Identifying Multi-word Expressions by Leveraging Morphological and Syntactic Idiosyncrasy

Hassan Al-Haj
Language Technologies Institute
Carnegie Mellon University
hhaj@cs.cmu.edu

Shuly Wintner
Department of Computer Science
University of Haifa
shuly@cs.haifa.ac.il

Abstract

Multi-word expressions constitute a significant portion of the lexicon of every natural language, and handling them correctly is mandatory for various NLP applications. Yet such entities are notoriously hard to define, and are consequently missing from standard lexicons and dictionaries. Multi-word expressions exhibit idiosyncratic behavior on various levels: orthographic, morphological, syntactic and semantic. In this work we take advantage of the morphological and syntactic idiosyncrasy of Hebrew noun compounds and employ it to extract such expressions from text corpora. We show that relying on linguistic information dramatically improves the accuracy of compound extraction, reducing over one third of the errors compared with the best baseline.

1 Introduction

Multi-word expressions (MWEs) are notoriously hard to define. They span a range of constructions, from completely frozen, semantically opaque idiomatic expressions, to frequent but morphologically productive and semantically compositional collocations. Various linguistic processes (orthographic, morphological, syntactic, semantic, and cognitive) apply to MWEs in idiosyncratic ways. Notably, MWEs blur the distinction between the lexicon and the grammar, since they often have some properties of words and some of phrases.

In this work we define MWEs as expressions whose linguistic properties (morphological, syntactic or semantic) are not directly derived from the properties of their word constituents. This is a functional definition, driven by a practical motivation: any natural language processing (NLP) application that cares about morphology, syntax or semantics must consequently store MWEs in the lexicon.

MWEs are numerous and constitute a significant portion of the lexicon of any natural language. They are a heterogeneous class of constructions with diverse sets of characteristics. Morphologically, some MWEs allow some of their constituents to freely inflect while restricting (or even preventing) the inflection of other constituents. MWEs may allow constituents to undergo non-standard morphological inflections that they would not undergo in isolation. Some MWEs contain words that never occur outside the context of the MWE. Syntactically, some MWEs appear in one rigid pattern (and a fixed order), while others permit various syntactic transformations. 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.

These morphological, syntactic and semantic idiosyncrasies make MWEs a challenge for NLP applications (Sag et al., 2002). They are even more challenging in languages with complex morphology, because of the unique interaction of morphological and orthographic processes with the lexical specification of MWEs (Oflazer et al., 2004; Alegria et al., 2004).

Because the idiosyncratic features of MWEs cannot be predicted on the basis of their component words, they must be stored in the lexicon of NLP applications. Handling MWEs correctly is beneficial for a variety of applications, including information retrieval, building ontologies, text alignment, and machine translation. Automatic identification and corpus-based extraction of MWEs is thus crucial for such (and several other) applications.

Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), pages 10–18,
Beijing, August 2010

In this work we describe an approach that leverages the morphological and syntactic idiosyncrasy of a certain class of Hebrew[1] MWEs, namely noun compounds, to help identify such expressions in texts. While the main contribution of this work is a system that can distinguish between MWE and non-MWE instances of a particular construction in Hebrew, thereby facilitating faster and more accurate integration of MWEs in a large-coverage lexicon of the language, we believe that it carries added value to anyone interested in MWEs. The technique that we propose here should be applicable in principle to any language in which MWEs exhibit linguistically idiosyncratic behavior.

We describe the properties of Hebrew noun-noun constructions in Section 2, and specify the irregularities exhibited by compounds. Section 3 presents the experimental setup and the main results. Compared with the best (collocation-based) baseline, our approach reduces over 30% of the errors, yielding accuracy of over 80%. We discuss related work in Section 4 and conclude with suggestions for future research.

2 Hebrew noun-noun constructions

We focus on Hebrew noun-noun constructions; these are extremely frequent constructions, and while many of them are fully compositional, others, called *noun compounds* (or just *compounds*) here, are clearly MWEs. We first discuss the general construction and then describe the peculiar, idiosyncratic properties of compounds.

2.1 The general case

Hebrew nouns inflect for number (singular and plural) and, when the noun denotes an animate entity, for gender (masculine and feminine). In addition, nouns come in three *states*: indefinite, definite and a *construct* state that is used in genitive constructions. Table 1 demonstrates the paradigm.

A noun-noun construction (henceforth NNC) consists of a construct-state noun, called *head* here, followed by a noun phrase, the *modifier* (Borer, 1988; Borer, 1996; Glinert, 1989).

