Computational Methods for the Analysis of Complementizer Variability in Language and Literature: The Case of Hebrew *še-* and *ki*

Avi Shmidman
Bar-Ilan University
Dicta: The Israel Center for Text Analysis
avi.shmidman@biu.ac.il

Aynat Rubinstein
The Hebrew University
of Jerusalem
aynat.rubinstein@mail.huji.ac.il

Abstract

We demonstrate a computational method for analyzing complementizer variability within language and literature, focusing on Hebrew as a test case. The primary complementizers in Hebrew are *še-* and *ki*. We first run a large-scale corpus analysis to determine the relative preference for one or the other of these complementizers given the preceding verb. On top of this foundation, we leverage clustering methods to measure the degree of interchangeability between the complementizers for each verb. The resulting tables, which provide this information for all common complement-taking verbs in Hebrew, are a first-of-its-kind lexical resource which we provide to the NLP community. Upon this foundation, we demonstrate a computational method to analyze literary works for unusual and unexpected complementizer usages deserving of literary analysis.

1 Introduction

Natural languages offer speakers a variety of means for expressing their sentiment and attitude toward events, be they actualized or unactualized. While the literature has traditionally focused on lexical items that convey sentiment and attitude (verbs, adjectives, nouns; underlined in (1)-(2)), it is well known that functional morphemes such as subordinating particles (*complementizers*; boldfaced in the examples) and mood inflection are also implicated in the expression of such meanings in certain languages (Mauri and Sansò, 2016).

(1) We {are proud / believe } **that** our athletes did their very best.

(2) ha-tiqva **še-/ki** taxzeru mexazeqet.
the-hope COMP you.will.return strengthens
'The hope that you all will return is emboldening.' (Hebrew)

Recent work on the interaction between attitude predicates and the grammatical forms they "select" in the embedded clause has pointed to subtle semantic effects of choosing one complementizer over another (in Greek; Giannakidou and Mari 2021), or one inflected form over another (in Romance languages; Portner and Rubinstein 2020; Mari and Portner 2021). In contemporary Hebrew, the language we focus on in this paper, the variation between *še-* and *ki* as complementizers has not been recognized as relating to the grammar of embedding, and has often been attributed to register: *ki* is viewed as being more formal (see Nir 2013).[1]

Understanding patterns of clausal complementation in a language and the allowed range of *variation* is crucial for both comprehension and production. The distinctions are subtle and may seem arbitrary. They are known to present substantial difficulty even for advanced second language (L2) learners (e.g,. Bartning and Schlyter 2004 on French; Kanwit and Geeslin 2018 on Spanish).

This paper presents the first attempt we know of to explore aspects of complementizer distribution and use from a computational perspective, in Hebrew but also more generally. As we survey in Section 2, earlier computational studies of clausal embedding in attitude contexts have focused on English or on curated annotations. Corpus-informed studies have been limited to languages with mood inflection (e.g., French; Petkovic and Rabiet 2016), leaving complementizer variability unexplored. The contributions of the present paper are as follows:

- Enriched lexical representations of clause-taking verbs in Hebrew, with corpus-based statistics both regarding overall tendencies, as well as regarding the degree to which these tendencies are exaggerated or overridden in marked contexts.

- Demonstration of how this data can be lever-

[1] *Ki* has an additional use as a subordinating conjunction of reason ('because'). We set it aside in what follows.

Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities, pages 294–307
November 16, 2024.

aged to reveal the characteristics of specific marked contexts which require selection of one or the other of the complementizers. The ability to identify these marked contexts is a key component for L2 instruction.

- Presentation of a language-agnostic method to identify unusual usages of subtle linguistic elements in literary corpora.
- Application of this method to a corpus of modern Hebrew literature, identifying unusual specimens that invite further literary analysis.

2 Related work

Large-scale datasets allowing for the investigation of clausal embedding have been developed within the MegaAttitude Project.[2] In particular, the MegaAcceptability dataset (White and Rawlins, 2020) provides acceptability judgments on the distribution of 1,000 attitude verbs in 50 syntactic frames in English. Additional datasets explore inferences patterns associated with attitude verbs and their interaction with elements such as negation and tense (Moon and White 2020; Kane et al. 2021).

Özyıldız et al. (2023) provide a database of theoretically informed syntactic and semantic properties of a set of 50 attitude verbs in 15 languages. The rich linguistic profile of each verb (including its complement types, factuality inferences, interaction with negation, focus sensitivity, gradability and more) is summarized in a table based on experts' judgments in response to a questionnaire. Hebrew is included in this database, but information about complementizer variability is lacking from its description. Moreover, the database provides information about a small set of verbs and is based on translation from English, not on language-internal distributions.

Computational resources for languages that have observable mood inflection in embedded environments include mood as a target of morphological annotation (e.g., the Romance Verbal Inflection Dataset by Beniamine et al. 2020; Romance languages in Özyıldız et al. 2023). Petkovic and Rabiet (2016) provide a corpus-based description of mood variation in embedded clauses in French. However, languages in which mood is marked on a subordinating particle have not yet received attention from a computational perspective.

In the NLP literature on Hebrew, Fadida et al. (2014) provide a corpus-informed dictionary of about 3,000 verbs along with the number and type of complements they tend to occur with, including clauses. The two complementizers *še-* and *ki* were treated as interchangeable in this earlier work. Our work extends this lexicon with detailed information about the complementizer variability characteristic of each verb.

The question of whether authors of Hebrew literature adhere to the same complementizer tendencies as found in other text types has not been previously explored. Nevertheless, we note that in one study of embedded clauses in Hebrew (Kuzar, 1993), examples of complementizer use from Hebrew literature are cited alongside those of newspapers, without any differentiation between the corpora.

3 Experimental Setup

3.1 Corpus-based Query for Complementizer Propensities

As noted above, different Hebrew verbs show propensities for different complementizer choices. However, these propensities have never before been investigated from a corpus perspective, due to the difficulty involved in running such a wide-scale query. Fortunately, such a query is now tractable. We use it to establish the overarching tendencies regarding complementizer use.

