Can You Tag the Modal? You Should.

Yael Netzer and **Meni Adler** and **David Gabay** and **Michael Elhadad**
Ben Gurion University of the Negev
Department of Computer Science
POB 653 Be'er Sheva, 84105, Israel
{yaeln,adlerm,gabayd,elhadad}@cs.bgu.ac.il

Abstract

Computational linguistics methods are typically first developed and tested in English. When applied to other languages, assumptions from English data are often applied to the target language. One of the most common such assumptions is that a "standard" part-of-speech (POS) tagset can be used across languages with only slight variations. We discuss in this paper a specific issue related to the definition of a POS tagset for Modern Hebrew, as an example to clarify the method through which such variations can be defined. It is widely assumed that Hebrew has no syntactic category of modals. There is, however, an identified class of words which are modal-like in their semantics, and can be characterized through distinct syntactic and morphologic criteria. We have found wide disagreement among traditional dictionaries on the POS tag attributed to such words. We describe three main approaches when deciding how to tag such words in Hebrew. We illustrate the impact of selecting each of these approaches on agreement among human taggers, and on the accuracy of automatic POS taggers induced for each method. We finally recommend the use of a "modal" tag in Hebrew and provide detailed guidelines for this tag. Our overall conclusion is that tagset definition is a complex task which deserves appropriate methodology.

1 Introduction

In this paper we address one linguistic issue that was raised while tagging a Hebrew corpus for part of speech (POS) and morphological information. Our corpus is comprised of short news stories. It includes roughly 1,000,000 tokens, in articles of typical length between 200 to 1000 tokens. The articles are written in a relatively simple style, with a high token/word ratio. Of the full corpus, a sample of articles comprising altogether 100,000 tokens was assembled at random and manually tagged for part of speech. We employed four students as taggers. An initial set of guidelines was first composed, relying on the categories found in several dictionaries and on the Penn treebank POS guidelines (Santorini, 1995). Tagging was done using an automatic tool[1]. We relied on existing computational lexicons (Segal, 2000; Yona, 2004) to generate candidate tags for each word. As many words from the corpus were either missing or tagged in a non uniform manner in the lexicons, we recommended looking up missing words in traditional dictionaries. Disagreement was also found among copyrighted dictionaries, both for open and closed set categories. Given the lack of a reliable lexicon, the taggers were not given a list of options to choose from, but were free to tag with whatever tag they found suitable. The process, although slower and bound to produce unintentional mistakes, was used for building a lexicon, and to refine the guidelines and on occasion modify the POS tagset. When constructing and then amending the guidelines we sought the best trade-off between

[1] http://wordfreak.sourceforge.net

Proceedings of the 5th Workshop on Important Unresolved Matters, pages 57–64,
Prague, Czech Republic, June 2007. ©2007 Association for Computational Linguistics

accuracy and meaningfulness of the categorization, and simplicity of the guidelines, which is important for consistent tagging.

Initially, each text was tagged by four different people, and the guidelines were revised according to questions or disagreements that were raised. As the guidelines became more stable, the disagreement rate decreased, each text was tagged by three people only and eventually two taggers and a referee that reviewed disagreements between the two. The disagreement rate between any two taggers was initially as high as 20%, and dropped to 3% after a few rounds of tagging and revising the guidelines.

Major sources of disagreements that were identified, include:

Prepositional phrases vs. prepositions In Hebrew, formative letters – ב,כ,ל,מ *b,c,l,m*[2] – can be attached to a noun to create a short prepositional phrase. In some cases, such phrases function as a preposition and the original meaning of the noun is not clearly felt. Some taggers would tag the word as a prepositional prefix + noun, while others tagged it as a preposition, *e.g.*, בעקבות *b'iqbot* (following), that can be tagged as ב-עקבות *b-iqbot* (in the footsteps of).

Adverbial phrases vs. Adverbs the problem is similar to the one above, *e.g.*, בדיוק *bdiyuq* (exactly), can be tagged as *b-diyuq* (with accuracy).

