Unit

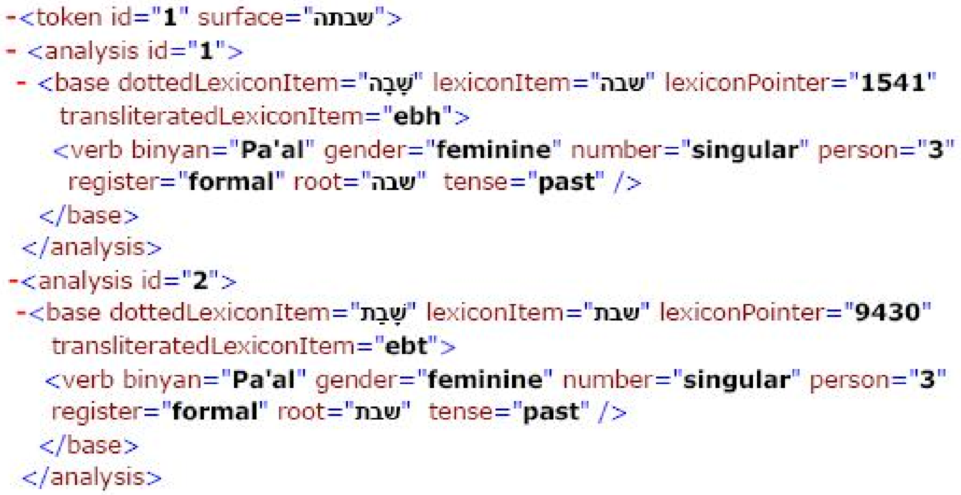
Tokenizer: The tokenization module operates on the input text (UTF-8 encoded raw data), and segments it into paragraphs, sentences and tokens. The output of the tokenizer (in XML format, discussed below) is fed into the morphological processor.

Morphological processor: The morphological processor strips possible prefixes (taken from a list of possible prefix particles) of each token and matches the remaining string against the database of inflected forms. When the match is successful, the prefix and remaining string are passed to the analyzer. The analyzer determines whether the combination of the prefix sequence and the inflected form is valid, in which case the analysis is fed to the XML wrapper.

XML wrapper and the corpus representation schema: The XML wrapper wraps all possible analyses of each token in XML and the XML document corresponding to the entire input text is returned. The XML document follows *The Hebrew Corpus XML Schema*,[10] which induces the following structure on the document. A document is a sequence of articles, each of which is a sequence of paragraphs which are sequences of sentences. A sentence is a sequence of tokens, and a token contains at least two attributes: *id* and *surface form* (the word in Hebrew script, UTF-8 encoded). In addition, a token may contain morphological analyses. A morphologically analyzed corpus contains all the analyses of a word (as produced by the morphological processor), regardless of context. Figures 5, 6, 7 depict all the analyses that are produced by the morphological analyzer for the form *šbth*. Each analysis consists of zero or more *prefix*es, a *base* and an optional *suffix*. The base specifies the properties of the lemma of the token, including its form (both in Hebrew and transliterated), part of speech and POS-dependent features (such as number, gender and nominal state in the case of nouns).

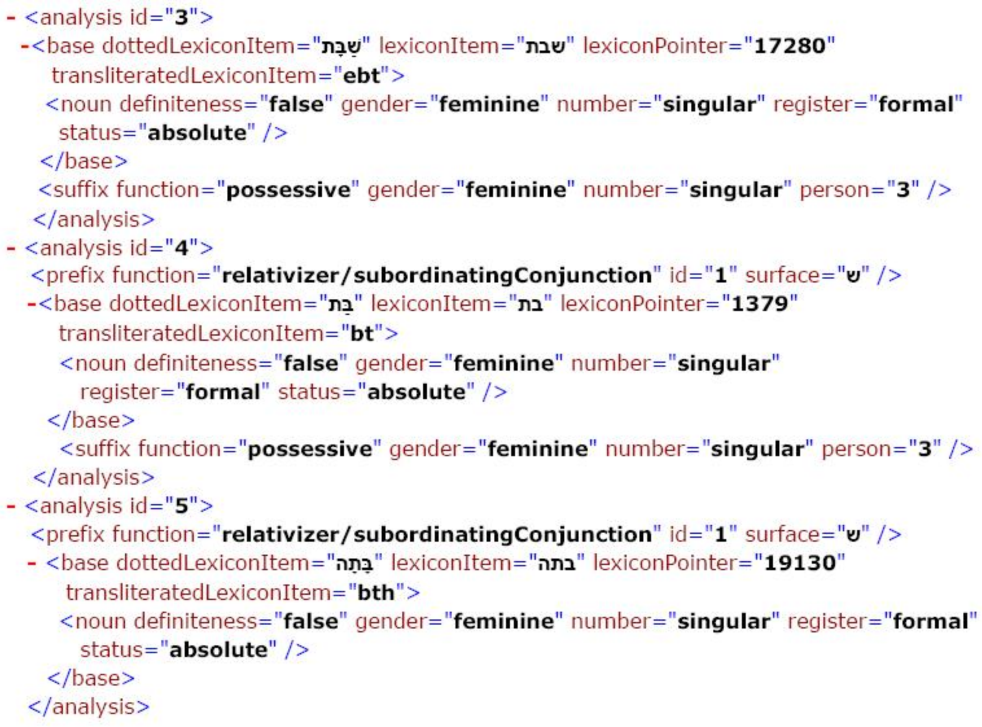
Note that the current morphological processor operates on a token-by-token basis, so a tokenization pre-processing step is usually required before morphological analysis. The tokenizer uses only blanks and punctuation to segment a text into tokens. In particular, it is completely independent of the lexicon. The lexicon includes single-word tokens only, and the morphological analyzer is unaware of MWEs. Next, we describe the changes we incorporated into the morphological processing system described above, which enables it to handle MWEs.

[10] Available at http://www.mila.cs.technion.ac.il/english/resources/standards/hebrewcorpus



```
-<token id="1" surface="שבתה">
 - <analysis id="1">
  - <base dottedLexiconItem="שָׁבָה" lexiconItem="שבה" lexiconPointer="1541"
      transliteratedLexiconItem="ebh">
     <verb binyan="Pa'al" gender="feminine" number="singular" person="3"
       register="formal" root="שבה"  tense="past" />
    </base>
  </analysis>
 -<analysis id="2">
  -<base dottedLexiconItem="שָׁבַת" lexiconItem="שבת" lexiconPointer="9430"
     transliteratedLexiconItem="ebt">
    <verb binyan="Pa'al" gender="feminine" number="singular" person="3"
     register="formal" root="שבת"  tense="past" />
    </base>
  </analysis>
```

Figure 5: Analysis 1-2 of *sbth*

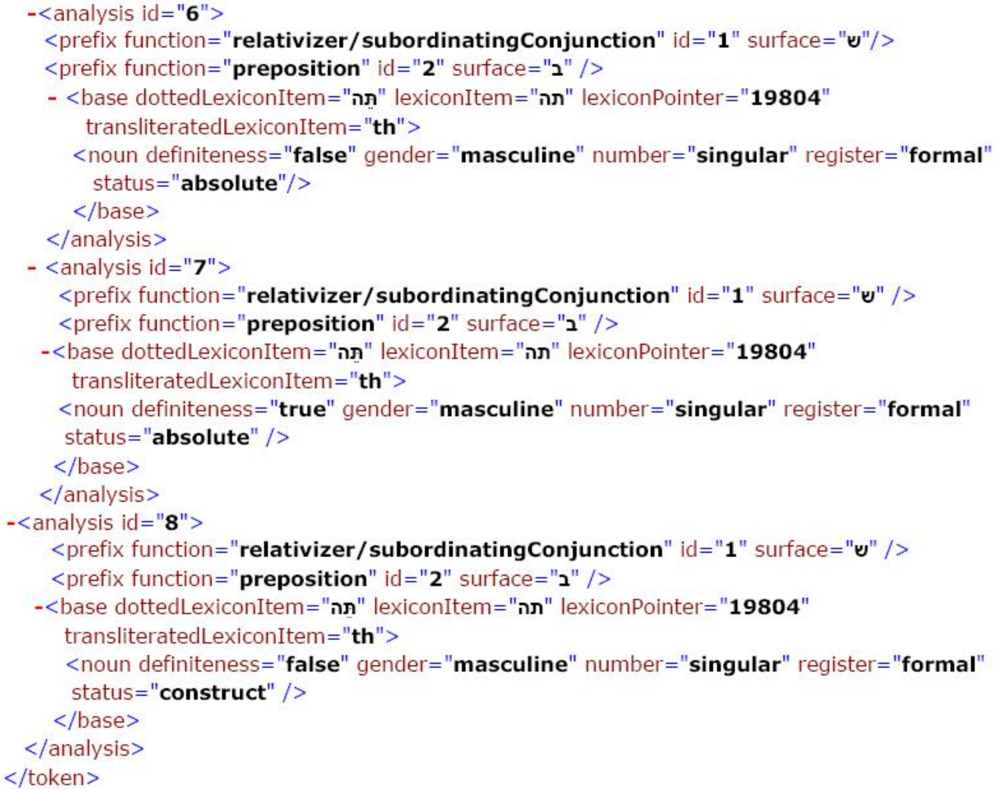


```
- <analysis id="3">
  -<base dottedLexiconItem="שֶׁבֶת" lexiconItem="שבת" lexiconPointer="17280"
     transliteratedLexiconItem="ebt">
    <noun definiteness="false" gender="feminine" number="singular" register="formal"
      status="absolute" />
   </base>
   <suffix function="possessive" gender="feminine" number="singular" person="3" />
  </analysis>
- <analysis id="4">
   <prefix function="relativizer/subordinatingConjunction" id="1" surface="ש" />
  -<base dottedLexiconItem="בַּת" lexiconItem="בת" lexiconPointer="1379"
     transliteratedLexiconItem="bt">
     <noun definiteness="false" gender="feminine" number="singular"
       register="formal" status="absolute" />
    </base>
     <suffix function="possessive" gender="feminine" number="singular" person="3" />
  </analysis>
- <analysis id="5">
   <prefix function="relativizer/subordinatingConjunction" id="1" surface="ש" />
  - <base dottedLexiconItem="בִּתָּה" lexiconItem="בתה" lexiconPointer="19130"
      transliteratedLexiconItem="bth">
     <noun definiteness="false" gender="feminine" number="singular" register="formal"
       status="absolute" />
    </base>
  </analysis>
```

Figure 6: Analysis 3-5 of *sbth*

4.2 Lexical representation of MWEs

The architecture of the upgraded morphological system is depicted in Figure 8. In this figure, componets that were changed are marked with a *, and new



```
  -<analysis id="6">
    <prefix function="relativizer/subordinatingConjunction" id="1" surface="ש"/>
    <prefix function="preposition" id="2" surface="ב" />
   - <base dottedLexiconItem="תֵה" lexiconItem="תה" lexiconPointer="19804"
        transliteratedLexiconItem="th">
      <noun definiteness="false" gender="masculine" number="singular" register="formal"
        status="absolute"/>
      </base>
    </analysis>
  - <analysis id="7">
      <prefix function="relativizer/subordinatingConjunction" id="1" surface="ש" />
      <prefix function="preposition" id="2" surface="ב" />
    -<base dottedLexiconItem="תֵה" lexiconItem="תה" lexiconPointer="19804"
       transliteratedLexiconItem="th">
      <noun definiteness="true" gender="masculine" number="singular" register="formal"
       status="absolute" />
     </base>
    </analysis>
-<analysis id="8">
     <prefix function="relativizer/subordinatingConjunction" id="1" surface="ש" />
     <prefix function="preposition" id="2" surface="ב" />
   -<base dottedLexiconItem="תֵה" lexiconItem="תה" lexiconPointer="19804"
      transliteratedLexiconItem="th">
      <noun definiteness="false" gender="masculine" number="singular" register="formal"
       status="construct" />
     </base>
  </analysis>
</token>
```

Figure 7: Analysis 6-8 of *sbth*

components are inside dashed boxes. This section discusses the modification introduced to the lexicon in order to represent MWEs. Our approach is to design a representation for MWEs that on one hand, is simple and consistent with the current lexicon, and on the other hand is expressive enough to account for any combination of syntactic and lexical properties exhibited by the various types of MWEs that we identified in Section 3. We adopted the current *Hebrew Lexicon XML Schema* with all its attributes and elements (retaining their values and functions), and further extended it by adding new elements and attributes.

4.2.1 Basics

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. The most fundamental change is that each *item* is now specified (in addition to its current attributes) for the following attributes:

type: This attribute can take one of two different values: *mwe* if this lexicon entry represents a MWE and *word* if it represents a single word.

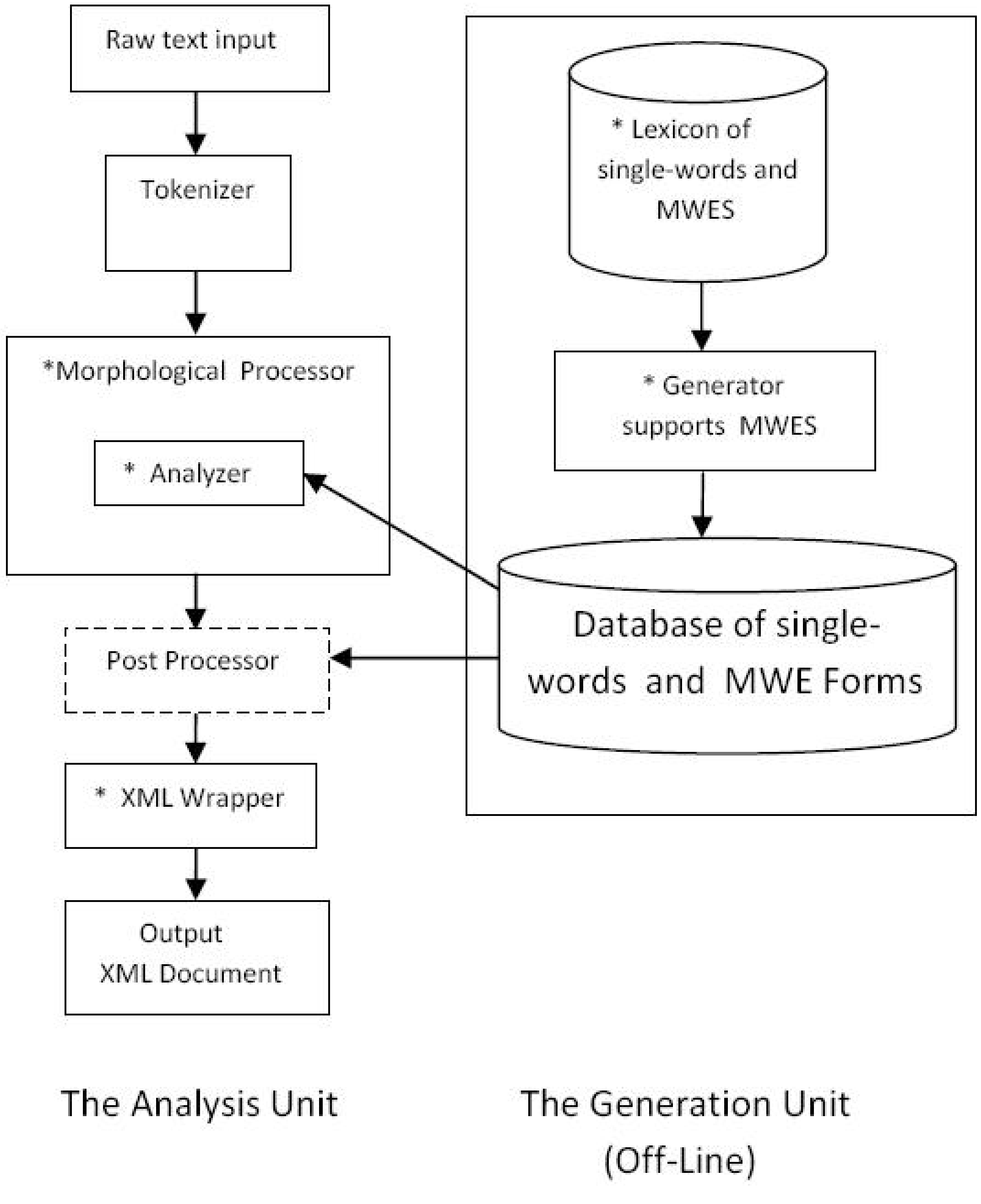
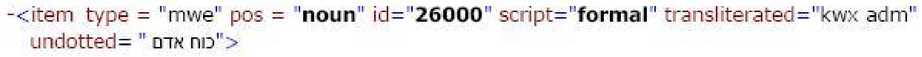


Figure 8: The architecture of the extended morphological system

pos: This attribute designates the part of speech of the entire MWE. In addition to the existing (lexical) POS categories, this attribute can have the values **NP** for noun phrase or **VP** for verb phrase, reflecting the observation that MWEs are sometimes phrases.