[1]To facilitate readability we use a transliteration of Hebrew using Roman characters; the letters used, in Hebrew lexicographic order, are *abgdhwzxTiklmns‘pcqršt*.

State	M/Sg	F/Sg	M/Pl	F/Pl
indefinite	*ild*	*ildh*	*ildim*	*ildwt*
definite	*hild*	*hildh*	*hildim*	*hildwt*
construct	*ild*	*ildt*	*ildi*	*ildwt*

Table 1: The noun paradigm, demonstrated on *ild* “child”

The semantic relation between the two is usually, but not always, related to possession (Levi, 1976). Construct-state nouns only occur in the context of NNC, and can never occur in isolation. When a NNC is definite, the definite article is expressed on its modifier (Wintner, 2000).

In the examples below, we explicitly indicate construct-state nouns by the morpheme ‘.CONST’ in the gloss; and definite nouns are indicated by the morpheme ‘the-’. We provide both a literal and a non-literal meaning of the MWE examples. Expressions that have a literal, but not the expected MWE meaning, are preceded by ‘#’.

Example 1 (Noun-noun constructions)

hxlTt *hw‘dh*
*decision.*CONST *the-committee*
“the committee decision”

‘wrk *h‘itwn*
*editor.*CONST *the-journal*
“the journal editor”

‘wrk *din*
*editor.*CONST *law*
“law editor” ⟹ lawyer

bti *xwlim*
*houses.*CONST *patients*
“patient houses” ⟹ hospitals

2.2 Noun compounds: Linguistic properties

While many of the NNCs are free, compositional combinations of words, some are not; we use the term *noun compounds* for the latter group. Compounds typically (but not necessarily) have non-compositional meaning; presumably due to their opaque, more lexical meaning, they also differ from other NNCs in their morphological and syntactic behavior. Some of these distinctive properties are listed below, to motivate the methodology that we propose in Section 3 to distinguish between compounds and non-MWE NNCs.

2.2.1 Limited inflection

When a NNC consists of two nouns, the second can typically occur in either singular or plural form. Compounds often limit the possibilities to only one of those.

Example 2 (No plural form of the modifier)

‘wrki	*h‘itwnim*
*editors-.*CONST	*the-journals*
“the journals’ editors”	

‘wrki	*hdin*
*editors.*CONST	*the-law*
“the law editors” ⟹ the lawyers	

#wrki	*hdinim*
*editors.*CONST	*the-laws*

Example 3 (No singular form of the modifier)

kiwwn	*hrwx*
*direction.*CONST	*the-wind*
“the wind’s direction”	

kiwwn	*hrwxwt*
*direction.*CONST	*the-winds*
“the winds’ direction”	

šwšnt	*h-rwxwt*
*lily.*CONST	*the-winds*
“lily of the winds” ⟹ compass rose	

#šwšnt	*h-rwx*
*lily.*CONST	*the-wind*

2.2.2 Limited syntactic variation

Since NNCs typically denote genitive (possessive) constructions, they can be paraphrased by a construction that uses the genitive preposition *šl* “of” (or, in some cases, other prepositions). These syntactic variants are often restricted in the case of compounds.

Example 4 (Limited paraphrasing)

h‘wrk	*šl*	*h‘itwn*
the-editor	*of*	*the-journal*
“the journal editor”		

#h‘wrk	*šl*	*hdin*
the-editor	*of*	*the-law*

Example 5 (Limited paraphrasing)

m‘il	*cmr*
*coat.*CONST	*wool*
“wool coat”	

m‘il	*mcmr*
coat	*from-wool*
“wool coat”	

cmr	*pldh*
*wool.*CONST	*steel*
“steel wool” ⟹ steel wool	

#cmr	*mpldh*
wool	*from-steel*

2.2.3 Limited syntactic modification

NNCs typically allow adjectival modification of either of their constituents. Since compounds tend to be more semantically opaque, it is often only possible to modify the entire compound, but not any of the constituents. In the following example, note that *‘wrkt* “editor” is feminine, whereas *‘itwn* “journal” is masculine; adjectives must agree on gender with the noun they modify.

