Combining grammatical features for dating texts in small-corpus languages

Evaluating a model for diachronic text classification in Biblical Hebrew

Camil Staps

Radboud University Nijmegen

Leiden University

January 26th, 2018

Table of Contents

Dating Biblical Hebrew / the Hebrew Bible

- Compiled from many sources from many contexts and times
- Knowing the context is vital for understanding meaning
- Few texts give historical references
- Can we use linguistic features to date texts?

Dating Biblical Hebrew (2)

- During the Babylonian exile, part of the population was taken captive
- They got Aramaic and Persian influences
- This leads to pre-exilic (early) and exilic/post-exilic (late) BH
- The extent to which this distinction is visible in texts is heavily debated (Young, Rezetko and Ehrensvärd 2008)

Dating Biblical Hebrew (3)

- The goal: combine features to get a better early/late distinction
- Use a transparent model to learn from its decisions

Table of Contents

Traditional evaluation approach

1. Split data in train and test set
2. Develop algorithm on train set (using cross validation)
3. Finally, test on the test set

Not suitable because:

- We don't want to lose data
- Splitting off part of a book in the test set still allows over-fitting on book-specific features

Inner evaluation

- Are the decisions 'reasonable' from a language development point of view?
- Is the model consistent in its predictions?

Table of Contents

Implementation

- Survey of properties (Young, Rezetko and Ehrensvärd 2008)
- Queries for those properties
- ETCBC database + Qumran (Van Peursen, Sikkel and Roorda 2015)
- 100-clause sliding windows
- Decision tree

Table of Contents

Results (outer evaluation)

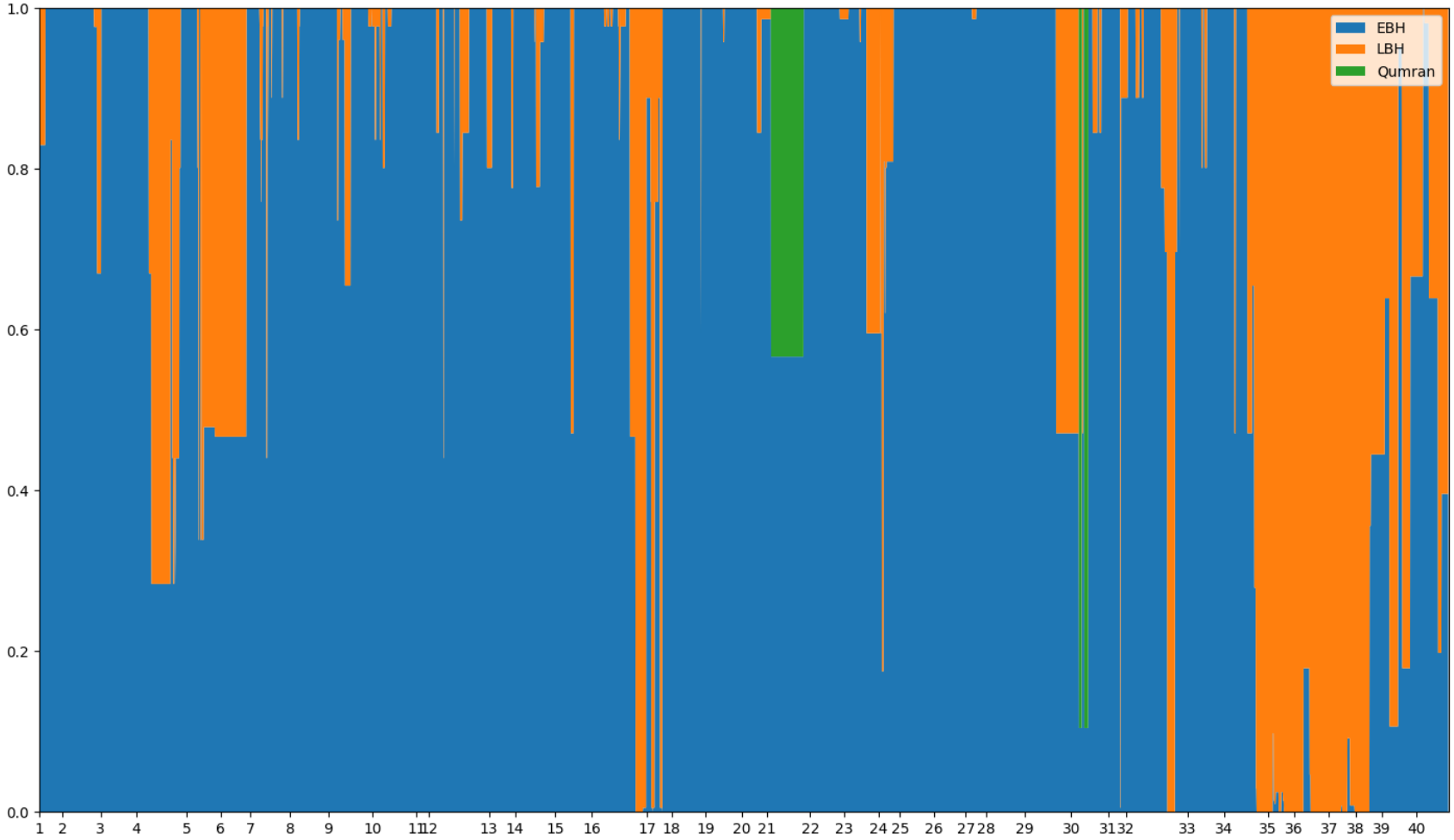
Texts	Clauses	Model	Permutation
Early, late, Q	30	0.601/0.133	0.330/0.005
	70	0.705/0.170	0.335/0.003
	100	**0.729**/0.221	0.329/0.006
Early, late	30	0.653/0.151	0.499/0.006
	70	0.696/0.168	0.501/0.007
	100	**0.693**/0.206	0.498/0.005