Hebrew corpus: We start with a corpus of 29 million modern Hebrew sentences, sourced from Hebrew news sources, Hebrew Wikipedia, Israeli Parliament Proceedings, and published Hebrew books (Table 1). We filter out sentences that are too short to be useful (under 6 words) or those that are overly long (over 60 words).[3]

Syntactic analysis: In order to identify cases of clausal complements, we obtain a syntactic parse for all of the sentences in our corpus. Currently, the two leading syntax parsers for modern Hebrew are the Levi-Tsarfaty parser (Levi and Tsarfaty, 2024) and DictaBERT-Parse (Shmidman et al., 2024). The former, though also achieving SOTA results, is not available to the public and is too

[2] http://megaattitude.io/

[3] In addition to this corpus, we also utilize a corpus of ten novels of the Hebrew novelist Amos Oz for a specific inquiry regarding Hebrew literature later in Section 6; see Appendix B for the list of books in that corpus.

Domain	Sentences (millions)
Hebrew Newspapers	22
Hebrew Wikipedia	5
Parliament Proceedings	1
Published books, Fiction	0.6
Published books, Non-fiction	0.4

Table 1: Sentences per domain in our Hebrew corpus.

slow to be tractable for a corpus as large as ours. We therefore used the latter. Parsing of the entire corpus required 73 hours on a single 4090 GPU.

In the parsed corpus, we examine all cases in which a verb has a "ccomp" dependency, indicating a clausal complement. We retrieve the complementizer by extracting the earliest token in the sentence within the scope of the clausal complement. The tabulation of the results of this query identifies the general tendencies of each verb in terms of complementizer choice.

3.2 Deviations from Complementizer Propensities

After calculating the overall complementizer propensities of each complement-taking verb, we wish to clarify the extent to which these propensities remain constant across usages of the verb, or whether there are specialized usages of the verb that exaggerate or override the general tendencies.

Especially interesting here is the question of whether certain usages or contexts entail exclusive use of one or the other of the complementizers. Received syntactic descriptions hold that either of the two complementizers can be used with virtually every clause-embedding attitude verb. However, in practice, there are cases where native Hebrew speakers will only find one of the two to be acceptable, while the other would sound unnatural.[4] The method we present here allows us to pinpoint such cases.

Prima facia, in order to identify cases in which only one of the complementizers is used in practice, we might have considered simply running a BERT Masked Language Model (MLM) to see whether only one of the complementizers is predicted for a given context. However, in practice, given a masked token in place of the complementizer, BERT will almost always provide both of the complementizers among its top predictions, because there simply aren't that many other options to fill the slot. That is, even if it would sound odd to a native speaker, if BERT's MLM head is pressed to choose a word to fill a complementizer function, and if the more usual complementizer has already been predicted, it will generally provide the other one, because, from a technical syntactic standpoint, both of them can theoretically function as a complementizer with any complement-taking verb. Instead, in order to gain a better sense of the extent of complementizer interchangeability, we examine contextualized embeddings for the complementizer positions, and we consider the extent to which the embeddings cluster into complementizer-specific sections, as follows:

Generating contextualized embeddings: For each complement clause identified in the previous step, we mask the complementizer, and we submit the sentence to BERT to generate a contextualized embedding for that masked token, independent of whether the complementizer was in fact *še-* or *ki*.[5]

2D Visualization: In order to visualize the interchangeability of the two complementizers across different contexts with the same governing verb, we reduce the 768-dimension space of the BERT embeddings using the t-SNE algorithm, and generate a two-dimensional plot of the embedding space for each complement-taking verb. We color the points based on the complementizer present in the corresponding sentence. As we demonstrate below, visual inspection of the relative distribution of the two colors across the plot allows us to easily and immediately identify areas of aberrations, representing specific contexts in which the tendency towards one complementizer or the other differs from the overall tendency within the corpus.

Clustering the embeddings: We add a clustering step to automatically isolate contexts with specialized complementizer tendencies. For each complement-taking verb, we collect the complementizer embeddings generated in the previous step (for practicality, we set a limit of 20,000 cases for any given verb; if the corpus contains more than this, then we randomly sample 20,000 cases

[4]For examples of such, see Section 6 below.

[5]In order to ensure that the embeddings are attuned to the nature of Hebrew prefixes (such as *še-*), we use a variation of the DictaBERT model. Leveraging the segmentation predictions of DictaBERT-Parse (Shmidman et al., 2024), we separate all prefixes in the DictaBERT training corpus into independent tokens, and then we run a new BERT pre-train based upon this prefix-separated corpus. The resulting BERT model is used to generate the embeddings for this step.

from across the corpus). We apply agglomerative clustering to these embeddings, with euclidean distance and average linking. We let the agglomerating continue until a majority of the samples have been clustered into the top three clusters. This ensures that the clustering process continues sufficiently long such that the majority usages of the word are clustered together in a few substantially-sized clusters, while still providing ample opportunity for specialized usages to occupy smaller individual clusters.

The key part of this clustering step is that neither the BERT embedding nor the clustering procedure has any information about the complementizer used in the sentence. This means that the algorithm cannot directly choose to cluster together sentences on the basis of the complementizer; rather, the clustering is based on the context alone. Thus, if a verb's tendencies regarding complementizer usage are context-independent, then we expect the resulting clusters to each contain a mixture of *še-* and *ki* cases, reflecting the overall tendency of the verb towards one or the other. However, if we find clusters that are highly divergent from the overall tendency, this indicates that the types of contexts included in those clusters entail specialized complementizer tendencies.

In order to automatically evaluate the degree to which a given cluster diverges from the overall norm for the governing verb, we calculate the Jensen–Shannon Divergence (JSD) for each resulting cluster (discarding tiny outlier clusters of under 100 sentences); we consider a cluster to reflect a divergent complementizer tendency if it bears a JSD score higher than 0.04. On this foundation, we calculate, for each governing verb, the percentage of sentences that were clustered into divergent clusters. The result provides a measure of the extent to which the verb's overall complementizer tendency holds true across the range of practical usages of the verb, in contrast with the extent to which the verb admits of specialized usages which affect its complementizer selection, and which an L2 learner would have to internalize in order to speak in a fully natural manner.

