# Modeling the *Dalet* Clitic in Historical Hebrew Texts: A New Prefix-Segmented BERT Model and Stylistic Analysis

**Rachel Tal,[2] Cheyn Shmuel Shmidman,[2] Avi Shmidman[1, 2]**
[1]Bar-Ilan University, Ramat Gan, Israel
[2]DICTA, Jerusalem, Israel
rachel.tal@mail.huji.ac.il
cheynshmuel@gmail.com avi.shmidman@biu.ac.il

## Abstract

The Aramaic proclitic *dalet*, widely used in historical Hebrew texts, serves two distinct grammatical functions: as a subordinating conjunction and as a possessive preposition. Because these functions are orthographically identical and no annotated resources exist for this task, large-scale computational analysis of their usage has previously been infeasible. In this paper we introduce a new BERT model for historical Hebrew in which all prefixes are segmented and encoded as independent tokens. This representation allows the model to evaluate proclitics directly and provides a probe-based unsupervised method for determining the grammatical role of the *dalet* clitic using masked language modeling predictions. We evaluate the approach on a manually annotated dataset drawn from historical Hebrew literature spanning multiple regions and historical periods, achieving over an average F1 score of over 0.89. Applying the method to a corpus of more than 300 million words of historical Hebrew texts, we conduct large-scale stylistic analyses of the choice between the Aramaic *dalet* and available Hebrew alternatives. The results reveal geographic and diachronic trends and identify distinct stylistic clusters within the corpus. The prefix-segmented model and annotated dataset are released for unrestricted use.

## 1 Introduction

The Aramaic clitic *dalet* - commonly integrated within historical Hebrew texts - admits to two separate grammatical functions: it may function as a subordinator, or as a possessive preposition.

The double usage of this clitic is attested in original Aramaic writing, such as the Babylonian Talmud, as in (1), as well as within post-Talmudic historical Hebrew works, as in (2).

We begin with an example from the Babylonian Talmud, containing three consecutive instances of the *dalet* clitic (bolded):

(1) wkêwān **d**dînêh **d**nokrî
and.since that.the.law.of.him of.gentile
**d**āzêl bātar ārbā...
that.follows according.to the.guarantor...
'And since the law of the gentile is that he follows the guarantor...'

The first and last instances function as a subordinator, while the one in the middle functions as a preposition.

In post-Talmudic historical Hebrew, this clitic continues to appear with great frequency, in both functions. Consider for example the following passage from Chiddushei Ha-Ritva, the commentary on the Talmud by Rabbi Yom Tov ben Avraham of Seville (Ha-Ritva, around c. 1250-1330 , Spain). This passage contains an intensive jumble of six consecutive occurrences of *dalet* clitic:

(2) whû haddîn dmāṣê limpārēk
and.it.is the.same that.he.can refute
mēemṣāît šebbkat rišônâ **d**bābā
from.the.middle of.group first of.latter
**d**sêpā **d**iqṭannôt **d**tištrê
of.clause of.small.ones that.will.be.permitted
**d**nêmā **d**lā miqqaryā qṭannâ
that.it.is.said that.not it.is.called a.small.one
'And, in fact, he could have raised a difficulty from the "middle one of the first group" that is [included] in the latter clause of the small ones, that it should be permitted. For it is stated that it is not called a small one.'

The first three instances function as prepositions, while the subsequent three function as subordinators.

Crucially, the writers of historical Hebrew could alternatively use Hebrew markers to express these same grammatical functions. Regarding subordination, the writers could avail themselves of the Hebrew subordinator *še-*, rather than the Aramaic *dalet*. And regarding the use of the *dalet* as a



*Proceedings of the 6th International Conference on Natural Language Processing for the Digital Humanities*, pages 121–131

preposition, where it marks possession, the writers could alternatively use the Hebrew prepositional possessive marker *šel*, or they could express the possessive connection via a Hebrew construct form. The key point is that the presence of the clitic *dalet* within historical Hebrew texts is a *stylistic choice*. In each case, writers may choose either the clitic *dalet* or pure Hebrew equivalents.

Almost all historical Hebrew texts mix and match these options, to greater or lesser degrees. For instance, here is a sentence from the book Shevet Halevi of Rav Shmuel Wosner (c. 1913-2015, Israel). This single sentence contains all of the aforementioned grammatical markers (bolded, in order: the Hebrew preposition *šel*, the *dalet* clitic as subordinator, a Hebrew construct form indicating a possessive, the Hebrew subordinator *šin*, another construct form indicating a possessive, and finally two instances of the *dalet* clitic as possessive preposition).

(3) būbdā byôm ṭôb **šel** šābûôt **d**'aḥar
in.the.case in.Yom Tov of Shavuos that.after
**gmar** tpillâ nizkar
conclusion prayer remembered
šā"ṣ **še**'āmar raq ḥăṣî hallēl
shaliach.tzibbur that.said only half Hallel
'im ṣārîk laḥăzōr, w'im ḥôzēr
whether he.should to.return and.if he.returns
'im ṣārîk lbārēk, ûb**dînê**
whether he.should to.bless and.in.laws
hapsāqâ **d**hallēl, wkammâ
interruption of.Hallel and.several
ysôdôt btaqqānâ **d**ehallēl
fundamentals in.institution of.Hallel
'In the case of the Yom Tov of Shavuos after the conclusion of the prayer, the cantor remembered that he had only said half of Prayer of Hallel, whether he should return, and if he returns whether he should bless, and the laws of interruption of Prayer of Hallel, and several fundamentals in the institution of Hallel.'