$F1$-scores for model and baseline

Results (inner evaluation)

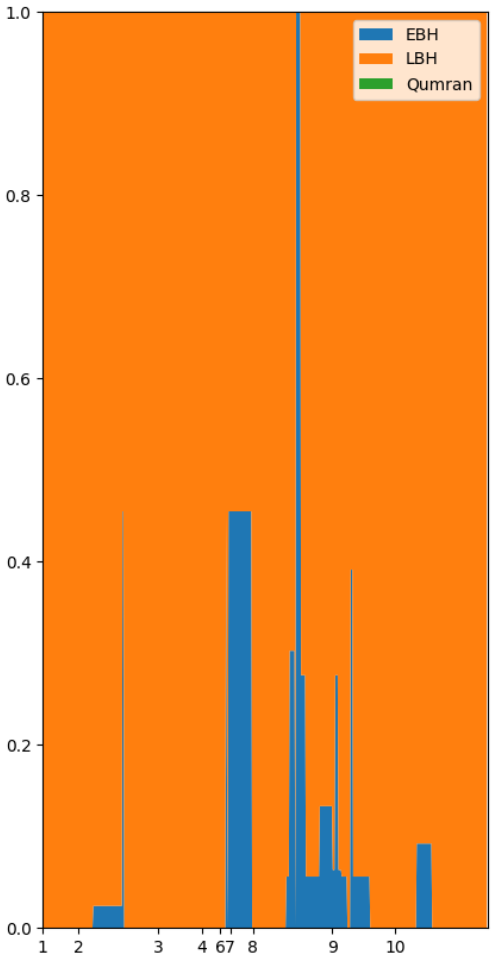
- The model only considers properties proposed in literature
- Hence, the decisions are 'reasonable' by construction!
- What about consistency?