3.3 Pinpointing unusual usages

Finally, we wish to leverage the foregoing infrastructure to identify cases in which literary authors deviate from normative usage by choosing an unexpected complementizer, inviting literary analysis of the unusual choice. In order to do so, we run a set of modern Hebrew novels through the process above, isolating all cases of complement clauses, and generating contextualized embeddings for the complementizer in each case. We then use a K-nearest-neighbor classifier (with k=3) in order to classify each one of these cases according to the clusters for the corresponding verb that we produced in the previous step, based on the full large-scale Hebrew corpus. The cluster assignment provides us with a sense of the expectations for complementizer selection, given both the specific verb and the specific context of use. If the cluster assignment indicates a context in which one of the two complementizers is expected with a probability of over 95% (that is, a context in which the complementizers are effectively not interchangeable, but rather one is blocked in practical usage), and if the author nevertheless chose the *other* complementizer, then we flag the sentence as reflecting an unusual and unexpected complementizer choice.

4 Enriched Verbal Lexicon

Table 2 shows the overall proportion of the two complementizers for the ten most frequent clause-embedding verbs among the sentences analyzed. Tables 4-6 in Appendix A provide this information for the 100 most frequent clause-embedding verbs, along with additional statistical measures to be described below. A visualization of complementizer proportions is provided in Figure 1 for nine verbs. The wide range of complementizer variability observed across different Hebrew verbs has never before been quantified.

Verb	*ki*	*še-*
amar ('said')	55%	45%
taʕan ('asserted')	65%	35%
xašav ('thought')	4%	96%
cuyan ('was mentioned')	86%	14%
hodiʕa ('informed')	74%	26%
qava ('decided')	72%	28%
yada ('knew')	12%	88%
heʕerix ('estimated')	70%	30%
siper ('told')	51%	49%
hevin ('understood')	17%	83%

Table 2: Ten most frequent verbs with *ki/še-* clausal complements.

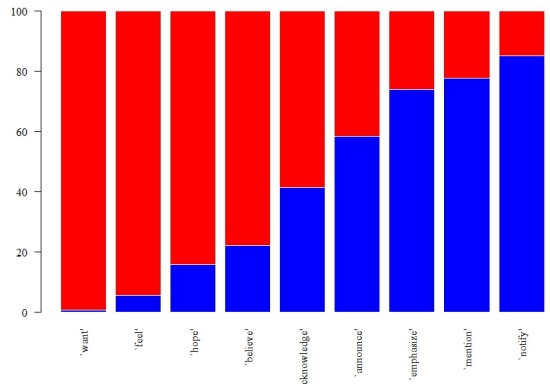

Figure 1: Proportion of *še-* (red) and *ki* (blue) complementizers for nine verbs. From left to right: *raca* ('want', 99% *še-*), *hirgiš* ('feel'), *qiva* ('hope'), *he?emin* ('believe'), *hoda* ('acknowledge'), *hixriz* ('announce'), *hidgiš* ('emphasize'), *ciyen* ('mention'), *masar* ('notify', 15% *še-*).

Previous literature noticed semantic trends in complementizer choice, but did not leave room for variation. For example, Zuckermann (2006) suggests the existence of a class of desire ("liking") verbs which categorically disallow *ki* (pp. 81-82). Our data shows that association with this complementizer in fact forms a scale among verbs that express desire and preference, with *roce* ('wants') on one end, *meqave* ('hopes') on the other end, and *maʕadif* ('prefers') somewhere in between (1%-8%-16% occurrence with *ki*).

Similarly, our LLM-based method uncovers more variation than is apparent from existing resources. An example is the verb *megale*, which Zuckermann (2006, 87) translates as 'discovers' and classifies as unlikely to occur with *ki*. In our corpus, a substantial 36% of the verb's occurrence with an overt complementizer are in fact with *ki*. The cluster that is most strongly associated with this complementizer uncovers a second use of the verb, shown in (3). In this use the verb is associated with inanimate subjects and conveys the meaning of 'reveals'.

(3) biqur be-yapan megale ki hakol
visit in-Japan reveals COMP everything
yaxasi.
relative
'A visit to Japan reveals that everything is relative.'

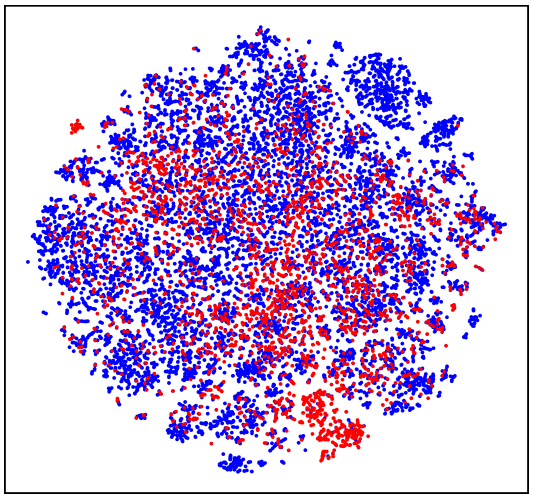

Figure 2: t-SNE plot visualizing complementizer interchageability for the verb עלי_קל *ʕala* ('arose').

5 Complementizer Propensities: Visualization and Analysis

As described above, in order to visualize the differences in complementizer variability across different types of usages of a given complement-taking verb, we generate t-SNE plots of the contextualized embeddings of the complementizers for a large set of sentences (up to 20K) for each verb.

For example, in Figure 2, we plot 20,000 instances of complementizers which open a subordinate clause for the verb *ʕala* (in the *qal* conjugation; see עלי_קל in Table 5). Each sentence is represented by a single dot; blue points represent the complementizer *ki*, and red points represent the complementizer *še-*. Crucially, the contextualized embeddings and the t-SNE plot were all computed with a mask over the complementizer. That is, those processes had no knowledge of the label for any given point; the colors were added afterward according to our ground truth labels. Thus, homogeneous sections of a single color on the plot reflect types of sentences which normatively are used with only one or the other of the complementizers.

To be sure, Figure 2 contains far more blue than red, indicating that on the whole, this verb is generally used with the complementizer *ki*. However, red points are interspersed throughout the plot, indicating that the complementizer *še-* is also attested as a practical option for the same types of sentences; i.e., in almost all cases of complement clauses with the verb *ʕala*, the two complementiz-

ers can be freely interchanged without worrying that the resulting sentence will sound odd or unnatural to present-day Hebrew readers.

