INTERNATIONAL JOURNAL OF TRANSLATION
Vol. XX, No. XX, XXXX XXXX

Hebrew WordNet: a Test Case of Aligning Lexical Databases across Languages

NOAM ORDAN
Bar Ilan University, Israel

SHULY WINTNER
University of Haifa, Israel

ABSTRACT

We report on the creation of a medium-scale WordNet for Hebrew. We address this task as an instance of building a lexical resource for a new language (Hebrew) in a setting where similar resources exist for other languages, and multilingual requirements call for an alignment of the new resource with the existing ones. We compare the two main paradigms, MultiWordNet and EuroWordNet, with an eye to other minority languages, who might lack, like Hebrew does, basic resources for carrying out such a task. As we show, the scales are tipped to the MultiWordNet paradigm for this very reason. Cast in this paradigm, the Hebrew WordNet is strictly aligned to the English lexicon. Consequently, the discrepancy between the languages has to be dealt with: on the one hand, the new resource has to be faithful to the linguistic data of the language for which it is created; on the other, it has to be aligned with existing resources for unrelated languages. We distinguish between contingent *and* systematic *cases of non-equivalence. For the former, we offer a corpus-based methodology that can be easily applied for any new language for which such a resource is planned. For the latter, we propose systematic solutions, focusing on the cases of gender, passive verbs, and antonyms. Where L2 is more specific in its semantic distinctions (as in the case of gender), we devise a solution which facilitates a full semantic inheritance. Where L2's distinctions are more general (as in passive verbs), our solution is partial and calls for further research. The case of antonyms is fully solved for most parts of speech, but it raises crucial questions regarding the typological bias of WordNet towards English (and other Indo-European languages), which may touch on both psycholinguistics and the feasibility of WordNet for such tasks as machine translation.*

1 INTRODUCTION

WordNet (Fellbaum, 1998) is a computational lexicographical resource which was motivated by psycholinguist concerns but turned out to be instrumental for a variety of computational tasks (Harabagiu, 1998). WordNet is used for information retrieval (Mandala *et al.*, 1998), word sense-disambiguation (Agirre and Rigau, 1996), text categorization (de Buenaga Rodríguez *et al.*, 1997), language generation (Jing, 1998), and semantic annotation (Fellbaum *et al.*, 2001), to name a few examples. Furthermore, the success of the original English WordNet boosted the preparation of similar resources for other languages, and there are currently at least fourty WordNet projects in other languages, completed or underway. There are obviously good reasons for compiling, maintaining and distributing WordNets for new languages. This paper reports on the creation of a medium-sized WordNet for Hebrew, the first Semitic language for which a substantial WordNet has been designed (two preliminary proposals for an Arabic WordNet are discussed by Diab (2004) and Black *et al.* (2006)).

In the next section we overview the general semantic design of WordNet and describe two options for synchronizing a WordNet in one language to a WordNet in another (Hebrew to English in this case). The challenges involved in this task are described in Section 3. We focus on three cases of systematic non-equivalence: *gender*, *passive* verbs, and *antonyms*. We then present our methodology, which is cast in the MultiWordNet paradigm, in Section 4. We conclude with a discussion and directions for future research.

The contribution of this work goes beyond the creation of the Hebrew resource, important as it might be. We believe that our insights will be valuable for a variety of languages. In particular, our methodology facilitates a systematic marking of semantic gaps across languages, and arms the lexicographer with tools for specifying semantic relations in spite of those gaps. This provides a way for using semantic inheritance mechanisms that are built into WordNet across languages.

2 LEXICAL-SEMANTIC RESOURCES IN A MULTILINGUAL WORLD

2.1 WordNet: an online lexical database

WordNet is based on insights taken from psycholinguistics on the one hand and the British school of structural/lexical semantics on the other (Lyons, 1963, 1977; Cruse, 1986). Both make the following points:

(1) There are semantic and lexical relations between lexical items which govern their organization and manifest their meaning; (2) These relations occur more often than not between words belonging to the same part of speech, thus nominal lexical items should be networked with other nominal lexical items, verbal lexical items with verbal ones, etc.

WordNet is not composed of entries in the traditional lexicographical sense. Its atoms, or lexical building blocks, are groups of synonymous lexical items, called *synsets* (synonym sets). WordNet assumes that synonyms grouped in synsets stand for concepts, and that most relations adhere to concepts rather than to single lexical items. For example, the synset {*car*, *auto*, *automobile*, *machine*, *motorcar*} stands for a concept which could be defined as "a motor vehicle with four wheels; usually propelled by an internal combustion engine". This concept is mapped onto another more generic concept defined as "a self-propelled wheeled vehicle that does not run on rails" which the language realizes in another synset, namely {*motor vehicle*, *automotive vehicle*}. Languages do not always realize concepts lexically, which is why WordNet resorts at times to artificial lexical items, like '*bad person*', when it wishes to use it as a node in its network.

WordNet defines several relations over words and over synsets (Miller *et al.*, 1990). We have mentioned the first relation, namely synonymy, which serves as the criterion for grouping lexical items into synsets. In addition, semantic relations hold between synsets, and lexical ones between single lexical items. *Hypernyms* adhere to nouns and verbs only. A hypernym is the generic term used to designate a whole class of specific instances. Y is a hypernym of X if X is a (kind of) Y. Synsets sharing the same immediate hypernym are considered *coordinate-terms* of each other. *Troponyms* adhere to verbs only. A troponym pertains to a verb that expresses a specific manner elaboration of another verb. X is a troponym of Y if to X is to Y in some manner. *Antonyms* adhere to all parts of speech. Other WordNet relations include *entailment*, *holonym--meronym* and *pertainym*.