In order to computationally investigate stylistic trends across historical Hebrew regarding the choice of the Aramaic clitic *dalet* versus the available Hebrew alternatives, we must first disambiguate the function of any given instance of the clitic *dalet*: is it filling the role of a subordinator, or is it filling the role of the possessive preposition? This is a challenging determination for computational methods, because the two functions of the clitic *dalet* are orthographically identical. Until now, no computational method has existed to make this determination, and thus, any large-scale computational investigation of these stylistic trends has been impossible.[1]

In this paper we present a novel solution to this challenge. The key contributions of this paper are:

- We introduce a new BERT model for historical Hebrew in which all prefixes are segmented and encoded as independent tokens, enabling contextual prediction at the level of the proclitic.
- We present an unsupervised method for determining the grammatical role of the *dalet* proclitic - whether subordinator or preposition - by probing the masked language modeling head of the prefix-segmented model.
- We construct and release the first annotated dataset labeling the grammatical function of the *dalet* clitic in historical Hebrew texts.
- We apply this method to a large corpus of historical Hebrew (over 300 million words) in order to analyze stylistic variation between the Aramaic *dalet* and its Hebrew alternatives.
- Using this large-scale analysis, we identify geographic, diachronic, and genre-based patterns in the use of the *dalet* clitic, as well as distinctive stylistic phenomena within the historical Hebrew corpus.

## 2 The Challenge

We seek to train a BERT classifier to determine whether a given instance of the *dalet* clitic functions as a subordinator or as a preposition.

Two BERT models currently exist for Historical Hebrew (Shmidman et al. (2022), Shmidman et al. (2024a)); however, both of them operate on a per-word basis, and do not provide any mechanism for independent encoding or prediction of the grammatical prefixes. Although it may be possible to

[1] To be sure, in certain situations, morpho-syntactic features can facilitate disambiguation. For example, interrogatives which contain "*kol* + *d-*" such as "*kol 'êmat d-*" (i.e. whenever, "*d-*" as a subordinator) or "*kol hêkā' d-*" (i.e. wherever, "*d-*" as a subordinator), see: Bar-Asher Siegal 2016. Additional guidelines regarding the determination of the *dalet* are discussed by Breuer 2007 and Asis 2010. However, rule-based grammatical determinations such as these only cover a small percentage of the actual cases of the *dalet* clitic, and thus they would be no more than a partial solution.

train a supervised classifier based on these models in order to determine the grammatical role of a given prefix, this would require a substantially-sized labeled training corpus of annotated prefixes, which does not exist for Historical Hebrew.

## 3 Our Solution: Prefix-Segmented BERT

In order to achieve unsupervised classification of the grammatical function of the *dalet* clitic, we propose a new pretrained BERT model, in which all proclitics are treated as independent tokens, separate from the words to which they are attached. This provides the BERT model with a new expressive capacity that does not exist in the current BERT models for historical Hebrew; specifically, it allows BERT to make contextual predictions on the level of the proclitic. As we will show in the upcoming sections, this approach provides an effective solution for unsupervised classification of the *dalet* clitic.

## 4 Training Details

### 4.1 Segmentation Classifier

In order to pretrain a BERT model in which prefixes are treated as independent tokens, we must first preprocess our full BERT pretraining corpus, adding a separator token after each prefix. Thus, we start by training a dedicated classifier to perform prefix segmentation on Rabbinic Hebrew text. We follow the same recipe introduced by Shmidman et al. (2024b) for this segmentation classifier. We release this segmentation classifier to the community: [2]

### 4.2 Corpus

For pretraining the BERT model, we use a corpus of 810 million words, collected from various sources and genres of Rabbinic Hebrew. We preprocess the corpus with the prefix segmentation classifier described in the previous section. All cases of prefixes are encoded as independent words, followed by prefix indicator mark.

### 4.3 Pretraining the new BERT model

With the newly preprocessed corpus, we pretrained our new BERT model. We started by training a new tokenizer based on the new prefix-segmented corpus, to ensure that all of the relevant prefix tokens are included in the token vocabulary. We then ran the full BERT pretrain on this preprocessed corpus. The technical details of the training are presented in Appendix B. We release this new BERT model for unrestricted use.[3]

## 5 Classification of the *dalet* clitic

### 5.1 Method

Given our new pretrained prefix-segmented BERT model, we can obtain an unsupervised prediction for any given *dalet* clitic, as follows: For each case of a word which opens with a *dalet* clitic, we query the model's masked language modeling (MLM) head at that position and obtain the top $k$ predicted tokens. This converts the classification problem into a masked language modeling prediction task, following the paradigm of pattern-based classification with language models (Schick and Schütze, 2021). From this ranked list we examine only two candidates corresponding to the Hebrew orthographic realizations of the two *dalet* grammatical roles: for *dalet.sub* (Subordinator), we look for the token ש and for *dalet.prep* (Preposition) we look for the token של. If neither candidate appears among the top-$k$ predictions, the instance is labeled *unknown*. If only one candidate appears, that form is selected. If both candidates appear, we choose one ranked higher.

Formally, let $x$ be an input sentence and $i$ the target position. The masked language model returns an ordered list of the top-$k$ predictions

$$\mathcal{P}_k(x, i) = (p_1, p_2, \ldots, p_k),$$

where $p_r$ is the token ranked at position $r$.

Let $c_{\text{shin}}$ denote the token corresponding to ש and $c_{\text{shel}}$ denote the token corresponding to של. Define their ranks in the prediction list as

$$r_1 = \begin{cases} r & \text{if } p_r = c_{\text{shin}} \text{ for some } 1 \le r \le k, \\ \infty & \text{otherwise,} \end{cases}$$

$$r_2 = \begin{cases} r & \text{if } p_r = c_{\text{shel}} \text{ for some } 1 \le r \le k, \\ \infty & \text{otherwise.} \end{cases}$$

The predicted form is then determined:

$$\hat{y}(x, i) = \begin{cases} \text{unknown} & \text{if } r_1 = \infty \text{ and } r_2 = \infty, \\ c_{\text{shin}} & \text{if } r_1 < r_2, \\ c_{\text{shel}} & \text{if } r_2 < r_1. \end{cases}$$