However, there are two substantially-sized homogeneous sections on the plot, which indicate specific uses of one or the other of the complementizers. One such section is the red cluster at the bottom of the graph. Inspection of the sentences represented by these points reveals that they all bear inflections of the idiomatic phrase **לא יעלה על הדעת** (*lo yaʕale ʕal ha-daʕat* 'it is inconceivable'; lit. 'it would not rise up upon the mind'). The implication, therefore, is that although this verb can generally be used with either complementizer, when it comes to its use within this idiomatic phrase, it is almost exclusively used with the complementizer *še-*.

Conversely, at the top right of the plot, we find a homogeneously blue section. Inspection of these sentences reveals what appears to be a typographic concern: all of the sentences contain a specification of a percentage statistic immediately after the complementizer, written out in digits and a percent sign as in (4). For instance:

(4) me-ha-duax ʕole ki 83%
from-the-report arises COMP 83%
me-ha-maʕasiqim ...
of-the-employers
'From the report one gleans that 83% of the employers ...'

In these cases, the orthographic distinction between the two complementizers comes into play. Whereas the two-letter complementizer *ki* is written as an independent word, the single letter complementizer *še-* is prefixed in print to the subsequent word. Typing a single Hebrew letter immediately adjacent to a sequence of numbers and the percent sign may result in jumbled text in some text editors (which have problems combining the right to left Hebrew text with numbers, which are written from left to right), or the visual anomaly of a single hanging letter in the text may lead writers to insist on the complementizer *ki* in such situations.

In contrast, the plot for the verb *hoxiax* (Figure 3) has no solid homogeneous clusters; rather, the red points are fairly evenly interspersed throughout the plot. This indicates that for this verb, both complementizers are accessible. Even though the use of *ki* is somewhat more frequent with this verb, it can optionally be switched out for *še-* without violating the expectations of native speakers.

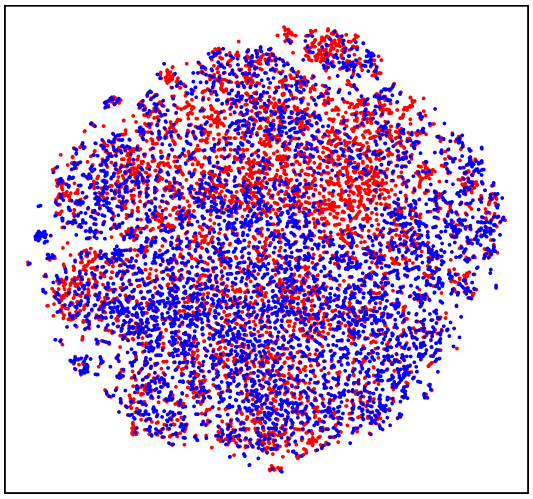

Figure 3: t-SNE plot visualizing complementizer interchageability for the verb **יכח_הפעיל** *hoxiax* ('proved').

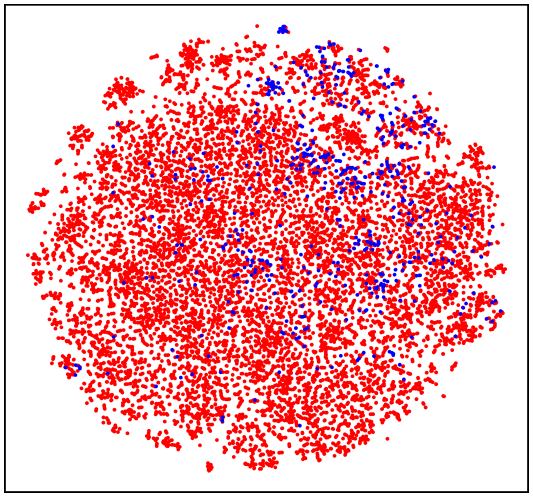

Figure 4: t-SNE plot visualizing complementizer interchageability for the verb **נגד_הפעיל** *higid* ('told').

A reverse phenomenon emerges from the plot for the verb *higid* (Figure 4), a suppletive form in the paradigm of 'said, told'. As opposed to the plot for *ʕala*, in which there were a few isolated homogeneous clusters, here the predominant majority of the plot is homogeneously red, reflecting the fact that this verb is used almost exclusively with the complementizer *še-* when it embeds a clause; the complementizer *ki* is generally not a practical option. However, there are a number of areas in the plot where we find blue points interspersed among the red. Inspection of these points reveals that they are sentences with a particular syntactic-semantic profile, exemplified in (5).

(5) a. nitan lehagid ki avodato šel
possible to.say COMP work.3MSG of

hoqusai ...
Hokusai
'It is possible to say that Hokusai's work (had a profound effect on popular themes in woodcut printmaking).'

b. yeš še-yagidu ki saxqaney
EXIST that-will.say COMP players
fifa ...
FIFA
'Some will say that FIFA players (are not considered real players).'

In these cases, the impersonal nature of the sentence is correlated with the deviation from the verb's general tendency of complementizer selection.

The three t-SNE plots that we have analyzed in this section demonstrate the value of inspecting divergent clusters: the inspection allows researchers to identify the characteristics of clusters which do not follow a verb's general embedding tendency. The ability to do this is of high importance both for NLP purposes and for L2 instruction. In both cases, knowing the general tendency is helpful: in NLP, this allows us to build stronger parsers which are informed by this expectation and which utilize the expectation when disambiguating the sentence tokens; and for L2 learners, knowing the general verb-complementizer pairings can allow them to formulate their sentences in accordance with the expectations of native speakers. However, knowledge of the overarching tendency of a verb is only useful up to a point. Given the clusters we have seen which highlight specialized usages, blind pursuit of the general tendency can lead one astray, whether for L2 language production or for NLP sentence parsing. Instead, the key is to both know the general tendencies of the verbs, and also to know how to identify sentences groups in which those general tendencies do not apply, and may even be flipped. Our t-SNE plots provide an effective method to inspect the usage of the verb overall, and to hone in on examples that constitute specialized usages.

Of course, to the extent possible, we prefer to automatically quantify the extent to which a verb's overall tendency regarding complementizer selection will apply consistently across the full range of its uses, without the need to resort to manual inspection of t-SNE plots. In order to do so, we run an automatic clustering routine on the sentences of a given verb, as described above, and we then calculate the Jensen-Shannon Divergence (JSD) of each substantially-sized cluster in order to identify divergences.