2.2 Multilingual WordNets

The two main paradigms for compiling multilingual WordNets, or aligning different WordNets to each other, are EuroWordNet (EWN) (Vossen, 1998, pp. 715–728) and MultiWordNet (MWN) (Bentivogli *et al.*, 2002). These paradigms express two different approaches for dealing with synchronization of networks across languages. In the former, a WordNet for each language is built from scratch, and aligning is done only thereafter via an Inter-Lingual Index (ILI), which is taken to be *language independent*. In the latter, all WordNets are aligned as strictly as possible to the American-English version of Princeton WordNet (PWN), under the assumption that most of the concepts are *universally shared*.

The working assumption of MWN is that there are two kinds of lexical idiosyncrasies relevant to representing different WordNets next to each other. The first pertains to *lexical gaps*, where L2 (the target language) lacks a lexical item for expressing a lexicalized concept existing in L1 (the source language). The other pertains to denotation differences, where a lexical item in L2 does exist, but is either more general or more specific than its equivalent in L1. The procedure devised by MWN for handling these idiosyncrasies enables the lexicographers to declare a GAP wherever necessary and then connect between the gapped empty synset of L2 to other synsets using WordNet's common semantic relations. Information is therefore inherited via this GAP to other synsets attached to it.

For example, the Hebrew[1] word '*gnb'* corresponds to the English '*thief*'. However, English distinguishes between a few kinds of thieves which are represented in WordNet accordingly as *hyponyms* of '*thief*', like '*snatcher'*, which denotes "a thief who grabs and runs". A translator may simply translate '*snatcher'* into the Hebrew '*gnb'*, or she may qualify this noun with a certain adjective or the action taken with an adverb, depending on the context. This may be considered a cross-linguistic hyponym (Bentivogli *et al.*, 2002). Crucially, a GAP is declared and a gloss in Hebrew accompanies this declaration, therefore

- The lexical idiosyncrasy is marked;
- The gloss of the gapped concept is specified;
- The information which belongs to synsets higher in the hierarchy can be inherited through the GAP down to lower synsets.

2.3 MultiWordNet vs. EuroWordNet

When a WordNet for a new language is constructed, and it is desirable to align it with existing WordNets for other languages, one is con-

[1] To facilitate readability we use a straight-forward transliteration of Hebrew using ASCII characters, where the characters (in Hebrew alphabetic order) are: abgdhwzxviklmnsypcqršt.

fronted with a choice of two paradigms. We discuss some considerations for making the right choice in this section.

MWN and EWN have more in common than meets the eye. Although the Inter-Lingual-Index is supposedly language-independent, a so-called unstructured superset of all the basic synsets in wordnets, its members were taken mainly from WordNet version 1.5 (Vossen, 1998). One could not tell in advance whether a given concept in ILI is missing in a language for which it wasn't planned from the outset.[2] In some particular cases, like WordNets for Spanish or Catalan (the first is part of EWN, whereas the latter is provided separately, see Farreres *et al.* (1998)), the resemblance to MWN's methodology is high, as bilingual dictionaries were used to map PWN's synsets to Spanish and Catalan candidate synsets (not dissimilar to the algorithm described here). Basically, then, EWN abstracts from English to meet up other languages on a higher level of abstraction, whereas MWN adheres to PWN more closely, although, as the Spanish/Catalan example shows, this could be only a matter of degree. On the other extreme, the WordNets for Romanian and German, for example, are much more "sensitive" to their respective language idiosyncrasies (Dutoit *et al.*, 1998; Tufiş *et al.*, 2004).

The MWN paradigm involves a potential risk, namely that the resulting WordNet be influenced by the structure of PWN. This risk is offset by devising a methodology to cope with it. We believe that MWN is a better option specifically for languages poor in resources. For WordNets built from scratch, following the EWN paradigm, developers used a multitude of resources, including:

Monolingual dictionaries A machine readable high-qualtity dictionary was used for automatic relation inference for the Spanish WordNet (Verdejo, 1999); a dictionary was used to extract relations such as synonyms and hyponyms in Dutch (Vossen *et al.*, 1999).

Specialized digital lexicons These were used to automatically enrich the lexicon of the British WordNet (Peters, 1998).

Bilingual dictionaries A machine-readable high-quality dictionary of Spanish was used for automatic generation of synset candidates (Farreres *et al.*, 1998); a similar resource for German was used in order to automate the linking of words to a shared ontology (Dutoit *et al.*, 1998).

[2] Anecdotally, we may note that one of the key-concepts up the hierarchy of EWN's ILI is *artifact*, a word which does not exist in Hebrew.

Monolingual corpora A corpus of 60 million tagged German words was used to compile a basic vocabulary and generate frequency lists (Dutoit *et al.*, 1998); an Italian corpus was used for extracting multi-word expressions (Alonge *et al.*, 1999).

Parallel corpora These were used to generate translation equivalents for a Romanian WordNet (Tufiş *et al.*, 2004) and to validate synsets in a Serbian WordNet (Krstev *et al.*, 2004).

Ontologies Existing ontological hierarchies were used to automatically draft the French WordNet (Dutoit *et al.*, 1998).

None of those resources existed for Hebrew. Even today, there is no open keyword in context (KWIC) utility for Hebrew with which lexicographers can search and browse occurrences of words in corpora in order to validate their documentation: listing the various senses of a word, accompanying each sense with examples taken from “real texts” as opposed to the traditional artificially constructed examples, and learning about the distribution of each word and sense within certain language domains (literary, journalistic, scientific, spoken, etc.) and across the whole language in general. For this deficiency in resources we had to resort to the Web and use it as a corpus for Hebrew. Although we used the Web as a corpus, it does have its limitations,[3] and a strict alignment of the Hebrew WordNet to PWN minimized the reliance on this unrepresentative corpus.