Figure 9 depicts a part of the lexicon entry of the MWE *kwx adm* (lit. “power man”) “manpower”. It depicts the element *item* along with all its attributes. All the elements and items described below refer to *items* with *type*=“mwe”.



```
-<item type = "mwe" pos = "noun" id="26000" script="formal" transliterated="kwx adm"
  undotted= " כוח אדם">
```

Figure 9: Lexicon entry of *kwx adm*

4.2.2 Accounting for morphological idiosyncrasy

We describe a set of attributes and elements that we added in order to account for the idiosyncratic morphological behavior of MWE constituents. Recall that constituents can appear in *frozen form*, or undergo *full inflection* or *partial inflection*. Constituents can also undergo *non-standard inflection*, or even be *Hapax legomena*. In the first three cases, the forms that a constituent can appear in are a subset of all the possible forms of its base form. So, for every constituent we keep a link (pointer) to its base form entry in the lexicon, and a propositional logic formula defined over the elements and attributes defined in the *Hebrew Corpus XML Schema*. The pointer together with the formula form a query that retrieves, from the *database of inflected forms*, all the forms the constituent can appear in. We do the same in case of *non-standard inflection*, but use additional elements and attributes, which together with the pointer and the formula define not just all the possible forms, but also non-standard ones. For the case of *Hapax legomena* we create a new lexicon entry for the constituent and specify in this entry that it can only appear in the frame of a MWE, and keep a pointer to this entry. Below we describe the new elements and attributes we use.

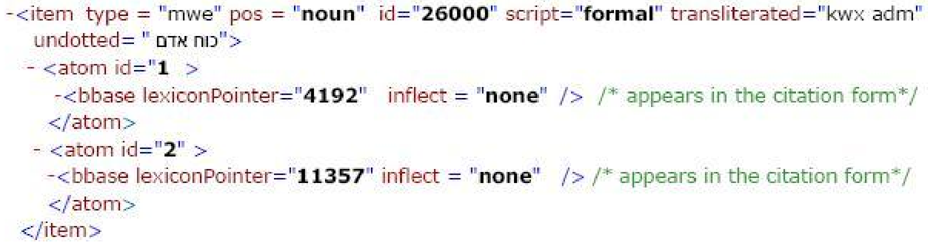
atom: The element *atom* defines a constituent along with all its possible forms (in the frame of the MWE). Every *atom* is specified for a unique *id* (within the *item* it is defined in). In addition, each *atom* has the following sub-elements:

bbase: This element defines all the possible forms (without a prefix) that the constituent can appear in. It is specified for the following attributes:

lexiconPointer: A pointer to the lexicon entry of the constituent's base (citation) form.

inflect This optional attribute, if it appears, can have one of two possible values. The first one is the value “none”, which means that this constituent appears in the canonical (citation) form only. The second possible value is a propositional logic formula defined over the elements and attributes defined in the *Hebrew Corpus XML Schema* (a formal definition of these formulas is given in Appendix A). If this attribute is not specified the constituent is assumed to appear in *all* of its possible forms.

Figure 10 depicts (a part of) the lexicon entry of the MWE *kwx adm*. Since both words are frozen in this expression, both atoms are specified as *inflect*="none".



```
-<item  type = "mwe" pos = "noun"  id="26000" script="formal" transliterated="kwx adm"
   undotted= " כוח אדם">
  - <atom id="1  >
     -<bbase lexiconPointer="4192"   inflect = "none"  />  /* appears in the citation form*/
    </atom>
   - <atom id="2" >
    -<bbase lexiconPointer="11357" inflect = "none"   /> /* appears in the citation form*/
    </atom>
  </item>
```

Figure 10: The lexicon entry of *kwx adm*

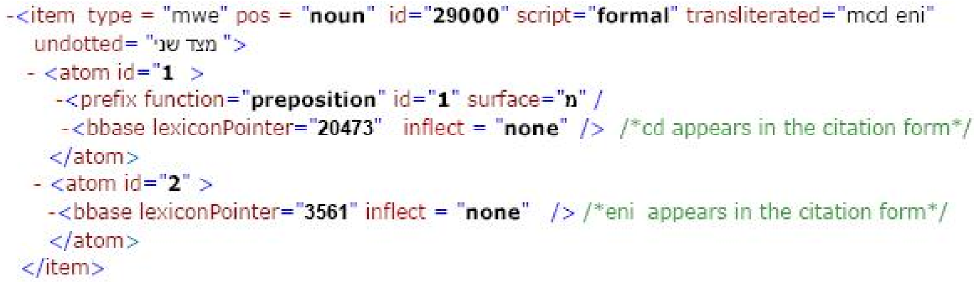
prefix: This element specifies a prefix that combines with the form (or forms) defined in *bbase*. This element is defined in the *Hebrew Corpus XML Schema*, where it is specified for the following attributes:

id: A unique ID within the *atom* it is defined in (an *atom* can have zero or more *prefix*es).

function: The function of the prefix (prefixes in Hebrew can have various functions, such as conjunction, preposition, subordinating conjunction, etc.)

surface: The surface form of the prefix.

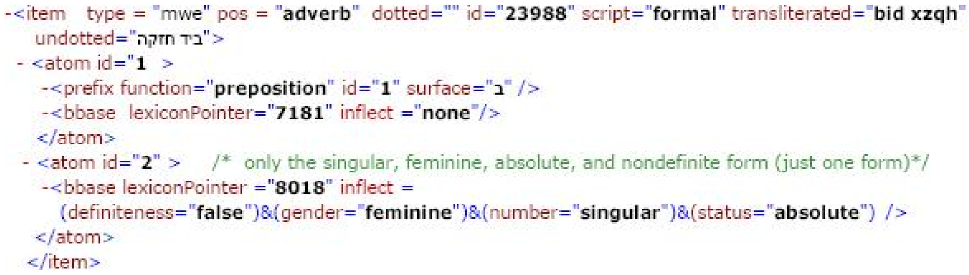
Figure 11 depicts (a part of) the lexicon entry of the MWE *mcd šni* (lit. "from side second") "on the other hand". Both words in this expression are frozen, but the first word is composed of a citation form *cd* combined with a prefix *m*.



```
-<item  type = "mwe" pos = "noun"  id="29000" script="formal" transliterated="mcd eni"
   undotted= "מצד שני ">
  - <atom id="1  >
     -<prefix function="preposition" id="1" surface="מ" /
      -<bbase lexiconPointer="20473"   inflect = "none"  />  /*cd appears in the citation form*/
    </atom>
   - <atom id="2" >
    -<bbase lexiconPointer="3561" inflect = "none"   /> /*eni  appears in the citation form*/
    </atom>
  </item>
```

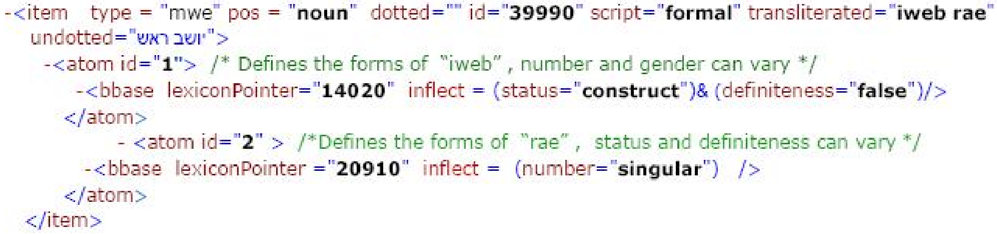
Figure 11: The lexicon entry of *mcd šni*

Figure 12 depicts (a part of) lexicon entry of the MWE *bid xzqh* (lit. "in hand strong") "forcefully". The value of "inflect" in the second atom reflects the fact that this frozen form is singular, feminine, indefinite, and absolute. Figure 13 depicts (a part of) the lexicon entry of the MWE *iwšb raš* (lit. "sitter head") "chairman". Note the possible inflected forms of both components. This entry yields forms such as: *iwebt rae* (feminine, singular, indefinite), *iweb hrae* (masculine, singular, definite), *iwebi rae* (a masculine, plural, indefinite), etc.



```
-<item   type = "mwe" pos = "adverb"  dotted="" id="23988" script="formal" transliterated="bid xzqh"
    undotted="ביד חזקה">
  - <atom id="1  >
     -<prefix function="preposition" id="1" surface="ב" />
     -<bbase  lexiconPointer="7181" inflect ="none"/>
    </atom>
  - <atom id="2" >     /*  only the singular, feminine, absolute, and nondefinite form (just one form)*/
     -<bbase lexiconPointer ="8018" inflect =
       (definiteness="false")&(gender="feminine")&(number="singular")&(status="absolute")  />
    </atom>
   </item>
```

Figure 12: Lexicon entry of *bid xzqh*



```
-<item   type = "mwe" pos = "noun"  dotted="" id="39990" script="formal" transliterated="iweb rae"
   undotted="יושב ראש">
     -<atom id="1">  /* Defines the forms of  "iweb" , number and gender can vary */
         -<bbase  lexiconPointer="14020"  inflect = (status="construct")& (definiteness="false")/>
       </atom>
              - <atom id="2" >  /*Defines the forms of  "rae" ,  status and definiteness can vary */
          -<bbase  lexiconPointer ="20910"  inflect =  (number="singular")   />
       </atom>
   </item>
```

Figure 13: Lexicon entry of *iweb rae*

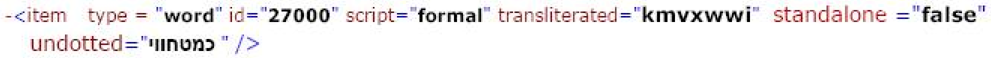
4.2.3 Hapax legomena

In the case of *Hapax legomena*, we follow the same approach that we use for partial and full inflection. Recall that *hapax legomena* include constituents that never occur outside the expression they appear in. This means that these constituents have no entry in the lexicon. So, for each one of these constituents we create an *item* in the lexicon which has the following attributes:

a) All the attributes that an *item* is specified for in the *Hebrew Lexicon XML Schema* (retaining their values and functions).

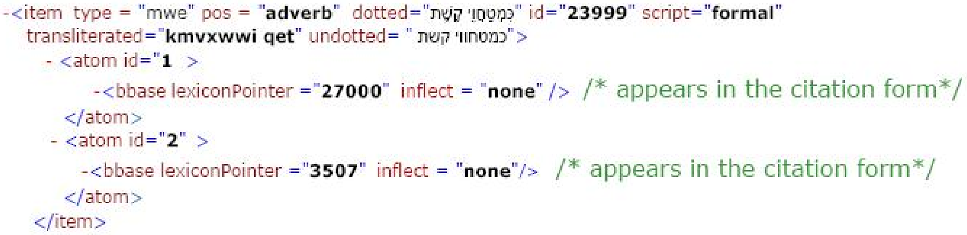
b) The attribute *standalone* that has the value **false** in case of a *hapax legomena* item. This attribute is not specified for "regular" single-word entries, for which the default value is **true**. This attribute is a directive to the generator to generate just one form for the item. This form has no morphological analysis, and is marked (in the Database of inflected forms) by a special flag.

Figure 14 depicts the lexicon entry of the hapax legomena *kmvxwwi*. As in the case of constituents that undergo *full inflection* or *partial inflection*, *hapax legomena* constituents are designated by the element *atom* with all its subelments and attributes (described above). The attribute *lexiconPointer* of *bbase* points to the lexicon entry of the constituent, and the attribute *inflect* is set to "none". Figure 15 depicts (part of) the lexicon entry of the MWE *kmTxwwi kšt* "stone's throw".



```
-<item   type = "word" id="27000" script="formal" transliterated="kmvxwwi"  standalone ="false"
  undotted="כמטחווי " />
```

Figure 14: Lexicon entry of *kmTxwwi*

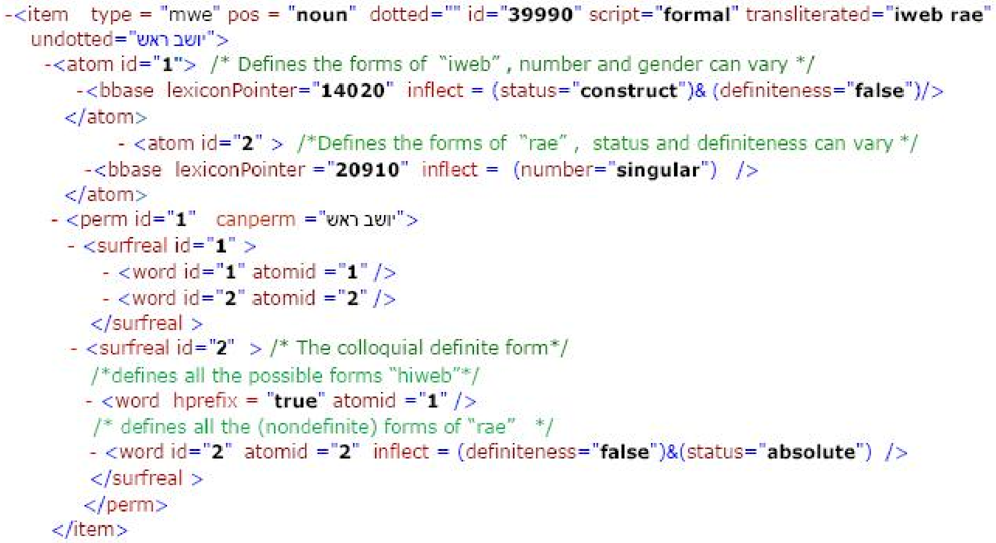


```
-<item  type = "mwe" pos = "adverb"  dotted="כִּמְטַחֲוֵי קֶשֶׁת" id="23999" script="formal"
  transliterated="kmvxwwi qet" undotted= " כמטחווי קשת">
   - <atom id="1  >
        -<bbase lexiconPointer ="27000"  inflect = "none" />  /* appears in the citation form*/
     </atom>
   - <atom id="2" >
       -<bbase lexiconPointer ="3507"  inflect = "none"/>  /* appears in the citation form*/
     </atom>
  </item>
```