Participles vs. Adjectives as both categories can modify nouns, it is hard to distinguish between them, *e.g*, מבט מאיים *mabaṭ m'ayem* (a threatening stare) - the category of מאיים *m'ayem* is unclear.

Another problem, on which the remainder of the article focuses, was a set of words that express modality, and commonly appear before verbs in the infinitive. Such words were tagged as adjectives or adverbs, and the taggers were systematically uncertain about them.

Beside the disagreement among taggers, there was also significant disagreement among Modern Hebrew dictionaries we examined, as well as computational analyzers and annotated corpora. Table 1 lists the various selected POS tags for these words, as determined by: (1) Rav Milim (Choueka et al., 1997), (2) Sapir (Avneyon et al., 2002), (3) Even-Shoshan (Even-Shoshan, 2003), (4) Knaani (Knaani, 1960), (5) HMA (Carmel and Maarek, 1999), (6) Segal (Segal, 2000), (7) Yona (Yona, 2004), (8) Hebrew Treebank (Sima'an et al., 2001).

As can be seen, eight different POS tags were suggested by these dictionaries: ad**J**ective (29.6%), adve**R**b (25.9%), **V**erb (22.2%), **A**uxilary verb (8.2%), **N**oun (4.4%), par**T**icle (3.7%), **P**reposition (1.5%), and **U**nknown (4.5%). The average number of options per word is about 3.3, which is about 60% agreement. For none of the words there was a comprehensive agreement, and the PoS of only seven words (43.75%) can be determinded by voting (*i.e.*, there is one major option).

In the remainder of the paper, we investigate the existence of a *modal* category in Modern Hebrew, by analyzing the characteristic of these words, from a morphological, syntactic, semantic and practical point of view. The decision whether to introduce a modal tag in a Hebrew tagset has practical consequences: we counted that over 3% of the tokens in our 1M token corpus can potentially be tagged as modals. Beyond this practical impact, the decision process illustrates the relevant method through which a tagset can be derived and fine tuned.

2 Modality in Hebrew

Semantically, *Modus* is considered to be the *attitude on the part of the speaking subject with regard to its content* (Ducrot and Todorov, 1972), as opposed to the *Dictum* which is the linguistic realization of a predicate. While a predicate is most commonly represented with a verb, modality can be uttered in various manners: adjectives and adverbs (*definitely*, *probable*), using thought/belief verbs, mood, intonation, or with modal verbs. The latter are recognized as a grammatical category in many languages (modals), *e.g.*, *can*, *should* and *must* in English.

From the semantic perspective, modality is coarsely divided into *epistemic modality* (the amount of confidence the speaker holds with reference to the truth of the proposition) and *deontic modality* (the degree of force exerted on the subject of the sentence to perform the action) views (de Haan, 2005).

Modal expressions do not constitute a syntactic class in Modern Hebrew (Kopelovich, 1982). In her work, Kopelovich reviews classic descriptive

[2]Transcription according to (Ornan, 2002)