Example 6 (Limited adjectival modification)

’wrkt	*h’itwn*
*editor-f.*CONST	*the-journal-m*
“the journal editor”	

‘wrkt	*h‘itwn*	*hxdšh*
*editor-f.*CONST	*the-journal-m*	*the-new-f*
“the new editor of the journal”		

‘wrkt	*h‘itwn*	*hxdš*
*editor-f.*CONST	*the-journal-m*	*the-new-m*
“the editor of the new journal”		

‘wrkt	*hdin*	*hxdšh*
*editor-f.*CONST	*the-law-m*	*the-new-f*
“the new law editor” ⟹ the new lawyer		

#‘wrkt	*hdin*	*hxdš*
*editor-f.*CONST	*the-law-m*	*the-new-m*

2.2.4 Limited coordination

Two NNCs that share a common head can be conjoined using the coordinating conjunction *w* “and”. This possibility is often blocked in the case of compounds.

Example 7 (Limited coordination)

mwsdwt	*xinwk*	*wbriawt*
*institutions.*CONST	*education*	*and-health*
“education and health institutions”		

bti	*spr*
*houses.*CONST	*book*
“book houses” ⟹ schools	

bti	*xwlim*	
*houses.*CONST	*patients*	
"patient houses" $\Longrightarrow$ *hospitals*		
#bti	*spr*	*wxwlim*
*houses.*CONST	*book*	*and-patients*

3 Identification 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. We capitalize on the morphological and syntactic irregularities of noun compounds described in Section 2.2.

Given a large monolingual corpus, the text is first morphologically analyzed and disambiguated. Then, all NNCs (candidate noun compounds) are extracted from the morphologically disambiguated text. For each candidate noun compound we define a set of features (Section 3.3) based on the idiosyncratic morphological and syntactic properties defined in Section 2.2. These features inform a support vector machine classifier which is then used to identify the noun compounds in the set of NNCs with high accuracy (Section 3.5).

3.1 Resources

We use (a subset of) the Corpus of Contemporary Hebrew (Itai and Wintner, 2008) which consists of four sub-corpora: The *Knesset* corpus contains the Israeli parliament proceedings from 2004-2005; the *Haaretz* corpus contains articles from the Haaretz newspaper from 1991; *The-Marker* corpus contains financial articles from the TheMarker newspaper from 2002; and the *Arutz 7* corpus contains newswire articles from 2001-2006. Corpora sizes are listed in Table 2.

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

Table 2: Corpus data

The entire corpus was morphologically analyzed (Yona and Wintner, 2008; Itai and Wintner, 2008) and POS-tagged (Bar-haim et al., 2008); note that no syntactic parser is available for Hebrew. From the morphologically disambiguated corpus, we extract all bi-grams in which the first token is a noun in the construct state and the second token is a noun that is not in the construct state, i.e., all two-word NNC *candidates*.

3.2 Annotation

For training and evaluation, we select the NNCs that occur at least 100 times in the corpus, yielding 1060 NNCs. These NNCs were annotated by three annotators, who were asked to classify them to the following four groups: compounds (+); non-compounds (–); unsure (0); and errors of the morphological processor (i.e., the candidate is not a NNC at all). Table 3 lists the number of candidates in each class.

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

Table 3: NNC classification by annotator

We adopt a conservative approach in combining the three annotations. First, we eliminate 204 NNCs that were tagged as errors by at least one annotator. For the remaining NNCs, a candidate is considered a compound or a non-compound only if all three annotators agree on its classification. This reduces the annotated data to 463 instances, of which 205 are compounds and 258 are clear cases of non-compound NNCs.[2]

3.3 Linguistically-motivated features

We define a set of features based on the idiosyncratic properties of noun compounds defined in Section 2.2. For each candidate NNC, we compute counts which reflect the likelihood of it exhibiting one of the linguistic properties.

Refer back to Section 2.2. We focus on the property of limited inflection (Section 2.2.1), and define features 1–8 to reflect it. To reflect limited syntactic variation (Section 2.2.2) we define features 9–10. Feature 11 addresses the phenomenon

[2]This annotated corpus is freely available for download.

of limited coordination (Section 2.2.4). To reflect limited syntactic modification (Section 2.2.3) we define feature 12. .