Figure 15: Part of the Lexicon entry of *kmTxwwi kšt*

4.2.4 irregular inflections

Recall from Section 3 that some MWEs (such as noun compounds) have a colloquial definite form in which the first word can appear in the definite construct-state as in “hiwšbi”. This form is ungrammatical outside of MWEs as it combines the definite article *h* with a construct state noun *iwšbi*. This form does not exsist in the *database of inflected forms*; to define such words we add to the element *word* a new attribute called *hprefix* which can be either **true** or **false**. This attribute is optional and its default value is **false**. If its value is **true**, it designates that the prefix *h* attaches to the constituent defined by the *word* element. Figure 16 depicts the lexicon entry of *iwšb raš*. Note that in the second *surfreal* the first word has *hprefix*=“true”.



```
-<item   type = "mwe" pos = "noun"  dotted="" id="39990" script="formal" transliterated="iweb rae"
  undotted="יושב ראש">
   -<atom id="1"> /* Defines the forms of  "iweb" , number and gender can vary */
      -<bbase  lexiconPointer="14020"  inflect = (status="construct")& (definiteness="false")/>
    </atom>
         - <atom id="2" >  /*Defines the forms of  "rae" ,  status and definiteness can vary */
      -<bbase  lexiconPointer ="20910"  inflect =  (number="singular")   />
    </atom>
  - <perm id="1"   canperm ="יושב ראש">
   - <surfreal id="1" >
       - <word id="1" atomid ="1" />
       - <word id="2" atomid ="2" />
     </surfreal >
   - <surfreal id="2"   > /* The colloquial definite form*/
     /*defines all the possible forms "hiweb"*/
    - <word  hprefix = "true" atomid ="1" />
     /* defines all the (nondefinite) forms of "rae"   */
     - <word id="2"  atomid ="2"  inflect = (definiteness="false")&(status="absolute")  />
     </surfreal >
    </perm>
  </item>
```

Figure 16: Lexicon entry of *iwšb raš*

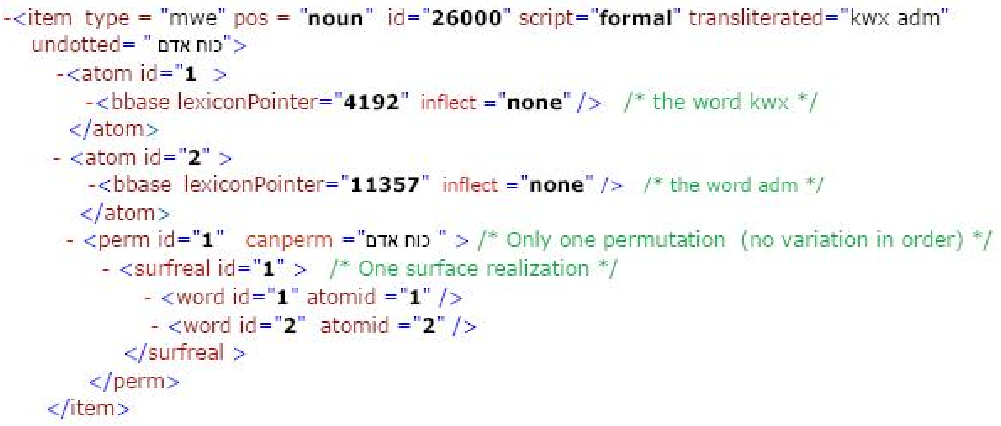
4.2.5 Accounting for syntactic flexibility

We describe here a set of attributes and elements that we added in order to account for the syntactic properties of MWEs. Recall from section 3.2 that MWEs allow variation in the order of their constituents. To encode this property we define *perm*utations. a MWE entry can include one or more *perm*s. Each *perm* element defines all the forms a MWE can appear in under a certain order (permutation) of the constituents. A *perm* element is specified for the following attributes:

id: a unique ID (within the *item* it is defined in).

canperm: The canonical form of the (forms defined under this) permutation.

The full lexicon entry of *kwx adm* is given in Figure 17 (the *surfreal* element will be explained presently). In this example, *word* elements specify the order of the *atoms* defined by the permutation. In this particular case, only one permutation is valid, in which the atoms occur in the canonical order (*atomid*= 1 followed by *atomid*= 2).



```
-<item  type = "mwe" pos = "noun"  id="26000" script="formal" transliterated="kwx adm"
   undotted= " כוח אדם">
     -<atom id="1  >
        -<bbase lexiconPointer="4192"  inflect ="none" />   /* the word kwx */
      </atom>
     - <atom id="2" >
        -<bbase  lexiconPointer="11357"  inflect ="none" />   /* the word adm */
       </atom>
      - <perm id="1"   canperm ="כוח אדם " > /* Only one permutation  (no variation in order) */
         - <surfreal id="1" >   /* One surface realization */
              - <word id="1" atomid ="1" />
               - <word id="2"  atomid ="2" />
            </surfreal >
         </perm>
    </item>
```

Figure 17: Lexicon entry of *kwx adm*

In some MWEs the possible inflections of one constituent may be dependent on the inflected forms of another constituent. For example, the MWE *milh nrdpt* (lit. "a word persecuted") "synonym" consists of a noun (*milh*) followed by an adjective (*nrdpt*). In Hebrew, an adjective that modifies a noun agrees with it on gender, number, and definiteness. In our case, the noun *mila* is feminine, and can inflect for number and definiteness, resulting in a total of four possible forms for this constituent. The adjective *nrdpt* agrees on the attributes of each of these forms, which results in a total of four possible forms for the whole MWE.

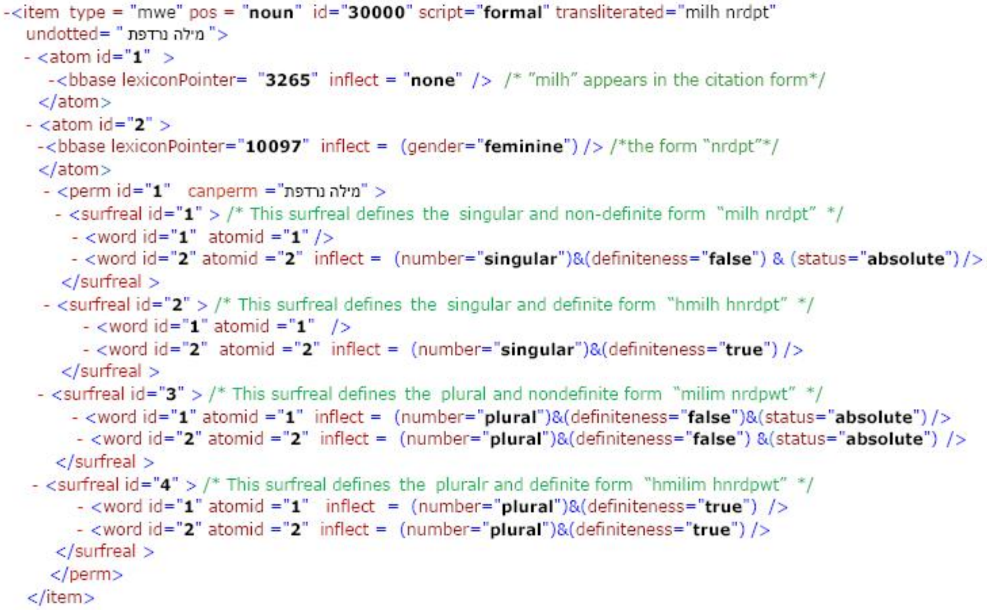
In our approach, each constituent and all its possible forms are defined in an *atom* independently of the other constituents.[11] In order to express a dependency between the attributes of different constituents under a certain

[11] We opted this approach for its simplicity. An alternative approach would have been to allow attributes (in the *inflect* formula) to have a variable as a value (e.g., *number*= **X**).

permutation, a sub-environment is defined, called *surfreal* (surface realization). Each *perm* consists of one or more *surfreal*s. Each *surfreal* defines a sub-group of all the forms (that agree on certain attributes) that a MWE can appear in under a certain permutation. A *surfreal* is specified for a unique *id* (within the permutation), and consists of one or more *word*s. A *word* is specified for the following attributes:

id: A unique ID (within the surfreal), which specifies its order defined by the permutation.

atomid: A pointer to one of the *atom*s defined above. For example, the lexicon entry of *milh nrdpt* is given in Figure 18. The expression *milh nrdpt* appears in one permutation, which consists of 4 surface realization (*surfreal*s). The first *surfreal* defines the indefinite singular form of the expression, and the second defines the definite singular form. The indefinite plural form is defined in the third *surfreal*, and the fourth *surfreal* defines the definite plural form.



```
-<item  type = "mwe" pos = "noun"  id="30000" script="formal" transliterated="milh nrdpt"
   undotted= " מילה נרדפת ">
  - <atom id="1"  >
    -<bbase lexiconPointer=  "3265"  inflect = "none"  />  /* "milh" appears in the citation form*/
   </atom>
  - <atom id="2" >
   -<bbase lexiconPointer="10097"  inflect =  (gender="feminine") /> /*the form "nrdpt"*/
   </atom>
   - <perm id="1"   canperm ="מילה נרדפת" >
    - <surfreal id="1" > /* This surfreal defines the  singular and non-definite form  "milh nrdpt"  */
      - <word id="1"  atomid ="1" />
      - <word id="2" atomid ="2"  inflect =  (number="singular")&(definiteness="false") & (status="absolute") />
     </surfreal >
   - <surfreal id="2" > /* This surfreal defines the  singular and definite form  "hmilh hnrdpt"  */
       - <word id="1" atomid ="1"   />
       - <word id="2"  atomid ="2"  inflect =  (number="singular")&(definiteness="true") />
     </surfreal >
  - <surfreal id="3" > /* This surfreal defines the  plural and nondefinite form  "milim nrdpwt"  */
      - <word id="1" atomid ="1"  inflect =  (number="plural")&(definiteness="false")&(status="absolute") />
       - <word id="2" atomid ="2"  inflect =  (number="plural")&(definiteness="false") &(status="absolute")  />
    </surfreal >
 - <surfreal id="4" > /* This surfreal defines the  pluralr and definite form  "hmilim hnrdpwt"  */
       - <word id="1" atomid ="1"   inflect  =  (number="plural")&(definiteness="true")  />
       - <word id="2" atomid ="2"  inflect =  (number="plural")&(definiteness="true") />
    </surfreal >
   </perm>
 </item>
```

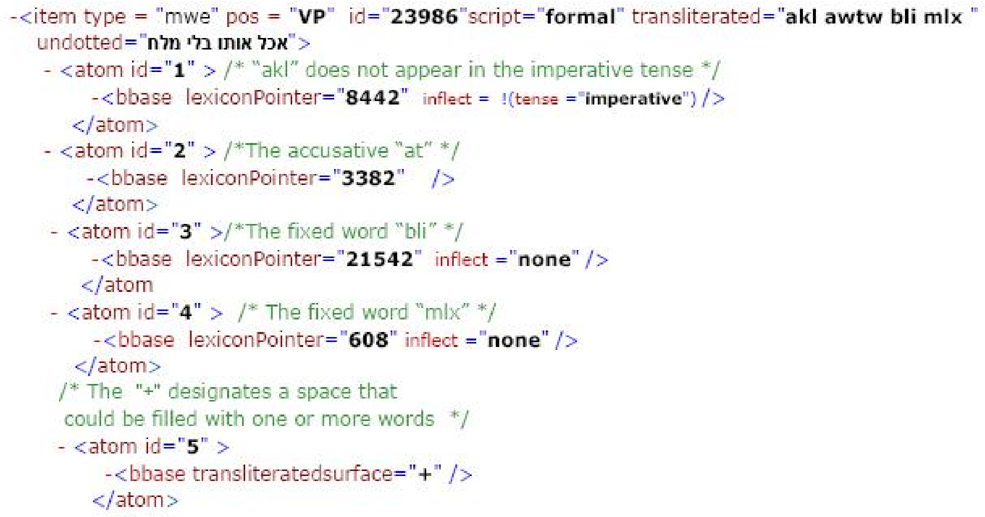
Figure 18: Lexicon entry of *milh nrdpt*

inflect: In some MWEs, the possible inflections that constituents can undergo change as a result of a change in their position (the permutation they appear in) within the expression. For example, the MWE *akl at — bli mlx* (lit. "eat — without salt") "defeated — easily" can also appear in the form *akl awtw bli mlx* (lit. "eat him without salt") "defeated him easily". While

This variable could then appear in the *inflect* formulas of different *atoms* (that define different constituents), forcing different constituent (under the same permutation) to agree on the value of some attribute. However, this approach is less straightforward for the lexicographer, and requires much more complex processing.

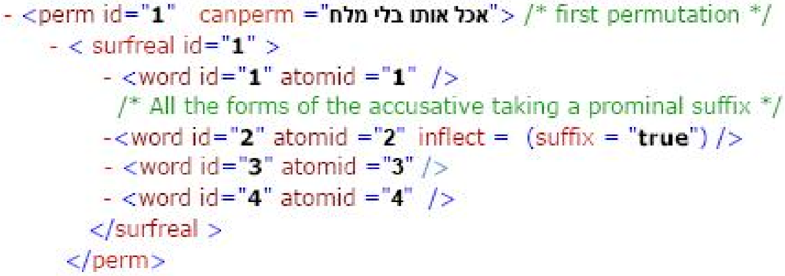
> in the first form the accusative preposition *at* appears in its citation form, in the second form it takes a pronominal suffix (singuar, masculine, third person). In order to account for the change of morphological inflections, we add an (optional) *inflect* attribute to *atomid* elements that can have a formula as a value which defines the inflections that the constituent can undergo in the frame of the *surfreal*. The actual inflections are the conjunction of those allowed by the *word* element with those allowed by the *atom* element (See Appendix B for detailed information).

As a concluding example, we depict the full lexicon entry of *akl at — bli mlx* in Figures 19 (the specification of the fifth atom is explained in Section 4.2.6), 20, 21, 22, 23. This lexicon entry contains most of the elements and attributes we describe above.



```
-<item type = "mwe" pos = "VP"  id="23986"script="formal" transliterated="akl awtw bli mlx "
   undotted="אכל אותו בלי מלח">
  - <atom id="1" > /* "akl" does not appear in the imperative tense */
       -<bbase  lexiconPointer="8442"  inflect =  !(tense ="imperative") />
     </atom>
  - <atom id="2" > /*The accusative "at" */
      -<bbase  lexiconPointer="3382"    />
     </atom>
   - <atom id="3" >/*The fixed word "bli" */
       -<bbase  lexiconPointer="21542"  inflect ="none" />
      </atom
   - <atom id="4" >  /* The fixed word "mlx" */
       -<bbase  lexiconPointer="608" inflect ="none" />
     </atom>
    /* The  "+" designates a space that
     could be filled with one or more words  */
    - <atom id="5" >
        -<bbase transliteratedsurface="+" />
       </atom>
```

Figure 19: Lexicon entry of *akl at — bli mlx*. The atoms

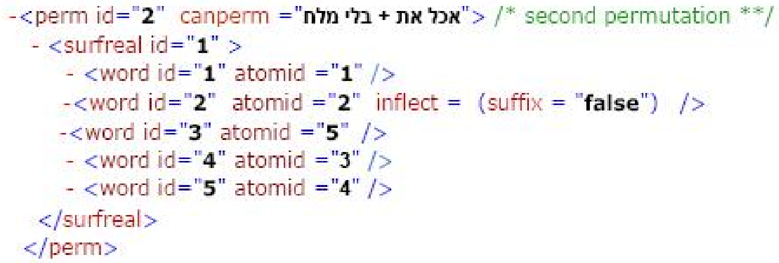


```
- <perm id="1"   canperm ="אכל אותו בלי מלח"> /* first permutation */
    - < surfreal id="1" >
        - <word id="1" atomid ="1"  />
         /* All the forms of the accusative taking a prominal suffix */
        -<word id="2" atomid ="2"  inflect =  (suffix = "true") />
        - <word id="3" atomid ="3" />
        - <word id="4" atomid ="4"  />
      </surfreal >
    </perm>
```

Figure 20: Lexicon entry of *akl at — bli mlx*. Permutation number 1

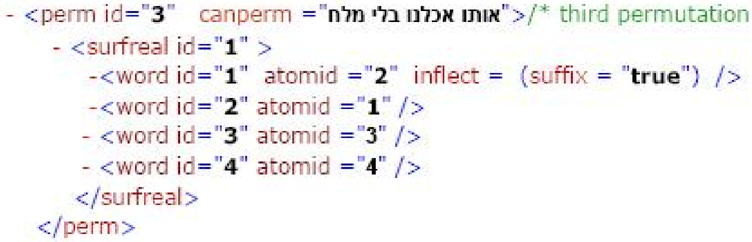
4.2.6 Open Slots

Recall from Section 3 that some MWEs contain *open slots*, which can be filled with complements. For example, consider the expression *akl at — bli mlx* (lit.