[2] https://huggingface.co/dicta-il/BEREL-seg

[3] https://huggingface.co/dicta-il/pre-BEREL

### 5.2 Test Corpus

To assess the reliability of our method for determining the grammatical role of the *dalet* clitic, we constructed an evaluation dataset consisting of eight rabbinic works spanning different historical periods and geographical regions. Each instance of the *dalet* clitic in the corpus is manually annotated as either *dal.sub* or *dal.prep*. Table 2 summarizes the works included in our dataset and the number of *dalet* instances of each class in each work.[4] This is the first dataset of its kind, and we release it to the community.[5]

### 5.3 Evaluation

We evaluate our method for a range of values of $k$, in both masked and unmasked scenarios. In the *masked* scenario, the *dalet* clitic is replaced with a `[MASK]` token and the model must predict the missing token. In the *unmasked* scenario, the clitic itself remains visible in the input and the model ranks candidate tokens for that position based on contextual likelihood. In both scenarios we retrieve the top-$k$ predictions from the MLM; these predictions are then evaluated using the ranking-based decision rule defined in Section 5.1.

Table 1 summarizes performance as a function of $k$. When the *dalet* clitic is masked, the relevant candidates often fall outside the top-$k$ predictions, producing a larger number of unknown classifications. Increasing $k$ reduces these cases and therefore substantially improves overall accuracy.

By contrast, in the unmasked scenario, when the clitic itself is visible, the model is given a strong morphological cue, which directly guides it toward these candidates, making them appear near the top of the prediction list and greatly reducing the number of unknown cases. However, in the unmasked scenario, the recall for the minority case (dal.prep) fails to climb above 71.8, indicating that the model is unfairly biased toward the majority case when the clitic is revealed. The masked scenario forces the model to make a decision without this bias, resulting in higher scores overall.

In the data experiments below, we adopt the highest performing approach from this evaluation, that is, we use the Masked scenario, with k=500.

## 6 The Use of the Clitic *dalet* in Historical Hebrew: New Data and a New Analysis

In the previous sections we demonstrated our new computational method to differentiate between the two roles of the proclitic *dalet*: whether as a subordinating conjunction or as a preposition. We now apply this method en masse to a large corpus of historical Hebrew, containing over 300 million words, from the 10th century to the 20th century. Table 4 in Appendix C shows the distribution of the corpus across chronological periods and geographical regions, and Table 5 in Appendix C shows the distribution of the texts across literary genres.

As noted above, the use of the clitic *dalet* in historical Hebrew writing is a stylistic choice. Although it is perfectly normative to interweave Aramaic prefixes within historical Hebrew texts, Hebrew alternatives do exist for both grammatical functions of the *dalet*. Essentially, we have before us two sets of stylistic minimal pairs; every writer of historical Hebrew could optionally decide whether to use the Aramaic *dalet* or the available Hebrew equivalents for every instance of subordination or possessive attachment. Our new computational method allows us to investigate how these stylistic choices come to the fore in historical Hebrew.

### 6.1 Mapping the use of the proclitic *dalet* across time and place

Our first experiment attempts to identify shifts and trends regarding the stylistic usage of the proclitic *dalet* within historical Hebrew literature.

#### 6.1.1 Experimental Setup

We begin by running each text through the prefix segmentation classifier (Section 4.1) in order to identify cases where the initial *dalet* of a word serves as a proclitic rather than as part of the primary lexeme. For each such case, we run the method described above (Section 5.1) in order to determine whether the *dalet* functions as a subordinator or as a preposition. We thus compute the total number of instances of *dalet*-as-subordinator and *dalet*-as-preposition within each composition.

[4] To be sure, these rabbinic works were also included in the pretraining corpus for the new prefix-segmented BERT model. Nevertheless, it should be emphasized that this does not mean that the results reflect memorization rather than genuine generalization. The unannotated digital Hebrew texts which are input to the BERT model contain no indication whatsoever as to whether a given *dalet* clitic functions as subordinator or as a preposition. Thus, memorization of the text provides no a priori advantage when it comes to classifying the function of the clitic.

[5] https://www.dropbox.com/scl/fi/446x13r4b2hn1h16w1ep9/DaletCorpus?rlkey=1rgpx2y5e1xh2gohx58um1v1z&st=i0oloyte&dl=0

| $k$ | Masked | | | | | | | Unmasked | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | dal.sub | | | dal.prep | | | Unk | dal.sub | | | dal.prep | | | Unk |
| | Prec | Recall | F1 | Prec | Recall | F1 | | Prec | Recall | F1 | Prec | Recall | F1 | |
| 10 | 97.9 | 71.0 | .823 | 90.3 | 46.2 | .611 | 31.2% | 97.8 | 70.2 | .818 | 93.8 | 47.5 | .631 | 31.8% |
| 25 | 96.1 | 84.2 | .898 | 89.9 | 63.7 | .745 | 15.3% | 95.6 | 92.3 | .939 | 93.4 | 63.7 | .757 | 8.4% |
| 50 | 96.2 | 90.9 | .935 | 89.9 | 72.2 | .801 | 8.0% | 94.6 | 97.8 | .962 | 91.9 | 70.8 | .800 | 1.2% |
| 100 | 95.7 | 95.5 | .956 | 89.4 | 75.3 | .818 | 3.0% | 94.3 | 98.3 | .963 | 91.4 | 71.3 | .801 | 0.3% |
| 150 | 95.4 | 96.8 | .961 | 89.5 | 76.2 | .823 | 1.4% | 94.3 | 98.5 | .963 | 91.4 | 71.8 | .804 | 0.1% |
| 200 | 95.3 | 97.1 | .962 | 89.5 | 76.7 | .826 | 0.9% | 94.3 | 98.6 | .964 | 91.4 | 71.8 | .804 | 0.0% |
| 300 | 95.2 | 97.4 | .963 | 89.5 | 76.7 | .826 | 0.6% | 94.3 | 98.6 | .964 | 91.4 | 71.8 | .804 | 0.0% |
| 400 | 95.2 | 97.6 | .964 | 89.5 | 76.7 | .826 | 0.4% | 94.3 | 98.6 | .964 | 91.4 | 71.8 | .804 | 0.0% |
| 500 | 95.2 | 98.0 | .966 | 89.5 | 76.7 | .826 | 0.1% | 94.3 | 98.6 | .964 | 91.4 | 71.8 | .804 | 0.0% |