In addition, the MWN paradigm provided several advantages in our case:

- Every WordNet necessitates a large number of subjective decisions, like deciding on the list of top categories under which all verbs and nouns should be organized, or making decisions pertaining to the appropriate design of semantic relations between different nodes within the network. Instead of forcing an alignment on two different subjective networks which are necessarily highly differential, the already existing network (PWN), subjective as it may be, is taken up by the added languages.

[3] See the special issue on the subject of Computational Linguistics, 29(3) (2003).

- There is no need to acquire the rights for a monolingual resource, as one can rely on the glosses given by PWN. There is only a handful of monolingual dictionaries of Hebrew, of which only two are available electronically. The chances to acquire the rights for using monolingual resource in most languages are low. Note that such a resource is often used to infer semantic relations.
- A reliance on a Hebrew corpus could be limited to ad-hoc decisions for solving particular cases, but is not needed in order to work the whole design of the network bottom-up. As Hebrew does not have an available representative large-scale corpus, this point is crucial.
- An existing user-interface is provided from which lexicographers can browse each synset separately, edit entries, offer translation equivalents, or look for evidence of each synset's instantiation in MultSemCor (see next item).
- The existence of MultiSemcor (see section 4.3) is a powerful tool for validating the matching between WordNets.
- Finally, as most of the work involved has to do with translation, it could be performed by experienced translators, a human resource more commonly available than lexicographers.

3 Non-equivalence Between Languages: The Challenge

Aligning two WordNets to each other is accompanied by the following dilemma: the more one wishes to represent the vocabulary of a language L1 in a way that reflects its inner structure, i.e., its morphology, semantics and their interrelations, the harder it is to align it to representations of L2, and consequently to construct multilingual databases. However, if one represents L1 in the first place according to models based on the morphology and semantics of L2, information might be lost. We distinguish between two kinds of non-equivalence: *contingent*, which is arbitrary and idiosyncratic; and *systematic*, which may be addressed in a more regular way.

3.1 Contingent non-equivalence

Languages differ in their lexicons, and matching lexical items from L1 to L2 is a non-trivial task. Bilingual dictionaries try to provide for every single lexeme, for each of its senses in L1, a list of all possible lexical items that can serve as its possible translation equivalents in L2 (and usually vice versa). No two lexicons can be mapped to each other using a one-to-one function. As an example of such a discrepancy, consider English ‘*honor*’ and ‘*respect*’, which are mapped to a single Hebrew lexeme (‘*kbwd*’) that denotes both senses. This kind of non-equivalence is a matter of contingency dependent on culturally-based lexical differences. In order to translate ‘*honor*’ or ‘*respect*’ into Hebrew one could use ‘*kbwd*’, but translating ‘*kbwd*’ into English would require a sensitivity to context in order to pick the right item (‘*honor*’ or ‘*respect*’). There are fuzzier cases where the decision of the translator or lexicographer is less predictable.

Consider a harder case of contingent non-equivalence. The verb ‘*get*’ is highly polysemous: WordNet lists no less than 37 senses for this lexeme. It is a member with ‘*acquire*’ in a synset the gloss of which is “come into the possession of something concrete or abstract”. The Hebrew equivalents are ‘*qibl*’ (also fairly polysemous: 15 senses) and ‘*hšig*’. The Rav-Milim Hebrew dictionary specifies two separate senses which are conflated in WordNet: ‘*to get something concrete*’ (sense 1), and ‘*to get something abstract*’ (sense 2). If we tried to align the Hebrew WordNet as strictly as possible to PWN, we would use ‘*qibl*’ (along with ‘*hšig*’, the equivalent of ‘*acquire*’) to denote both senses.

3.2 Systematic non-equivalence

Sometimes, however, non-equivalence is systematic: a subset of the lexicon in L1 is different from its semantically similar subset in L2 in a predictable way. Three such cases are discussed below.

3.2.1 *Gender: a case of consistent alignment*

One example of systematic non-equivalence has to do with the gender of nouns and is described by Ordan and Wintner (2005). Hebrew, like other languages (e.g., Arabic or Italian), marks gender on nouns in a fairly regular manner, both for animate and inanimate entities. Although English does not have grammatical gender for non-animate entities, it does mark the gender of many animate entities, some of which are marked morphologically (like using the suffix ‘*-ess*’ as in ‘*princess*’) and most are marked on grounds of meaning entailing pronoun agreement (like using *he* and *she* for ‘*king*’ and ‘*queen*’, respectively).

Due to cultural reasons (politics, gender inequality and social organization), English lexicalizes animate entities in various structures. Encoding them in PWN is inconsistent. Ordan and Wintner (2005) identify

6 different classes for animate nouns: nouns denoting both sexes (‘*citizen*’), nouns referring to females only (‘*midwife*’), nouns referring to males only (‘*womanizer*’), a class of nouns in which a single noun refers to both sexes (‘*parent*’) as well two separate nouns for each sex (‘*mother*’ and ‘*father*’), a class of nouns where no gender neutral noun exists but a noun for each sex does (‘*prince*’, ‘*princess*’), and a class of two nouns, one gender neutral (‘*actor*’), the other female specific (‘*actress*’). WordNet uses various lexical structures to handle these classes (one rare case is not discussed here):

- Listing all of them in one synset.
- Listing female and male specific synsets as coordinate terms with a gender neutral hypernym.
- Listing female and male specific synsets as indirect coordinate terms, both sharing a common hypernymical synset somewhere up the hierarchy.
- Listing the female as a hyponym of the male.