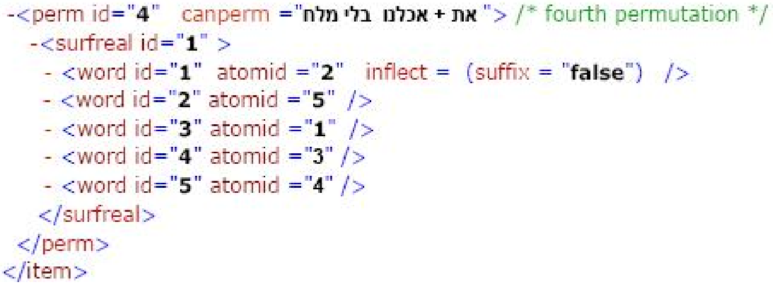
```
-<perm id="2"  canperm ="אכל את + בלי מלח"> /* second permutation **/
  - <surfreal id="1" >
     - <word id="1" atomid ="1" />
    -<word id="2"  atomid ="2"  inflect =  (suffix = "false")  />
    -<word id="3" atomid ="5"  />
     - <word id="4" atomid ="3" />
     - <word id="5" atomid ="4" />
   </surfreal>
  </perm>
```

Figure 21: Lexicon entry of *akl at — bli mlx*. Permutation number 2



```
- <perm id="3"   canperm ="אותו אכלנו בלי מלח">/* third permutation */
   - <surfreal id="1" >
      -<word id="1"  atomid ="2"  inflect =  (suffix = "true")  />
      -<word id="2" atomid ="1" />
     - <word id="3" atomid ="3" />
     - <word id="4" atomid ="4" />
     </surfreal>
  </perm>
```

Figure 22: Lexicon entry of *akl at — bli mlx*. Permutation number 3



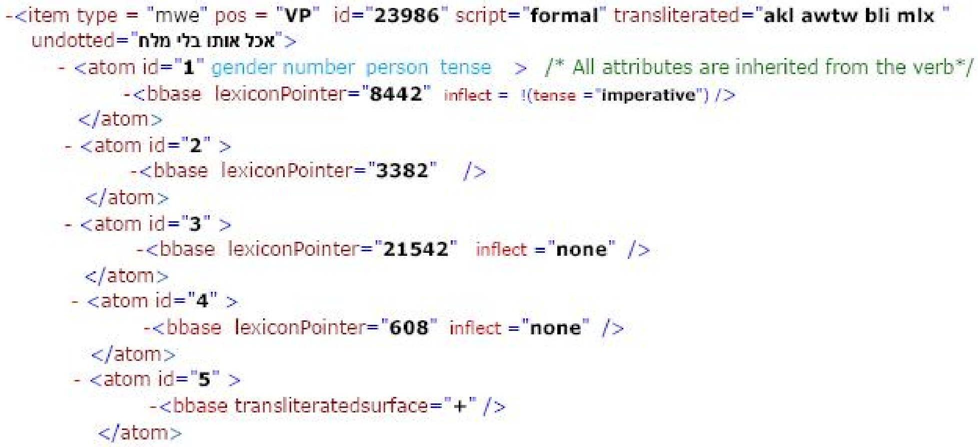
```
-<perm id="4"   canperm ="את + אכלנו  בלי מלח "> /* fourth permutation */
  -<surfreal id="1" >
   - <word id="1"  atomid ="2"   inflect =  (suffix = "false")  />
   - <word id="2" atomid ="5"  />
   - <word id="3" atomid ="1"  />
   - <word id="4" atomid ="3" />
   - <word id="5" atomid ="4" />
   </surfreal>
  </perm>
</item>
```

Figure 23: Lexicon entry of *akl at — bli mlx*. Permutation number 4

"eat — without salt") "defeated — easily". The open slot in this expression can be filled by a noun phrase, such as *akl at iribiw hxzqim bli mlx* (lit. "eat his powerful oponents without salt") "defeated his powerful opponents easily". Similar to words, each *open slot* is defined by an *atom*, whose *bbase* element has only one attribute, *transliteratedsurface*. This attribute can have different characters (wildcards) as value, which determine the number or the part of speech of the complement that can fill the *open slot*. The possible values are "+" for one or more words, and "*" for zero or more words. Figure 19 depicts part of the lexicon entry of *akl at — bli mlx*. Note the *transliteratedsurface*="+" in *atom* number 5, which defines an *open slot* that can be complemented by one or more words.

4.2.7 MWE attributes

Like single words, MWEs have attributes that are determined by their POS. For example, the MWE *iwšbi raš* (which is a noun) is plural, masculine, and indefinite. In most cases, the attributes of the whole MWE are inherited from the attributes of its constituents. To express this propery in the lexicon, each *atom* designates all the attributes that the expression inherits from it. Figure 24 defines (part of) the lexicon entry of *akl at — bli mlx*. Note that all the attributes of the expression (*gender*, *number*, *person*, *tense*) are inherited from the first word.



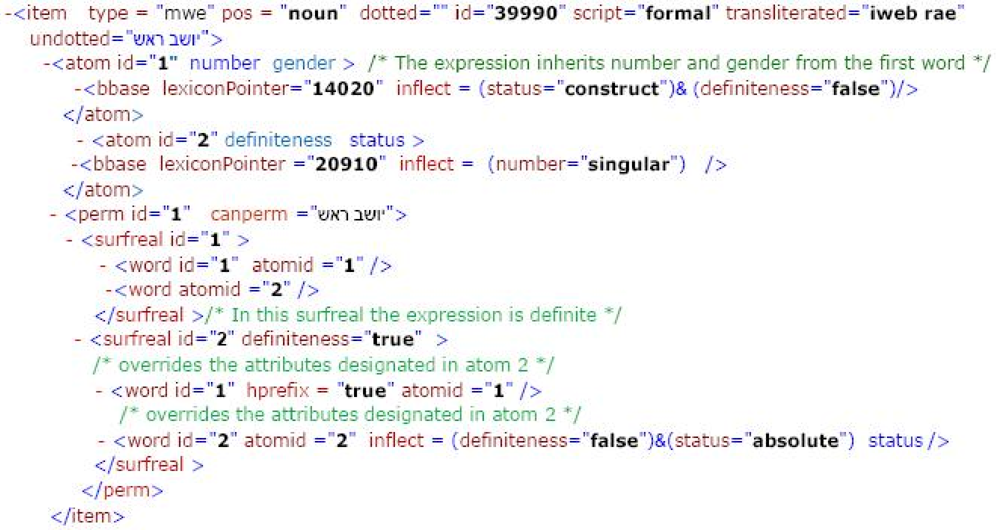
```
-<item type = "mwe" pos = "VP"  id="23986" script="formal" transliterated="akl awtw bli mlx "
   undotted="אכל אותו בלי מלח">
      - <atom id="1" gender number  person  tense   >  /* All attributes are inherited from the verb*/
              -<bbase  lexiconPointer="8442"  inflect =  !(tense ="imperative") />
         </atom>
       - <atom id="2" >
               -<bbase  lexiconPointer="3382"    />
          </atom>
       - <atom id="3" >
                -<bbase  lexiconPointer="21542"   inflect ="none"  />
          </atom>
        - <atom id="4" >
                 -<bbase  lexiconPointer="608"  inflect ="none"  />
           </atom>
        - <atom id="5" >
                  -<bbase transliteratedsurface="+" />
            </atom>
```

Figure 24: Part of the lexicon entry of *akl at — bli mlx*

In some cases the attributes that an expression inherits from a certain constituent depend on the permutation or the surface forms it appears in. To account for that, we also enable a *word* element to designate the attributes that the expression inherits from it (under this surface realization). The attributes designated in the *word* element override the attributes designated in the *atom*. Figure 25 revisits the lexicon entry of *iweb rae*. In the first *surfreal* the expression inherits its definiteness and status from the second word, whereas in the second *surfreal* it only inherits its status from the second word. In some cases, MWEs can have attributes that are not inherited from any of the constituents. We designate these attributes as a pair of attribute and value for the *surfreal* element. For example, in Figure 25 (second *surfreal*), the definiteness of the expression is not inherited from any of the constituents. It is designated as an attribute of the second *surfreal*.

4.2.8 MWEs constructions and templates

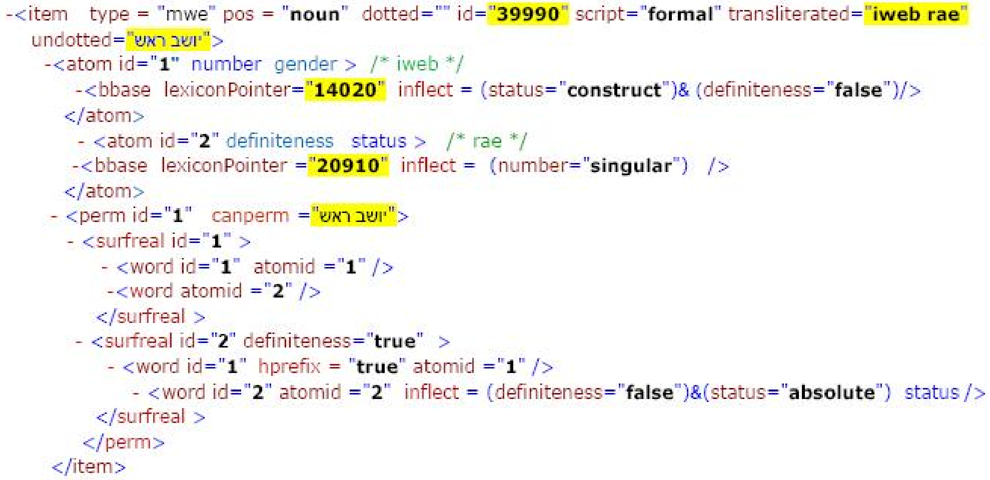
As shown in 3.5, noun compounds are MWE constructions which share an idiosyncratic behavior. While many of their linguistic properties are peculiar and uniqe to MWEs, they still behave in a general, predictable way. For example, Figures 26 , 27 depict the lexicon entries of the two noun compound MWEs *iweb rae* and *ywrk din*, respectively. The differing components are marked in



```
-<item   type = "mwe" pos = "noun"  dotted="" id="39990" script="formal" transliterated="iweb rae"
   undotted="יושב ראש">
    -<atom id="1"  number  gender >  /* The expression inherits number and gender from the first word */
        -<bbase  lexiconPointer="14020"  inflect = (status="construct")& (definiteness="false")/>
      </atom>
        - <atom id="2" definiteness   status >
        -<bbase  lexiconPointer ="20910"  inflect =  (number="singular")   />
      </atom>
     - <perm id="1"   canperm ="יושב ראש">
       - <surfreal id="1" >
            - <word id="1"  atomid ="1" />
             -<word atomid ="2" />
           </surfreal >/* In this surfreal the expression is definite */
         - <surfreal id="2" definiteness="true"   >
           /* overrides the attributes designated in atom 2 */
            - <word id="1"  hprefix = "true" atomid ="1" />
               /* overrides the attributes designated in atom 2 */
            - <word id="2" atomid ="2"  inflect = (definiteness="false")&(status="absolute")  status />
           </surfreal >
         </perm>
     </item>
```

Figure 25: The lexicon entry of *iweb rae*

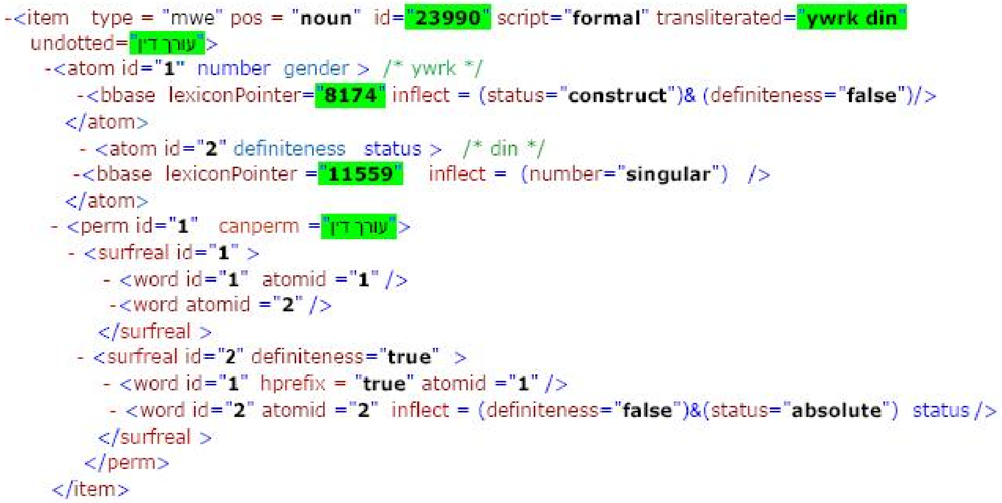
yellow in Figure 26 and in green in Figure 27. The remaining components of both lexicon entries are similar.



```
-<item   type = "mwe" pos = "noun"  dotted="" id="39990" script="formal" transliterated="iweb rae"
   undotted="יושב ראש">
    -<atom id="1"  number  gender >  /* iweb */
        -<bbase  lexiconPointer="14020"  inflect = (status="construct")& (definiteness="false")/>
      </atom>
        - <atom id="2" definiteness   status >   /* rae */
        -<bbase  lexiconPointer ="20910"  inflect =  (number="singular")   />
      </atom>
     - <perm id="1"   canperm ="יושב ראש">
       - <surfreal id="1" >
            - <word id="1"  atomid ="1" />
             -<word atomid ="2" />
           </surfreal >
         - <surfreal id="2" definiteness="true"   >
            - <word id="1"  hprefix = "true" atomid ="1" />
               - <word id="2" atomid ="2"  inflect = (definiteness="false")&(status="absolute")  status />
           </surfreal >
         </perm>
     </item>
```

Figure 26: Lexicon entry of *iweb rae*

To represent such MWE constructions we define *templates*. A template defined for a certain MWE construction takes as parameters the variable parts of the construction, and is compiled into the full lexical specification of the MWE. Figure 28 depicts a template for noun compounds. It takes 5 parameters: *id*, *transliterated*, *undotted*, *canperm*, *lexiconPointer1*, and *lexiconPointer2*. This template also has the attribute *construction*, which designates the name of the MWE construction that the template defines. To define an entry of a noun

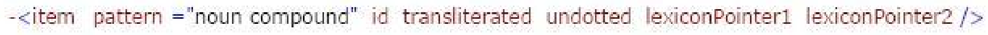


```
-<item   type = "mwe" pos = "noun"  id="23990" script="formal" transliterated="ywrk din"
   undotted="עורך דין">
    -<atom id="1"  number  gender >  /* ywrk */
        -<bbase  lexiconPointer="8174" inflect = (status="construct")& (definiteness="false")/>
      </atom>
        - <atom id="2" definiteness   status >   /* din */
       -<bbase  lexiconPointer ="11559"    inflect =  (number="singular")   />
      </atom>
     - <perm id="1"   canperm ="עורך דין">
       - <surfreal id="1" >
           - <word id="1"  atomid ="1" />
            -<word atomid ="2" />
          </surfreal >
        - <surfreal id="2" definiteness="true"   >
           - <word id="1"  hprefix = "true" atomid ="1" />
            - <word id="2" atomid ="2"  inflect = (definiteness="false")&(status="absolute")  status />
          </surfreal >
        </perm>
    </item>
```

Figure 27: Lexicon entry of *ywrk din*

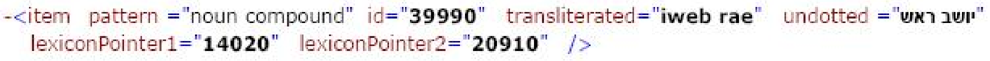
compound MWE, one has to set the values of the parameters in this template. Figure 29 depicts the lexicon entry of the expression *iweb rae*, using the template of Figure 29. This lexicon entry is compiled into the full specification of the MWE, as in Figure 26.

