**Distribution**

The grammars commonly describe adverbials as "optional" arguments in the clause.[[1]](#footnote-1) This seems to imply that some functions are used more sporadically than others. Does this intuition match the distribution of arguments throughout the HB? An item's distribution across various portions of text can be measured using a formula called the Degree of Dispersion (DP).[[2]](#footnote-2) The relative size of each text portion, in this case each book, constitutes an expected proportion. For instance, the book of Psalms accounts for 7.4% of all phrases in the dataset, while Obadiah only accounts for 0.1%. Thus, if a given function were perfectly distributed throughout the corpus, we could expect 7.4% of its total instances to occur in Psalms and 0.1% in Obadiah. Any observed percentage above or below these expected values are considered a deviation from the expected proportion. DP takes a list of deviations (corresponding to individual sections), sums their absolute value, and divides by 2 to obtain a score between 1 and 0.[[3]](#footnote-3) 1 represents a hypothetical "least evenly distributed," whereas 0 represents a hypothetical "most evenly" distributed. In addition to books, I also test genre[[4]](#footnote-4), chapter, and 10-clause cluster segmentations[[5]](#footnote-5), with a resulting DP score for each function. The values reported in this section are calculated from the coarse-grained dataset (N=99,426).[[6]](#footnote-6)

Figure 2: Degree of Dispersion (DP) for six sampled phrase functions, lower is more evenly distributed