We utilize this method to calculate corpus-based complementizer statistics for the 100 most frequent complement-taking verbs in contemporary Hebrew. The first part of the table is presented here (Table 3); the full table is presented in Appendix A.[6] We provide this as a lexical resource for future Hebrew NLP work.

For each verb, we first present the overall statistical tendency towards one or the other of the complementizers, and we note the number of sentences that the statistics were based upon. Additionally, we measure the extent to which the statistical tendency holds true across the corpus, and the extent to which we find specialized usages of the verb in which the tendency is exaggerated or flipped. Regarding many of the verbs, the last three columns contain zeroes, indicating that the overall balance between the two complementizers remains stable for the given verb across the corpus. In contrast, other verbs reveal specialized clusters to greater or lesser degrees. For instance, for the verb *biqeš* ('asked'), the first few columns indicate that in general *še-* is far more likely; at the same time, in 10% of the cases, there is a much stronger affinity to *ki*. On the flipside, regarding *he?ešim* ('accused'), the preference is generally towards *ki*, but 14% of the sentences cluster into groups in which the tendency is flipped toward *še-* instead. Finally, some verbs split in both directions. For instance, for *amar* ('said'), the overall statistics point to balanced usage between the two complementizers, but the divergence columns indicate that, in fact, the usage is often not balanced at all: 21% of the sentences are in clusters that show a specific preference for *ki*, and 16% of the sentences are in clusters that show a preference for *še-*. In other words, in over a third of the corpus, it is not the case that the two complementizers are equally interchangeable, but rather, the varying contexts in which *amar* occurs determine which complementizer is expected.

6 Unusual Complementizers in Hebrew Literature

As explained above in Section 3.3, we propose the use of complementizer clusterings in order

[6] Translations provided for each verb represent its most salient meaning as a clause-embedding predicate; other translations may be more appropriate in specific contexts.

Root	Samples	*ki*	*še-*	Cases in Divergent Clusters	Divergence toward *ki*	Divergence toward *še-*
אות_פיעל *otet* ('signal')	3532	57%	43%	0%	0%	0%
אים_פיעל *iyem* ('threaten')	5330	56%	44%	0%	0%	0%
אמן_הפעיל *he?emin* ('believe')	20000	22%	78%	3%	0%	3%
אמר_קל *amar* ('said')	20000	55%	45%	37%	21%	16%
אשם_הופעל *ho?ašam* ('was accused')	3175	63%	37%	5%	5%	0%
אשם_הפעיל *he?ešim* ('accused')	4328	62%	38%	14%	0%	14%
אשר_פיעל *išer* ('confirmed')	16066	70%	30%	0%	0%	0%
בהר_הפעיל *hivhir* ('clarified')	20000	64%	36%	4%	4%	0%
בון_הפעיל *hevin* ('understood')	20000	17%	83%	0%	0%	0%
בחן_הפעיל *hivxin* ('noticed')	5011	40%	60%	0%	0%	0%
בטח_הפעיל *hivtiax* ('promised')	20000	34%	66%	1%	0%	1%
בקש_פיעל *biqeš* ('asked')	20000	16%	84%	10%	10%	0%

Table 3: Complementizer Propensities and Divergencies (Initial part of the table; full table appears in Appendix A)

to identify places in which literary authors deviate from the norm. We analyze all instances of clausal complements within a corpus of novels by the modern Israeli author Amos Oz (see Table 7 in Appendix B for the list of books in this corpus). For each case, we extract the verb which governs the complement, and we then run a K-nearest-neighbor routine to classify the sentence within one of the clusters of that verb (as per the clustering from Section 3.2). We then query this data for cases in which the relevant clusters are highly homogeneous - indicating a preference 95% or higher for one specific complementizer - yet the novelist deliberately chooses the other option. Effectively, in these case, the novelist subtly undermines the reader's expectations.

Our first example, in (6), exemplifies an unexpected use of *ki* in a story by Oz (Oz, 1976, p. 57):

(6) biršuta agid la ki
with.her.permission will.say.1SG to.her COMP
lo beit marzeax kan
not tavern here
'With her permission, I will tell her that it is not a tavern here.'

The use of the complementizer *ki* with the verb 'told' is exceedingly rare in general. As we saw above, the one cluster in which this verb is naturally used with *ki* is when the statement is impersonal, with the subject generally unspecified. Yet, the statement in (6) could hardly be more personal; it is phrased in the first person, with a personal plea at the beginning ('please my lady' introduces the sentence we see here). The use of *ki* as a complementizer in this context conflicts with the reader's expectations, and characterizes the statement as subtly unusual. And, indeed, in this paragraph, Oz wishes to paint this character - described in the book as an "elderly poet" - as one who interacts with fairly archaic Hebrew expression. In addition to the originally biblical *ki*, in the continuation of the paragraph, this character uses a number of other archaic (Biblical or Talmudic) words and phrases, such as the negative interrogative **הלא** *hălō*, as well as **אימתי** *ʔeymatay* ('when'), and **מי אנוכי כי אדע** *mī ʔānōḵī kī ʔēdaʕ* ('who am I to know'). The coupling of the verb 'told' with the complementizer *ki*, while exceedingly unusual for contemporary Hebrew, is in fact well-attested in Biblical Hebrew (e.g. Genesis 3:11, Genesis 31:20, Psalms 92:16, and more). Oz's selection of this complementizer is thus clearly deliberate, serving to help characterize the Biblical idiom of the "elderly poet".

The complementizer *še-* is the more general of the two complementizers (no verb exclusively selects *ki*, as can be seen in the table in Appendix A) and is often thought to be a general-purpose complementizer in Hebrew. However, not all uses of *še-* are equally felicitous. Example (7) is highlighted by our procedure as an unexpected use of the complementizer *še-* in Oz's prose (Oz, 1986, p. 168):

(7) ve-od raciti lehodiaʕaxa
and-more wanted.1SG to.inform.you
še-me-ha-mixtav še-šalaxta lanu
that-from-the-letter that-sent.2MSG to.us

imxa baxta bi-dmaʕot
your.mother cried in-tears
'And I also wanted to inform you that your mother wept in tears from the letter you sent us.'