Hebrew, on the other hand, systematically marks gender on animate entities, with rare exceptions such as the loan-word ‘*prwpswr*’ (‘*professor*’). Since this difference marks a systematic non-equivalence that has to do with structural differences between the two languages, a systematic solution for aligning the two lexicons is preferred, for purposes of both parsimony and consistency.

3.2.2 *Passive voice: represented or ignored?*

English uses *syntax* (via the auxiliary ‘*be*’) in order to produce the passive voice in a systematic way, and hence there is no need to represent it in English dictionaries. However, the case of Hebrew is more complex and less regular, and therefore, a reliable representation of the Hebrew lexicon cannot forgo a specification of the passive voice.

Like other Semitic languages, Hebrew word formation is based on root and pattern morphology, whereby consonantal roots are combined with patterns in a non-concatenative way. There are seven verbal patterns in Hebrew, traditionally called *binyanim*, and while—like other derivational morphological processes—their semantics is sometimes idiosyncratic, several generalizations can be made. In particular, the relation between the patterns CiCCeC and CuCCaC and between hiCCiC and huCCaC is almost uniquely a relation of active–passive. Additionally, the pattern niCCaC is often the passive form of the counterpart CaCaC, but here the relation is more arbitrary.

Another justification for representing the passive voice, especially that of niCCaC, is that the latter, unlike its equivalents in English, can be conjugated in the imperative. Thus, the form '*nbdq*' ('*be checked*') can be conjugated as '*hibdq*' ('*get yourself checked*'). This is not a rare case in Hebrew. The matter can get even more complicated in languages like Malagasy, where the passive imperative may adhere to the objects which undergo the action, like '*Sasao ny lamba!*' "be.washed the clothes" to denote "wash the clothes!" (Van Valin, 2001, p. 41).

3.2.3 *Aligning pairs of antonyms*

Antonyms are long debated in the literature and do not lend themselves easily to a formal definition (Cruse, 1986, pp. 204-206), yet most people have no problem identifying them as they encounter them (Miller *et al.*, 1990). The most typical examples belong to adjectives and adverbs, e.g., '*big*' vs. '*little*' or '*rarely*' vs. '*often*'. These two are pairs of unrelated word forms, but more often than not English antonyms are marked by a prefix and relate to each other both in form and in meaning, be it in adjectives ('*interesting*/*uninteresting*'), adverbs ('*willingly*/*unwillingly*'), verbs ('*satisfy*/*dissatisfy*, *respect*/*disrespect*'), or nouns ('*belief*/*unbelief*'). As we shall see, antonyms are very dominant in the organization of WordNet, but since they are lexical rather than semantic relations, they depend on the morphology of English and other related languages, and hence pose a serious problem when aligning English to, say, a Semitic language like Hebrew or Arabic.

In English (and one can extend this claim to other Indo-European languages), as opposed to Hebrew (and other Semitic languages), antonymy is usually realized by means of affixation, mostly prefixation. In other words, morphology facilitates the generation of antonyms. We assume that this interrelation between morphology and semantics makes antonymy more central to the organization of the English lexicon and less central to Hebrew, and consequently we encounter problems in aligning the two lexicons when it comes to antonyms.

Hebrew has a few prefixes equivalent to '*un*-, *in*-, *dis*-', etc, although they are not as productive as their English counterparts. In addition, the use of negation prefixes in Hebrew is not as strict as the English one, that is, given a certain adjective there is no single preferred prefix with which to negate it. Whereas in English '*uninteresting*' is the antonym of '*interesting*', in Hebrew '*mynin*' "interesting" can have as its antonyms ‘*la-mynin*’ or ‘*blti-mynin*’, both sound acceptable. A search in a corpus, however, reveals that their distribution is not completely random: for example, in many cases the latter is preferred in the expression ‘*la-blti-mynin*’ “not uninteresting”, most probably in order to avoid a duplication of the prefix/negation-word ‘*la*’.

The representation of antonyms in PWN is based mainly on the English morphology means of using morphological antonym markers (like ‘*dis-*’). Out of 2125 antonym pairs for adjectives, more than 1686 pairs (i.e., nearly 80%) are pairs of an unmarked adjective (like ‘*possible*’) with its morphologically derived marked equivalent (‘*impossible*’).[4] Most of these adjectives contain highly productive affixes (like ‘*un-*, *in-*’ or its allomorph ‘*im-*, *non-*, *dis-*’ and ‘*-less*’, see Table 1) along with some less productive ones (‘*mis-*’, or ‘*under-*’). Similar data are available for other parts of speech (Table 2). This serves as a good example for the typographical bias with which PWN was designed. What is presented in PWN as semantic universals is more typical to English (and other Indo-European languages) than Hebrew.

Table 1: number of morphologically marked pairs in WordNet version 2.0

Affixes	Adjectives	Adverbs	Verbs	Nouns	Total
prefix ‘*un-*’	885	139	92	176	1322
prefix ‘*under-*’	6	1	19	4	30
prefix ‘*in-*’ or ‘*im-*’	384	101	0	243	728
prefix ‘*non-*’	186	3	0	45	234
prefix ‘*dis-*’	74	18	103	84	279
infix/suffix ‘*-less’*	151	27	0	41	219
total:	1686	289	214	593	2782

[4] Our database queries yielded only coarse results and do not provide a 100 percent reflection of the situation discussed. For example, when looking at adjectives beginning with the prefix ‘*dis-*’, the query also retrieved the pair ‘*close–distant*’.