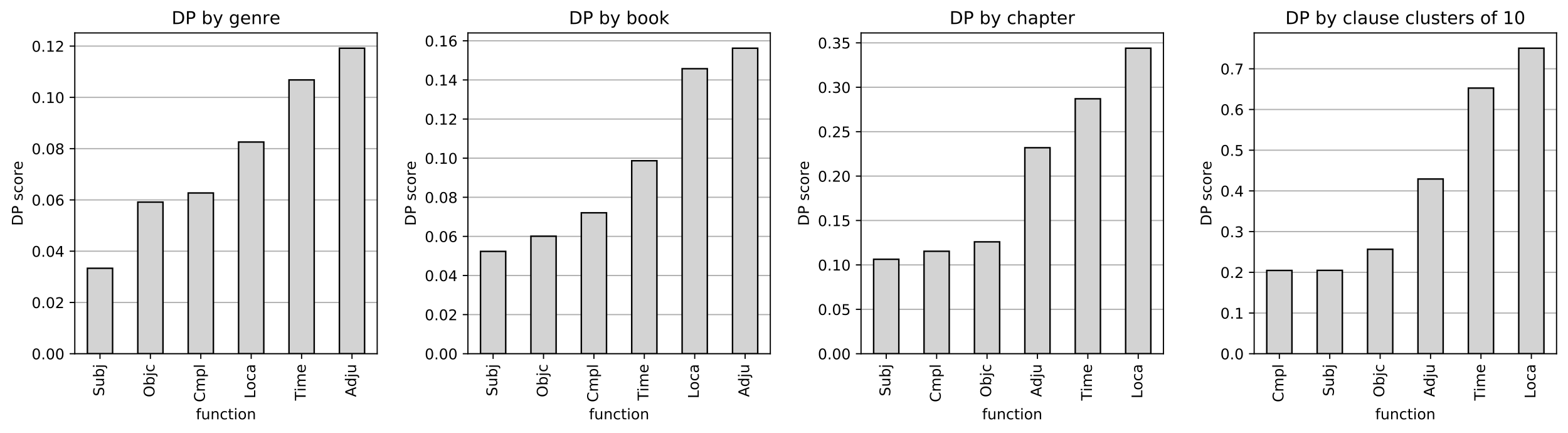


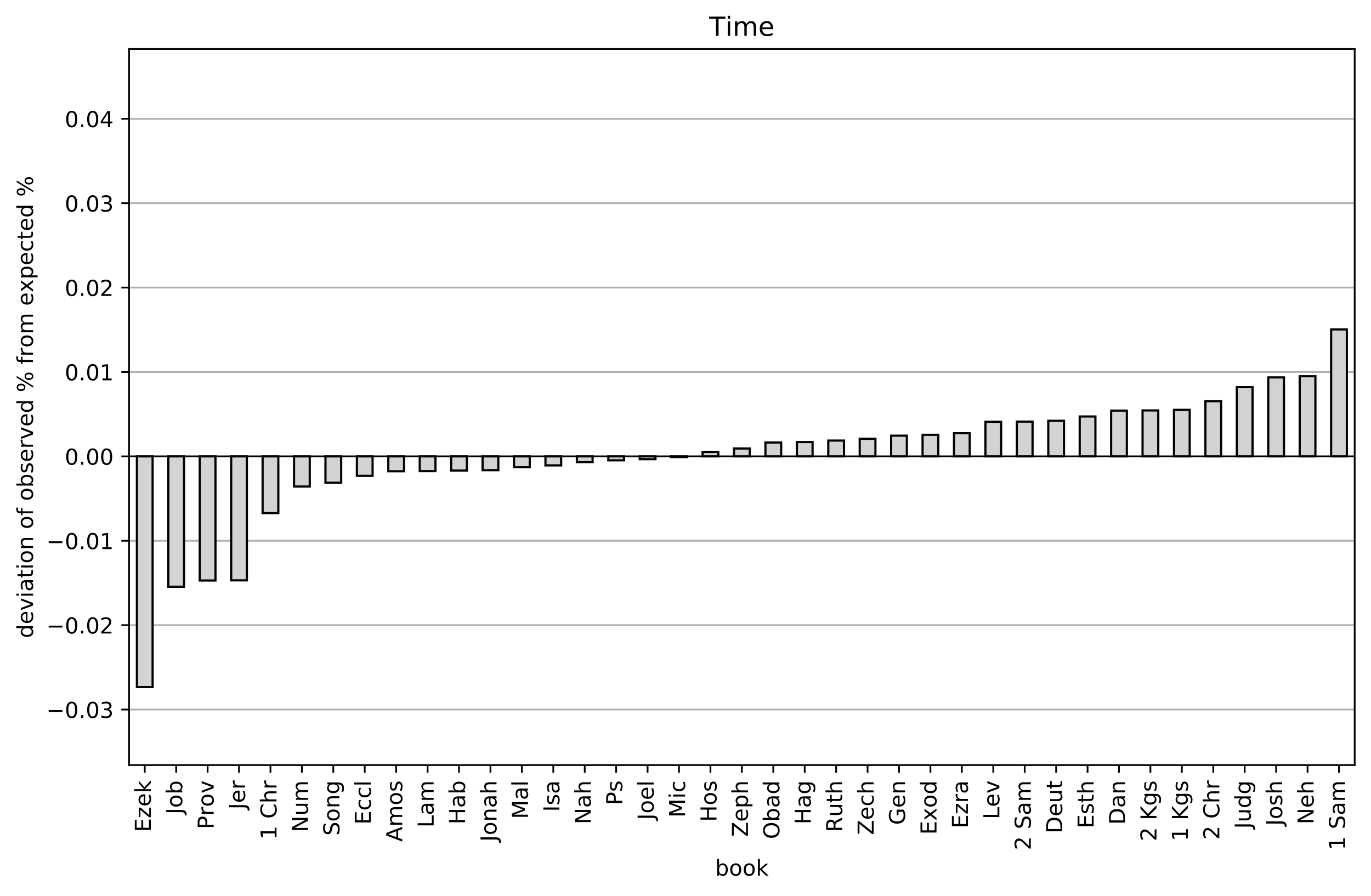
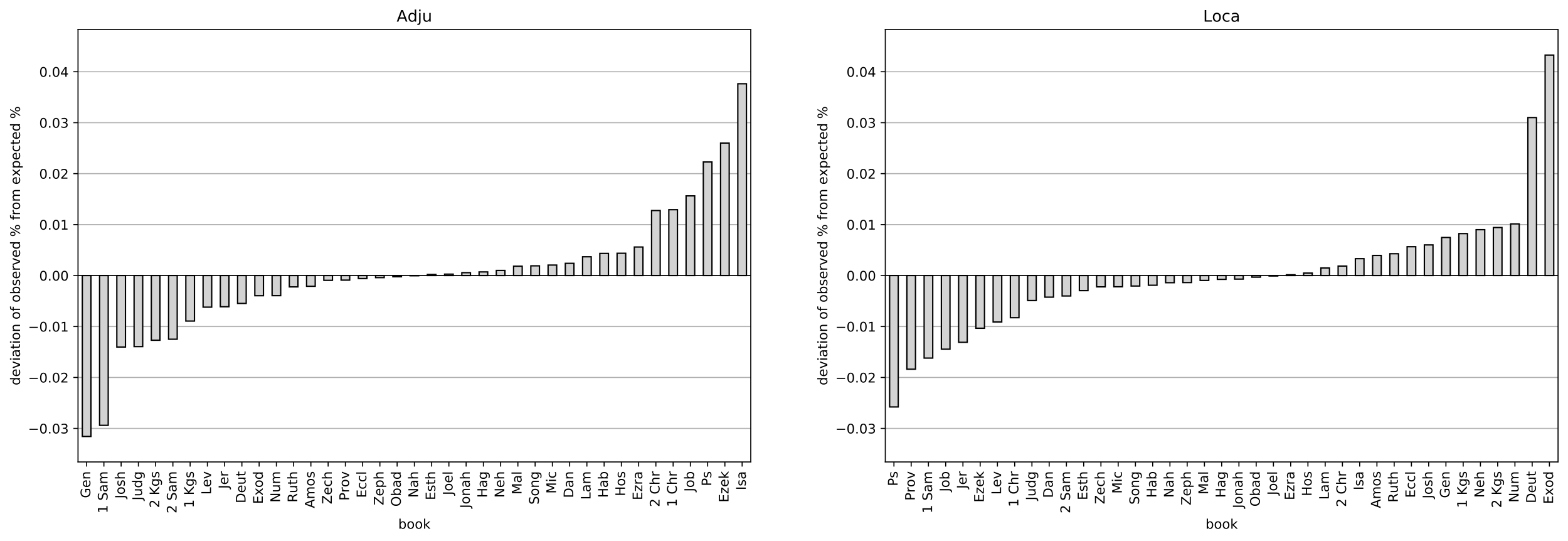
Table 3: DP scores by function, sorted by sum across rows

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | genre | book | chapter | 10-clauses |
| *Subj* | 0.03 | 0.05 | 0.11 | 0.20 |
| *Cmpl* | 0.06 | 0.07 | 0.12 | 0.20 |
| *Objc* | 0.06 | 0.06 | 0.13 | 0.26 |
| *Adju* | 0.12 | 0.16 | 0.23 | 0.43 |
| *Time* | 0.11 | 0.10 | 0.29 | 0.65 |
| *Loca* | 0.08 | 0.15 | 0.34 | 0.75 |

The DP scores mostly match what intuition expects. Subj, Objc, and Cmpl are grouped together as more evenly distributed in all three segmentations. Notably, the smaller segmentations produce a stronger separation between more and less optional arguments. Time is typically somewhere in between other kinds of adjuncts and location phrases.