Word	Example	1	2	3	4	5	6	7	8
יש ***yeš*** **should**	יש לשים לב לניסוח *yeš laśim leb lanisuḥ* Attention should be paid to the wording	R	N	N	R	N	A	R	V
אין ***'ein*** **shouldn't**	אין לשים לב לניסוח *'ein laśim leb lanisuḥ* Attention should not be paid to the wording	R	U	N R	U	P	P	R	V
חייב ***ḥayab*** **must**	הציבור חייב להבין את העניין *hacibur ḥayab lhabin 'et ha'inyan* The public should be made aware of this issue	J	J	J	J	J	J	J	V
מותר ***mutar*** **allowed**	מותר לה לצאת לטיול *mutar lah lacet lṭiyul* She is allowed to go on a trip	R	N	J	R	J	A	V	J
אסור ***'asur*** **forbidden**	אסור לה לצאת לטיול ביום ראשון *'asur lah lacet lṭiyul byom rišon* She is not allowed to go on a trip on Sunday	R	R	R	R	J	A	J	J V
אפשר ***'epšr*** **may**	אפשר לרמוז רמזים *'epšr lirmoz rmazim* Giving hints is allowed	U	R	R	R	T	A	R	V
אמור ***'amur*** **supposed**	נשים אמורות ללבוש רעלות *našim 'amurot lilboš r'alot* Women are supposed to wear veils	J	A	J	J	J	A	J	V
צריך ***carik*** **should**	במו"מ צריך לעמוד על שלך *bmw"m carik la'amod 'al šelka* In negotiation you should keep strong	J	J	R	J	J	A	J	V
ניתן ***nitan*** **can**	ניתן לפתור בעיה מכאיבה זו *nitan liptor b'ayah mak'ibah zo* This troublesome problem can be solved	U	V	V	V	V	V	V	V
עלול ***'alul*** **may**	הכלב עלול לנשוך *hakeleb 'alul linšok* The dog may bite	J	J	J	J N	J	A	J	V
כדאי ***kda'y*** **worthwhile**	כדאי לשאול האם הדלת עשויה היטב *kda'y liš'ol ha'im hadelet 'aśuyah heiṭeb* It is worth asking whether the door is well built	R	R	R	R	J	A	R	J
מוטב ***mutab*** **better**	מוטב להיות בשקט ולהנות *mutab lihyot bešeqeṭ wulhnot* Better to keep quiet and enjoy	R	R	R	R	T	T	V	V
מסוגל ***msugal*** **able**	הוא היה מסוגל לראותו בבית הלבן *hu' hayah msugal lir'oto babait halaban* He could envision him sitting in the White House	J	R	J	J	J V	A	J	V
יכול ***yakol*** **can**	אנשים יכולים לתרום תרומות *'anašim ykolim litrom trumot* People can make contributions	V	V	V	J	V	A	V	V
אכפת ***'ikpat*** **care/mind**	אכפת לך ללכת? *'ikpat lka laleket?* Do you mind going	U	V R	V R	U	T	T	R	V
ראוי ***ra'uy*** **should**	ראוי לשלם על שרות זה *ra'uy lšalem 'al šerut zeh* This service deserves to be paid for	R	R	R	R	J	V J	R	J

Table 1: Parts of speech of selected words

publications on the syntax of Hebrew and claims that these works (Ornan, Rubinstein, Azar, Rosen, Aronson-Berman and Maschler)[3] do not provide a satisfying description or explanation of the matter. In this section we review three major approaches to modality in Hebrew - the first is semantic (Kopelovich), the second is semantic-syntactic (Zadka) and the third is purely morphologico-syntactic (Rosen).

Kopelovich provides three binary dimensions that describe the modal system in Hebrew: Personal - Impersonal, Modality - Modulation and Objective - Subjective plane. The Personal-Impersonal system is connected to the absence or presence of a surface subject in the clause. A personal modal has a grammatical subject:

(1) דוד צריך להסיע את אמו
dawid carik lhasi‘ ’et ’imo
David should to-drive ACC mother-POSS
David should drive his mother

An impersonal modal has no grammatical subject, and modality predicates the entire clause.

(2) צריך להסיע את אמו לעבודה
carik lhasi‘ ’et ’imo la‘abodah
should to-drive ACC mother-POSS to-the-work
His mother should be driven to work

Kopelovich makes no distinction between the various syntactic categories that the words may belong to, and interchangeably uses examples of words like מותר, יש, אפשר *mutar, yeš, ’epšar* [adverb, existential, participle respectively].

The Modality-Modulation plane, according to the functional school of Halliday (Halliday, 1985), refers to the interpersonal and ideational functions of language: Modality expresses the speaker's own mind (epistemic modality - possibility, probability and certainty) עלול לרדת גשם מחר *‘alul laredet gešem maḥar* (it may rain tomorrow). Modulation participates in the content clause expressing external conditions in the world (deontic modality - permission, obligation, ability and inclination): אתה יכול להתחיל עכשיו *’ata yakol lhatḥil ‘akšaw* (you can start now). Modality does not carry tense and cannot be negated, while modulation can be modified by tense and can be negative.