For each NNC candidate N_1 N_2, the following features are defined:

1. The number of occurrences of the NNC in which both constituents are in singular.
2. The number of occurrences of the NNC in which N_1 is in singular and N_2 is in plural.
3. The number of occurrences of the NNC in which N_1 is in plural and N_2 is in singular.
4. The number of occurrences of the NNC in which both constituents are in plural.
5. The number of occurrences of N_1 in plural outside the expression.
6. The number of occurrences of N_1 in singular outside the expression.
7. The number of occurrences of N_2 in plural outside the expression.
8. The number of occurrences of N_2 in singular outside the expression.
9. The number of occurrences of N_1 *šl* N_2 "N_1 of N_2" in the corpus.
10. The number of occurrences of N_1 *m* N_2 "N_1 from N_2" in the corpus.
11. The number of occurrences of N_1 N_2 *w* N_3 "N_1 N_2 and N_3" in the corpus, where N_3 is an indefinite, non-construct-state noun.
12. The number of occurrences of N_1 N_2 *Adj* in the corpus, where the adjective *Adj* agrees with N_2 on both gender and number, while disagreeing with N_1 on at least one of these attributes.

We also define four features that represent known collocation measures (Evert and Krenn, 2001): Point-wise mutual information (PMI); T-Score; log-likelihood; and the raw frequency of N_1 N_2 in the corpus.[3]

[3] A detailed description of these measures is given by Manning and Schütze (1999, Chapter 5); see also `http://www.collocations.de/`, where several other association measures are discussed as well.

3.4 Training and evaluation

For each NNC in the annotated set of Section 3.2 we create a vector of the 16 features described in Section 3.3 (12 linguistically-motivated features plus four collocation measures). We obtain a list of 463 instances, of which 205 are positive examples (noun compounds) and 258 are negative. We use this set for training and evaluation of a two class soft margin SVM classifier (Chang and Lin, 2001) with a radial basis function kernel. We experiment below with different combinations of features, where for each combination we use 10-fold cross-validation over the 463 NNCs to evaluate the classifier. We report Precision, Recall, F-score and Accuracy (averaged over the 10 folds).

3.5 Results

The results of the different classifiers that we trained are given in Table 4. The first four rows of the table show the performance of classifiers trained using each of the four different collocation measure features alone. Both PMI and Log-likelihood outperform the other collocation measures, with an F-score of 60, which we consider our baseline. We also report the performance of two combinations of collocation measures, which yield small improvement. The best combinations provide accuracy of about 70% and F-score of 63.

The remaining rows report results using the linguistically-motivated features (LMF) of Section 3.3. These features alone yield accuracy of 77.75% and an F-score of 76. Adding also Log-likelihood improves F-score by 1.16 and accuracy by 1.29%. Finally, using Log-likelihood with a subset of the LMF consisting of features 1-2, 4-6, 9-10 and 12 (see below) yields the best results, namely accuracy of over 80% and F-score of 78.85, reflecting a reduction of over one third in classification error rate compared with the baseline.

3.6 Optimizing feature combination

We search for the combination of linguistically-motivated features that would yield the best performance. Training a classifier on all possible feature combinations is clearly infeasible. Instead, we follow a more efficient greedy approach, whereby we start with the best collocation mea-

Features	Accuracy	Precision	Recall	F-score
PMI	67.17	64.97	**56.09**	**60.20**
Frequency	60.47	60.00	32.19	41.90
T-Score	61.98	59.86	42.92	50.00
Log-likelihood	**69.33**	**71.42**	51.21	59.65
T-score+Log-likelihood	**70.62**	**71.42**	56.09	62.84
PMI+Log-likelihood	69.97	68.96	**58.53**	**63.32**
LMF	77.75	71.98	81.46	76.43
LMF+PMI	77.32	71.18	**81.95**	76.19
LMF+Log-likelihood	79.04	73.68	**81.95**	77.59
Log-likelihood+LMF[1-2,4-6,9-10,12]	**80.77**	**76.85**	80.97	**78.85**

Table 4: Results: 10-Fold accuracy, precision, recall, and F-score for classifiers trained using different combinations of features. *LMF* stands for linguistically-motivated features

sure, Log-likelihood, and add other features one at a time, in the order in which they are listed in Section 3.3. After adding each feature the classifier is retrained; the feature is retained in the feature set only if adding it improves the 10-fold F-score of the current feature set.

Table 5 lists the results of this experiment. For each feature set the difference in the 10-fold F-score compared to the previous feature set is listed in parentheses. The results show that the best feature combination improves the F-score by 1.26, compared with using all features. This experiments shows that features 3, 7, 8 and 11 turn out not to be useful, and the classifier is more accurate without them. We also tried this approach with PMI as the starting feature, with very similar results.