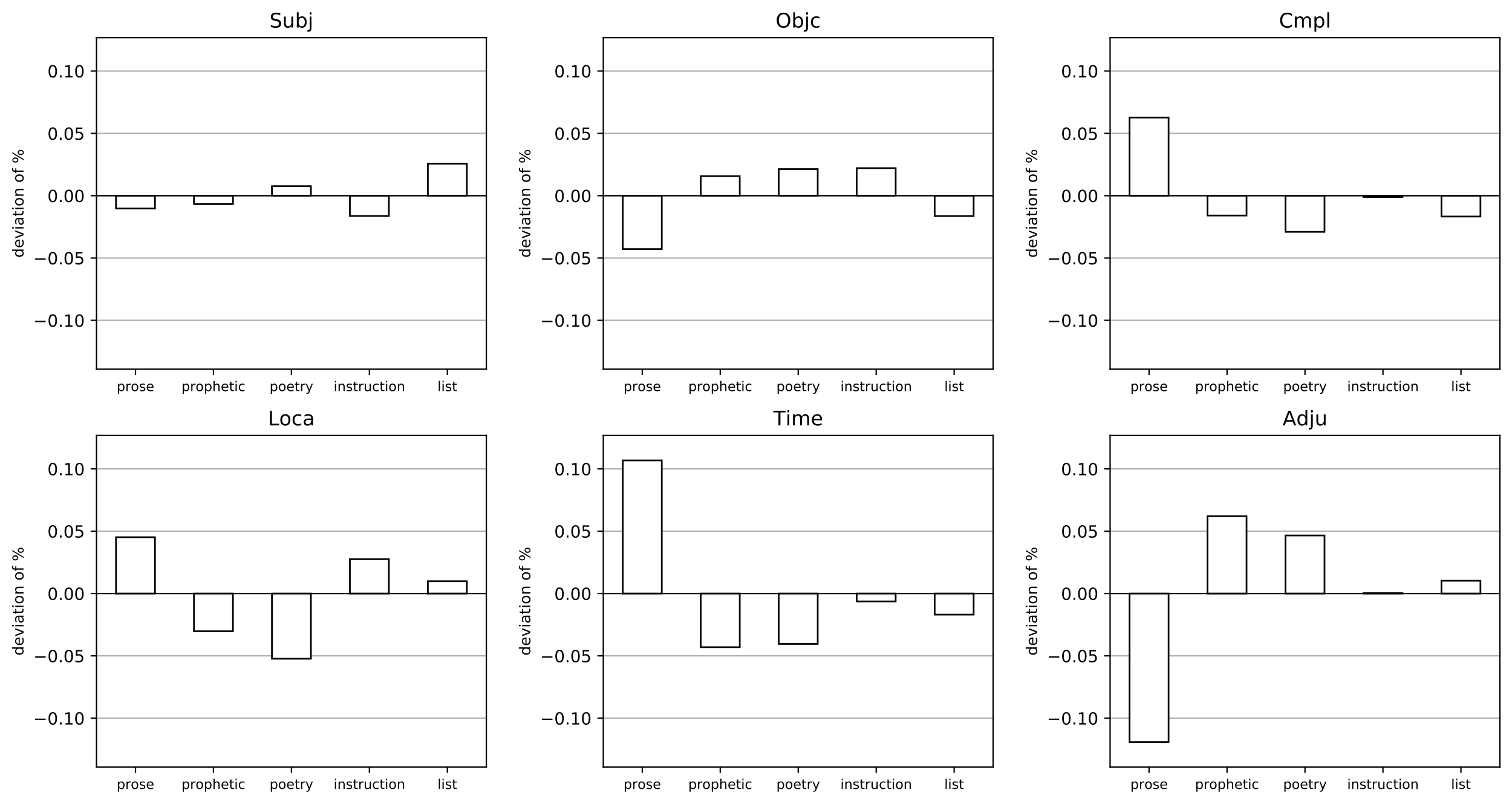
There is a surprise, however, in the book segmentation with Time at 0.10, which is closer to Cmpl (0.07) than Loca (0.15) and Adju (0.16). The data suggest that Adju and Loca are more affected by book content than Time is. The book-by-book breakdown of the difference of expected and observed proportions is provided on the next page for Time, Adju and Loca.

Figure 3: Difference between expected and observed proportions by book for Time, Adju and Loca



The plots show a wider range of variation by book for Adju and Loca than for Time. Loca is notably overrepresented in Exod (+4.3%) and Deut (+3.1%); Adju appears to be overrepresented in more poetic books. Time's strongest deviation is underrepresentation in Ezek (-2.7%). Further work is needed to determine why. Loca and Time share a general trend towards higher representation in narratival books.

Figure 4: Deviation of proportions by function across genre[[7]](#footnote-7)



As seen with books, Time shows strong preference for prose (+10%); Loca similarly reflects a +5% overrepresentation. Adju, as noted above, is underrepresented in prose with plusses in prophecy and poetry. As seen in the next section, this may be due to the Adju label being applied to a large number of adverb-type phrases (e.g. **כֵּן**) more prominent in poetic speech.

**Head Word Semantic Specialization**

Intuition predicts that adjuncts would deploy head words that are semantically specialized to communicate those roles. The term 'semantic specialization' is used here to refer to a head word and argument type that co-occur with such frequency that they are mutually associated. Two quantitative methods in particular can help investigate this hypothesis. The first is a measure of lexical diversity using a normalized ratio of unique head lexemes per function. For instance, Objc occurs with 2,214 different lexical heads per its 16,684 occurrences (fine-grained dataset).[[8]](#footnote-8) This can be normalized per every 100 uses by dividing 2,214 by 16,684 and multiplying the result by 100.[[9]](#footnote-9) The result, 13.27, shows that Objc has about 13 different head lexemes for every 100 sampled cases.

The second method is a measure of statistical significance. The concept of statistical significance is too broad to fully introduce here but is used throughout the remainder of this study.[[10]](#footnote-10) Statistical significance is a measure of dependence between two categorical variables.[[11]](#footnote-11) For instance, suppose one variable is the verb category and another variable is the predicate phrase function. The presence of a verb obviously increases the likelihood of a predicate phrase also being present. A measure of statistical significance can tell whether the co-occurrence frequency of two variables happens more frequently than by coincidence. Statistical significance scores often reveal relationships that cannot be detected with simple percentages.[[12]](#footnote-12)

There are two varieties of significance tests relevant for this study: unidirectional and bidirectional.[[13]](#footnote-13) A unidirectional test such as ΔP tells how strongly a given variable predicts the presence of another variable, regardless of whether the association is reciprocated.[[14]](#footnote-14) For instance, a noun such as 'snow' is likely to predict a word like 'white', whereas 'white' would probably not be as predictive of 'snow'. Bidirectional measures like Fisher's Exact, on the other hand, test association without regard for directionality and can be useful where sample sizes are small.[[15]](#footnote-15) Both ΔP and Fisher's are utilized in this study where appropriate. In the case of semantic specialization, a unidirectional measure makes sense since some head lexemes are rare and the goal is to find heads and arguments that are mutually attractive.

A diversity ratio and a significance test for all head lexemes are calculated for the six phrase functions using the fine-grained sample (N=73,120). The Fisher test returns a score, or 'p-value', where lower than 0.05 is the conventionally accepted threshold for statistical significance, showing a <5% likelihood the observed value is due to random chance.[[16]](#footnote-16) For every unique head lexeme in the sample, a count is made between it and any of the 6 functions it appears in. An excerpt of the co-occurrence table is shown below.