Table 1: Per-class performance as a function of the number of retrieved MLM candidates ($k$), contrasting Masked and Unmasked scenarios. Unk is the percentage of instances where neither candidate appeared in the top-$k$ predictions.

| Author | Birth Year | Birth Place | Book | dal.sub | dal.prep |
|---|---|---|---|---|---|
| Rabbi Yoseph ibn Migash | 1077 | Seville, Spain | Responsa of Ri Migash | 141 | 23 |
| Rabbi Israel Isserlein | 1390 | Wiener Neustadt, Lower Austria | Terumat HaDeshen | 128 | 29 |
| Rabbi Joseph Karo | 1488 | Toledo, Spain | Shulhan Arukh | 141 | 30 |
| Rabbi Shlomo Luria | 1510 | Poznan, Poland | Hokhmat Shlomo | 125 | 27 |
| Rabbi Yechezkel Landau | 1713 | Opatów, Poland | Shut Noda B'Yehuda | 141 | 18 |
| Rabbi Avraham Danzig | 1748 | Danzig, Poland | Hayei Adam | 136 | 14 |
| Rabbi Yosef Hayyim | 1835 | Baghdad, Iraq | Ben Ish Hai – Derashot | 102 | 65 |
| Rabbi Ben-Zion Meir Hai Uziel | 1880 | Jerusalem, Israel | Mishpetei Uziel | 145 | 17 |

Table 2: The historical Hebrew texts included in the annotated dataset. For each book we report the raw counts of the two classes of the *dalet* clitic: *dal.sub* (used as subordinator) and *dal.prep* (used as preposition).

Furthermore, we compute the total number of times that each composition contains the Hebrew alternatives for these roles. Regarding subordination, we tally the number of times that the Hebrew šin character appears as a proclitic; and regarding the possessive connection, we tally the number of times that the text contains the Hebrew preposition šel, and the number of times that the text contains Hebrew construct forms.[6]

Finally, for each text, we compute two statistics, for each of the two grammatical roles, in order to represent the proportion of cases in which the writer chose the Aramaic *dalet* rather than the Hebrew stylistic alternative. Regarding subordination, we take the count of *dalet*-as-subordinator, and we divide it by the total cases of subordination in the text, whether Hebrew or Aramaic. A value of 1 indicates that the author always chose the Aramaic stylistic alternative for subordination; a value of 0 indicates that the author always chose the Hebrew alternative; and a value of 0.5 indicates that the author chooses evenly between the two alternatives. Regarding the possessive connection we do the same, dividing the number of instances of *dalet*-as-preposition by the total possessive connections, whether Hebrew or Aramaic.

We then visualize these author-level statistics in four interpolated geographic maps, in order to identify diachronic and spatial changes in this stylistic preference. We generate separate maps for the two grammatical roles and, within each role, separate maps for medieval authors (11th–15th centuries) and modern authors (16th–20th centuries). Each map plots author-level values according to the longitude and latitude of the author's birthplace, and the background surface is obtained by inverse-distance-weighted (IDW) interpolation over geographic space.[7]

[6] In order to identify words in the construct state, we use the Hebrew morphological analyzer provided at https://morphology.dicta.org.il.

[7] For technical details regarding the creation of the maps see Appendix A.

### 6.1.2 Results

We display the results of this experiment in Figures 1 and 2.[8]. In each map, lighter regions indicate higher interpolated values for the Aramaic variant relative to the Hebrew alternative.

We pay particular attention to the two major literary spheres of European Jewry: Ashkenaz and Spain.[9] In Figures 1 and 2, the Ashkenaz region appears primarily in the Franco-German area of Central Europe, approximately east of $4^\circ$ longitude and north of $47^\circ$ latitude, whereas the Spanish region lies in the Iberian Peninsula, approximately west of $0^\circ$ longitude and south of $44^\circ$ latitude.[10]

Geographically, the clearest observed contrast lies between Ashkenaz and Spain in the medieval period. The division between the two geographical regions here is relatively sharp. In Spain there is a far higher tendency toward the Hebrew stylistic alternative for both subordination and possessive connections, while the Ashkenaz region uses the *dalet* proclitic with far more frequency.

From the diachronic perspective, we cannot meaningfully compare Ashkenaz to Spain, because the modern Spanish corpus contains only a limited number of examples, due to the disruption and eventual dissolution of Jewish communal life in Christian Spain following the expulsions and forced conversions of the late fifteenth century.[11] Instead, we point to a diachronic change within the Ashkenazic corpus itself, that is, medieval writers in Ashkenaz versus modern writers in Ashkenaz. The observed corpus-level patterns suggest that Medieval Ashkenaz is characterized by a consistently strong tendency toward the Aramaic alternatives for both subordination and for the genitive connection, while modern Ashkenaz demonstrates a much more heterogenous distribution: within the same region, we find many writers tending toward the Hebrew alternatives, alongside those who tend toward the Aramaic ones. It is possible that the reason underlying this shift is an increased communal migration in the modern period, which would have led to increased contact between linguistic traditions and the blending of mother tongues and writing conventions.[12]

## 6.2 Author Variability

Our second experiment investigates the question: how strong are individual author tendencies regarding the stylistic usage of the dalet clitic? How consistent are authors in their usage of the dalet?