[3]For reference see (Kopelovich, 1982), see below for Rosen's analysis

The Objective-Subjective plane is what Kopelovich calls the *perception of the world*. Objectivity is achieved using different tenses of *to-be* in conjunction with the modal (including tense of modal if it is a verb), and their order subjective vs. objective:

(3) דוד היה צריך לנסוע לתל אביב
dawid haya carik lisw‘ ltel ’abib
David was have to-drive to-Tel Aviv
David had to drive to Tel Aviv

(4) כדי להעביר את ההחלטה, צריך היה לכנס את כל העובדים
kdei lha‘abir ’et hahaḥlata,
carik haya lkanes ’et kol ha‘obdim
In-order to-pass ACC the-decision,
should to-assemble ACC all the-employees
In order to obtain a favorable vote on this decision,
all of the employees had to be assembled.

Zadka (1995) defines the class of *single-argument bridge verbs*[4], *i.e.*, any verb or pro-verb that can have an infinitive as a subject or object, and that does not accept a subordinate clause as subject or object:

(5) אסור לעשן **[subject]**
’sur l‘ašen
Forbidden to-smoke
It is forbidden to smoke

(6) הוא רצה/התחיל לשחק **[object]**
hua racah/htḥil lšaḥeq
He wanted/started to-play
He wanted/started to play

(7) יוסף התחיל/עמד/מסוגל לקרוא את הדו"ח במלואו
Yosep hitḥil/’amad/msugal liqro’
’et hado”ḥ bimlo’o.
Yosef began/is-about/is-capable to-read
ACC the report entirely.
Yosef began/is-about/(is-capable) to read (of reading)
the report entirely.

(8) *יוסף התחיל/עמד/מסוגל שיקרא את הדו"ח במלואו
**Yosep hitḥil/’amad/msugal šiqra’*
’et hado”ḥ bimlo’o.
*Yosef started/was-about/is-capable that-he-read
ACC the report entirely.

Zadka classifies these verbs into seven semantic categories: (1) Will (2) Manner (3) Aspect (4) Ability (5) Possibility/Certainty (6) Necessity/Obligation

[4]"Ride Verb" in Zadka's terminology, פועל רוכב חד-מושאי

and (7) Negation. Categories 1, 4, 5, 6 and 7 are considered by Zadka to include *pure modal verbs*, *e*.g., *alethic* and *deontic* verbs.

In his paper, Zadka defines classification criteria that refer to syntactic-semantic properties such as: can the infinitive verb be realized as a relative clause, are the subject of the verb and its complement the same, can the infinitive be converted to a gerund, animacy of subject; deep semantic properties – argument structure and selectional restrictions, the ability to drop a common subject of the verb and its complement, factuality level (factual, non-factual, counter-factual); and morphological properties.

Will, Manner and Aspectual verbs as Zadka defines are not considered modals by Kopelovich since they can be inflected by tense (with the exceptions of אמור *'amur* (supposed),עתיד *'atid* (should). Ability verbs are יכול *yakol* (can), מסוגל *msugal* (can,capable) [participle]. They have both an animate actor as a subject and an infinitive as a complement, with the same deep subject. These verbs are counter-factual.