Feature set	F-score	
Log-likelihood	59.65	
Log-likelihood,1	60.34	(+0.68)
Log-likelihood,1-2	65.42	(+5.08)
Log-likelihood,1-3	64.87	(-0.54)
Log-likelihood,1-2,4	66.66	(+1.78)
Log-likelihood,1-2,4-5	70.00	(+3.33)
Log-likelihood,1-2,4-6	74.37	(+4.37)
Log-likelihood,1-2,4-7	73.78	(–0.58)
Log-likelihood,1-2,4-6,8	73.58	(–0.79)
Log-likelihood,1-2,4-6,9	78.72	(+4.35)
Log-likelihood,1-2,4-6,9-10	78.83	(+0.10)
Log-likelihood,1-2,4-6,9-11	77.37	(–1.46)
Log-likelihood,1-2,4-6,9-10,12	78.85	(+0.02)

Table 5: Optimizing the set of linguistically-motivated features

4 Related work

There has been a growing awareness in the research community of the problems that MWEs pose, both in linguistics and in NLP (Villavicencio et al., 2005). Recent works address the definition, lexical representation and computational processing of MWEs, as well as algorithms for extracting them from data.

Focusing on acquisition of MWEs, early approaches concentrated on their collocational behavior (Church and Hanks, 1989). Pecina (2008) compares 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. Our results show that this is indeed the case, but the contribution of collocation methods is limited, and more information is needed in order to distinguish frequent collocations from bona fide MWEs.

Other works show that adding linguistic information to collocation measures can improve identification accuracy. Several approaches rely on the semantic opacity of MWEs; but very few semantic resources are available for Hebrew (the Hebrew WordNet (Ordan and Wintner, 2007), the only lexical semantic resource for this language, is small and too limited). Instead, we capitalize on the morphological and syntactic irregularities that MWEs exhibit, using computational resources that are more readily-available.

Ramisch et al. (2008) evaluate a number of association measures on the task of identifying English Verb-Particle Constructions and German Adjective-Noun pairs. They show that adding linguistic information (mostly POS and POS-sequence patterns) to the association measure yields a significant improvement in performance over using pure frequency. We follow this line of research by defining a number of syntactic patterns as a source of linguistic information. In addition, our linguistic features are much more specific to the phenomenon we are interested in, and the syntactic patterns are enriched by morphological information pertaining to the idiosyncrasy of MWEs; we believe that this explains the improved performance compared to the baseline.

Several works address the *lexical fixedness* or *syntactic fixedness* of (certain types of) MWEs in order to extract them from texts. 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 literal expression. Syntactically fixed expressions prohibit (or restrict) syntactic variation.

For example, Van de Cruys and Villada Moirón (2007) use lexical fixedness to extract Dutch Verb-Noun idiomatic combinations (VNICs). Bannard (2007) uses syntactic fixedness to identify English VNICs. 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). While these approaches are in line with ours, they require lexical semantic resources (e.g., a database that determines semantic similarity among words) and syntactic resources (parsers) that are unavailable for Hebrew (and many other languages). Our approach only requires morphological processing, which is more readily-available for several languages.

Another unique feature of our work is that it computationally addresses Hebrew (and, more generally, Semitic) MWEs for the first time. Berman and Ravid (1986) define the *dictionary degree* of noun compounds in Hebrew as their closeness to a single word from a grammatical point of view, as judged by the manner in which they are grasped by language speakers. 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 pragmatic uses of the noun compound and the semantic relation between its constituents define the dictionary degree of the compound. Not having access to semantic and pragmatic knowledge, we are trying to approximate it using morphology.

Attia (2005) proposes methods to process fixed, semi-fixed, and syntactically-flexible *Arabic* MWEs (adopting the classification and the terminology of Sag et al. (2002)). Fabri (2009) 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. He also provides morphological, syntactic, and semantic properties of this group which distinguishes them from other non-compound constructions. Automatic identification of MWEs is not addressed in either of these works.

5 Conclusions and future work

We described a system that can identify Hebrew noun compounds with high accuracy, distinguishing them from non-idiomatic noun-noun constructions. The methodology we advocate is based on careful examination of the linguistic peculiarities of the construction, followed by corpus-based approximation of these properties via a general machine learning algorithm that is fed with features based on the linguistic properties. While our application is limited to a particular construction in a particular language, we are confident that it can be equally well applied to other constructions and other languages, as long as the targeted MWEs exhibit a consistent set of irregular features (especially in the morphology).

This work can be extended in various directions. Addressing other constructions is relatively easy, and requires only a theoretical linguistic investigation of the construction. We are currently interested in extending the system to cope also with Adjective-Noun, Noun-Adjective and Verb-Preposition constructions in Hebrew.