### 6.2.1 Experimental Setup

Of course, in order to assess this question, we must have a set of multiple works authored by the same writer. Therefore, we restricted this part of the analysis to authors in our corpus with at least three distinct books. Applying this threshold yields a set of 62 authors, each represented by between three and twenty-six texts in the corpus. Table 3 presents a representative subset of authors who satisfied this criterion. For each of the books of each of these authors, we calculate the proportions of the use of the *dalet* clitic for each of its two grammatical functions, as compared with the Hebrew alternatives, using the procedure described above (6.1.1). Further, we calculate the standard deviation for each of the two *dalet* proportions across all of the books of each author, in order to examine the consistency of the author's usage of the *dalet* clitic.

We plot these statistics in Figure 3, using the standard deviation of the dalet-as-subordinator proportion as the x-coordinate, and the standard deviation of the dalet-as-preposition proportion as the y-coordinate.

[8] The color scale is identical in both maps. Although the values among modern writers are slightly higher — reflecting a greater reliance on the Aramaic proclitic *dalet* both for subordination and also for expressing a possessive connection — a unified scale was retained in order to facilitate direct visual comparison between the two distributions. In addition, while inverse-distance interpolation is visually compelling, it may create misleadingly smooth patterns that risk overstating spatial coherence. Accordingly, the maps should be interpreted with caution: they serve primarily to visualize broad tendencies for the reader, whereas the substantive analysis and conclusions derive from the statistical evidence.

[9] In its most basic definition, "Ashkenaz" refers to the Jewish communities of medieval Franco-Germany, centered in the Rhineland and northern France. For a historical definition of Ashkenaz and its cultural-linguistic scope, see Marcus 2010. This regional distinction is not merely geographical but also reflects differences in Jewish legal tradition, communal organization, and literary style.

[10] Regarding other regions—such as modern Hebrew writers in North Africa—the available data in the present corpus is insufficient to support statistically robust conclusions.

[11] On the historical development and decline of Jewish communities in Christian Spain, see Yitzhak Baer, Baer 1958

[12] Such migrations of Jewish communities are attested at the end of the fifteenth century following the expulsion of the Jews from Spain. Many Jews from Spain settled in Italy (Beinart 1988); others migrated to the Balkan regions of the Ottoman Empire (Benbassa and Rodrigue 2001). Additionally, beginning in the late Middle Ages and continuing through the fifteenth century, repeated expulsions from German territories and other regions of Western Europe led to large-scale migration of Jewish communities eastward. These movements ultimately contributed to the emergence of major Jewish centers in Eastern Europe, particularly in

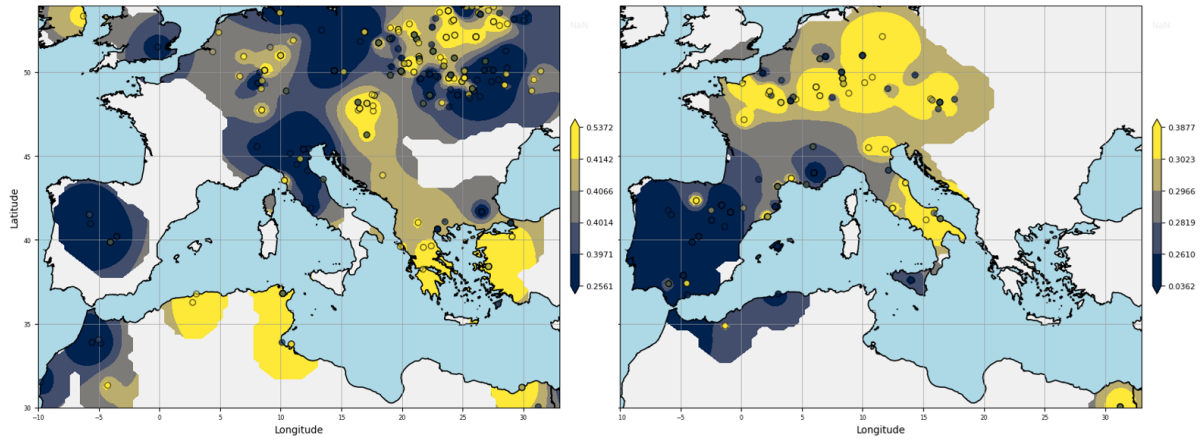


Figure 1: Geographic distribution of the stylistic choice of grammatical subordinator among medieval writers (right) and modern writers (left). Higher values (lighter colors) indicate a tendency towards the Aramaic clitic *dalet*, whereas lower values (darker colors) indicate preference for Hebrew *šin*. Authors are plotted according to their birthplace. The background surface shows inverse-distance-weighted interpolation across geographic space.

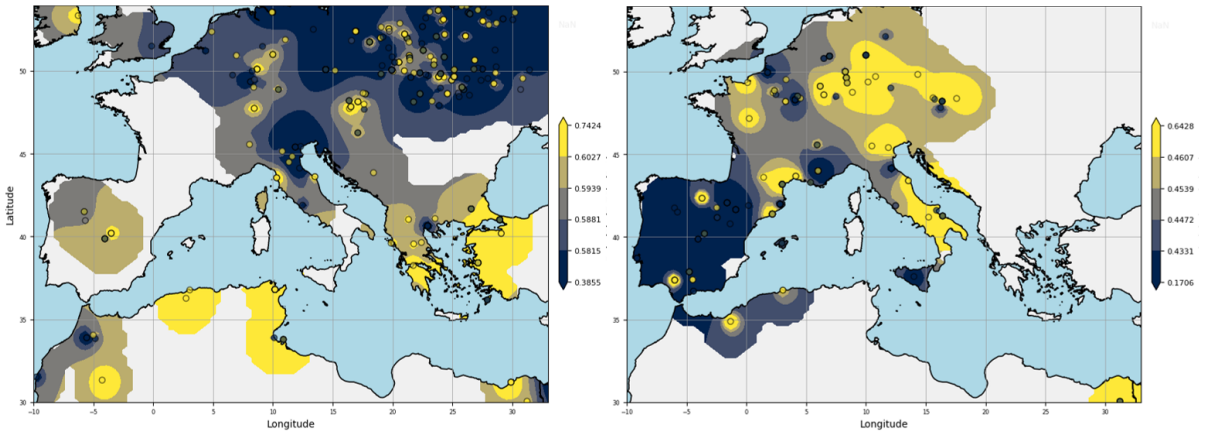


Figure 2: Geographic distribution of the stylistic choice of possessive marker among medieval writers (right) and modern writers (left). Higher values (lighter colors) indicate a tendency towards the Aramaic clitic *dalet*, whereas lower values (darker colors) indicate a preference towards Hebrew alternates. Authors are plotted according to their birthplace, and the background surface shows inverse-distance-weighted interpolation across geographic space.