Using templates for defining MWE lexicon entries has several practical advantages: templates form a *high level* language, which enables the lexicographer to add new MWE entries without the need to master the detailed *low level* language defined in Section 4.2. Using templates also facilitates automatic generation of lexicon entries: an automatic extraction procedure (see Section 5) that identifies MWEs in corpora can generate template instances rather than the detailed specification described above.



```
-<item   pattern ="noun compound"  id  transliterated  undotted  lexiconPointer1  lexiconPointer2 />
```

Figure 28: A template for noun compounds



```
-<item   pattern ="noun compound"  id="39990"   transliterated="iweb rae"   undotted ="יושב ראש"
   lexiconPointer1="14020"   lexiconPointer2="20910"   />
```

Figure 29: The lexicon entry of *iweb rae* defined using a template

4.3 Morphological generation and storage of MWEs

The MWE generator follows in principle the method used to generate inflected forms of single words described in Section 4.1.1. The *elements* and the values of the *attributes* described in Section 4.2 serve as directives for generating, *off-line*, all possible forms of the given MWE including those with irregular inflections. All the generated forms are stored in the *Database of single words and MWE forms*. For each MWE form the database stores the following information:

Information for the whole expression:

1. The POS of the expression.
2. The morphological analysis of the whole expression.
3. The lexical ID of the MWE entry in the lexicon
4. The number of constituents in the expression.

Information for each constituent:

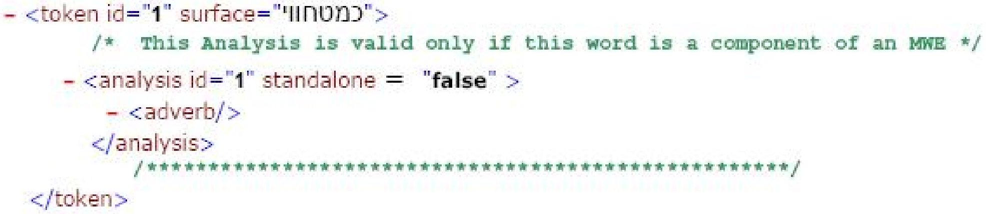
1. The surface form of the constituent (without the prefix).
2. The prefix that attaches to base form (if available).
3. A pointer to the lexicon entry of the constituent's citation form (the *lexiconPointer*).

4.4 Morphological Analysis Of MWEs

Recall that current morphological processor operates on a token-by-token basis, and is totally unaware of MWEs. We describe here the changes we incorporated into the current morphological system in order to identify and produce analyses for MWEs appearing in the input text. The first question that we had to answer is the stage in which the MWEs should be processed. We considered the following possibilities:

Preprocessing Preprocessing involves adding a new component between the Tokenizer and the Morphological processor in Figure 1. The first task in processing MWEs (even in the simplest case of fixed expressions) would require this component to strip possible prefixes of each token and check whether the remaining string matches the first constituent of a MWE in the *Database of single words and MWE forms*. This means that this component would have to perform a task that is already done by the current Morphological processor, reduplicating the effort, and resulting in a non-modular solution.

Inprocessing The upgraded morphological processor works in the same fashion that we mentioned in Section 4.1, but does the following steps to take care of MWEs: after stripping a possible prefix of the given token it matches the remaining string against the *Database of single words and MWE forms* to determine whether there exists a MWE that starts with this string. When the match is successful, the prefix and POS of the candidate MWE are sent to the analyzer, which determines whether the combination of the prefix sequence and the POS is valid (see Section 1.2), in which case the token is marked as a potential start of a MWE (henceforth head). In case the token is a *Hapax legomena* or a word that undergone a non-standard inflection, the processor generates an analysis for it with the *standalone* attribute set to **false**. This designates that the analysis is valid only in a frame of a MWE. Figure 30 depicts the full analysis of the *Hapax legomena kmTxwwi*. Note that *standalone*= **false**, which means that this analysis is valid only if the word will turn out to be a constituent of a MWE. The next step in processing MWEs would require the processor to "look ahead" into the stream of tokens produced by the tokenizer, and check whether these tokens together with the current token (the head) match an existing MWE. We decided to do this work outside the morphological processor in a *Postprocessing* stage for two main reasons. First, this would involve introducing a new task for the current morphological processor which could be done, efficiently (as we show below), outside of the morphological processor. Second, having a separate unit would enable to easily switch off the MWE analysis feature in case only the single word morphological analysis is needed.



```
- <token id="1" surface="כמטחווי">
      /*  This Analysis is valid only if this word is a component of an MWE */
    - <analysis id="1" standalone =  "false" >
       - <adverb/>
      </analysis>
        /*******************************************************/
  </token>
```

Figure 30: The full analysis of *kmTxwwi*

Post Processing We build a new component called the *post processor*. This component receives the output produced by the Morphological processor, which for each token includes all its possible analyses, and whether the token is a potential start of a MWE (henceforth head). For each head token the post processor looks up all the MWEs in *Database of single words and MWE forms* that start with this token. All the matched MWEs (henceforth candidates) are concisely represented by a finite state machine (FSM). The edges of this FSM are marked by the constituents of the MWEs, and a state *qf* is a final state if there exists a path from the initial state *q0* to *qf* where the sequence of the words on the edges comprise a candidate MWE. The post processor checks if one of the candidate MWE appears in the input by simply finding out if the FSM accepts an input consisting of the current and the following tokens. When a MWE is detected in the input, the post processor produces an analysis for the first token specifying it as a head of the MWE, and an analysis for the following tokens as constituents of the MWE. In case the given token is a *Hapax legomena* or a word that undergone a non-standard inflection, the post processor discards its *non-standalone* analysis if the token is not a constituent of any MWE. Then, the analysis produced by the post processor is fed (along with the analyses produced by the morphological processor) to the XML wapper.

4.5 XML representation of morphologically analyzed text

The upgraded XML wrapper receives all possible analyses of each token from the post processor, wraps them in XML, and returns the XML document corresponding to the entire input text. In order to represent the MWE analyses of a given token, we add new *attribute*s to *The Hebrew Corpus XML Schema.* The following are the attributes and the information that a MWE analysis includes:

mweid: A unique ID (within the document it appears in) assigned for every instance of a MWE.

wanalysisid: A pointer to one of the possible analyses of the token (as a single word), which is the analysis of this constituent of the MWE. We will later discuss the cases of *Hapax legomena* and non-standard inflections.

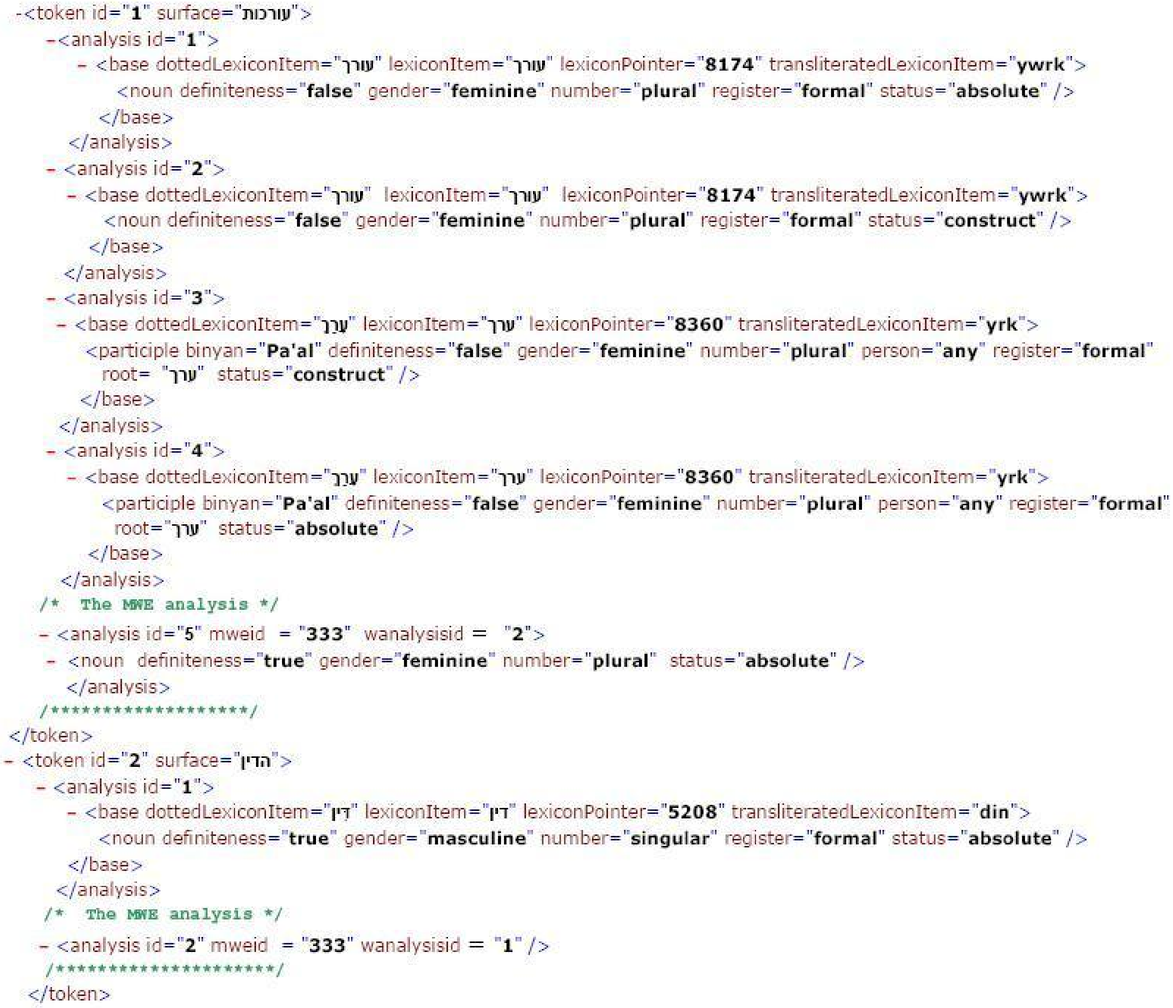
islot: An optional boolean attribute whose default value is **false**. If its value is **true**, then it designates that this token fills an *open slot* in the MWE. In case the token is the first constituent of a MWE, its MWE analysis also includes the POS of the expression and the attributes that are determined by its POS (e.g., *gender*, *number*, *definiteness*, etc.)

In case the token is the first constituent of a MWE, its MWE analysis will also include the POS of the whole expression along with all its other attributes that are determined by its POS (see Section 4.2.6).

Figure 31 depicts the full analysis of the expression *ywrkwt hdin* as produced by the *Post Processor.* In this example the token *ywrkwt* has 5 analyses: the first 4 analyses as a single word, and the last analysis as the first constituent of the expression *ywrkwt hdin.* Note that in the last analysis *wanalysisid*="2", which points to the correct analysis of the constituent *ywrkwt.* This analysis also designates that the POS of the MWE is a noun, which is *plural*, *feminine*, and *definite.* The second analysis describes of the token *hdin* is its analysis as a constituent of the MWE. Note that it has the same *mweid* value as *ywrkwt*, which designates that both are constituents of the same MWE *ywrkwt hdin.* Appendix C includes an example of the full analysis of the sentence *akl at kl iribiw bli mlx* (lit. "(he) ate all his opponents without salt") "(he) defeated all his opponents easily", which is an instance of the MWE *akl at — bli mlx.*

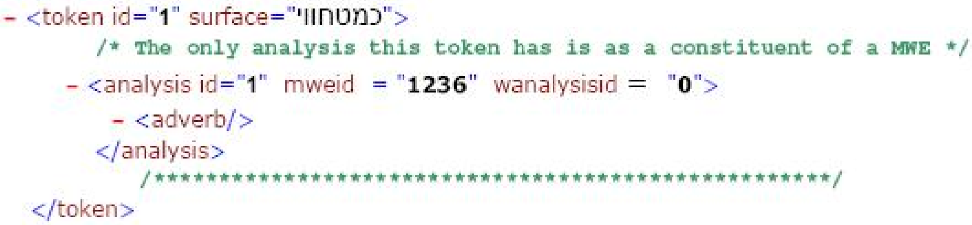
Wrapping hapax legomenas: Recall from Section 4.2.3 that a *hapax legomena* has no analysis in isolation. However, when it appears inside a MWE, the *post processor* generates for it a MWE analysis similar to that of "regular" constituents. Figure 32 depicts the full analysis of the *hapax legomena kmTxwwi* as a constituent of the MWE *kmTxwwi kšt* "a stone's throw". Note that the *kmTxwwi* has only one anlysis, which is a MWE analysis with *wanalysisid*="0".

Wrapping words with non-standard inflections: Recall from Section 4.2.4 that some MWE constituents can appear in an inflected form which is ungrammatical outside the MWE. We followthe notations introduced in Section 4.2.4 to produce analyses for constituents with irregular inflections. We keep the analyses as similar as possible to the analyses of "regular" constituents, by using



```
-<token id="1" surface="עורכות">
   -<analysis id="1">
      - <base dottedLexiconItem="עורך" lexiconItem="עורך" lexiconPointer="8174" transliteratedLexiconItem="ywrk">
           <noun definiteness="false" gender="feminine" number="plural" register="formal" status="absolute" />
        </base>
     </analysis>
   - <analysis id="2">
     - <base dottedLexiconItem="עורך"  lexiconItem="עורך"  lexiconPointer="8174" transliteratedLexiconItem="ywrk">
          <noun definiteness="false" gender="feminine" number="plural" register="formal" status="construct" />
        </base>
     </analysis>
   - <analysis id="3">
     - <base dottedLexiconItem="עָרַך" lexiconItem="ערך" lexiconPointer="8360" transliteratedLexiconItem="yrk">
         <participle binyan="Pa'al" definiteness="false" gender="feminine" number="plural" person="any" register="formal"
           root= "ערך"  status="construct" />
        </base>
     </analysis>
   - <analysis id="4">
     - <base dottedLexiconItem="עָרַך" lexiconItem="ערך" lexiconPointer="8360" transliteratedLexiconItem="yrk">
          <participle binyan="Pa'al" definiteness="false" gender="feminine" number="plural" person="any" register="formal"
            root="ערך"  status="absolute" />
        </base>
     </analysis>
   /*  The MWE analysis */
   - <analysis id="5" mweid  = "333"  wanalysisid =  "2">
    - <noun  definiteness="true" gender="feminine" number="plural"  status="absolute" />
     </analysis>
   /*******************/
 </token>
- <token id="2" surface="הדין">
   - <analysis id="1">
     - <base dottedLexiconItem="דִין" lexiconItem="דין" lexiconPointer="5208" transliteratedLexiconItem="din">
          <noun definiteness="true" gender="masculine" number="singular" register="formal" status="absolute" />
     </base>
    </analysis>
   /*  The MWE analysis */
   - <analysis id="2" mweid  = "333" wanalysisid = "1" />
    /*********************/
  </token>
```

Figure 31: The full analysis of *ywrkwt hdin*

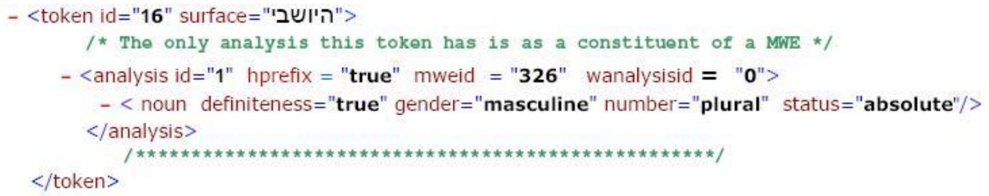


```
- <token id="1" surface="כמטחווי">
        /* The only analysis this token has is as a constituent of a MWE */
     - <analysis id="1"  mweid  = "1236"  wanalysisid =  "0">
        - <adverb/>
        </analysis>
           /*****************************************************/
  </token>
```

Figure 32: The full analysis of *kmTxwwi* as a constituent of the MWE *kmTxwwi kšt*

the same format and attributes and introducing new attributes where needed. Figure 33 depicts the full analysis of the token "hiwšbi" as a constituent of the

MWE *iwšb raš*. Note the attribute *hprefix* which has the value **true**. This attribute has the same functionality described in Section 4.2.4 . Note also that, as in the case of *hapax legomenas*, *wanalysisid*="0".



```
- <token id="16" surface="היושבי">
        /* The only analysis this token has is as a constituent of a MWE */
      - <analysis id="1"  hprefix = "true"  mweid  = "326"  wanalysisid =  "0">
         - < noun  definiteness="true" gender="masculine" number="plural"  status="absolute"/>
        </analysis>
           /****************************************************/
  </token>
```

Figure 33: The full analysis of *hiwšbi* as a constituent of the MWE *hiwšbi raš*

5 Identification and extraction of noun compounds

In this section we describe a system that identifies noun compounds in Hebrew text, and extracts them in order to extend the lexicon. The text is first morphologically analyzed and disambiguated. Then, all NNCs (see Section 3.5) are extracted from the morphologically disambiguated text. For each candidate noun compound we define a set of features based on the idiosyncratic morphological and syntactic prorperties defined in Section 3.5. These features are fed to a support vector machine classifier which is then used to identify the noun compounds in the list of NNCs.