Certainty verbs include מוכרח *mukrak* (must), צריך *carik* (should), נאלץ *ne'elac* (be forced to), יכול *yakol* (can), הכרחי *hekreḥi* (necessary), עשוי *'aśuy* (may), עלול *'alul* (might), צפוי *capuy* (expected). They represent the alethic and epistemic necessity or possibility of the process realized in the clause. All of them cannot be inflected morphologically. The modal predicates the whole situation in the proposition, and may be subjective (epistemic) or objective (alethic). The subject of these verbs coreferences with the subject of the modal:

(9) אני מוכרח לקנות מכונית
'ani mukrak liqnot mkonit
I must to-buy car
I must buy a car

Necessity/Obligation includes adjectives – *e.g.*, חייב *ḥayab* (must), רשאי *raša'y* (allowed), gerunds – מוכרח *mukraḥ* (must), אסור *'asur* (forbidden), מותר *mutar* (allowed) and the verb יכול *yakwl* (can) [5]. Necessity verbs/proverbs present deontic modality, and all clauses share, in Zadka's view - a causing participant that is not always realized in the surface.

[5]as well as nouns and prepositions - among them יש and אין *yeš* and *'ein* - according to Zadka

From the morphological point of view, one may characterize impersonals by a non-inflectional behavior, *e.g.*, יש *yeš*, אין *'ein*, מותר *mutar*, אסור *'asur*, אפשר *'epšar*,אכפת *'ikpat*. All of these words do not inflect in number and gender with their argument. But this criterion leaves out all of the gender-quantity inflected words, *e.g.*, ראוי *ra'uy*, מסוגל *msugal*, עלול *'alul*, אמור *'amur*, צריך *carik*, יכול *yakol*, which are all classified as modals by Zadka. On the other hand, including all the gender-quantity inflected words with infinite or relative clause complements as modals, will include certain adjectives, *e.g.*, מוסמך *musmak* (certified), nouns, *e.g.*, זכות *zkut* (credit), and participles, *e.g.*, נמנע *nimna'* (avoid), as well. It appears that Zadka's classification relies primarily on semantic criteria.

Rosen (1977, pp. 113-115) defines a syntactic category he calls *impersonals*. Words in this category occur only as the predicative constituent of a sentence with an infinitive or a subordinate clause argument. Past and future tense is marked with the auxilary verb היה *hayah* (to-be). In addition, impersonals cannot function as predicative adjectives: כדאי *kda'i* (worthwhile), מוטב *mutab* (better), אכפת *'ikpat* (care/mind).

Personal reference can be added to the clause (governed by the infinitive) with the ל *l* dative preposition:

(10) כדאי לי לשתות
kda'y li lištot
worthwhile to-me to-drink
It is worthwhile for me to drink

2.1 Criteria to Identify Modal-like Words in Hebrew

We have reviewed three major approaches to categorizing modals in Hebrew:
Semantic - represented mostly in Kopelovich's work, modality is categorized by three dimensions of semantic attributes. Since her claim is that there is no syntactic category of modality at all, this approach 'over-generates' modals and includes words that from any other syntactic or morphologic view fall into other parts of speech.
Syntactic-semantic - Zadka classifies seven sets of verbs and pro-verbs following syntactic and semantic criteria. His claim is that modality actually is marked by syntactic characteristics, which can be

identified by structural criteria. However, his evaluation mixes semantics with syntactic attributes.
Morphological-syntactic - Rosen's definition of *Impersonals* is strictly syntactic/morpholgical and does not try to characterize words with modality. Consequently, words that are usually considered modals, are not included in his definition אסור *'asur* (forbidden), מותר *mutar* (allowed), יכול *yakol* (can).

3 Proposed Modal Guidelines to Identify

The variety of criteria proposed by linguists reflects the disagreements we identified in lexicographic work about modal-like words in Hebrew. For a computational application, all words in a corpus must be tagged. Given the complex nature of modality in Hebrew, should we introduce a modal tag in our tagset, or instead, rely on other existing tags? We have decided to introduce a modal tag in our Hebrew tagset. Although there is no distinct syntactic category for modals in Hebrew, we propose the following criteria: (i) They have an infinitive complement or a clausal complement introduced by the binder ש *š*. (ii) They are NOT adjectives. (iii) They have irregular inflections in the past tense, *i.e.*, רציתי לדעת *raciti lada'at* (I wanted to know) is not a modal usage.

