Towards an Integrated Methodology of Dating Biblical Texts: The Case of the Book of Jeremiah

Martijn Naaijer, Aren Wilson-Wright
University of Zurich, Switzerland
{martijn.naaijer, aren.wilson-wright}@uzh.ch

Abstract

In this paper we describe our research project on dating the language of the Book of Jeremiah using a combination of traditional biblical scholarship and machine learning. Jeremiah is a book with a long history of composing and editing, and the historical background of many of the sections in the book are unclear. Moreover, redaction criticism and historical linguistics are mostly separate fields within the discipline of Biblical Studies. With our approach we want to integrate these areas of research and make new strides in uncovering the compositional history of Book of Jeremiah.

1 Introduction

In this paper we present an overview of the research that we will conduct in the following years. The goal of this research is to develop an integrated approach to dating the biblical book of Jeremiah using a combination of traditional biblical scholarship and machine learning.

Dating texts from the Hebrew Bible is a notoriously difficult task. We know that its books are the product of the first millennium BCE, but their exact date within this time span remains debated.

2 The book of Jeremiah and its background

At almost 30,000 words, the Book of Jeremiah is the longest book in the Hebrew Bible by word count. It consists of 52 chapters and contains texts from a variety of different genres, the historical background of which is not always clear.

The book itself is set in the turbulent final decades of the 7th century and the first half of the 6th century BCE, during which the kingdom of Judah came under the control of various regional superpowers.

For most of the 7th century BCE Judah was a vassal state of the Neo-Assyrian empire. When the Neo-Assyrian empire began to decline in the latter half of the 7th century, however, Judah enjoyed a brief period of relative independence. Judah's autonomy came to an end in 609 BCE when the Egyptian army under Pharaoh Nekau II killed king Josiah, and brought Judah under Egyptian vassalage. In 605 BCE, control of Judah changed hands, when the Babylonians defeated the Egyptian army at Carchemish. King Jehoiakim stopped paying tribute to king Nebuchadnezzar in 601 BCE, after the Babylonian king suffered heavy losses trying to invade Egypt. Nebuchadnezzar subsequently plundered Jerusalem in 597 BCE, and deported part of the Judean population to Babylon. Zedekiah succeeded Jehoiakim as king, but later revolted against Nebuchadnezzar by withholding tribute and allying himself with Pharaoh Apries sometime in 587 BCE. Nebuchadnezzar then returned and destroyed the city of Jerusalem and its temple in 586 BCE. Another part of the Judean population was deported. The year 586 BCE marks the start of the so-called Babylonian exile, which lasted until 539 BCE, when the Persian king Cyrus conquered Babylon, and allowed the Judean exiles to return home (Crouch, 2021).

The Book of Jeremiah paints a portrait of the prophet with the same name, who receives his prophecies from God. He has a scribe, Baruch the son of Neriah (e.g. Jer. 36:4) who records these prophecies. In the book, we read that the prophet is imprisoned (ch. 37), but is later released (ch. 39)

Proceedings of the Second Ancient Language Processing Workshop associated with NAACL 2025, pages 59–64
May 4, 2025. ©2025 Association for Computational Linguistics

and travels to Egypt (ch. 43). Despite this apparent biographical information, it is difficult to say much with certainty about the historical prophet Jeremiah, and to what extent events in the Book of Jeremiah can be related to phases in his life (Leuchter, 2021).

A comparison of the Hebrew version of Jeremiah as preserved in the Masoretic Text (MT) with the later Greek translation, called the Septuagint, reveals several discrepancies between the two text traditions. The Greek text is approximately 8% shorter than its Hebrew counterpart, and locates some passages in a different place in the book.

According to most researchers, the Greek version of Jeremiah reflects an earlier stage in the redaction of the book than MT Jeremiah. An important piece of evidence for this is that the additional material in the MT contains a lot of very specific vocabulary that is absent from the Greek version (Stipp, 2021).

There are strongly varying opinions as to when the book was composed. According to Holladay (1986), the book dates back to the lifetime of the historical prophet (7th–6th century BCE) but others date the book later. According to Fischer (2005), for instance, the book was written in the 4th century BCE.