### 6.2.2 Results

The plot reveals a clear positive correlation between the two measures (Pearson r = 0.79, $R^2$ = 0.62), indicating that authors who display greater variability in the use of one construction tend also to exhibit greater variability in the other. The regression line added to the plot highlights this general tendency and suggests that the variability of the two constructions is not independent but part of a broader stylistic continuum. At the same time, it is apparent that there is much more stylistic variation regarding the use of the *dalet.sub*, as compared with *dalet.prep*; this may be because *dalet.prep* is often linked to fixed collocations from Talmudic literature, and thus its free variation may be more limited.

Several authors cluster in the bottom left area of the graph, indicating particularly low variability for both types of dalet. These authors include Abravanel, Ramhal, Moshe ibn Habib, Rabbi Eliyahu Israel, and Ramhal (Rabbi Moshe Hayyim Luzzatto). Indeed, for each of these authors, their multiple compositions belong largely to the same genre. For instance, the four books by Ramhal in our corpus are all books of Jewish Thought. Such genre homogeneity appears to produce relatively stable stylistic behavior.

By contrast, authors located at the upper right of the graph display markedly higher variability

Poland–Lithuania (see Kaplan 2021).

| Author | Birth Year | Birth Place | Books | Mean | | Median | | Std. Dev. | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | dal.sub | dal.prep | dal.sub | dal.prep | dal.sub | dal.prep |
| Rabbi Shlomo Yitzḥaki (Rashi) | 1040 | Troyes, France | 9 | 0.265 | 0.114 | 0.268 | 0.107 | 0.148 | 0.057 |
| Rabbi Abraham ben David of Posquières (Raavad) | 1120 | Narbonne, France | 6 | 0.337 | 0.184 | 0.304 | 0.172 | 0.134 | 0.039 |
| Rabbi Moses ben Nahman (Naḥmanides / Ramban) | 1194 | Girona, Spain | 12 | 0.267 | 0.128 | 0.272 | 0.140 | 0.238 | 0.099 |
| Rabbi Meir of Rothenburg (Maharam) | 1215 | Worms, Germany | 5 | 0.412 | 0.206 | 0.399 | 0.219 | 0.169 | 0.056 |
| Rabbi Nissim ben Reuven of Girona (Ran) | 1290 | Barcelona, Spain | 6 | 0.413 | 0.199 | 0.518 | 0.230 | 0.223 | 0.079 |
| Rabbi Yoseph Karo | 1488 | Toledo, Spain | 4 | 0.412 | 0.234 | 0.366 | 0.205 | 0.288 | 0.155 |
| Rabbi Judah Loew ben Bezalel (Maharal of Prague) | 1520 | Poznań, Poland | 13 | 0.210 | 0.091 | 0.068 | 0.077 | 0.236 | 0.063 |
| Rabbi Hayyim ibn Attar | 1696 | Salé, Morocco | 6 | 0.413 | 0.170 | 0.412 | 0.168 | 0.314 | 0.130 |
| Rabbi Jonathan Eybeschütz | 1696 | Pinczów, Poland | 6 | 0.599 | 0.188 | 0.698 | 0.189 | 0.266 | 0.061 |
| Rabbi Yosef ben Meir Teomim (Pri Megadim) | 1727 | Szczuczyn, Poland | 6 | 0.604 | 0.215 | 0.597 | 0.221 | 0.094 | 0.042 |
| Rabbi Pinḥas Horowitz | 1730 | Chortkiv, Ukraine | 5 | 0.570 | 0.185 | 0.607 | 0.214 | 0.147 | 0.071 |
| Rabbi Jacob Lorberbaum | 1770 | Zbarazh, Ukraine | 9 | 0.376 | 0.144 | 0.465 | 0.142 | 0.183 | 0.059 |
| Rabbi Naftali Zvi Yehuda Berlin (Netziv) | 1817 | Mir, Belarus | 5 | 0.378 | 0.131 | 0.422 | 0.164 | 0.218 | 0.079 |
| Rabbi Tzadok HaKohen Rabinowitz of Lublin | 1823 | Krizburg, Lithuania | 20 | 0.272 | 0.166 | 0.252 | 0.155 | 0.149 | 0.043 |
| Rabbi Yisrael Meir HaKohen (Hafetz Hayyim) | 1839 | Radin, Poland | 26 | 0.129 | 0.061 | 0.103 | 0.051 | 0.100 | 0.044 |

Table 3: Evaluation of author consistency. We display a sample of the authors for whom our corpus contains three or more books each. We report the mean, median, and standard deviation of the proportions indicating how often the Aramaic *dalet* clitic is chosen rather than its Hebrew counterparts for the two grammatical functions under investigation: for subordination (*dal.sub*) and for possessive marking (*dal.prep*).