Results (inner evaluation)

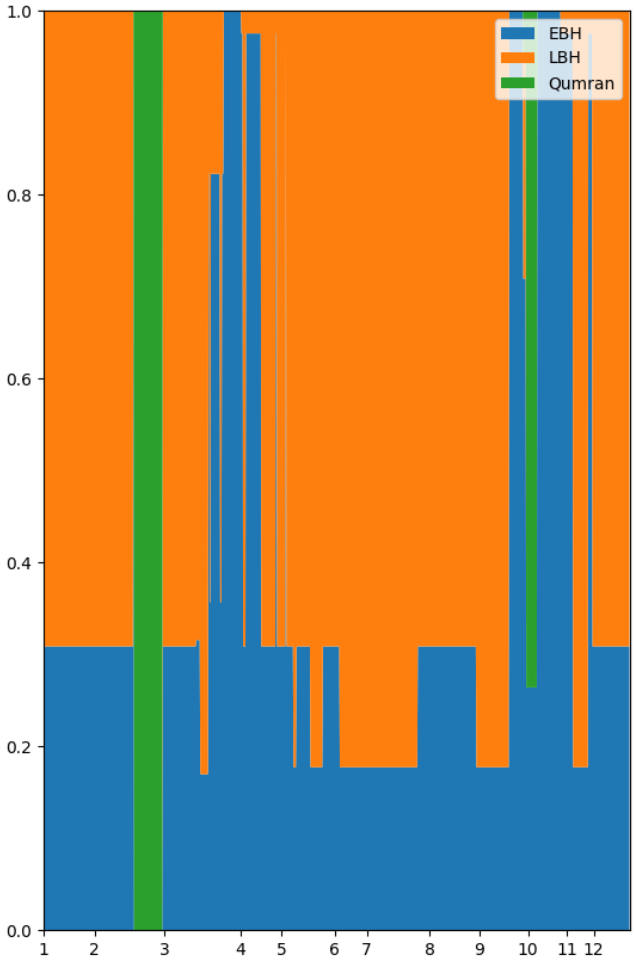


Early, late and Qumran distribution on Exodus

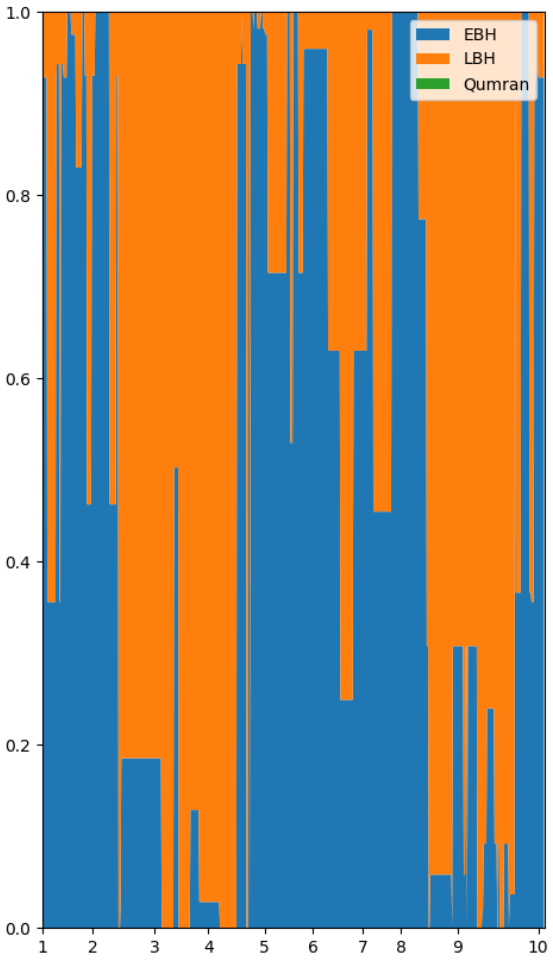
Results (inner evaluation)