In general, the preferred complementizer with the verb *hodiʕa* ('inform') is *ki* - 74% across the whole corpus. Moreover, this particular sentence is classified as part of a cluster in which the preference for *ki* is far more extreme: over 95%. The cluster includes multiple specimens of the verb with a second-person pronominal suffix, as we find in Oz's novel. Such phrases are typical of formal and legal documents, which dish out an objective and impersonal ruling; hence the overwhelming preference for *ki*. Yet, Oz's context is not legal at all. Rather, it is from a letter written by a woman's second husband, in which he struggles to connect with his step-son, a boy portrayed as unruly and rough, both in character and in his use of language. Oz's formulation reflects the letter-writer's struggle in this endeavor. On the one hand, the sentence begins with the highly formal legalese "to inform you" - a phrasing that normally creates a distanced atmosphere. Yet, the unexpected choice of the less-formal complementizer can be seen as an attempt to step back and make the message more accessible to the boy, more personal and more sensitive.

In sum, our method identifies cases where a Hebrew literary master makes a complementizer choice that goes against the grain of how attitudes are usually expressed, inviting further literary analysis to suggest what may have motivated the oddity.

7 Conclusion

This paper demonstrates a language-agnostic method to run a large-scale corpus-based investigation of complementizer variability. We show how this method can be used to isolate cases where authors deviate strikingly from an expected complementizer; such aberrations may well reflect a deliberate literary choice, and invite literary analysis. We apply this method to contemporary Hebrew.

This is the first time that Hebrew complementizers have been investigated from a large-scale corpus-based perspective. Contra the perceived view about Hebrew, we find that the language does have grammatical marking of mood: not in the verbal morphology, but in its subordinating particles. Moreover, there is not a true subset relation between the uses of the two complementizers; there are verbs that strongly prefer *ki* and allow *še-* only under highly specific contexts. We provide results for complementizer selection regarding the top 100 clause-embedding verbs in contemporary Hebrew. We expect that this first-of-its-kind lexical resource will comprise a helpful resource both for L2 learners, as well as for Hebrew NLP researchers.

Limitations

We demonstrate the ability to identify specialized usages of a verb whose complementizer tendencies differ from the general use of the verb. However, because this method depends on the existence of deviant clusters which highlight the specialized usages, it is inherently limited to usages that are sufficiently well-attested. If a specialized usage only occurs in a few dozen sentences in the corpus, then the exceedingly small cluster that they form will not be sufficient to provide a robust statistic about their complementizer tendencies.

Another limitation inherent in our method is that although we succeed in automatically isolating clusters with specialized complementizer tendencies, we do not currently possess the ability to automatically identify what it is that uniquely characterizes the sentences in that cluster. Rather, once a specialized cluster is identified, it requires human inspection in order to extract the generalized property of the sentences therein.

Ethics Statement

Hebrew data are provided with transliteration and translation as well as in standard Hebrew script in order to increase the accessibility of the paper to native speakers and readers of Hebrew. Transliteration follows the widely used guidelines of the Encyclopedia of Hebrew Language and Linguistics (Khan, 2013).

Acknowledgements

The work of the first author has been funded by the Israel Science Foundation (Grant No. 2617/22) and by the European Union (ERC, MiDRASH, Project No. 101071829), for which we are grateful. Views and opinions expressed are, however, those of the authors only and do not necessarily reflect those of the European Union or the European

Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

References

Inge Bartning and Suzanne Schlyter. 2004. Itinéraires acquisitionnels et stades de développement en français L2. *Journal of French Language Studies*, 14(3):281–299.

Sacha Beniamine, Martin Maiden, and Erich Round. 2020. Opening the Romance verbal inflection dataset 2.0: A CLDF lexicon. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3027–3035, Marseille, France. European Language Resources Association.

Hanna Fadida, Alon Itai, and Shuly Wintner. 2014. A Hebrew verb–complement dictionary. *Language Resources & Evaluation*, 48:249–278.

Anastasia Giannakidou and Alda Mari. 2021. *Truth and Veridicality in Grammar and Thought: Mood, Modality, and Propositional Attitudes*. University of Chicago Press, Chicago.

Benjamin Kane, Will Gantt, and Aaron Steven White. 2021. Intensional gaps: Relating veridicality, factivity, doxasticity, bouleticity, and neg-raising. In *Proceedings of Semantics and Linguistic Theory (SALT) 31*, pages 570–605.

Matthew Kanwit and Kimberly L. Geeslin. 2018. Exploring lexical effects in second language interpretation: The case of mood in Spanish adverbial clauses. *Studies in Second Language Acquisition*, 40(3):579–603.

Geoffrey Khan, editor. 2013. *Encyclopedia of Hebrew Language and Linguistics*. Brill.

Ron Kuzar. 1993. Nominal clauses in Israeli Hebrew. *Hebrew Linguistcs*, 36:71–89. In Hebrew.

Danit Yshaayahu Levi and Reut Tsarfaty. 2024. A truly joint neural architecture for segmentation and parsing.

Alda Mari and Paul Portner. 2021. Mood variation with belief predicates: Modal comparison and the raisability of questions. *Glossa: a journal of general linguistics*, 40(1).

Caterina Mauri and Andrea Sansò. 2016. The Linguistic Marking of (Ir)Realis and Subjunctive. In *The Oxford Handbook of Modality and Mood*. Oxford University Press.

Ellise Moon and Aaron Steven White. 2020. The source of nonfinite temporal interpretation. In *Proceedings of the 50th Annual Meeting of the North East Linguistic Society (NELS)*, pages 11–24.

Bracha Nir. 2013. Complementizer. In Geoffrey Khan, editor, *Encyclopedia of Hebrew Language and Linguistics*. Brill, Leiden.

Amos Oz. 1976. *The Hill of Evil Council*. Keter Publishing House.

Amos Oz. 1986. *Black Box*. Am Oved Publishers.

Deniz Özyıldız, Ciyang Qing, Floris Roelofsen, Maribel Romero, and Wataru Uegaki. 2023. A crosslinguistic database for combinatorial and semantic properties of attitude predicates. In *Proceedings of the 5th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 65–75, Dubrovnik, Croatia. Association for Computational Linguistics.