Table 4: Excerpt of co-occurrence table for head lexemes and phrase functions

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Adju | Cmpl | Loca | Objc | Subj | Time |
| 'God'**אֱלֹהִים** | 34 | 203 | 0 | 100 | 434 | 0 |
| 'day'**יוֹם** | 47 | 34 | 0 | 81 | 166 | 1353 |
| 'mountain'**הַר** | 13 | 162 | 72 | 29 | 48 | 0 |

A Fisher's score is calculated for every head lexeme × function. For each function, the proportion of statistically attracted head words (where p < 0.05) is calculated. Those data are shown below alongside the lexical diversity ratios.

Figure 5: Head lexeme diversity and attraction

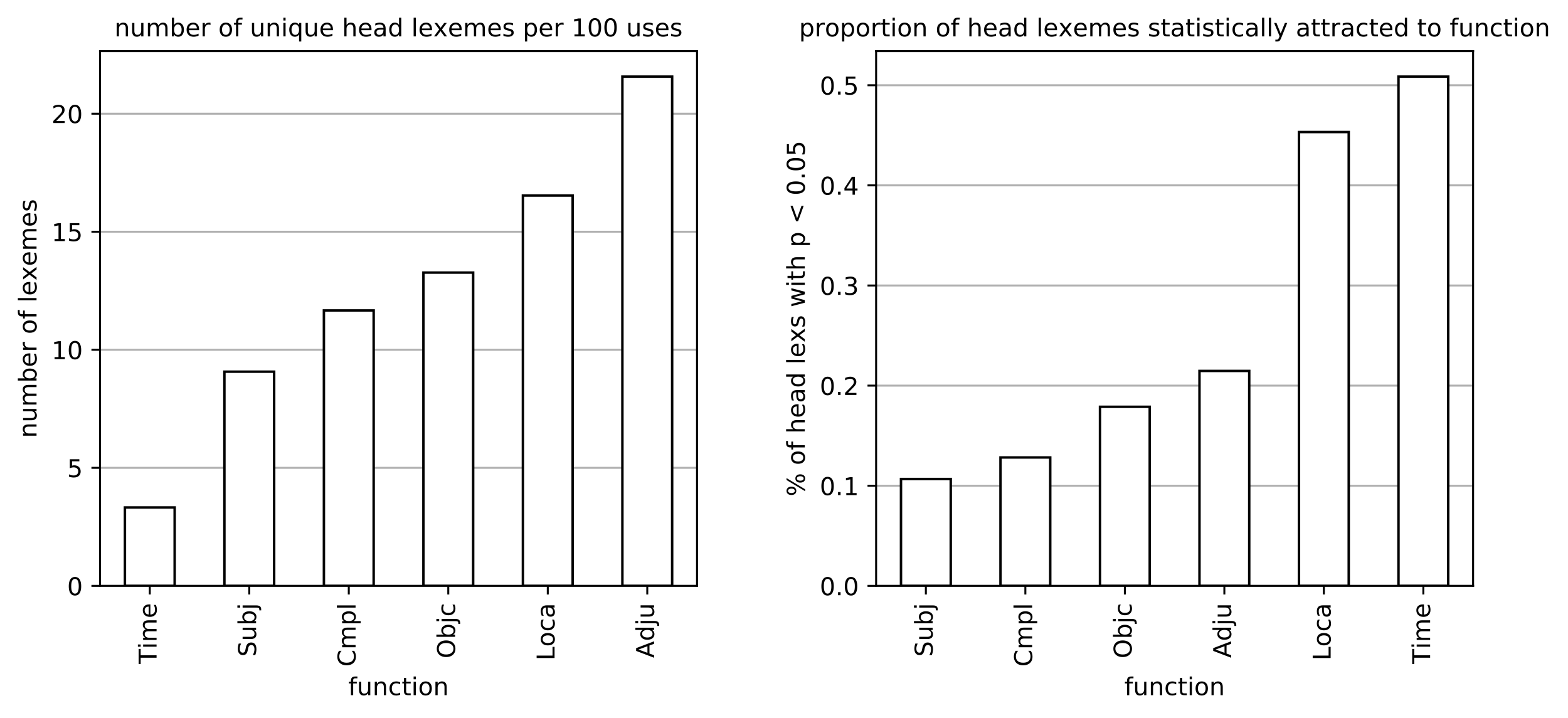


Table 5: Head lexeme diversity and attraction values

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Adju | Cmpl | Loca | Objc | Subj | Time |
| *lex per 100* | 21.57 | 11.67 | 16.53 | 13.27 | 9.07 | 3.32 |
| *% of assoc. lex* | 0.21 | 0.13 | 0.45 | 0.18 | 0.11 | 0.51 |

Time displays extremes in both measures. It possesses the least diversity in lexemes that serve as its head (3.3 for every 100 uses), and 51% of its attested head words are statistically attracted to it. To put it another way, the head words for Time are uniquely specialized both in their homogeneity and their preference for the temporal function.

The top 20 most attracted head words are shown below per function. In this case, since the p-values are too small to show, they are modified by applying *±log10*, as is commonly done in language studies with Fisher scores.[[17]](#footnote-17) The higher the value, the stronger the attraction between the function and head lexeme. A value of *inf* represents a maximum score (p=0), and thus maximum attraction.