5.1 Resources

We use the Corpus of Contemporary Hebrew[12] (Itai and Wintner, 2008) which can be partitioned according to its source to four parts:

Knesset corpus: The corpus contains the Knesset (Israeli parliament) session protocols from 2004-2005.

Harretz corpus: The corpus contains articles from the Haaretz newspaper from 1991.

The Marker corpus The corpus contains financial articles from the TheMarker newspaper from 2002.

Arutz 7 corpus The corpus contains newswire articles from the Arutz 7 news channel from 2001-2006.

Table 1 shows the number of tokens in each corpus and the total number of tokens in the corpora that we use.

Corpus	Number of tokens
Knesset	12,742,879
Harretz	463,085
The Marker	684,801
Arutz 7	7,714,309
All corpora	21,605,074

Table 1: Number of tokens in each corpus.

The entire corpus was morphologically analyzed (Yona and Wintner, 2007) and disambiguated[13] (Bar-haim, Sima'an, and Winter, 2008). From the morphologically disambiguated corpus, we extract all bigrams in which the first token is a noun in the construct state and the second token is a noun in the absolute state, i.e. all NNCs (all nouns with the same citation form are considered identical).

[12] Available via http://www.mila.cs.technion.ac.il/english/resources/corpora/.

[13] We actually use a part of speech tagger rather than a morphological disambiguator. For a given token, all the analyses with the correct part of speech (chosen by the tagger) are considered valid.

5.2 Morphological features

We define a set of features based on the idiosyncratic properties of noun compounds defined in Section 3.5. For each NNC we collect counts which indicate how many times this NNC exhibits the idiosyncratic property in the corpora. Below we list 16 features that we defined for extracting construct nominals. The features are grouped according to the idiosyncratic property that they reflect.

I. The first 8 features reflect the second idiosyncratic morphological property mentioned in Section 3.5.1:

   (1) The number of occurrences of the NNC in which both constituents are in the singular form.

   (2) The number of occurrences of the NNC in which the first constituent is in the singular form and the second constituent is in the plural form.

   (3) The number of occurrences of the NNC in which the first constituent is in the plural form and the second constituent in the singular form.

   (4) The number of occurrences of the NNC in which both constituents are in the plural form.

   (5) The number of occurrences of the head of the NNC in the plural form outside the expression.

   (6) The number of occurrences of the head of the NNC in the singular form outside the expression.

   (7) The number of occurrences of the modifier of the NNC in the plural form outside the expression.

   (8) The number of occurrences of the modifier of the NNC in the singular form outside the expression.

II. The following 2 features reflect the first idiosyncratic syntactic property mentioned in Section 3.5.2:

   (9) Given the NNC N_1 N_2 this feature counts the number of times N_1 šl N_2 appears in the corpus.

   (10) Given the NNC N_1 N_2 this feature counts the number of times N_1 mN_2 appears in the corpus.

III. The following 2 features reflect the second idiosyncratic syntactic property mentioned in Section 3.5.2:

   (11) Given the NNC N_1 N_2 this feature counts the number of times N_1 N_2 wN_3 appears in the corpus, where the noun N_3 is in the indefinite, absolute form.

   (12) Given the NNC N_1 N_2 this feature counts the number of times N_1 N_2 Adj appears in the corpus. The adjective Adj is in the absolute

form and agrees with N_2 (the modifier) on gender and number, while disagreeing with at least one of these attributes with N_1.[14]

We also defined the following 4 features that represent 4 known collocation measures:[15]

13. Pointwise mutual information association measure (PMI).
14. The T-Score association measure.
15. The log-likelihood association measure.
16. Raw frequency of N_1 N_2 in the corpora.

5.3 Methodology

5.3.1 Annotation

We extract all NNCs from the corpus, and filter out all NNCs which occur less than 100 times. The remaining 1060 NNCs were annotated by three annotators. Each annotator had to tag each NNC with one of the 4 following tags:

1. **+ :** A noun compound (MWE).
2. **- :** Not a noun compound.
3. **0 :** Can't decide.
4. **err :** This expression is not a NNC (an error of the morphological disambiguator).

Table 2 summarizes the annotation of the three annotators.

Anotator	+	-	0	err
1	314	332	238	176
2	335	403	179	143
3	400	630	16	14

Table 2: Annotation by different annotators.

We chose a conservative approach in combining the three annotations. First, we eliminated 204 NNCs that were tagged as *err* by at least one annotator. The annotation for the remaining NNCs in the list was combined using the *consensus approach*, i.e., a NNC is tagged only if all annotators agree on one of the tags.

[14] Recall that in Hebrew the adjective agrees on gender, number, and definiteness with the modified noun. We did not check here if the adjective agrees with N_2 on definiteness because as we mentioned in Section 4.2.7 the whole expression inherits definiteness from the modifier. So, checking the agreement on definiteness is not helpful as in our case we are trying to find the cases where the adjective modifies N_2 but not the whole expression.

[15] Definitions and formulas for these 4 collocation measures can be found at http://www.collocations.de/AM/contents.html.

The consensus annotation is given in Table 3. As can be seen from Table 3, the agreement percentage among annotators was 54.09%, indicating that the task is hard (and probably not sufficiently well-defined).

Tag	Count
+	205
-	258
All	463

Table 3: consensus annotations.

5.3.2 Training and evaluation

For each NNC on the annotated list of Section 5.3.1 we create a vector of the 16 features described in Section 5.2. We obtain a list of 463 instances, of which 205 are *positive* examples (noun compounds) and 258 are negative. We use the whole set to train a two class soft margin SVM classifier (Chang and Lin, 2001) with a radial basis function (RBF) kernel. We experiment with different combinations of features, where for each combination a classifier is trained to optimize the 10-fold F-score on the development set. First, we describe how the classification accuracy, precision, recall and F-score are computed, then we show the results for different classifiers.

In order to compute the accuracy, precision, recall, and F-Score we combine the classification results of the 10-folds into one classification of the whole data. Then we define four different quantities, and using these quantities we define the measures:

1. **MPP:** Number of positive instances that are classified as positive.
2. **MPN:** Number of positive instances that are classified as negative.
3. **MNP:** Number of negative instances that are classified as positive.
4. **MNN:** Number of negative instances that are classified as negative.

$(1) Precision = \frac{MPP}{MPP+MNP} \cdot 100$
$(2) Recall = \frac{MPP}{MPP+MPN} \cdot 100$
$(3) Fscore = 2 \cdot \frac{Precision \cdot Recall}{Precision+Recall}$
$(4) Accuracy = \frac{MPP+MNN}{MPP+MPN+MNP+MNN} \cdot 100$
We also define accuracy variance (V) over the 10-Folds as:

$$V^2 = \frac{1}{9} \cdot \sum_{i=1}^{10} (A_i - \overline{A})$$

A_i is the accuracy of the i$^{\text{th}}$ Fold and $\overline{A}$ is the average accuracy of all folds.

5.3.3 Results

The results of the different classifiers that we trained are given in Table 4. For each classifier the 10-Fold accuracy, precision, recall, and F-score are given. *PMI* stands for pointwise mutual information, while *IMSF* stands for *Idiosyncratic morphological and syntactic features*, which are the features 1-12 from Section 5.2. The first four rows of the table give the results of classifiers trained using the four different collocation measure features. From the results it is apparent that the *Frequency* classifier performs the worst with an F-Score of 41.90%. The *PMI* classifier performs the best among the collocation measures with an F-Score of 60.21%, so we set it as our baseline.

For all other classifiers the gain or loss, in any of the four measures, compared with our baseline, is given inside parentheses. We obtain some improvement over the baseline using combinations of collocation measures. Training a classifier using both T-score and log-liklehood (T-score+LogLikle) gives us a 2.6321% improvment in F-score over the baseline, and an imporvment of 3.1151% in the case of PMI and log-liklehood (PMI+LogLikle).

Using the idiosyncratic morphological and syntactic features (IMSF) (without the collocation measures) we obtain a significant improvement of 16.22% over the baseline. Further improvement of 1.2% is achieved by combining IMSF with log-liklehood. Using the IMSF with PMI reduces F-score by 0.24%. The approach that we use for finding the feature set of the last classifier in Table 4, which yields the best F-Score, is described in Section 5.3.4. In summary, combining our linguistically-informed classifier with a naive collocation measure results in an accuracy of over 80%, reflecting a reduction of 36.16% relative in the classification error rate compared with the baseline.

Features	Accuracy	Variance	Precision	Recall	F-score
PMI	67.17	6.73%	64.97	56.09	60.20
Frequency	60.47(-6.69)	7.06%	60.00(-4.97)	32.19(-23.90)	41.90(-18.30)
T-Score	61.98(-5.18)	9.27%	59.86(-5.10)	42.92(-13.17)	50.00(-10.20)
Log-liklehood	69.33(+2.16)	8.49%	71.42(+6.45)	51.21(-4.87)	59.65(-0.55)
T-score+LogLikle	70.62(+3.45)	7.72%	71.42(+6.45)	56.09(0)	62.84(+2.63)
PMI+LogLikle	69.97(+2.80)	5.55%	68.96(+3.99)	58.53(+2.43)	63.32(+3.11)
IMSF	77.75(+10.58)	7.38%	71.98(+7.01)	81.46(+25.36)	76.43(+16.22)
IMSF+PMI	77.32(+10.15)	6.11%	71.18(+6.21)	81.95(+25.85)	76.19(+15.98)
IMSF+LogLik	79.04(+11.87)	7.34%	73.68(+8.71)	81.95(+25.85)	77.59(+17.38)
LogLikle,1-2,4-6,9-10,12	80.77(+13.60)	5.90%	76.85(+11.88)	80.97(+24.87)	78.85(+18.65)

Table 4: The 10-Fold accuracy, precision, recall, and F-Score for classifiers trained using different combinations of features

5.3.4 Finding the optimal feature combination

One way of finding the feature combination with the best 10-Fold F-Score is using the brute force approach, i.e., training a classifier on all possible feature combinations, 65536 in total, and choosing the feature combination with the maximal 10-Fold F-Score. This approach is clearly intractable. So, we followed a more efficient greedy approach, whereby we start by training and optimizing the 10-Fold F-score using one of the collocation measures. Then, other features are added one at a time. After adding the feature the classifier is retrained and the 10-Fold F-score is optimized on the training set. The added feature is kept in the feature set only if adding it improves (or does not decrease) the 10-Fold F-Score of the current feature set. Otherwise, the feature is skipped and we move to the next feature. Table 5 lists the results of our approach when we start with the log-liklehood feature. We chose log-liklehood as our collocation measure as it gave the highest 10-Fold F-score combined with the IMSF features (see Table 4). In Table 5 LogLikle stands for the log-liklehood feature. For each feature set the increase or decrease in the 10-Fold F-score compared to the previous feature set is given inside parentheses. The results show that the best feature combination gives an improvement of +1.260% over IMSF+LogLik, which is the best classifier in Table 4. We also tried this approach starting from the PMI feature. This showed an improvment over IMSF+PMI from Table 4 but did not beat the IMSF+LogLik feature combination.

Features set	F-score
LogLikle	59.65
LogLikle,1	60.34(+0.68)
LogLikle,1-2	65.42(+5.08)
LogLikle,1-3	64.87(-0.54)
LogLikle,1-2,4	66.66(+1.78)
LogLikle,1-2,4-5	70.0(+3.33)
LogLikle,1-2,4-6	74.37(+4.37)
LogLikle,1-2,4-7	73.78(-0.58)
LogLikle,1-2,4-6,8	73.58(-0.79)
LogLikle,1-2,4-6,9	78.72(+4.35)
LogLikle,1-2,4-6,9-10	78.83(+0.10)
LogLikle,1-2,4-6,9-11	77.37(-1.46)
LogLikle,1-2,4-6,9-10,12	78.85(+0.02)

Table 5: Finding the optimal feature combination using log-liklehood as the starting feature

5.4 Error analysis

We investigated the nature and the properties of errors that the classifiers make. We wanted to see what type of expressions, negative or positive, are harder for the classifiers to predict, and what type of errors does the IMSF help to reduce. The first two columns of Table 6 list the MNP and MPN (see Section 5.3.2) for each of the classifiers. The relative change in the MNP and MPN for the different classifiers compared to PMI baseline is given inside parentheses. The third coloumn lists the percent of the positive candidates that are mispredicted, i.e., positive error rate (PER), and the fourth column shows the percent of negative expressions that are mispredicted, i.e, negative error rate (NER). From the results in Table 6, it is clear that for the baseline PMI classifier, predicting positive examples is harder than predicting negative ones. The error rate for predicting positive examples is 43.90% ,while the error rate for predicting negative expressions is 24.03%. Adding in the IMSF reduces the MNP error significantly by more than 58% while decreasing the MPN error by 19.34%. This indicates that IMSF are more effective against the "hard" cases, the positive expressions. As can be seen in coloum three in Table 6, adding the IMSF reduced the positive prediction error rate from 43.90% to 19.02%, resulting in a calssifier (last line in Table 6) with a symmetric error rate for both types of expressions.

Features set	MPN	MNP	PER	NER
PMI	90	62	43.90%	24.03%
IMSF	38(-57.77%)	65(+4.83%)	18.53%	25.19%
IMSF+PMI	37(-58.88%)	68(+9.67%)	18.04%	26.35%
IMSF+Log	37(-58.88%)	60(-3.22%)	18.04%	23.25%
LogLikle,1-2,4-6,9-10,12(BEST)	39(-56.666%)	50(-19.35%)	19.02%	19.37%

Table 6: MPN, MNP, PER, NER for different classifiers

6 Conclusions and Future Works

In this thesis we investigated the morphological, syntactic, and semantic properties of Hebrew MWEs. Based on the linguistic investigation, we developed an architecture for lexical representation of MWEs, accompanied by a specification of the integration of MWEs into a morphological processor of Hebrew. The architecture can represent MWEs in the lexicon along with their morphological and syntactic properties. We further developed a system that can identify MWEs in texts, and provide morphological analyses for the MWEs using XML. Finally, we developed a system that extracts noun compounds from Hebrew raw text, based on their idiosyncratic morphological and syntactic properties. We showed that combining our linguistically-informed features with collocation measures yields a significant improvement in noun compound classification accuracy over the baseline. Our best linguistically-informed classifier results in a classification accuracy of over 80%, reflecting a reduction of 36.16% in the classification error rate compared with the best collocation measure baseline classifier.