The accuracy of MWE acquisition systems can be further improved by combining our morphological and syntactic features with semantically informed features such as translational entropy computed from a parallel corpus (Villada Moirón and Tiedemann, 2006), or features that can capture the local linguistic context of the expression using latent semantic analysis (Katz and Giesbrecht, 2006). We are currently working on the former direction (Tsvetkov and Wintner, 2010b), utilizing a small Hebrew-English parallel corpus (Tsvetkov and Wintner, 2010a).

Finally, we are interested in evaluating the methodology proposed in this paper to other languages with complex morphology, in particular to Arabic. We leave this direction to future research.

Acknowledgments

This research was supported by THE ISRAEL SCIENCE FOUNDATION (grants No. 137/06, 1269/07). We are grateful to Alon Itai for his continuous help and advice throughout the course of this project, and to Bracha Nir for very useful comments. We also wish to thank Yulia Tsvetkov and Gily Chen for their annotation work.

References

Alegria, Iñaki, Olatz Ansa, Xabier Artola, Nerea Ezeiza, Koldo Gojenola, and Ruben Urizar. 2004. Representation and treatment of multiword expressions in Basque. In Tanaka, Takaaki, 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.

Bar-haim, Roy, Khalil Sima'an, and Yoad Winter. 2008. Part-of-speech tagging of Modern Hebrew text. *Natural Language Engineering*, 14(2):223–251.

Berman, Ruth A. and Dorit Ravid. 1986. Lexicalization of noun compounds. *Hebrew Linguistics*, 24:5–22. In Hebrew.

Borer, Hagit. 1988. On the morphological parallelism between compounds and constructs. In Booij, Geert and Jaap van Marle, editors, *Yearbook of Morphology 1*, pages 45–65. Foris publications, Dordrecht, Holland.

Borer, Hagit. 1996. The construct in review. In Lecarme, Jacqueline, Jean Lowenstamm, and Ur Shlonsky, editors, *Studies in Afroasiatic Grammar*, pages 30–61. Holland Academic Graphics, The Hague.

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

Evert, Stefan and Brigitte Krenn. 2001. Methods for the qualitative evaluation of lexical association measures. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 188–195, Morristown, NJ, USA. Association for Computational Linguistics.

Fabri, Ray. 2009. Compounding and adjective-noun compounds in Maltese. In Comrie, Bernard, Ray Fabri, Elizabeth Hume, Manwel Mifsud, Thomas Stolz, and Martine Vanhove, editors, *Introducing Maltese Linguistics*, volume 113 of *Studies in Language Companion Series*. John Benjamins.

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.

Glinert, Lewis. 1989. *The Grammar of Modern Hebrew*. Cambridge University Press, Cambridge.

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

Katz, Graham and Eugenie Giesbrecht. 2006. Automatic identification of non-compositional multiword expressions using latent semantic analysis. In *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, pages 12–19, Sydney, Australia, July. Association for Computational Linguistics.

Levi, Judith N. 1976. A semantic analysis of Hebrew compound nominals. In Cole, Peter, editor, *Studies in Modern Hebrew Syntax and Semantics*, number 32 in North-Holland Linguistic Series, pages 9–55. North-Holland, Amsterdam.

Manning, Christopher D. and Hinrich Schütze. 1999. *Foundations of statistical natural language processing*. The MIT Press, Cambridge, Mass.

Oflazer, Kemal, Özlem Çetinoğlu, and Bilge Say. 2004. Integrating morphology with multi-word expression processing in Turkish. In Tanaka, Takaaki, 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.

Ordan, Noam and Shuly Wintner. 2007. Hebrew WordNet: a test case of aligning lexical databases across languages. *International Journal of Translation, special issue on Lexical Resources for Machine Translation*, 19(1).

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*.

Sag, Ivan, Timothy Baldwin, 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.

Tsvetkov, Yulia and Shuly Wintner. 2010a. Automatic acquisition of parallel corpora from websites with dynamic content. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, pages 3389–3392. European Language Resources Association (ELRA), May.

Tsvetkov, Yulia and Shuly Wintner. 2010b. Extraction of multi-word expressions from small parallel corpora. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, August.

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.

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.

Wintner, Shuly. 2000. Definiteness in the Hebrew noun phrase. *Journal of Linguistics*, 36:319–363.

Yona, Shlomo and Shuly Wintner. 2008. A finite-state morphological grammar of Hebrew. *Natural Language Engineering*, 14(2):173–190, April.