Table 6: Top 20 statistically attracted head lexemes per function

*score =* *±log10(Fisher), 1.3 = p(0.05)*

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Adju | | Cmpl | | Loca | | Objc | | Subj | | Time | |
|  | lex | score | lex | score | lex | score | lex | score | lex | score | lex | score |
| ranks |  |  |  |  |  |  |  |  |  |  |  |  |
| 1 | כֵּן | 115.76 | אֶרֶץ | 82.29 | שָׁם | 292.97 | מָה | 51.19 | הוּא | inf | יֹום | inf |
| 2 | רֹב | 31.29 | בַּיִת | 76.25 | אֶרֶץ | 80.81 | דָּבָר | 50.49 | אֲנִי | inf | עַתָּה | inf |
| 3 | פֶּה | 30.41 | יְרוּשָׁלִַם | 71.55 | מִדְבָּר | 63.61 | בֶּגֶד | 50.13 | יְהוָה | 298.71 | עֹוד | inf |
| 4 | חֶרֶב | 26.44 | שָׁם | 54.5 | הַר | 40.84 | בְּרִית | 43.77 | אַתָּה | 283.6 | שָׁנָה | 269.43 |
| 5 | אַיִן | 23.69 | יָד | 48.51 | שָׂדֶה | 22.76 | לֶחֶם | 40.07 | אָנֹכִי | 151.07 | עֹולָם | 213.34 |
| 6 | בֶּטַח | 23.5 | מִצְרַיִם | 39.71 | בַּיִת | 19.78 | פָּנֶה | 34.44 | הִיא | 139.92 | אָז | 161.0 |
| 7 | עַיִן | 22.15 | עִיר | 29.91 | יְרוּשָׁלִַם | 19.6 | חֵן | 27.78 | אַתֶּם | 110.52 | עֵת | 156.73 |
| 8 | מִשְׁפָּט | 15.32 | בָּבֶל | 26.76 | מָקֹום | 19.41 | זָהָב | 22.88 | אִישׁ | 94.35 | לַיְלָה | 144.99 |
| 9 | מִשְׁפָּחָה | 14.17 | אֹהֶל | 26.68 | פֹּה | 18.33 | נֶפֶשׁ | 20.77 | הֵמָּה | 93.6 | בֹּקֶר | 121.29 |
| 10 | שָׁלֹום | 13.96 | מָקֹום | 25.3 | קֶרֶב | 17.99 | כֶּסֶף | 20.31 | הֵם | 83.07 | עֶרֶב | 91.78 |
| 11 | עֲבוּר | 13.4 | הַר | 23.57 | שֶׁמֶשׁ | 17.41 | עֶרְוָה | 19.88 | מִי | 72.43 | חֹדֶשׁ | 72.86 |
| 12 | רִאשֹׁון | 12.91 | קֶרֶב | 17.76 | חוּץ | 16.31 | דָּם | 19.72 | אֲדֹנָי | 67.1 | תָּמִיד | 61.23 |
| 13 | שְׁגָגָה | 12.36 | תָּוֶךְ | 17.5 | שַׁעַר | 15.19 | עֹלָה | 17.76 | אֵלֶּה | 57.81 | פַּעַם | 55.7 |
| 14 | כֹּל | 12.06 | רֵעַ | 17.23 | פֶּתַח | 15.1 | שֵׁם | 16.74 | דָּוִד | 45.43 | מָחָר | 53.14 |
| 15 | מַעַל | 12.05 | יִשְׂרָאֵל | 16.73 | שָׁמַיִם | 14.28 | קְטֹרֶת | 16.06 | אֲנַחְנוּ | 43.42 | יֹומָם | 31.81 |
| 16 | עֹור | 11.95 | מַחֲנֶה | 15.58 | חֶבְרֹון | 13.33 | יָהּ | 14.26 | כֹּהֵן | 40.54 | מָחֳרָת | 31.68 |
| 17 | רָצֹון | 11.57 | אָב | 14.9 | גְּבוּל | 13.23 | בָּשָׂר | 13.7 | בֵּן | 31.83 | מָתַי | 29.31 |
| 18 | מְאֹד | 10.05 | אֵשׁ | 13.77 | חֹרֵב | 11.51 | שְׁבוּת | 13.53 | זֶה | 30.46 | פִּתְאֹם | 29.26 |
| 19 | נְעוּרִים | 9.64 | דֶּרֶךְ | 13.73 | צִיֹּון | 11.19 | טֹוב | 13.38 | אֱלֹהִים | 29.57 | נֵצַח | 26.66 |
| 20 | שֵׁבֶט | 9.47 | פֶּתַח | 13.06 | שֹׁמְרֹון | 9.63 | אֲרֹון | 12.71 | מֶלֶךְ | 27.67 | שֵׁנִי | 23.7 |

Several observations are pertinent. Firstly, we can validate that the statistical significance scores are meaningful since the headwords align with functions one would expect intuitively. Time attracts words such as **יוֹם** 'day' and **עַתָּה** 'now'. Loca has location words like **שָׁם** 'there' and **אֶרֶץ** 'land'. Subj, Objc, and Cmpl are interesting since one may not expect for these functions to have specialized head lexemes. Subj attracts many pronouns, as well as major actants in the HB (**יְהוָה**, **דָּוִד**, **אֱלֹהִים**). Objc attracts the pronoun **מָה** 'what', conceptual objects like **דָּבָר** 'word', **בּרִית** 'covenant', and physical objects **בֶּגֶד** 'clothes', **לֶחֶם** 'bread', etc. Cmpl has much overlap with the Loca function, apparently due to the argument's association with movement verbs. Adju appears to be a mixed bag of semantic classes. This is not unexpected since the Adju role currently serves as a collection of "other" adjunct arguments besides Time and Loca, a current shortcoming of the dataset.

Though Time is not the only argument that becomes associated with particular heads, it is unique in the homogeneity and selectiveness of its head lexemes. Head words such as **יוֹם** 'day', **לַיְלָה**, 'night', **בֹּקֶר** 'morning', based on movements of the sun, are likewise pervasive in time adverbials of other world languages.[[18]](#footnote-18) These objects provide natural reference points for temporal expressions. Aside from noun-based terms, Time also contains several deictic adverbs (based on speech time) such as **עתָּה** 'now', **אָז** 'then', **מָחָר** 'tomorrow'.

The strong associations of these terms with Time confirms that the importance of semantic specialization. Surprisingly, this dynamic also appears to be important for non-optional arguments (e.g. Subj and **הוּא**).

Python algorithms are utilized to further parse phrase structures into the data necessary for the study. In the interest of open science and reproducibility, all of the data and code utilized by this study are available for download at the footnoted link.\*

Five other argument types were selected for the study, based on their comparability with time adverbials and attestation in the dataset: subject, complement, object, location, and other adjunct phrases (henceforth Subj, Cmpl, Objc, Loca, Adju).[[19]](#footnote-19) Each phrase is parsed for all of the relevant features used in the study. For the sake of reliable data, some restrictions are made to the complexity of phrases selected. Only phrases with a single profiled head element and without any phrase-external relations (e.g. **אָשֶׁר** modifiers) are selected.[[20]](#footnote-20) The results of the various selections are illustrated in the charts below; the last chart reflects all of the criteria together and what proportion of each function is kept in the sample. An average of 84% of all eligible phrase functions in the HB are kept, a total of 84,044 phrases.