This work can be extended in various directions. The linguistically-informed acquisition system that we demonstrated on noun compounds can be straightforwardly expanded to identify other types of Hebrew MWEs. Examples of such MWEs are Adj-N and N-Adj expressions. Our system can be extended to extract such MWEs by providing the idiosyncratic morphological and syntactic properties, specific for these MWEs, as features for training and testing the classifier. We also believe that the MWE acquisition classification accuracy can be further improved by combining our linguistically-informed features with features such as *translational entropy* defined over aligned parallel copora as in Villada Moirón and Tiedemann (2006), or features that can capture the local linguistic context of the expression using latent semantic analysis as in Katz and Giesbrecht (2006).

References

Alegria, Iñaki, Olatz Ansa, Xabier Artola, Nerea Ezeiza, Koldo Gojenola, and Ruben Urizar. 2004. Representation and treatment of multiword expressions in Basque. In Takaaki Tanaka, Aline Villavicencio, Francis Bond, and Anna Korhonen, editors, *Second ACL Workshop on Multiword Expressions: Integrating Processing*, pages 48–55, Barcelona, Spain, July. Association for Computational Linguistics.

Attia, Mohammed A. 2005. Accommodating multiword expressions in an lfg grammar. The ParGram Meeting, Japan September 2005, September. Mohammed A. Attia The University of Manchester School of Informatics mohammed.attia@postgrad.manchester.ac.uk.

Bannard, Colin. 2007. A measure of syntactic flexibility for automatically identifying multiword expressions in corpora. In *Proceedings of the Workshop on A Broader Perspective on Multiword Expressions*, pages 1–8. Association for Computational Linguistics.

Bannard, Colin, Timothy Baldwin, and Alex Lascarides. 2003. A statistical approach to the semantics of verb-particles. In Diana McCarthy Francis Bond, Anna Korhonen and Aline Villavicencio, editors, *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 65–72.

Baptista, Jorge, Anabela Correia, and Graça Fernandes. 2004. Frozen sentences of portuguese: Formal descriptions for nlp. In Takaaki Tanaka, Aline Villavicencio, Francis Bond, and Anna Korhonen, editors, *Second ACL Workshop on Multiword Expressions: Integrating Processing*, pages 72–79, Barcelona, Spain, July. Association for Computational Linguistics.

Bar-haim, Roy, Khalil Sima'an, and Yoad Winter. 2008. Part-of-speech tagging of Modern Hebrew text. *Natural Language Engineering*. To appear.

Berman, Ruth A. and Doreet ravid. 1986. Dictionary degree of construct state nominals. *Hebrew Linguistics*, 24.

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

Chang, Chih-Chung and Chih-Jen Lin, 2001. *LIBSVM: a library for support vector machines.* Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

Church, K. W. and P. Hanks. 1989b. Word association norms, mutual information and lexicography (rev). *Computational Linguistics*, 19(1):22–29.

Church, K.W. and R.L. Mercer. 1993. Introduction to the special issue on computational linguistic using large corpora. *Computational Linguistics*, 19:1–24.

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

Dowdle, Harold L. 1967. Observations on the uses of a and de in spanish. *Hispania*, 50(2):329–334, May.

Dunning, T. 1993. Accurate methods for the statistics of surprise and coincidence. computational linguistics. *Computational Linguistics*, 19:61–74.

Erman, Britt and Beatrice Warren. 2000. The idiom principle and the open choice principle. *Text*, 1(1):29–62, March.

Fabri, Ray. 2007. Compounding in maltese. October.

Fazly, Afsaneh and Suzanne Stevenson. 2006. Automatically constructing a lexicon of verb phrase idiomatic combinations. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 337–344.

Fellbaum, Christiane, editor. 1998. *WordNet: An Electronic Lexical Database.* Language, Speech and Communication. MIT Press.

Gadish, Ronit, editor. 2001. *Klalei ha-Ktiv Hasar ha-Niqqud.* Academy for the Hebrew Language, 4th edition. In Hebrew.

Grégoire, Nicole. 2007. Design and implementation of a lexicon of dutch multiword expressions. In *Proceedings of the Workshop on A Broader Perspective on Multiword Expressions*, pages 17–24, Prague, Czech Republic, June. Association for Computational Linguistics.

Hashimoto, Chikara and Daisuke Kawahara. 2008. Construction of an idiom corpus and its application to idiom identification based on wsd incorporating idiom-specific features. In *EMNLP '08: Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 992–1001, Morristown, NJ, USA. Association for Computational Linguistics.

Itai, Alon and Shuly Wintner. 2008. Language resources for Hebrew. *Language Resources and Evaluation*, 42:75–98, March.

Jackendoff, Ray. 1997. *The Architecture of the Language Faculty.* MIT Press, Cambridge, USA.

Katz, Graham and Eugenie Giesbrecht. 2006. Automatic identification of non-compositional multi-word expressions using latent semantic analysis. In *MWE '06: Proceedings of the Workshop on Multiword Expressions*, pages 12–19, Morristown, NJ, USA. Association for Computational Linguistics.

Oflazer, Kemal, Özlem Çetinoğlu, and Bilge Say. 2004. Integrating morphology with multi-word expression processing in Turkish. In Takaaki Tanaka, Aline Villavicencio, Francis Bond, and Anna Korhonen, editors, *Second ACL Workshop on Multiword Expressions: Integrating Processing*, pages 64–71, Barcelona, Spain, July. Association for Computational Linguistics.

Paltiely, Rachel and Michal Ephrat. 2006. Idioms in hebrew- properties and characteristics. Master's thesis, University of Haifa, Mount Carmel, Haifa.

Pecina, Pavel. 2005. An extensive empirical study of collocation extraction methods. In *Proceedings of the ACL Student Research Workshop*, pages 13–18, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Pecina, Pavel. 2008. A machine learning approach to multiword expression extraction. In *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions*.

Ramisch, Carlos, Paulo Schreiner, Marco Idiart, and Alline Villavicencio. 2008. An evaluation of methods for the extraction of multiword expressions. In *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions*.

Resnik, P. 1993. Selection and information: A class-based approach to lexical relationships.

Sag, Ivan, Timothy, Francis Bond, Ann Copestake, and Dan 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)*, pages 1–15, Mexico City, Mexico.

Shimron, Joseph, editor. 2003. *Language Processing and Acquisition in Languages of Semitic, Root-Based, Morphology*. Number 28 in Language Acquisition and Language Disorders. John Benjamins.

Van de Cruys, Tim and Begoña Villada Moirón. 2007. Semantics-based multiword expression extraction. In *Proceedings of the Workshop on A Broader Perspective on Multiword Expressions*, pages 25–32, Prague, Czech Republic, June. Association for Computational Linguistics.

van der Vlist, Eric. 2002. *XML Schema*. O'Reilly.

Villada Moirón, Begoña and Jörg Tiedemann. 2006. Identifying idiomatic expressions using automatic word alignment. In *Proceedings of the EACL 2006 Workshop on Multi-word-expressions in a multilingual context*. Association for Computational Linguistics.

Villavicencio, Aline, Francis Bond, Anna Korhonen, and Diana 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, Aline, Ann Copestake, Benjamin Waldron, and Fabre Lambeau. 2004. Lexical encoding of mwes. In Takaaki Tanaka, Aline Villavicencio, Francis Bond, and Anna Korhonen, editors, *Second ACL Workshop on Multiword Expressions: Integrating Processing*, pages 80–87, Barcelona, Spain, July. Association for Computational Linguistics.

Yona, Shlomo and Shuly Wintner. 2007. A finite-state morphological grammar of Hebrew. *Natural Language Engineering*. To appear.

A Well-formed formulas

Below are the syntax and semantics of the formulas that specify possible inflected forms of MWE components.

Syntax:

1. Every pair *attr*=**val** or *elem*= {**false or true**}. The string *attr* is a name of an attribute, **val** is one of its possible values, and *elem* is an element. They are all defined in *Hebrew Corpus XML Schema*.
2. If α is a formula then $\mathbf{!}\alpha$, and $\mathbf{(}\alpha\mathbf{)}$ are formulas.
3. If α and β are formulas, then $\alpha \wedge \beta$, $\alpha \vee \beta$ are formulas.

Semantics: A formula evaluated on a given word form *wf* (stored in the *database of inflected forms*) may have one of two different values, **true** or **false**. The value of a formula on a word form *wf* (which includes its encoded properties in the database) is computed as follows:

1. The formula *attr*=**val** is true for the form *wf* iff the value of *attr* attribute of *wf* is equal to **val**. The formula *elem* = **true** is true for the form *wf* iff the element *elem* is a sub-element of the (encoded) analysis of *wf*. The formula *elem* = **false** is true for the form *wf* iff the element *elem* is **not** a sub-element of the analysis of *wf*.
2. The unary operators defined in 2, and the binary operators defined in 3 have the same semantics that they have in propositional logic.

B Variation in the morphological inflection

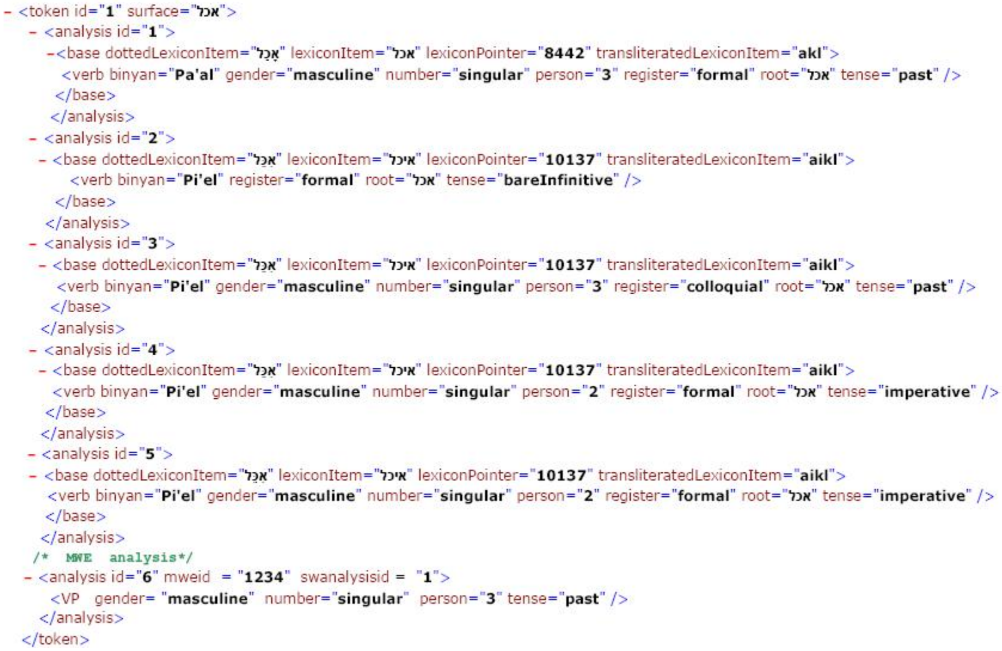
As mentioned in Section 4.2.5, the morphological inflections that a constituent undergoes can vary as a result of the change in its place within the MWE. In order to account for this variation, we add an attribute *inflect* to *atomid* elements. This (optional) attribute was defined earlier as an attribute of *bbase*. Let *Winflect* be this attribute, and let *Binflect* be the *inflect* attribute of the *atom* specified by *atomid*. The values of *Winflect* and *Binflect* together define all possible forms of the constituent. The attribute *Winflect*, if it appears, must have a formula as a value (it cannot have the value **none**). If *Winflect* does not appear then the possible forms are defined by the value of *Binflect*. If *Winflect* does appear, then let a formula *formula1* be its value. In this case, the possible forms depend on the value of *Binflect* as follows:

- If *Binflect* does not appear, or has the value **none**, the forms are defined by *formula1*.
- If *Binflect* appears, then let *formula2* be its value. The forms are defined by the conjunction of the two formulas *formula1*$\wedge$*formula2*.

The lexicon entry of *milh nrdpt* is given in Figure 18. The first atom defines the citation form of *milh*, and the second atom defines all the feminine forms of *nrdpt* (definiteness, status, number can have any value). In the first *surfreal*, the word *milh* is not specified for the attributes *inflect*, which means that it appears in the citation form (defined by the first *atom*). The second word in this *surfreal* has an *inflect* attribute with a formula as a value. The conjunction of this formula with the formula *gender*="feminine" (appearing in the second *atom*) defines all the possible forms of this word, which are the indefinite, absolute, singular, and feminine form of the word *nrdpt*.

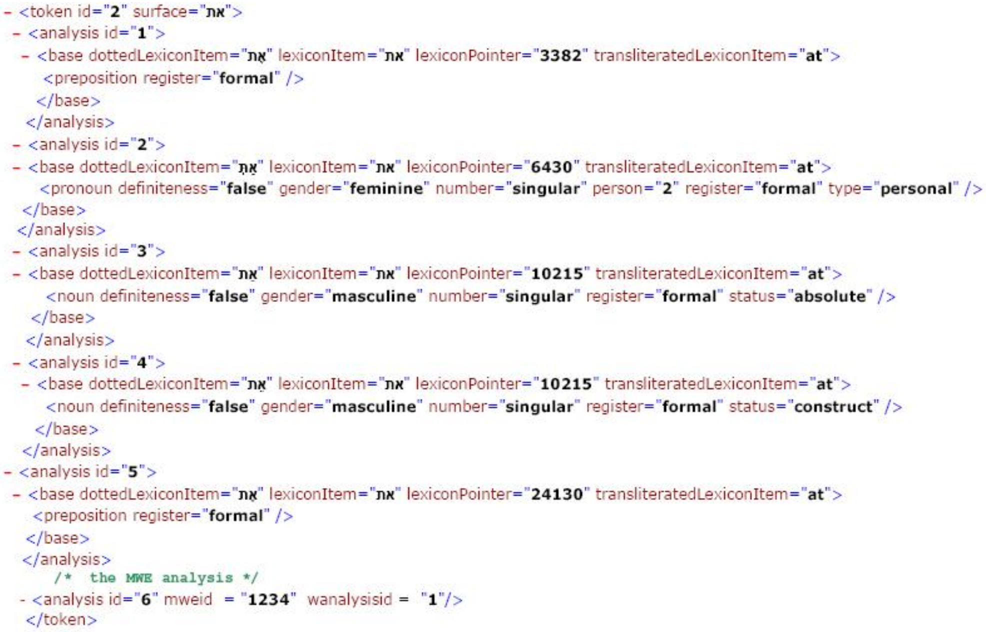
C More examples

Figures 34, 35, 36, 37, 38, and 39 depict the full analysis of the phrase *akl at kl iribiw bli mlx* (lit. "(he) ate all his opponents without salt") "(he) defeated all his opponents easily", which is an instance of the MWE *akl at — bli mlx*. Note in Figure 36 that *islot*="true". This designates that both words *kl*, *irbiw* complement an *open slot* in the MWE.