The tests to distinguish modal from non-modal usages are:

- יש and אין which can be also existential, are used as modals if they can be replaced with צריך.
- Adjectives are gradable and can be modified by מאוד *m'od* (very) or יותר *yoter* (more).
- Adjectives can become describers of the nominalized verb: קל להרוס ⇒ ההריסה קלה מאוד *qal laharos* ⇒ *haharisah qala m'od* (easy to destroy ⇒ the destruction is easy).
- In all other cases where a verb is serving in to convey modality, it is still tagged as a verb, *e.g.*, מובן שיוסי הוא המנצח *muban šyosi hu' hamnaceḥ* (it is clear that Yossi is the winner).

We first review how these guidelines help us address some of the most difficult tagging decisions we had to face while building the corpus, we then indicate quantitatively the impact of the modal tag on the practical problem of tagging.

3.1 "What do I care" מה אכפת לי

One of the words tagged as a modal in our corpus - the word אכפת *'ikpat* – is not considered thus far to be a modal. However, at least in some of its instances it fits our definition of modal, and it can also be interpreted as modality according to its sense. The only definition that is consistent with our observation is Rosen's *impersonals*.

Looking back at its origins, we checked the Historical Lexicon of the Hebrew Language[6], the word אכפת was used in the medieval period in the Talmud and the Mishna, where it only appears in the following construction:

(11) מה אכפת לך
mah 'ikpat lk
what care to-you
what do you care

Similarly, in the Ben Yehuda Project – an Israeli version of the Guttenberg project[7] which includes texts from the Middle Ages up to the beginning of the 20th century – we have found 28 instances of the word, with the very same usage as in older times.
While trying to figure its part of speech, we do not identify אכפת as a NOUN - as it cannot have a definite marker ה[8], and is not an adjective[9].

Traditional Hebrew Dictionaries consider אכפת to be an intransitive verb (Kohut, 1926; Even-Shoshan, 2003; Avneyon et al., 2002) or an adverb. Some dictionaries from the middle of the 20th century (Gur, 1946; Knaani, 1960), as well as recent ones (Choueka et al., 1997) did not give it a part of speech at all.

In our corpus we found 130 occurences of the word אכפת of which 55 have an infinitive/relative clause complement, 35 have null complement, and 40 have מ *m* PP complement אכפת לו מהמדינה *'ikpat lo mehamdina* (he cares for the country). The latter has no modal interpretation. We claim that in this case it should be tagged as a participle (בינוני). The test to tell apart modal and participle is:

[6] http://hebrew-treasures.huji.ac.il/ an enterprise conducted by the Israeli Academy of the Hebrew Language.

[7] http://www.benyehuda.org, http://www.gutenberg.org

[8] Although we found in the internet clauses as נסתמו לי נקבוביות האכפת *nistmu li naqbubiyot ha'ikpat* (My caring pores got blocked).

[9] Only its derivatives אכפתי, אכפתיות *'ikpati*, *ikpatiyut* (caring, care) allows adjectival usage.

(12) אכפת לו לשטוף כלים ⇒
*הוא אכפתי כלפי שטיפת כלים
ikpat lo lištop kelim ⇒
**hu' 'ikpati klapei šṭipat kelim*
mind him to-wash dishes ⇒
*he concerned for washing dishes
He minds washing dishes ⇒
*He is concerned about washing dishes

(13) אכפת לו מהעניים ⇒
הוא אכפתי כלפי העניים
'ikpat lo meha'aniyim ⇒
hu' ikpati klapei ha'aniyim
care him of-the-poor-people ⇒
he caring for the-poor-people
He cares for the poor people ⇒
He is caring for the poor people

All other tests for modality hold in this case: (1) Infinitive/relative clause complement, (2) Not an adjective, (3) Irregular inflection (no inflection at all). To conclude this section, our proposed definition of modals allows us to tag this word in a systematic and complete manner and to avoid the confusion that characterizes this word.