The statistical analyses depend on human-labeled data that have been reviewed by the author for consistency. Human-labeled data are always subject to bias and error; on the other hand, good data analysis should seek to establish the internal consistency and reliability of the data. Often inconsistencies are revealed in the data exploration process. The dataset used herein remains a work in progress, but overall it is a good representation of the language tendencies. [Insert note from Stefanowitsch about the difference of data in large-scale corpus analytic studies\*] Furthermore, good data analysis is not theory-neutral, but seeks to continually test theory against data. Stefanowitsch\*

The definite article itself can be used in a number of ways. Most generally it indicates that an item is somehow "identifiable."[[21]](#footnote-21) Bekins notes four main sources of identifiability:

1. anaphoric, to refer to an item introduced earlier in the discourse
2. immediate situation, to refer to an item immediately present for the speaker and hearer
3. global situation, to refer to things well-known from world knowledge (e.g. 'the sun')[[22]](#footnote-22)
4. frame, to refer to things associated to a particular semantic frame that has been evoked contextually (e.g. 'the meal' in the frame of 'dinner')

The use of the definite article with these generic nouns suggests the "global use" of the article. In order to test this hypothesis, I annotated the cases of [preposition + definite + head] (333 for Time and 505 for Loca). Four primary categories of the article were tagged: anaphoric (anap), immediate (imme), global (glob), and frame. I added the category 'name' for the numerous cases in Loca where the head-word is a proper noun (e.g. **הַיָרדֵּן** 'the Jordan'\*).[[23]](#footnote-23) The annotations are rough and general, as the primary goal is to test global versus non-global uses of the definite article.[[24]](#footnote-24) The results of the annotations are visualized on the subsequent page as ratios

In order to force parts of speech onto unruly constructions, the classical model must deploy long lists of exceptions and complex rules. Cognitive construction grammar provides an attractive alternative by distinguishing semantic categories from language-specific categories.\* Semantic categories are determined by human concepts, many of which are found to be pervasive across world languages.\* Language-specific categories are contextualized patterns that users have paired with a given semantic meaning. This form-meaning link provides an elegant and powerful explanation for the diversity and commonality in languages. Rather than language categories deriving from an unseen, and as yet unprovable, language machine in the brain,[[25]](#footnote-25) they can be explained as fully diverse cultural artifacts; universal semantic categories, on the other hand, can be explained by the global training data provided to humans by the outside world.\*

In this model, the traditional parts of speech are semantic categories which derive from the way humans tend to organize their knowledge of the world. Croft recognizes two axes of this knowledge space, that of pragmatic function and semantic class:[[26]](#footnote-26)

Table 13: Croft's two-dimensional parts of speech space

|  |  |  |  |
| --- | --- | --- | --- |
|  | reference | modification | predication |
| objects | *core nouns* | genitives,  adjectivals,  PP modifiers | predicate nominals |
| properties | abstract de-adjectival nouns | *core adjectives* | predicate adjectives |
| events | nominalizations, infinitives, gerunds, complements | participles, relative clauses | *core verbs* |

The italicized items in the diagonal correspond with the prototypical parts of speech categories. The two-dimensional axis allows one to see how the prototypical and non-prototypical parts of speech exist on a spectrum of pragmatic and semantic properties. The individual cells in the space function more as regions, with some boxes potentially becoming interconnected as polysemous patterns span multiple meanings.[[27]](#footnote-27) Using this model allows the possibility for constructions to exist along a continuum of semantic poles.

The particle עָתּה 'now' occurs by itself in 354/346 (99.4%) cases. But in 2 intriguing passages it is used alongside the demonstrative.

|  |  |  |
| --- | --- | --- |
| reference | sentence | translation\* |
| *1 Kgs 17:24* | **עַתָּה֙ זֶ֣ה יָדַ֔עְתִּי כִּ֛י אִ֥ישׁ אֱלֹהִ֖ים אָ֑תָּה וּדְבַר־יְהוָ֥ה בְּפִ֖יךָ אֱמֶֽת׃** |  |
| *2 Kgs 5:22* | **הִנֵּ֣ה עַתָּ֡ה זֶ֠ה בָּ֣אוּ אֵלַ֧י שְׁנֵֽי־נְעָרִ֛ים מֵהַ֥ר אֶפְרַ֖יִם מִבְּנֵ֣י הַנְּבִיאִ֑ים** |  |

These cases demonstrate that particle-like words can be construed into nouns by setting them inside larger constructions (e.g. definite noun phrase).

The table below contains the full sample set for **רֶגַע**, a relatively small set (n=7). **רֶגַע** is placed slightly on the particle side of the graph.

Figure 7: Sampled sentences for *רֶגַע*

|  |  |  |
| --- | --- | --- |
| reference | sentence | translation |
| *Num 17:10* | **וַאֲכַלֶּ֥ה אֹתָ֖ם כְּרָ֑גַע** | And I destroyed them in a moment. |
| *Isa 54:8* | **בְּשֶׁ֣צֶף קֶ֗צֶף הִסְתַּ֨רְתִּי פָנַ֥י רֶ֨גַע֙ מִמֵּ֔ךְ** | In a flood of anger I hid my face—in a moment—from you. |
| *Ps 73:19* | **אֵ֤יךְ הָי֣וּ לְשַׁמָּ֣ה כְרָ֑גַע סָ֥פוּ תַ֝֗מּוּ מִן־בַּלָּהֹֽות׃** | How they have become a ruin in a moment. They are ended; they are finished from their terror. |
| *Exod 33:5* | **רֶ֧גַע אֶחָ֛ד אֶֽעֱלֶ֥ה בְקִרְבְּךָ֖** | In a single moment I will go up in your midst. |
| *Isa 26:20* | **חֲבִ֥י כִמְעַט־רֶ֖גַע** | Hide quickly!\* |
| *Isa 27:3* | **לִרְגָעִ֖ים אַשְׁקֶ֑נָּה** | \* |
| *Ezra 9:8* | **כִּמְעַט־רֶגַע֩ הָיְתָ֨ה תְחִנָּ֜ה מֵאֵ֣ת׀ יְהוָ֣ה אֱלֹהֵ֗ינוּ לְהַשְׁאִ֥יר לָ֨נוּ֙ פְּלֵיטָ֔ה וְלָתֶת־לָ֥נוּ יָתֵ֖ד בִּמְקֹ֣ום קָדְשֹׁ֑ו לְהָאִ֤יר עֵינֵ֨ינוּ֙ אֱלֹהֵ֔ינוּ וּלְתִתֵּ֛נוּ מִֽחְיָ֥ה מְעַ֖ט בְּעַבְדֻתֵֽנוּ׃** | In just a short moment\* |