Table 2: number of antonym pairs per POS in WordNet version 2.0

POS	Antonym Pairs	Percentage
Adjectives	2125	42.13
Nouns	1412	28.66
Verbs	906	18.39
Adverbs	483	9.80
total POS	4926	100.00

Aligning pairs of adjective antonyms, therefore, is enabled due to the availability of antonym morphological markers in Hebrew. Their availability is a matter of recent history, both in Hebrew and in Arabic. The case may be different for other Semitic languages which did not have such a close contact with the West, like Amharic or Tigrinya. In addition, the unavailability of similar morphological markers for other parts of speech poses a serious problem for alignment.

3.2.4 *Verb relations in Turkish*

The above observations, of course, are not specific to any pair of languages. In Turkish, for example, there are many suffixes that bear regular semantic effect on base lexemes to which they are suffixed. English does not have these morphological markers. However, as soon as English (or any other language) is aligned to Turkish, these semantic relations can be revealed in English as well. For example, the suffix '*-lAş*' denotes a '*become*' relation, whereas '*-DHr*' denotes a '*cause*' relation. In the Turkish WordNet there are 763 pairs of the former and 782 pairs of the latter. Thus, the '*become*' relation yields in English such pairs as '*good–improve*, silent–*hush*' etc., and the '*cause*' relation yields pairs such as '*dress–wear*, *dissuade–give up*', etc. (Bilgin *et al.*, 2004). While the following discussion uses Hebrew and English to demonstrate the main issues, most of the problems will have similar instances in any language pair.

4 A METHODOLOGY FOR ALIGNING TWO WORDNETS

In this section we discuss our approach for creating a WordNet for Hebrew, aligned with the Princeton English WordNet. We opted for casting it within MultiWordNet for the reasons discussed in Section 2.3. Our methodology consists of three steps. First, we bootstrap a preliminary assignment of Hebrew words to PWN synsets (Section 4.1). We then (Section 4.2) propose systematic solutions for consistently aligning some of the more regular cases of non-equivalence discussed in Section 3. The contingent non-equivalence is of course resolved through manual lexicographic work. We offer a corpus-based methodology to carry out this work. Finally, we discuss in Section 4.3 a validation scheme for the alignments, based on a semantically annotated corpus.

4.1 Assignment algorithm

The construction of Hebrew WordNet was crucially based on two automatic procedures. The first is the *assign-procedure*: given a Hebrew word sense, the procedure selects a weighed list of the most likely corresponding PWN synsets. This list is then used by lexicographers to actually build the Hebrew synsets. The second procedure supports the detection of lexical gaps, which are cases when a lexical concept of a language is expressed through a free combination of words in another language (see Section 2.2). The essentials of the assignment algorithms are described in much detail by Bentivogli *et al.* (2002); we provide here a brief summary.

Both procedures use, as a crucial linguistic resource, the electronic version of a Hebrew-English bilingual dictionary. Given the limited resources available for a language such as Hebrew, this was a major obstacle. We ended up using a small-scale, low-quality printed dictionary for which we could obtain the rights (Dahan, 1997).

Following the MWN model, our aim is to build, whenever possible, Hebrew synsets which are synonymous with the PWN synsets. When this is not possible, a gap is declared. Hebrew synsets can be built following different strategies. The first strategy is based on English-to-Hebrew translation equivalents (TEs). For each PWN synset *S*, we look for the Hebrew TEs which are cross-linguistic synonyms of the English words of *S*. The union of such TEs is the Hebrew synonymous synset of *S*. If this set is empty, we have found an English-to-Hebrew lexical idiosyncrasy. The second strategy is based on Hebrew-to-English TEs. For each sense *s* of a Hebrew word *w*, we look for a PWN synset *S* including at least one English TE of *w*; and we establish a link between *w* and *S*. When the procedure has been applied to all Hebrew word senses, we build the equivalence class of all sets of Hebrew words which have been linked to the same PWN synset. Each set in the equivalence class is the Hebrew synset synonymous with some PWN synset. If, for a set of Hebrew synonyms there is no PWN synonymous synset, we have found a Hebrew-to-English lexical idiosyncrasy.

4.2 Consistent alignment of systematic non-equivalences

The application of the assignment algorithm to our small bilingual dictionary yielded many assignment candidates, which we then had to manually confirm. However, in some cases of systematic non-equivalence, a more general approach was possible, which we delineate below.

4.2.1 *Gender*

A solution to the gender issue presented in Section 3.2.1, in the MWN paradigm, is offered by Ordan and Wintner (2005). They propose a neutral structure that allows for all lexical possibilities to occur: a gender neutral synset is declared as a hypernym of both gender specific synsets, irrespective of whether or not lexicalization for each exists (see Figure 1).

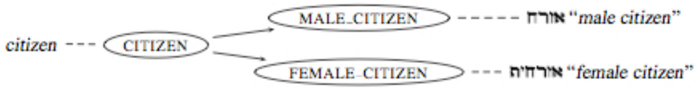

Figure 1: The structural organization of gendered nouns.

If a gender synset is not realized in Hebrew, a GAP on the Hebrew side is created, and relations from this GAP to existing synsets ensure inheritance of information. On the other hand, if a gender specific synset exists in Hebrew but not in English, a new Hebrew synset is mapped to a GAP on the English side, and this GAP in turn is mapped onto other synsets using lexical relations.