3.2 "It's really hard" קשה לי

Some of the words tagged as modals are commonly referred to as *adjectives*, such as אסור, מותר *'asur, mutar* (allowed, forbidden), though everyone agrees - and tags these words as adverbs or participals (see table 1). However, questions are raised of how to tell apart modals as such from adjectives that show very similar properties: קשה לי ללכת *qaše li laleket* (it is hard for me to walk). Ambar (1995) analyzes the usage of adjectives in modal contexts, especially of *ability* and *possibility*. In sentences such as קשה לנו להסתגל לרעש *qaše lanu lhistagel lara'aš* (it is hard for us to get used to the noise) the adjective is used in a modal and not an adverbial meaning, in the sense that meaning of the adverbial בקושי *bqwši* (with difficulty) and the modal יכול *yakwl* (can) are unified into a single word קשה. Similarly, the *possibility* sense of קשה is unified with the modal אפשר *'epšar*. In any usage of the adjective as the modal, it is not possible to rephrase a clause in a way that the adjective modifies the noun, *i.e.*, the range is the action itself and not its subject.

(14) קשה לבצע את ההסכם
qaše lbace' 'et haheskem
hard to-perform PREP the-agreement
It is hard to perform the agreement

(15) *ההסכם קשה
haheskem kaše
the-agreement hard
The agreement is hard

However, following Ambar, there are cases where the usage of קשה ל *qaše le* is not modal, but an emotional adjective:

(16) קשה/נעים לשוחח איתו
qaše/na'im lšoḥeḥ 'ito
hard/pleasant to-chat with-him
It is hard/pleasant to chat with him

Berman (1980) classifies subjectless constructions in Modern Hebrew, and distinguishes what she calls *dative-marked experientials* where (mostly) adjective serves as a predicate followed by a dative-marked nominal

(17) קשה לרינה בחיים
qaše le-rinah baḥayim
hard for-Rina in-the-life
It is hard for Rina in life

Adjectives that allow this construction are circumstantial and do not describe an *inner state*: רינה עצובה *rina 'acuba* (Rina is sad) vs. עצוב לרינה *'acub lrina* (it is sad for Rina). Another recognized construction is the *modal expressions* that include sentences with dative marking on the individuals to whom the modality is imputed אסור לנו לדבר ככה *'asur lanu ldaber kakah* (we are not allowed to talk like this); Berman suggests that the similarity is due to the perception of the experiencer as recipient in both cases; This suggestion implies that Berman does not categorize the modals (*'asur, mutar*) as adjectives. Another possible criterion to allow these words to be tagged as modals (following Zadka) is the fact that for Necessary/Obligation modals there exists an 'outside force' which is the agent of the modal situation. Therefore, if אסור לנו לדבר ככה *'asur lanu ldaber kakah* (we are not allowed to talk like this), this is because someone forbids us from talking, while if קשה לרינה בחיים *qaše lrinah baḥayim* (It is hard for Rina in life) then no "outside force" is obliged to be the agent which makes her life hard. To

conclude - we suggest tagging both *'asur* and *mutar* as modals, and we recommend allowing modal tagging for other possible adjectives in this syntactic structure.

4 Conclusion

We recommend the introduction of a modal POS tag in Hebrew, despite the fact that the set of criteria to identify modal usage is a complex combination of syntactic and morphological constraints. This class covers as many as 3% of the tokens observed in our corpus.

Our main motivation in introducing this tag in our tagset is that the alternative (no modal tag) creates confusion and disagreement: we have shown that both traditional dictionaries and previous computational resources had a high level of disagreement over the class of words we tag as modals. We have confirmed that our guidelines can be applied consistently by human taggers, with agreement level similar to the rest of the tokens (over 99% pairwise). We have checked that our guidelines stand the test of the most difficult disagreement types identified by taggers, such as "care to" and "difficult for".