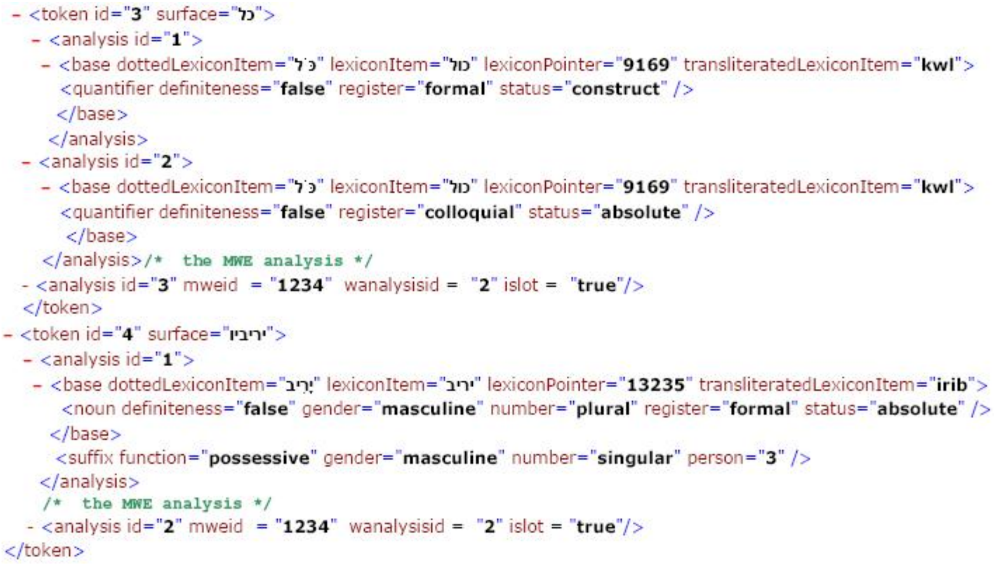
```
- <token id="1" surface="אכל">
  - <analysis id="1">
    -<base dottedLexiconItem="אָכַל" lexiconItem="אכל" lexiconPointer="8442" transliteratedLexiconItem="akl">
      <verb binyan="Pa'al" gender="masculine" number="singular" person="3" register="formal" root="אכל" tense="past" />
     </base>
    </analysis>
  - <analysis id="2">
   - <base dottedLexiconItem="אִכֵּל" lexiconItem="איכל" lexiconPointer="10137" transliteratedLexiconItem="aikl">
       <verb binyan="Pi'el" register="formal" root="אכל" tense="bareInfinitive" />
     </base>
    </analysis>
  - <analysis id="3">
   - <base dottedLexiconItem="אִכֵּל" lexiconItem="איכל" lexiconPointer="10137" transliteratedLexiconItem="aikl">
      <verb binyan="Pi'el" gender="masculine" number="singular" person="3" register="colloquial" root="אכל" tense="past" />
     </base>
    </analysis>
  - <analysis id="4">
   - <base dottedLexiconItem="אִכֵּל" lexiconItem="איכל" lexiconPointer="10137" transliteratedLexiconItem="aikl">
      <verb binyan="Pi'el" gender="masculine" number="singular" person="2" register="formal" root="אכל" tense="imperative" />
     </base>
    </analysis>
  - <analysis id="5">
  - <base dottedLexiconItem="אִכֵּל" lexiconItem="איכל" lexiconPointer="10137" transliteratedLexiconItem="aikl">
     <verb binyan="Pi'el" gender="masculine" number="singular" person="2" register="formal" root="אכל" tense="imperative" />
     </base>
    </analysis>
   /*  MWE  analysis*/
  - <analysis id="6" mweid  = "1234"  swanalysisid =  "1">
     <VP   gender= "masculine"  number="singular"  person="3" tense="past" />
    </analysis>
 </token>
```

Figure 34: The full analysis of *akl*



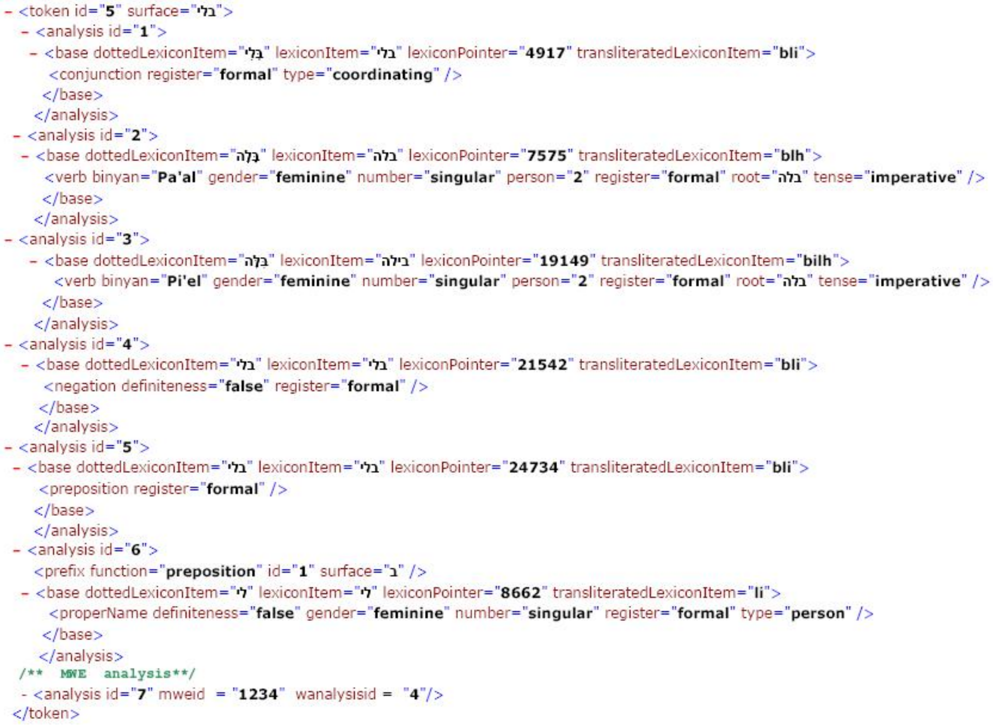
```
- <token id="2" surface="את">
 - <analysis id="1">
  - <base dottedLexiconItem="אֵת" lexiconItem="את" lexiconPointer="3382" transliteratedLexiconItem="at">
     <preposition register="formal" />
    </base>
   </analysis>
 - <analysis id="2">
 - <base dottedLexiconItem="אַתְּ" lexiconItem="את" lexiconPointer="6430" transliteratedLexiconItem="at">
    <pronoun definiteness="false" gender="feminine" number="singular" person="2" register="formal" type="personal" />
  </base>
 </analysis>
 - <analysis id="3">
 - <base dottedLexiconItem="אֵת" lexiconItem="את" lexiconPointer="10215" transliteratedLexiconItem="at">
     <noun definiteness="false" gender="masculine" number="singular" register="formal" status="absolute" />
   </base>
   </analysis>
 - <analysis id="4">
  - <base dottedLexiconItem="אֵת" lexiconItem="את" lexiconPointer="10215" transliteratedLexiconItem="at">
     <noun definiteness="false" gender="masculine" number="singular" register="formal" status="construct" />
    </base>
  </analysis>
- <analysis id="5">
 - <base dottedLexiconItem="אֵת" lexiconItem="את" lexiconPointer="24130" transliteratedLexiconItem="at">
   <preposition register="formal" />
   </base>
  </analysis>
     /*  the MWE analysis */
 - <analysis id="6" mweid  = "1234"  wanalysisid =  "1"/>
  </token>
```

Figure 35: The full analysis of *at*



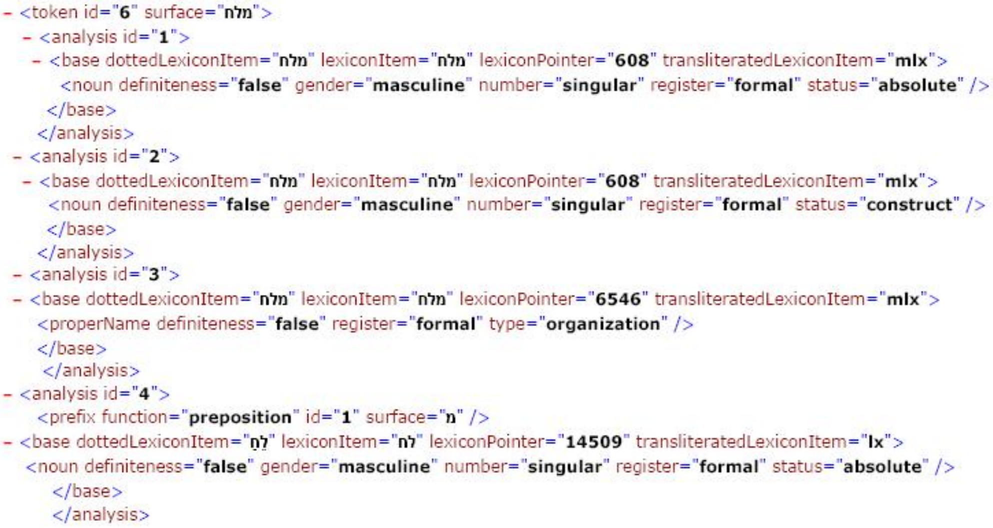
```
 - <token id="3" surface="כל">
   - <analysis id="1">
    - <base dottedLexiconItem="כֹּל" lexiconItem="כול" lexiconPointer="9169" transliteratedLexiconItem="kwl">
       <quantifier definiteness="false" register="formal" status="construct" />
       </base>
      </analysis>
  - <analysis id="2">
    - <base dottedLexiconItem="כֹּל" lexiconItem="כול" lexiconPointer="9169" transliteratedLexiconItem="kwl">
       <quantifier definiteness="false" register="colloquial" status="absolute" />
        </base>
     </analysis>/*  the MWE analysis */
  - <analysis id="3" mweid  = "1234"  wanalysisid =  "2" islot =  "true"/>
  </token>
- <token id="4" surface="יריביו">
  - <analysis id="1">
   - <base dottedLexiconItem="יָרִיב" lexiconItem="יריב" lexiconPointer="13235" transliteratedLexiconItem="irib">
       <noun definiteness="false" gender="masculine" number="plural" register="formal" status="absolute" />
     </base>
      <suffix function="possessive" gender="masculine" number="singular" person="3" />
    </analysis>
    /*  the MWE analysis */
   - <analysis id="2" mweid  = "1234"  wanalysisid =  "2" islot = "true"/>
</token>
```

Figure 36: The full analysis of *kl* and *iribiw*



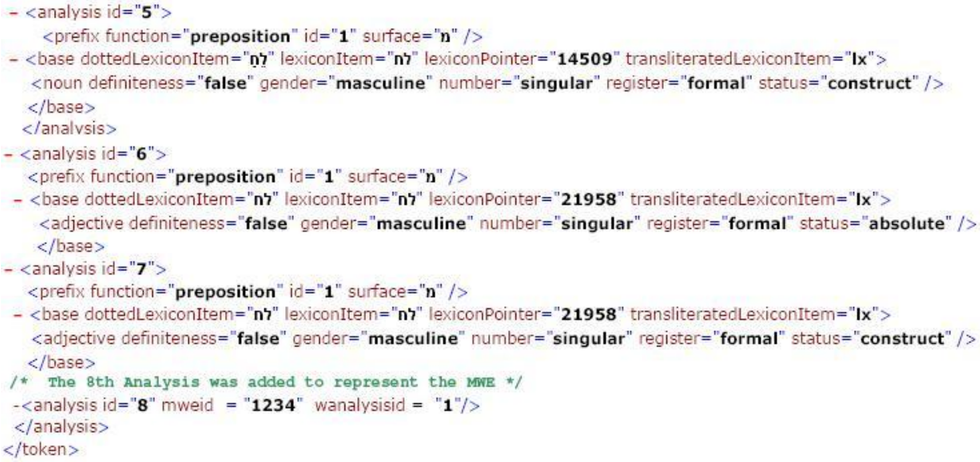
```
- <token id="5" surface="בלי">
  - <analysis id="1">
    - <base dottedLexiconItem="בְּלִי" lexiconItem="בלי" lexiconPointer="4917" transliteratedLexiconItem="bli">
        <conjunction register="formal" type="coordinating" />
      </base>
     </analysis>
   - <analysis id="2">
    - <base dottedLexiconItem="בָּלָה" lexiconItem="בלה" lexiconPointer="7575" transliteratedLexiconItem="blh">
       <verb binyan="Pa'al" gender="feminine" number="singular" person="2" register="formal" root="בלה" tense="imperative" />
      </base>
     </analysis>
- <analysis id="3">
    - <base dottedLexiconItem="בִּלָּה" lexiconItem="בילה" lexiconPointer="19149" transliteratedLexiconItem="bilh">
        <verb binyan="Pi'el" gender="feminine" number="singular" person="2" register="formal" root="בלה" tense="imperative" />
      </base>
     </analysis>
- <analysis id="4">
   - <base dottedLexiconItem="בלי" lexiconItem="בלי" lexiconPointer="21542" transliteratedLexiconItem="bli">
       <negation definiteness="false" register="formal" />
      </base>
     </analysis>
- <analysis id="5">
  - <base dottedLexiconItem="בלי" lexiconItem="בלי" lexiconPointer="24734" transliteratedLexiconItem="bli">
      <preposition register="formal" />
     </base>
     </analysis>
  - <analysis id="6">
      <prefix function="preposition" id="1" surface="ב" />
   - <base dottedLexiconItem="לי" lexiconItem="לי" lexiconPointer="8662" transliteratedLexiconItem="li">
        <properName definiteness="false" gender="feminine" number="singular" register="formal" type="person" />
      </base>
      </analysis>
  /**  MWE  analysis**/
   - <analysis id="7" mweid  = "1234"  wanalysisid =  "4"/>
  </token>
```

Figure 37: The full analysis of *bli*



```
- <token id="6" surface="מלח">
  - <analysis id="1">
    - <base dottedLexiconItem="מלח" lexiconItem="מלח" lexiconPointer="608" transliteratedLexiconItem="mlx">
        <noun definiteness="false" gender="masculine" number="singular" register="formal" status="absolute" />
      </base>
     </analysis>
 - <analysis id="2">
   - <base dottedLexiconItem="מלח" lexiconItem="מלח" lexiconPointer="608" transliteratedLexiconItem="mlx">
      <noun definiteness="false" gender="masculine" number="singular" register="formal" status="construct" />
      </base>
     </analysis>
 - <analysis id="3">
 - <base dottedLexiconItem="מלח" lexiconItem="מלח" lexiconPointer="6546" transliteratedLexiconItem="mlx">
     <properName definiteness="false" register="formal" type="organization" />
     </base>
      </analysis>
- <analysis id="4">
     <prefix function="preposition" id="1" surface="מ" />
- <base dottedLexiconItem="לַח" lexiconItem="לח" lexiconPointer="14509" transliteratedLexiconItem="lx">
   <noun definiteness="false" gender="masculine" number="singular" register="formal" status="absolute" />
      </base>
      </analysis>
```

Figure 38: Analysis 1-4 of *mlx*



```
- <analysis id="5">
     <prefix function="preposition" id="1" surface="מ" />
 - <base dottedLexiconItem="לַחַ" lexiconItem="לח" lexiconPointer="14509" transliteratedLexiconItem="lx">
    <noun definiteness="false" gender="masculine" number="singular" register="formal" status="construct" />
   </base>
  </analvsis>
- <analysis id="6">
   <prefix function="preposition" id="1" surface="מ" />
 - <base dottedLexiconItem="לח" lexiconItem="לח" lexiconPointer="21958" transliteratedLexiconItem="lx">
     <adjective definiteness="false" gender="masculine" number="singular" register="formal" status="absolute" />
    </base>
- <analysis id="7">
   <prefix function="preposition" id="1" surface="מ" />
 - <base dottedLexiconItem="לח" lexiconItem="לח" lexiconPointer="21958" transliteratedLexiconItem="lx">
    <adjective definiteness="false" gender="masculine" number="singular" register="formal" status="construct" />
   </base>
 /*  The 8th Analysis was added to represent the MWE */
 -<analysis id="8" mweid  = "1234"  wanalysisid =  "1"/>
 </analysis>
</token>
```

Figure 39: Analysis 5-8 of *mlx*