For example, for the synset {*citizen*} two Hebrew synsets are created, *female* and *male*, and both are hyponyms of a GAP of the English gender-neutral synset. In addition, each of the new gender-specific synsets is further mapped onto the Hebrew equivalents to '*citizen*', namely '*female-citizen*' and '*male-citizen*'. Although the general policy of PWN is not to create more than two hypernyms per single synset (Miller *et al.*, 1990), this solution does burden the system with many such cases, as for almost every gender specific synset in Hebrew, two hypernymical relations are added. This is the price paid for aligning the gender system of Hebrew to the English one, the advantage being keeping the richness of semantic information on the Hebrew part intact.[5]

4.2.2 *Voice*

To solve the passive issue discussed in Section 3.2.2, we propose to add a special lexical relation for Hebrew, namely *passive-form-of*, that is mapped onto the active form of the verb. WordNet's representation of verbs is highly structured and includes unique relations for verbs (Fellbaum, 1998). Therefore, the most important question with regard to applying a "passive-form-of relation" is whether it can inherit semantic relations via the active form of the verb. For example, the troponyms of '*bend*' (in the sense of "form a curve"), like '*crook*', '*arch*', '*retroflex*' or '*crouch*', all have passive voice, which could be listed as troponyms of '*be bent*' (the passive voice of bend). How it reflects on other kinds of entailment is yet to be researched. The *causation* relation, for example, is left unimpaired: for if '*teach*' causes '*learn*', then '*being learned*' is the result of '*being taught*'.

4.2.3 *Antonyms*

In order to address cross-lingual gaps in the *antonym* relation, it is important to realize that antonyms are not merely semantic oppositions. For example, while '*respect*' and '*honor*' are synonyms, '*disrespect*' is only an antonym of the former, not of the latter.

Our solution to the discrepancy discussed in Section 3.2.3 is based on WordNet's notion of *similarity*. Since Hebrew does not have a lexical counterpart for '*disrespect*', this concept is represented as a gap in the Hebrew network. However, we take advantage of the fact that the *antonym* relation is a relation between lexical items, not between synsets.

In the case of adjectives the matter is straightforward and can be easily exemplified: the English '*interested*' is an antonym of '*uninterested*'. Each of these antonyms is located in the center of a cluster of *near synonymous* words, related to it in a relation of *similarity*. The first has only one related word, namely '*concerned*', whereas the second has various related synsets, among them {*apathetic*, *indifferent*} and {*blase*,

[5] Note that having expressed PWN's reservation, the new version of WordNet (2.1) does contain more than 2000 synsets that have more than one hypernym, ranging between 2-6 hypernyms per single synset.

bored}. As Hebrew does not have a counterpart to '*uninterested*', we declare a GAP on this synset; however, we connect this GAP to other synsets related to it by a relation of similarity, like '*adiš*', 'šwh-*npš*' "indifferent" and '*apvi*' "apathetic". Looking for the antonyms of the Hebrew equivalent to '*interested*', '*mywnin*', one would find several options *similar* to its non-existent *antonym*.

The case for other parts of speech is different. Consider nouns: antonyms are marked on one member within a given synset, thus relating two lexical items of two respective synsets and leaving the other lexical items untouched, although related. For example, in the synset {*esteem*, *regard*, *respect*} only the noun '*esteem*' is represented as the antonym of '*disesteem*', a single membered sysnet. Hebrew has three possible synonyms to '*disesteem*', namely '*zlzwl*', '*qls*' and '*bwz*'. Into this potential sysnet we introduce a zero lexical-item, namely the GAP of '*disesteem*'. Again, on looking for the antonyms of the Hebrew equivalent to '*esteem*', one would find similar related words to the non-existent antonym. It is important to note that conceptual relations are not inherited via the lexical relation of *antonymy*.

Out of the 483 adverb antonym pairs in WordNet, 372 pairs (77%) are produced by the bound morpheme '-*ly*' on both sides of the pair: 'slowly ↔ *quickly*'. Of these, only few are lexicalized in Hebrew ('*lav*' ↔ '*mhr*'), while the rest, like most adverbs, require the prepositional phrase "in the manner of", which cannot be considered a lexical item. Using GAPs for adverb antonyms would entail almost zero encoding of adverbs in Hebrew. We find it a shortcoming of the Hebrew WordNet, but also, admittedly, a bias on the part of lexical semantics and psycholinguistics, from which PWN stems, towards Indo-European languages.

4.3 Using a semantically annotated corpus to validate alignments

In order to address contingent non-equivalences we use MultiSemCor (Bentivogli and Pianta, 2005), a corpus in which each token is mapped onto a synset in WordNet, to validate our choices. We exemplify this approach on the case of the verb '*get*' discussed in Section 3.1.

The total number of occurrences of the synset {*get*, *acquire*} is 94, and indeed, many times we can use '*qibl*' or '*hšig*' as perfect translation equivalents for {*get*, *acquire*}. We therefore decided to use these for '*get*'. However, in a non-negligible number of occurrences, especially where '*get*' is used abstractly, the members of {*qibl*, *hšig*} cannot replace {*get*, *acquire*} (see examples in Table 3). Note that in all the occurrences given in Table 3, '*get*' is used metaphorically and is 'applied' to familiar objects one can ‘get’ (relief, happiness, value), where we face the arbitrariness and contingency of idiomatic expressions. The decision to use ‘*hšig*’ and ‘*qibl*’ for ‘*get*’, as has been shown by the corpus lookup, is not perfect, but in this case, it is a reasonable compromise. In other cases, the lexicographer could have decided otherwise. This kind of non-equivalence is not worked out systematically but rather one by one on the go. In other words, when contingent non-equivalence is concerned, one cannot devise an overall solution that would pertain to all cases, but rather let lexicographers work on them manually one by one. MultSemCor provides a semantic environment to handle such a task.

Table 3: Examples of contingent non-equivalence of English-Hebrew pairs.

English original	Hebrew translation	Literal back translation into English
get relief	‘xš hqlh’	feel relief
get happiness	‘zkh b-/mca awšr’	win/find happiness
acquire tact	‘sigl lycmw xwš vqv’	adapted for himself a sense of tact

5 DISCUSSION

Hebrew is the first Semitic language for which a substantial WordNet has been designed. The Hebrew WordNet currently contains 5261 synsets, with an average of 1.47 synonyms per synset, where nouns are much more frequent than other parts of speech (almost 78 percent, see Table 4).