Divna Petkovic and Victor Rabiet. 2016. La polysémie lexicale et syntaxique de l'alternance modale indicatif/subjonctif – perspectives TAL (lexical and syntactic polysemy of the modal alternation indicative/subjunctive – NLP perspectives). In *Actes de la conférence conjointe JEP-TALN-RECITAL 2016. volume 3 : RECITAL*, pages 80–93, Paris, France. AFCP - ATALA.

Paul Portner and Aynat Rubinstein. 2020. Desire, belief, and semantic composition: variation in mood selection with desire predicates. *Natural Language Semantics*, pages 343–393.

Shaltiel Shmidman, Avi Shmidman, Moshe Koppel, and Reut Tsarfaty. 2024. MRL parsing without tears: The case of Hebrew. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 4537–4550, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Aaron White and Kyle Rawlins. 2020. Frequency, acceptability, and selection: A case study of clause-embedding. *Glossa: a journal of general linguistics*, 5(1):105.

Ghil'ad Zuckermann. 2006. Complement clause types in Israeli. In Dixon R. M. W. and Aikhenvald Alexandra Y., editors, *Complementation: A Cross-Linguistic Typology*, Vol. 3. Oxford University Press.

A Appendix: Full Statistics for the Top 100 Verbs

As described above, we calculate corpus-based statistics regarding complementizer propensities for the top 100 complement-taking verbs in the Hebrew language, presented here in a table across the next three pages. For a detailed explanation of the fields in this table, see above, end of Section 5.

Root	Samples	*ki*	*še-*	Cases in Divergent Clusters	Divergence toward *ki*	Divergence toward *še-*
אות_פיעל *otet* (‘signal’)	3532	57%	43%	0%	0%	0%
אים_פיעל *iyem* (‘threaten’)	5330	56%	44%	0%	0%	0%
אמן_הפעיל *heʔemin* (‘believe’)	20000	22%	78%	3%	0%	3%
אמר_קל *amar* (‘said’)	20000	55%	45%	37%	21%	16%
אשם_הופעל *hoʔašam* (‘was accused’)	3175	63%	37%	5%	5%	0%
אשם_הפעיל *heʔešim* (‘accused’)	4328	62%	38%	14%	0%	14%
אשר_פיעל *išer* (‘confirmed’)	16066	70%	30%	0%	0%	0%
בהר_הפעיל *hivhir* (‘clarified’)	20000	64%	36%	4%	4%	0%
בון_הפעיל *hevin* (‘understood’)	20000	17%	83%	0%	0%	0%
בחן_הפעיל *hivxin* (‘noticed’)	5011	40%	60%	0%	0%	0%
בטח_הפעיל *hivtiax* (‘promised’)	20000	34%	66%	1%	0%	1%
בקש_פיעל *biqeš* (‘asked’)	20000	16%	84%	10%	10%	0%
ברר_התפעל *hitbarer* (‘turned out’)	20000	43%	57%	0%	0%	0%
בשר_פיעל *biser* (‘apprised’)	3191	43%	57%	0%	0%	0%
גלי_התפעל *hitgala* (‘was discovered’)	3665	58%	42%	0%	0%	0%
גלי_פיעל *gila* (‘discovered, revealed’)	20000	36%	64%	1%	1%	0%
גרס_קל *garas* (‘held’)	10602	64%	36%	2%	2%	0%
דאג_קל *daʔag* (‘ensure’)	6472	7%	93%	0%	0%	0%
דגש_הפעיל *hidgiš* (‘emphasized’)	20000	74%	26%	0%	0%	0%
דוח_פיעל *diveax* (‘reported’)	20000	83%	17%	0%	0%	0%
דרש_קל *daraš* (‘demanded’)	15093	37%	63%	3%	0%	3%
ודא_פיעל *vide* (‘confirmed’)	14358	17%	83%	1%	0%	1%
זהר_הפעיל *hizhir* (‘warned’)	16870	71%	29%	17%	9%	8%
זכר_הפעיל *hizkir* (‘reminded’)	19241	44%	56%	0%	0%	0%
זכר_נפעל *nizkar* (‘recalled’)	3295	17%	83%	0%	0%	0%
זכר_קל *zaxar* (‘remembered’)	20000	21%	79%	24%	0%	24%
חוש_קל *xaš* (‘sensed’)	10731	27%	73%	0%	0%	0%
חזי_קל *xazi* (‘predicted’)	2929	65%	35%	0%	0%	0%
חכי_פיעל *xika* (‘waited for, anticipated’)	2407	1%	99%	0%	0%	0%
חיב_התפעל *hitxayev* (‘obligated oneself’)	4573	53%	47%	3%	0%	3%
חלט_הופעל *huxlat* (‘was decided’)	2542	62%	38%	0%	0%	0%
חלט_הפעיל *hexlit* (‘decided’)	20000	32%	68%	22%	16%	6%
חשב_קל *xašav* (‘thought’)	20000	4%	96%	0%	0%	0%
חשד_קל *xašad* (‘suspected’)	10822	47%	53%	6%	0%	6%
חשף_קל *xasaf* (‘exposed’)	5956	80%	20%	0%	0%	0%
חשש_קל *xašaš* (‘worried’)	20000	31%	69%	2%	0%	2%
טען_קל *taʕan* (‘asserted’)	20000	65%	35%	3%	3%	0%
ידי_הפעיל *hoda* (‘acknowledged’)	20000	41%	59%	8%	5%	3%
ידע_הפעיל *hodiʕa* (‘informed’)	20000	74%	26%	8%	2%	6%
ידע_נפעל *noda* (‘became known’)	10019	73%	27%	21%	0%	21%
ידע_קל *yada* (‘knew’)	20000	12%	88%	0%	0%	0%

Table 4: Table of top 100 verbs (part 1; the table continues on the following pages)