Finally, the immediate context of modals includes a high proportion of infinitive words. Infinitive words in Hebrew are particularly ambiguous morphologically, because they begin with the letter ל *l* which is a formative letter, and often include the analysis le+ participle, *e.g.* לשמור can be interpreted, depending on context, as *lišmwr* (to guard), *le-šamur* (to a guarded), or *la-šamur* (to the guarded). Other ambiguities might occur too, *e.g.*, לשיר can be interpreted as *lašir* (to sing), *le-šir* (to a song), or as *la-šir* (to the song). We have measured that on average, infinitive verbs in our expanded corpus can be analyzed in 4.9 distinct manners, whereas the overall average for all word tokens is 2.65. The identification of modals can serve as an anchor which helps disambiguate neighboring infinitive words.

References

Ora Ambar. 1995. From modality to an emotional situation. *Te'udah*, 9:235–245. (in Hebrew).

Eitan Avneyon, Raphael Nir, and Idit Yosef. 2002. *Milon sapir: The Encyclopedic Sapphire Dictionary*. Hed Artsi, Tel Aviv. (in Hebrew).

Ruth Berman. 1980. The case of (s)vo language: Subjectless constructions in Modern Hebrew. *Language*, 56:759–776.

David Carmel and Yoelle S. Maarek. 1999. Morphological disambiguation for Hebrew search systems. In *Proceeding of NGITS-99*, pages 312–326.

Yaacov Choueka, Uzi Freidkin, Hayim A. Hakohen, , and Yael Zachi-Yannay. 1997. *Rav Milim: A Comprehensive Dictionary of Modern Hebrew*. Stimatski, Tel Aviv. (in Hebrew).

Ferdinand de Haan. 2005. Typological approaches to modality in approaches to modality. In William Frawley, editor, *Approaches to Modality*, pages 27–69. Mouton de Gruyter, Berlin.

Oswald Ducrot and Tzvetan Todorov. 1972. *Dictionnaire encyclopédique des sciences du langage*. Éditions de Seuil, Paris.

Avraham Even-Shoshan. 2003. *Even Shoshan's Dictionary - Renewed and Updated for the 2000s*. Am Oved, Kineret, Zmora-Bitan, Dvir and Yediot Aharonot. (in Hebrew).

Yehuda Gur. 1946. *The Hebrew Language Dictionary*. Dvir, Tel Aviv. (in Hebrew).

M. A. K. Halliday. 1985. *An introduction to functional grammar*. Edward Arnold, USA, second edition.

Yaakov Knaani. 1960. *The Hebrew Language Lexicon*. Masada, Jerusalem. (in Hebrew).

Alexander Kohut. 1926. *Aruch Completum auctore Nathane filio Jechielis*. Hebraischer Verlag - Menorah, Wien-Berlin. (in Hebrew).

Ziona Kopelovich. 1982. *Modality in Modern Hebrew*. Ph.D. thesis, University of Michigan.

Uzi Ornan. 2002. Hebrew in Latin script. *Lěšonénu*, LXIV:137–151. (in Hebrew).

Haiim B. Rosen. 1977. *Contemporary Hebrew*. Mouton, The Hague, Paris.

Beatrice Santorini. 1995. Part-of-speech tagging guidelines for the Penn Treebank Project. 3rd revision;. Technical report, Department of Computer and Information Science, University of Pennsylvania.

Erel Segal. 2000. Hebrew morphological analyzer for Hebrew undotted texts. Master's thesis, Technion, Haifa, Israel. (in Hebrew).

Khalil Sima'an, Alon Itai, Alon Altman Yoad Winter, and Noa Nativ. 2001. Building a tree-bank of modern Hebrew text. *Journal Traitement Automatique des Langues (t.a.l.)*. Special Issue on NLP and Corpus Linguistics.

Shlomo Yona. 2004. A finite-state based morphological analyzer for Hebrew. Master's thesis, Haifa University.

Yitzhak Zadka. 1995. The single object "rider" verb in current Hebrew: Classification of modal, adverbial and aspectual verbs. *Te'udah*, 9:247–271. (in Hebrew).