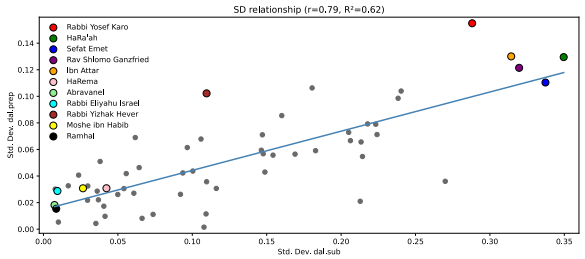


Figure 3: Author-level stylistic variability in the use of the Aramaic *dalet* clitic across multiple works. Each point represents an author with $\geq 3$ books in the corpus; the x-axis gives the standard deviation of the *dalet*-as-subordinator proportion and the y-axis gives the standard deviation of the *dalet*-as-preposition proportion. The fitted regression line indicates a strong positive correlation between the two measures (Pearson $r = 0.79$).

in both constructions. Indeed, the works in the corpus written by these authors are spread across several different genres, which naturally increases stylistic variability. For instance, the works of Ibn Attar (cc. 1696-1743, Morocco), represented on the plot by the orange circle, belong to three different literary genres: Biblical Commentary, Legal Responsa, and Jewish Legal Codes and Customs.

### 6.3 Outliers and Distinctive Stylistic Clusters

The data that we generated regarding all instances of the *dalet* proclitic across the full corpus of hundreds of millions of words also allows us to identify striking outliers and anomalous stylistic clusters. We present two such cases here.

The first stylistic cluster that we examine is found in a group of medieval writers from Provence, including Jonathan ben David ha-Cohen from Lunel (cc. around 1135-1211) and Menachem HaMeiri (cc. 1249–1315). In their writings we find an overwhelming tendency toward the Hebrew options for both subordination and possessive marking, with exceptionally low frequencies of the Aramaic *dalet*. In fact, within their writings, we even find the Aramaic *dalet* missing from Talmudic collocations in which the Aramaic *dalet* is standard and expected. For instance, instead of "gəzērâ dərabbâ" ("a decree of Rabba") - a common expression in Jewish legal literature - HaMeiri writes "gəzērâ dərabbâ" (same meaning, but achieved via a Hebrew construct state). So too, instead of "rôtēaḥ dəṣālî" ("boiling of roasted meat"), he uses "rôtēaḥ šel ṣālî", replacing the Aramaic *dalet* with the Hebrew *šel*. Indeed, these Provençal writers are known to have been influenced by Maimonides (see Ben-Shalom 2008), who was particularly vocal in his insistence on pure Hebrew writing, to the exclusion of Aramaic.

The second anomalous stylistic cluster highlighted by our data involves a group of modern writers in Eastern Europe who exhibit a strong Hebrew tendency when it comes to the possessive marker, yet at the same time demonstrate a preference for the Aramaic *dalet* when it comes to subordination. Within this group are authors such as Rabbi Shimon Shkop, who in his commentary on the Babylonian Talmud employs *dalet* as a subordinator at a rate above the corpus average (52.98%), yet almost entirely avoids *dalet* as a

preposition (8.43%). Similarly, Rabbi El‘azar ha-Kohen of Pułtusk, in his composition Ḥiddushei Maharakh, uses *dalet* as a subordinator more than the corpus average (50.02%), while avoiding its use as a preposition (8.11%). The third author within this group is Samuel ben Uri Shraga Phoebus (52.79% subordinator, 11.08% preposition). This stylistic cluster is quite exceptional, because, as we have seen, when it comes to tendencies towards or away from the Aramaic *dalet*, the two grammatical roles tend to go hand in hand: writers who avoid one also avoid the other, and vice versa. Yet here we have a group of writers who eagerly adopt the *dalet* proclitic for only one of its grammatical roles, while preferring the Hebrew option for the other role.[13]

## 7 Future Work

The large-scale analyses presented above reveal geographic, diachronic, genre-based, and author-specific patterns in the use of the Aramaic *dalet* clitic relative to its Hebrew alternatives. At the same time, the present study should be understood as an initial corpus-wide exploration of these patterns rather than as a final causal model of the factors governing them. In particular, the geographic and diachronic visualizations are based on author-level aggregate statistics, and although they are useful for identifying broad trends, they do not fully disentangle the potentially overlapping effects of period, region, genre, author identity, and corpus size.

In future work, we plan to extend our analysis by modeling the data using mixed-effects regression. This will make it possible to estimate the contribution of period, genre, region, and text length, while separating author-specific stylistic tendencies from broader historical, geographic, and generic effects. We also plan to add robustness checks for the main geographic and diachronic claims, in order to assess how stable these patterns remain under alternative corpus weightings, regional groupings, and chronological divisions.

## 8 Conclusion

We have produced a new BERT model, pre-trained from scratch on prefix-segmented historical Hebrew, capable of performing unsupervised proclitic disambiguation in Hebrew historical texts. We demonstrate how this new model allows us to run big-data experiments across the full corpus of historical Hebrew, gaining new insights into the stylistics of historical Hebrew writing; such experiments were previously all but impossible. We are pleased to release the new BERT model to the community for unrestricted use, both for this specific disambiguation task, as well as for any other downstream task which would benefit from token prediction at the level of the segmented proclitic.

[13] No reverse behavior was identified; we did not find authors who make extensive use of the *dalet* as a preposition while strongly avoiding its usage as a subordinator.

## 9 Acknowledgments

This work has been funded by the Israel Science Foundation (grant No. 2617/22) and by the European Union (ERC, MiDRASH, Project No. 101071829; principal investigators: Avi Shmidman, Bar-Ilan University; Daniel Stökl, EPHE-PSL; Nachum Dershowitz, Tel Aviv University; and Judith Olszowy-Schlanger, EPHE-PSL), 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

Moshe Asis. 2010. *A Treasury of Jerusalemite Expressions*. The Jewish Theological Seminary of America, New York and Jerusalem. Hebrew; published in 2010 (Hebrew year 5770).

Yitzhak Baer. 1958. *Toledot ha-Yehudim bi-Sefarad ha-Notzrit*. Am Oved, Tel Aviv. 2nd rev. and expanded ed.; Hebrew.

Elitzur A. Bar-Asher Siegal. 2016. *Introduction to the Grammar of Jewish Babylonian Aramaic*, 2 edition. Ugarit-Verlag, Münster. References: §§ 10.1, 4.3.