3 of the text examples are used without any kind of nominal modifier, 3 employ quantifiers, and 1 has a plural ending. From an intuitional point of view, the quantifiers and plural ending is enough to classify **רֶגַע** as more noun-like even though the algorithm has placed it slightly on the particle side. This example highlights a shortcoming of the classification method: a lack of nominalizing collocations is not in itself evidence for or against a noun category (nouns are frequently zero-marked). It also highlights the importance of considering sample size in the part of speech model.[[28]](#footnote-28)

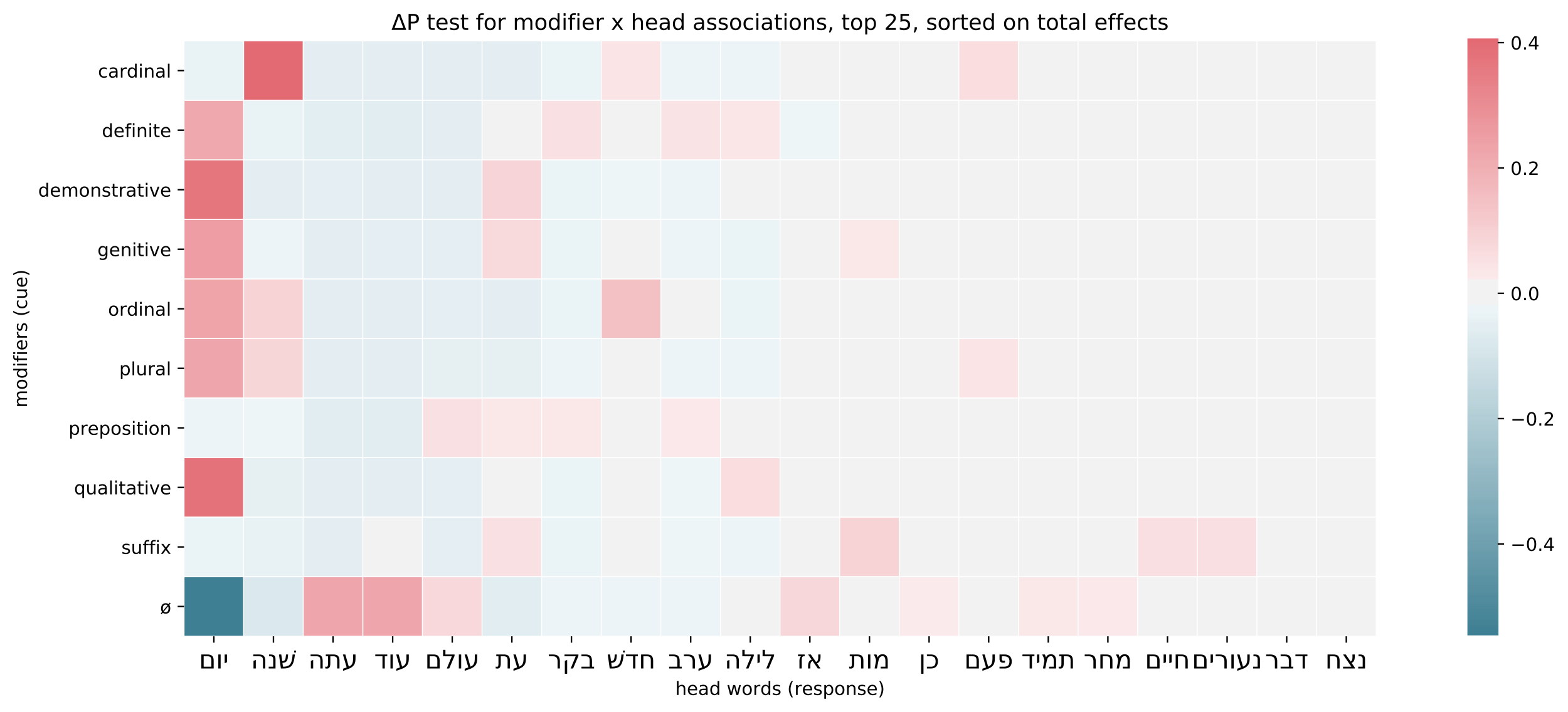
The parts of speech space above is only a subset of semantic categories which are mapped onto constructions. Aside from the three major pragmatic categories of reference, modification, and predication, Croft notes a number of minor categories: constructions used to categorize entities (e.g. lexical roots, morphemes); situate entities in a physical space (e.g. demonstratives); and situate entities in a mental space (e.g. definite articles). These categories and their respective BH constructions are reflected in the diagram below.

\*diagram

***Modifiers with head lexemes***

The appearance of cardinal numbers and ordinals with Time makes sense intuitively since time adverbials are frequently used to orient duration or location using certain time units. A better picture for Time can be seen by observing the ΔP predictability of a modifier for a given head-word. The heatmap below shows which modifiers predict which lexemes.

Figure 8: ΔP test for Time, modifier as cue, head word as response



**יוֹם** 'day' clearly dominates the chart, likely due to its very high representation of Time as a whole. In fact, **יוֹם** accounts for 34% of all Time phrases with a valid head (1351/3999). As others have observed, **יוֹם** is the prototypical time word in BH.[[29]](#footnote-29) But the same is true for the languages of the world, which privilege nouns connected to the natural cycles of the Earth's movement.[[30]](#footnote-30)

The chart helps to see which nouns have similar profiles to **יוֹם** and how. שָׁנָה 'year' shares associations with ordinal and plural indicators, but more strongly prefers cardinal modification. חֹדֶשׁ 'month' likewise shares the preference for ordinal and cardinal modification. **עֵת** 'time', shares similarities with יוֹם (+demonstrative, +genitive), but is different in its lack of attraction to plural, cardinal, and ordinal modification.