Ezra

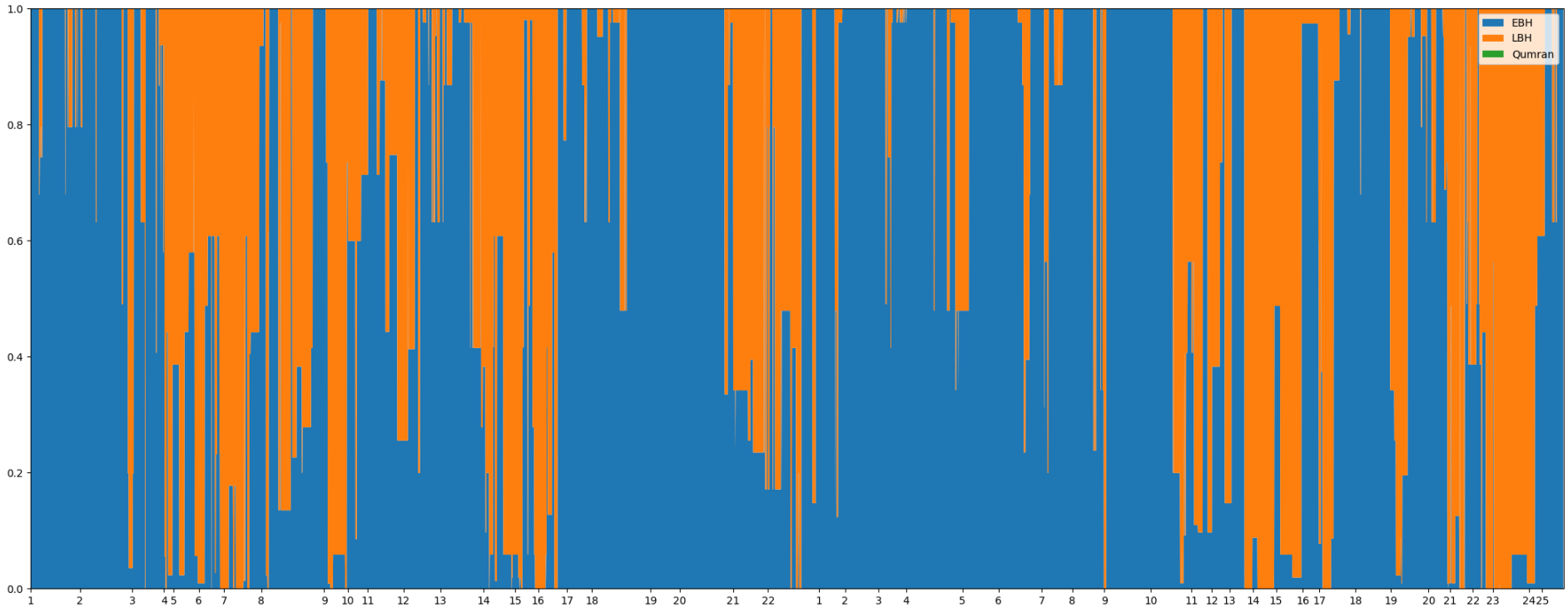


Ecclesiastes

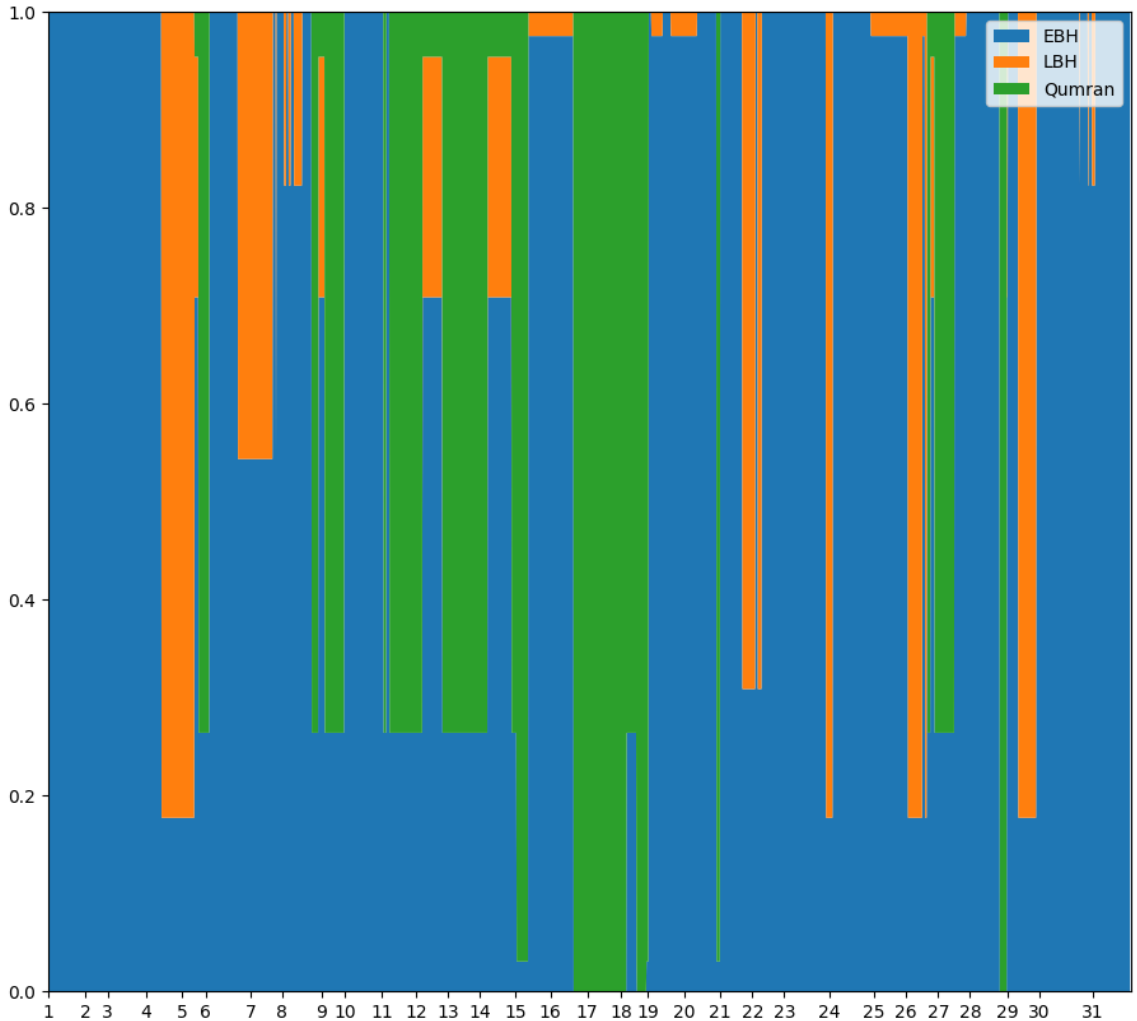


Esther

Results (inner evaluation)



Results (inner evaluation)



Proverbs

Table of Contents

Conclusions

- This model does not perform well enough
- Adding more features may help
- The real issue is the lack of training data

- Outer evaluation results seemed OK...
- ...but inner evaluation shows the model should not be trusted!
- Standard evaluation is not suitable for the HB / small corpora
- Inner evaluation:
  - Sensibility of decision-making process
  - Prediction consistency

References

Jacobs, Jarod. 2016. 'The balance of probability: statistics and the diachronic study of ancient Hebrew'. *Journal for Semitics* 25 (2): 927–960.

Van Peursen, W. T., M. Sc. C. Sikkel and D. Roorda. 2015. 'Hebrew Text Database ETCBC4b'. doi:10.17026/dans-z6y-skyh.

Young, Ian, Robert Rezetko and Martin Ehrensvärd. 2008. *Linguistic Dating of Biblical Texts: An Introduction to Approaches and Problems.* London: Equinox Publishing Ltd.

Appendices

List of features

Property	Features
1	*qaṭal* without *waw*; *yiqṭol* without *waw*; *weyiqṭol*; *wayyiqṭol*
2	*wayehi*; *wayehi* with *b*; *wayehi* with *k*
6	*'az* with perfect tense
8	Direct speech after verbs of speaking
	l and infinitive construct after verbs of speaking
9	*l* with infinitive construct in final clause of sentence
10	*b* with infinitive construct
	k with infinitive construct
11	*'eyn* with infinitive construct
	lō with finite verb
	lebiltî with finite verb
	lō l with infinitive construct
12	Direct object before infinitive construct
	Direct object after infinitive construct
17	Infinitive absolute followed by verb of same root

Ground truth

EBH	Genesis, Exodus, Leviticus, Numbers, Deuteronomy, Joshua, Judges, Samuel, Kings
LBH	Esther,Ezra,Nehemiah,Chronicles
Qumran	1QM, 1QS

- Adding Job core, Daniel and Ecclesiastes helps with standard deviation of the $F1$-score, not with the mean.

Dendrogram (cf. Jacobs 2016)