Table 4: current state of the Hebrew WordNet.

POS	Number of synsets
Nouns	4090
Verbs	609
Adjectives	779
Adverbs	151
total	5261

Principally, using GAPs is a methodology which allows different levels of semantic inheritance. Both cases of non-equivalence, systematic and contingent, are of concern here. For both, the crucial question is whether L2 introduces a more specific or a more general lexical concept/relation. Whenever a more general synset in relation to L1 is introduced, we keep partial inheritance. When L2 is more specific, a full semantic inheritance can be safely kept.

Consider first the contingent cases. The Hebrew lexeme '*kbwd*' is a cross-language hypernym with respect to English, for it can be considered the hypernym of both '*honor*' and '*respect*'. Following our methodology we declare a GAP on both '*honor*' and '*respect*' (its hyponyms), create a new sysnet for '*kbwd*' in Hebrew, and relate the new Hebrew synset to both gaps. Ideally, it could enrich our network semantically, as it would make our network denser, bringing the synsets related to '*honor*' and '*respect*' closer to each other. For example, it would relate '*reputation*', a hyponym of '*respect*', directly to '*honor*'. Taking such a decision for each contingent case may require substantial effort, and therefore we advocate, at least for short-term projects, the solution of a partial inheritance, to the effect that gapped synsets should relate to each other only on a local level. In our example, '*kbwd*' would relate '*honor*' and '*respect*' directly and locally, and then either would run along its semantic route keeping the same semantic distance they had in PWN. On the other hand, in the '*snatcher*' example, where Hebrew's '*gnb*' is a more general concept, a full inheritance is maintained.

The same holds for systematic cases of non-equivalence. As we saw in the case of gender, a full inheritance is maintained due to the fact that Hebrew gender distinctions are more specific in Hebrew, and can therefore be nested within English synsets. However, in passive verbs, the Hebrew distinction is "external" to English, it cannot be nested within it, and therefore we maintain only a partial inheritance, between the Hebrew passive verbs of the niCCaC pattern and their respective active voiced forms.[6] Admittedly, for systematic cases it is worthwhile to further explore the possibility of a full inheritance (we checked it only for the *causation* relation, which seems to hold between active and passive verbs). As was exemplified in the Turkish WordNet (see Section 3.2.4), the more general distinctions of L2 enriched L1's network (Turkish and English, respectively). New kinds of relations and con-

[6] The Hebrew niCCaC pattern is remarkably similar to the seventh verb pattern in Arabic, inCaCaCa. Whereas in Arabic (almost) every verb pattern has a passive version (generated by vocalization change), this seventh verb pattern is irregular: it may indicate the passive form of the first verb pattern, sometimes in addition to, and not instead of, the passive counterpart of this pattern. As inCaCaCa is irregular we maintain that it should be specified in the Arabic WordNet, including where it manifests passive forms.

cepts in L2 may introduce new and unexpected relations in L1. Indeed, this calls for further research.

The problems we have raised touch on more general questions, practical and theoretical. The first relates to lexicographical work. Classical A-Z bilingual dictionaries present isolated pairs of lexical items in two languages, and therefore lexical mismatches are always solved on a local basis. In a highly structured and complex lexical representation like WordNet, topological variance should be taken into account in order to design consistent solutions on a large scale. In our project we offered a modest solution for such a design when gendered nouns are concerned. A solution for passive voice is only sketched. The case of antonyms was solved for all parts of speech, except adverbs, mainly due an unbridgeable typological variance between Hebrew and English.

Another consequence relates to psycholinguistics. WordNet's structuring of adjectives relies heavily on English morphology. Although organization occurs on the word level, PWN does consider them as concepts using the term lexicalized concepts (Fellbaum, 1998). One option is to reconsider the organization of WordNet according to parts of speech; another would be to further research the way the mental lexicon is represented: it may be the case that it is differently organized in speakers from different language families. The response of English speakers to one word with its antonym might be accounted for by a combination of semantics and morphology. It would be interesting to discuss and research this topic in relation to languages where no equivalent to adjectives is found and modification is produced in other ways (cf. Dixon (1982)).

As more and more WordNets are being compiled, one would have to dedicate more efforts to lexicographical design on the multilingual level. Simple multilingual encodings would complicate the task of using multilingual WordNets for the task of lexical transfer in MT systems. One would either need to improve the multilingual encoding, or add extra rules in order to use what multilingual WordNet-based databases have to offer. One measure for a good design is a multilinguistic representation that remains true to both L1 and L2's internal idiosyncrasies and therefore enables a relatively accurate lexical transfer under changing inter- and intra-textual environments.

Acknowledgments This research was supported by THE ISRAEL SCIENCE FOUNDATION (grant No. 137/06); by the Israel Internet Association; by the Knowledge Center for Processing Hebrew; and by the Caesarea Rothschild Institute for Interdisciplinary Application of Computer Science at the University of Haifa. The first author is grateful to the Hertie Foundation for a scholarship which allowed him to conduct this research. We are indebted to Emanuele Pianta, Luisa Bentivogli, Christian Girardi, Bernardo Magnini, Oliviero Stock and Iris Eyal for their help and support. We are grateful for the encouragement and comments of Cristiane Fellbaum and her group in the Berlin-Brandenburgische Akademie der Wissenschaften, where an earlier version of this paper was presented. Thanks are also due to Sabri Elkateb for consulting on Arabic and addressing typological issues.