What of the areas where **יוֹם** does not dominate? The dark blue square above **יוֹם** shows that null (Ø) modification is negatively associated (-55% less likely).

**comp comp**

A traditional grammatical inventory of these modifiers as found in many Hebrew grammars gives the impression that they are independent, movable pieces. To give an analogy, the items are portrayed like legos, which can be attached or detached freely and in any combination. The data, however, shows that these components function less like legos and more like magnets, which can still be freely detached, but have a natural tendency to attract.

1. E.g. Van der Merwe, Naudé, and Kroeze, *BHRG*, §33.1. [↑](#footnote-ref-1)
2. Gries, “Dispersions and Adjusted Frequencies.” [↑](#footnote-ref-2)
3. \*Add formula and Gries ref. [↑](#footnote-ref-3)
4. The genre values are taken from the ETCBC's syntactic variation project. The genres are intended to follow a basic, intuitional labeling. The categories are shown further below. [↑](#footnote-ref-4)
5. There are 8,813 10-clause clusters. [↑](#footnote-ref-5)
6. 'N' is used throughout to refer to the sample size, as is standard in statistics. [↑](#footnote-ref-6)
7. The genre labels were made manually during the ETCBC's recent syntactic variation project. They are intended to be coarse-grained. Books receive an overall label (e.g. 'prose' for Gen); that label is then overridden in specific stretches (i.e. 'prose' becomes 'list' in Gen 5, and 'poetry' in Gen 49). See the discussion in \*github issues \*syntactic variation ref [↑](#footnote-ref-7)
8. Note that this sum is lower than Subj's total representation in the sample. This is because we exclude cases which do not have a word-based head. This includes PPs where a suffix serves as a subject. [↑](#footnote-ref-8)
9. This normalization is adapted from word-count normalizations as explained by “Normalizing Word Counts,” *The Grammar Lab* (blog), accessed July 2, 2020, https://web.archive.org/web/20200702101532/http://www.thegrammarlab.com/?nor-portfolio=normalizing-word-counts. [↑](#footnote-ref-9)
10. For Hebraists, see the helpful introduction by Jarod Jacobs, “The Balance of Probability: Statistics and The Diachronic Study of Ancient Hebrew,” *JSem* 25, no. 2 (2016): 927–60. For a broader introduction see Natalia Levshina, *How to Do Linguistics with R: Data Exploration and Statistical Analysis* (Amsterdam: John Benjamins, 2015), 199–239. [↑](#footnote-ref-10)
11. Levshina, *Linguistics with R*, 199–200. [↑](#footnote-ref-11)
12. This is because a percentage between a given construction A and B, calculated from A's total uses, does not reflect the percentage that A represents in B's total uses. Raw percentages also do not account for the sample size, and thus the significance of a co-occurrence in relation to all other co-occurrences. [↑](#footnote-ref-12)
13. Levshina, *Linguistics with R*, 224–25. [↑](#footnote-ref-13)
14. Lorraine G. Allan, “A Note on Measurement of Contingency between Two Binary Variables in Judgment Tasks,” *Bulletin of the Psychonomic Society* 15, no. 3 (March 1980): 147–49, https://doi.org/10.3758/BF03334492; Nick C. Ellis, “Language Acquisition as Rational Contingency Learning,” *Applied Linguistics* 27, no. 1 (March 1, 2006): 10–12, https://doi.org/10.1093/applin/ami038. [↑](#footnote-ref-14)
15. Stefanowitsch and Gries, “Collostructions,” 218. [↑](#footnote-ref-15)
16. Levshina, *Linguistics with R*, 9–13. [↑](#footnote-ref-16)
17. A positive sign is applied if the observed frequency is higher than the expected frequency, whereas a negative sign means the observed frequency is lower than expected. \*Levshina. [↑](#footnote-ref-17)
18. Martin Haspelmath, *From Space to Time: Temporal Adverbials in the World’s Languages*, LINCOM Studies in Theoretical Linguistics 2 (Münchn: Lincom Europa, 1997), 25. [↑](#footnote-ref-18)
19. Other phrase functions such as predicates or conjunctions are also interesting to compare with time adverbials, but ultimately made less sense for this study which primarily focus on noun-like phrase components. Other functions such as modifiers (Modi), Intj (interjection), etc. were too rare to be considered for the analysis. [↑](#footnote-ref-19)
20. Heads are determined algorithmically using a combination of lexical sets and rules based on syntactic independence. BHSA encodes parts of complex phrases as 'atoms', which themselves can be smaller phrases. Thus phrases with only 1 atom are selected. External relations such as אָשֶׁר clauses are connected to phrases via a "mother" relationship. Thus, only phrases with 0 mother relations (outgoing or incoming) are selected. [↑](#footnote-ref-20)
21. \*Add additional literature cited by Bekins. Bekins, 226–27. [↑](#footnote-ref-21)
22. Bekins provides the example of the 'sun', which is interesting since this term appears in the Loca head list. This word seems to be restricted to Ecclesiastes’ construction תָּחַת הָשֶׁמֶשׁ 'under the sun'.\* [↑](#footnote-ref-22)
23. These cases did not seem to quite fit any of the four main categories. [↑](#footnote-ref-23)
24. In some cases it was difficult to choose between the 'frame' or 'anaphora category. I thus followed a rather strict definition of anaphora to only mark it when the exact head-word appears earlier in the chapter. Anaphora beyond the chapter limits are not recognized in my tagging. [↑](#footnote-ref-24)
25. Give Chomsky reference.\* [↑](#footnote-ref-25)
26. Copied from Croft, “A Conceptual Framework,” 248. [↑](#footnote-ref-26)
27. For instance, in English nominalization can be added property or action words. Thus the two bottom left regions have some overlap, as reflected in Croft, *Radical Construction Grammar*, 99. [↑](#footnote-ref-27)
28. This example may be an indicator that a measure of statistical significance, rather than a simple ratio, should be considered in the future for normalizing the counts. [↑](#footnote-ref-28)
29. See especially DeVries, *Yesterday, Today and Tomorrow*. [↑](#footnote-ref-29)
30. Haspelmath, *From Space to Time*, 25. [↑](#footnote-ref-30)