Haim Beinart. 1988. Jewish–converso relations between spain and italy. In Haim Beinart, editor, *Jews in Italy: Studies Dedicated to the Memory of U. Cassuto on the 100th Anniversary of His Birth*, pages 275–288. Magnes Press, Jerusalem.

Ram Ben-Shalom. 2008. *The Jews of Provence and Languedoc*. Magnes Press, Jerusalem.

Esther Benbassa and Aron Rodrigue. 2001. *The Jews of the Balkans: The Judeo-Spanish Community, 15th–20th Centuries*. The Zalman Shazar Center for Jewish History, Jerusalem. Hebrew translation. Originally published in French.

Yochanan Breuer. 2007. The babylonian aramaic in tractate karetot: According to ms oxford. *Aramaic Studies*, 5(1):1–45.

Yosef Kaplan. 2021. Jews on the move: Early modern jewish migration. The Posen Library of Jewish Culture and Civilization.

Ivan G. Marcus. 2010. Ashkenaz. YIVO Encyclopedia of Jews in Eastern Europe, accessed 2026-03-12.

Timo Schick and Hinrich Schütze. 2021. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online. Association for Computational Linguistics.

Avi Shmidman, Joshua Guedalia, Shaltiel Shmidman, Cheyn Shmuel Shmidman, Eli Handel, and Moshe Koppel. 2022. Introducing berel: Bert embeddings for rabbinic-encoded language. *Preprint*, arXiv:2208.01875.

Avi Shmidman, Ometz Shmidman, Hillel Gershuni, and Moshe Koppel. 2024a. MsBERT: A new model for the reconstruction of lacunae in Hebrew manuscripts. In *Proceedings of the 1st Workshop on Machine Learning for Ancient Languages (ML4AL 2024)*, pages 13–18, Hybrid in Bangkok, Thailand and online. Association for Computational Linguistics.

Shaltiel Shmidman, Avi Shmidman, and Moshe Koppel. 2023. Dictabert: A state-of-the-art bert suite for modern hebrew. *Preprint*, arXiv:2308.16687.

Shaltiel Shmidman, Avi Shmidman, Moshe Koppel, and Reut Tsarfaty. 2024b. 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.

## A Appendix: Geographic Interpolation Procedure

The geographic visualizations in Figures 1 and 2 were constructed from author-level observations consisting of birthplace coordinates (latitude and longitude) and a scalar feature value representing the relative frequency of the linguistic form under study. These observations were interpolated over a regular grid spanning longitude $-10^\circ$ to $33^\circ$ and latitude $30^\circ$ to $54^\circ$, sampled at a resolution of $100 \times 100$ cells. For each grid cell, great-circle distances to all observed author locations were computed using the haversine formula. Interpolation was then performed by inverse distance weighting (IDW), with each observation weighted by the inverse of its distance from the grid cell. Cells for which no observation occurred within 250 kilometers were omitted from the interpolated surface; when this condition was satisfied, all observations within 5000 kilometers were included in the calculation. Distances were lower-bounded by a small positive constant in order to prevent division by zero. The interpolated values were rendered as filled contours using quantile-based levels (0, 20, 40, 60, 80, and 100 percent) and the `cividis` color map, while the original author locations were superimposed as scatter points on the same scale.

## B Appendix: Training Details

We trained the model on a DGX-A100 with 4xA100 40GB cards. The training was done with the fused lamb optimizer combined with AMP (Automatic Mixed Precision). A polynomial warmup learning rate scheduler was used to warm up for a portion of the training steps and then decay the learning rate over the total steps. We followed the training recipe introduced by (Shmidman et al., 2023) with regard to the training objectives & the construction of the training examples.

We used the HuggingFace framework wrapped with NVIDIA libraries[14] which are highly optimized for training compute-heavy machine learning models on NVIDIA hardware. We pretrained the model for a total of 36,200 iterations, each iteration consisting of 8,192 examples. The first 23,200 iterations were done with sequences of up to 256 tokens, followed by 13,000 iterations with sequences of up to 512 tokens. Total training time was 7.28 days.

## C Appendix: Corpus Distribution

| Period | Region | % of Corpus |
|---|---|---|
| Medieval Writers | Spain | 7.68 |
| | Ashkenaz | 3.74 |
| | Provence | 2.80 |
| | France | 2.70 |
| | Italy | 1.14 |
| Modern Writers | Ashkenaz | 53.99 |
| | Modern Hebrew | 10.00 |
| | Arabic language sphere | 6.85 |
| | Ottoman Empire | 4.15 |
| | Balkans | 3.22 |
| | Land of Israel | 2.70 |
| | Italy | 0.93 |
| | Yemen | 0.10 |

Table 4: Distribution of the investigated corpus by historical period and geographical region.

[14] https://github.com/NVIDIA/DeepLearningExamples/tree/master/PyTorch/LanguageModeling/BERT

| Group | Literary Genre | % of Corpus |
|---|---|---|
| **Exegetical and Commentarial Literature on Canonical Texts** | Biblical Commentaries | 6.19 |
| | Commentaries on the Mishnah and Halakhic Midrash | 1.25 |
| | Commentaries and Halakhic Works on the Babylonian Talmud | 17.09 |
| | Commentaries and Halakhic Works on the Jerusalem Talmud | 1.14 |
| **Halakhic Literature** | Responsa Literature | 21.47 |
| | Commentaries and Works on the Shulḥan Arukh | 7.80 |
| | Commentaries on the Tur | 1.25 |
| | Commentaries on Maimonides | 13.43 |
| | Halakhic Codes and Customs | 10.33 |
| | Commandment Literature | 0.91 |
| **Intellectual and Homiletic Literature** | Jewish Thought and Ethical Literature | 8.14 |
| | Hasidic Literature | 7.10 |
| **Systematic and Meta-Halakhic Literature** | Legal Principles and Chronological Works | 1.37 |
| | Thematic and Conceptual Halakhic Works | 2.53 |

Table 5: Distribution of the investigated corpus by literary genre (percentage of total words).