REFERENCES

Eneko Agirre and German Rigau.Word sense disambiguation using conceptual density. In *Proceedings of the 16th conference on Computational linguistics*, pages 16–22, Morristown, NJ, USA, 1996. Association for Computational Linguistics.

A. Alonge, F. Bertagna, N. Calzolari, and A. Roventini. The Italian WordNet. Deliverable d032d033, EuroWordNet, 1999.

Luisa Bentivogli and Emanuele Pianta. Exploiting parallel texts in the creation of multilingual semantically annotated resources: the MultiSemCor corpus. *Natural Language Engineering*, 11(3):247–261, 2005.

——, ——, and Christian Girardi. MultiWordNet: developing an aligned multilingual database. In *Proceedings of the First International Conference on Global WordNet*, Mysore, India, January 2002.

Orhan Bilgin, Özlem Çetinoğlu, and Kemal Oflazer. Morphosemantic relations in and across WordNets: A study based on Turkish. In *Proceedings of the Second Global WordNet Conference*, Brno, Czech Republic, January 2004. GWC.

William Black, Sabri Elkateb, Horacio Rodriguez, Musa Alkhalifa, Piek Vossen, Adam Pease, and Christiane Fellbaum. Introducing the Arabic WordNet project. In *Proceedings of the Third Global WordNet Meeting*. GWC, January 2006.

Alan D. Cruse. *Lexical Semantics*. Cambridge University Press, 1986.

Hiya Dahan. *Hebrew–English English–Hebrew Dictionary*. Academon, Jerusalem, 1997.

Manuel de Buenaga Rodríguez, José María Gómez Hidalgo, and Belén Díaz-Agudo. Using wordnet to complement training information in text categorization. In *Proceedings of the 2nd International Conference on Recent Advances in Natural Language Processing*, 1997.

Mona Diab. The feasibility of bootstrapping an Arabic WordNet leveraging parallel corpora and an EnglishWordNet. In *Proceedings of the Arabic Language Technologies and Resources*, Cairo, Egypt, September 2004. NEMLAR.

Robert M. W. Dixon. Where have all the adjectives gone? In R. M.W. Dixon, editor, *Where Have All the Adjectives Gone? and Other Essays in Semantics and Syntax*, pages 1–62. Mouton, Berlin-Amsterdam-NewYork, 1982.

Dominique Dutoit, Laurent Catherin, and Andreas Wagner. Specification of French and German WordNets. Deliverable 2d002, EuroWordNet, July 1998.

Xavier Farreres, German Rigau, and Horacio Rodr´ıguez. Using WordNet for building WordNets. In Sanda Harabagiu, editor, *Usage of WordNet in Natural Language Processing Systems: Proceedings of the Coling-ACL 1998 Workshop*, pages 65–72. Association for Computational Linguistics, 1998.

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

——, Martha Palmer, Hoa Trang Dang, Lauren Delfs, and Susanne Wolf. Manual and automatic semantic annotation with WordNet. In *Proceedings of WordNet and Other Lexical Resources Workshop*, 2001.

Sanda Harabagiu, editor. *Usage of WordNet in Natural Language Processing Systems: Proceedings of the Coling-ACL 1998 Workshop*. Association for Computational Linguistics, Montreal, Canada, 1998.

Hongyan Jing. Usage ofWordNet in natural language generation. In Sanda Harabagiu, editor, *Usage of WordNet in Natural Language Processing Systems: Proceedings of the Coling-ACL 1998 Workshop*, pages 128–134. Association for Computational Linguistics, 1998.

Cvetana Krstev, Gordana Pavlovic-Lazetic, Duško Vitas, and Ivan Obradovic. Using textual and lexical resources in developing the Serbian Wordnet. *Romanian Journal of Information Science and Technology*, 7(1-2):147–161, 2004.

John Lyons. Structural Semantics. Oxford, 1963.

——. *Semantics*. Cambridge University Press, Cambridge, 1977.

Rila Mandala, Takenobu Tokunaga, Hozumi Tanaka, Akitoshi Okumura, and Kenji Satoh. Ad hoc retrieval experiments using wordnet and automatically constructed thesauri. In *TREC*, pages 414–419, 1998.

George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller. Five papers on WordNet. *International journal of lexicography*, 3(4), 1990.

Noam Ordan and Shuly Wintner. Representing natural gender in multilingual lexical databases. *International Journal of Lexicography*, 18(3):357–370, September 2005.

Wim Peters. The English WordNet. Deliverable d032d033, EuroWordNet, 1998.

Dan Tufiş, Dan Cristea, and Sofia Stamou. BalkaNet: Aims, methods, results and perspectives. A general overview. *Romanian Journal of Information Science and Technology*, 7(1-2):9–43, 2004.

R. D. Van Valin. *An introduction to syntax*. Cambridge University Press, 2001.

Felisa M. Verdejo. The Spanish Wordnet. Deliverable d032d033, EuroWordNet, Madrid, Spain, 1999.

Piek Vossen, editor. *EuroWordNet: a multilingual database with lexical semantic networks for European Languages*. Kluwer, Dordrecht, 1998.

——, Laura Bloksma, and Paul Boersma. The DutchWordnet. Technical report, University of Amsterdam, The Netherlands, 1999.

NOAM ORDAN
DEPARTMENT OF TRANSLATION AND INTERPRETING STUDIES
BAR ILAN UNIVERSITY, ISRAEL
E-MAIL: <NOAM.ORDAN@GOOGLEMAIL.COM>

SHULY WINTNER
DEPARTMENT OF COMPUTER SCIENCE
UNIVERSITY OF HAIFA, ISRAEL
E-MAIL: <SHULY@CS.HAIFA.AC.IL>