Root	Samples	*ki*	*še-*	Cases in Divergent Clusters	Divergence toward *ki*	Divergence toward *še-*
יכח_הפעיל *hoxiax* (‘proved’)	20000	33%	67%	0%	0%	0%
יסף_הפעיל *hosif* (‘added’)	20000	85%	15%	20%	20%	0%
יצא_קל *yatsa* (‘emerged’)	2537	8%	92%	0%	0%	0%
יצע_הפעיל *hiciaʕ* (‘suggested’)	11491	32%	68%	22%	0%	22%
כחש_הפעיל *hixxiš* (‘denied’)	8314	54%	46%	0%	0%	0%
כרז_הפעיל *hixriz* (‘announced’)	16230	58%	42%	7%	7%	0%
כתב_נפעל *nixtav* (‘was written’)	2930	77%	23%	18%	18%	0%
כתב_קל *katav* (‘wrote’)	20000	64%	36%	43%	32%	11%
לון_התפעל *hitlonen* (‘complained’)	6493	51%	49%	0%	0%	0%
למד_פיעל *limed* (‘taught’)	16000	41%	59%	4%	4%	0%
למד_קל *lamad* (‘learned’)	8428	27%	73%	0%	0%	0%
מלץ_הפעיל *himlic* (‘recommended’)	2628	53%	47%	0%	0%	0%
מסר_נפעל *nimsar* (‘was reported’)	7335	84%	16%	4%	0%	4%
מסר_קל *masar* (‘notified, provided statement’)	20000	85%	15%	0%	0%	0%
מצא_קל *maca* (‘found’)	20000	57%	43%	3%	0%	3%
נגד_הפעיל *higid* (‘told’)	20000	3%	97%	0%	0%	0%
נוח_הפעיל *heniax* (‘assumed’)	20000	20%	80%	1%	1%	0%
נסק_הפעיל *hisiq* (‘concluded’)	5877	43%	57%	0%	0%	0%
סבר_הפעיל *hisbir* (‘explained’)	20000	55%	45%	15%	5%	10%
סבר_התפעל *histaber* (‘turned out’)	3167	28%	72%	0%	0%	0%
סבר_קל *savar* (‘opined’)	20000	45%	55%	0%	0%	0%
סכם_הפעיל *hiskim* (‘agreed’)	14670	35%	65%	9%	0%	9%
סכם_פיעל *sikem* (‘agreed upon’)	3684	59%	41%	11%	6%	5%
ספר_פיעל *siper* (‘told’)	20000	51%	49%	0%	0%	0%
עדף_הפעיל *heʕedif* (‘preferred’)	3631	8%	92%	0%	0%	0%
עוד_הפעיל *heʕid* (‘testified’)	20000	57%	43%	1%	1%	0%
עור_הפעיל *heʕir* (‘commented’)	4172	59%	41%	11%	11%	0%
עלי_הפעיל *heʕela* (‘revealed’)	13213	67%	33%	13%	0%	13%
עלי_קל *ʕala* (‘arose’)	20000	85%	15%	2%	0%	2%
עני_קל *ʕana* (‘replied’)	6696	43%	57%	12%	10%	2%
עקש_התפעל *hitʕaqeš* (‘insisted’)	5885	32%	68%	5%	0%	5%
ערך_הפעיל *heʕerix* (‘estimated’)	20000	70%	30%	0%	0%	0%
פסק_קל *pasaq* (‘ruled’)	7979	68%	32%	11%	0%	11%
פרסם_פיעל *pirsem* (‘advertised’)	4417	79%	21%	0%	0%	0%
צהר_הפעיל *hichir* (‘declared’)	20000	64%	36%	7%	6%	1%
ציין_פועל *cuyan* (‘was mentioned’)	3718	86%	14%	0%	0%	0%
ציין_פיעל *ciyen* (‘mentioned’)	20000	78%	22%	6%	5%	1%
צפי_פיעל *cipa* (‘expected’)	13468	16%	84%	1%	0%	1%
צפי_קל *cafa* (‘predicted’)	20000	56%	44%	9%	1%	8%
קבע_קל *qava* (‘decided’)	20000	72%	28%	0%	0%	0%
קוי_פיעל *qiva* (‘hoped’)	20000	16%	84%	0%	0%	0%
קרי_קל *qara* (‘happened’)	4268	2%	98%	0%	0%	0%

Table 5: Table of top 100 verbs, part 2 (continuation from previous page)

Root	Samples	*ki*	*še-*	Cases in Divergent Clusters	Divergence toward *ki*	Divergence toward *še-*
ראי_הפעיל *herʔa* (‘showed’)	20000	51%	49%	15%	7%	8%
ראי_נפעל *nirʔa* (‘seemed’)	5712	20%	80%	0%	0%	0%
ראי_קל *raʔa* (‘saw’)	20000	28%	72%	1%	0%	1%
רגש_הפעיל *hirgiš* (‘felt’)	20000	6%	94%	0%	0%	0%
רמז_קל *ramaz* (‘hinted’)	9200	54%	47%	4%	0%	4%
רצי_קל *raca* (‘wanted’)	20000	1%	99%	0%	0%	0%
רשם_התפעל *hitrašem* (‘got the impression’)	3715	34%	66%	0%	0%	0%
שוב_הפעיל *hešiv* (‘replied’)	11062	67%	33%	12%	12%	0%
שים_קל+לב *sam lev* (‘noticed’)	2492	13%	87%	0%	0%	0%
שכח_קל *šaxax* (‘forgot’)	6168	12%	88%	0%	0%	0%
שכנע_התפעל *hištaxneʕa* (‘became convinced’)	3221	31%	69%	0%	0%	0%
שכנע_פועל *šuxna* (‘was convinced’)	15479	22%	78%	2%	1%	1%
שכנע_פיעל *šixneʕa* (‘convinced’)	9204	27%	73%	0%	0%	0%
שמח_קל *samax* (‘was happy’)	6309	4%	96%	0%	0%	0%
שמע_קל *šama* (‘heard’)	12454	17%	83%	0%	0%	0%
שער_פיעל *šiʕer* (‘assumed’)	7254	30%	70%	0%	0%	0%
תרע_הפעיל *hitriʕa* (‘warned’)	3852	77%	23%	0%	0%	0%

Table 6: Table of top 100 verbs, part 3 (continuation from previous page)

B Appendix: Corpus of Novels by Amos Oz

Book Name	Word Count	Year of Publication
My Michael	57K	1968
Unto Death	28K	1971
Touch the Water, Touch the Wind	36K	1973
The Hill of Evil Counsel	51K	1976
Black Box	66K	1986
To Know a Woman	62K	1989
Panther in the Basement	30K	1995
Suddenly in the Depth of the Forest	18K	2005
Rhyming Life and Death	25K	2007
A Tale of Love and Darkness	184K	2010

Table 7: List of novels by Amos Oz which we analyzed for complementizer usage. All books were originally penned in Hebrew.