3 State of the art: Biblical Studies

3.1 General

One of the main goals of this research is to combine redaction criticism and historical linguistics. In Biblical Studies, these modes of inquiry are usually kept separate, and their different presuppositions and methods often lead to contradicting results. Here, we introduce both fields briefly.

3.2 Linguistics

Even though the Hebrew Bible was written and edited throughout the first millennium BCE, its language, Biblical Hebrew is relatively homogeneous. But it does exhibit some variation. In the literature, the most important explanation for this linguistic variation is diachrony—the change of the language over time. Biblical scholarship distinguishes roughly between Classical (or Early) Biblical Hebrew (CBH or EBH) and Late Biblical Hebrew (LBH).

The CBH corpus consists of the Pentateuch and the Former Prophets[1], and the core LBH corpus contains the books Esther, Daniel, Ezra, Nehemiah and Chronicles. Some scholars also include the book of Qoheleth among the LBH books, but not everyone does so (e.g. Young, 1993, 140–156). Other texts and books are more controversial. For instance, Rendsburg (2012) considers the book of Haggai to be written in LBH, and Paul (2012) observes that there is a concentration of late features in Isaiah 40–66.

The Babylonian Exile (587/6–539 BCE) serves as the dividing line between CBH and LBH with CBH reflecting the written variety of Hebrew used prior to 587/6 BCE and LBH reflecting that used after 539 BCE. LBH differs from CBH in terms of phonology, morphology, syntax, lexicon, and style (Fassberg, 2016, 8). Some of these differences are the result of internal developments within the Hebrew language, while others are the result of language contact between Hebrew and Aramaic, the chancery language of the Persian empire. Fassberg (2016, 11–14) mentions various Aramaic features in LBH. On page 14 he gives some Persian loanwords as well.

Several scholars working on the linguistic dating of Hebrew take it as axiomatic that every text written in CBH must have been written at an early date (e.g. Joosten, 2016, 336). This is a controversial point of view. On the one hand, we know that late texts are late because they deal with late (political) events. However, CBH texts dealing with early events could have been composed at a later date.

Based on the distinction between CBH and LBH, scholars have tried to date biblical texts of unknown date with the help of their language. A prominent scholar who developed a method for linguistic dating is Avi Hurvitz. Hurvitz has published many papers and books on this topic; some of his most important works are Hurvitz (1974) and Hurvitz (2014). According to him a late linguistic feature can be identified on the basis of three criteria. The first is distribution. A late feature should occur predominantly or exclusively in late texts. The second criterion is contrast. A late feature should have a semantic equivalent which occurs in early texts. The third criterion is extra-biblical attestation. A linguistic feature is only a late feature if it is used more broadly than in a single text, because then it could be an idiosyncrasy. If these three conditions are

[1] The Pentateuch or Torah consists of the books Genesis, Exodus, Leviticus, Numbers and Deuteronomy, while the books of Joshua, Judges, Samuel and Kings comprise the Former Prophets.

satisfied, one can say that a feature is late. Finally, for an entire text to be considered late, it must contain an accumulation of late features, because one or two late features could be just a coincidence (e.g. Hurvitz, 2014, 9–10).

In 2014, Aaron Hornkohl published a monograph on the language of the Book of Jeremiah from a linguistic dating perspective. Hornkohl found clear signs of late language in the Book of Jeremiah, but not in such a concentration as in the core LBH books. He describes the language as Jeremiah as a mix of CBH and LBH features that is not found in the early *or* late books (Hornkohl 2014, 59). For some features, Jeremiah uses the early variant, but for others, it uses a mix of early and late language (Hornkohl, 2014, 59–62).

Hornkohl points out that the book of Jeremiah has a complex history of composition and editing (Hornkohl, 2014, 65), and he observes that some parts may contain a higher concentration of late features than others. However, none of these parts have a concentration that is as high as core LBH texts (Hornkohl, 2014, 66).

Various scholars have contested the idea that it is possible to date biblical texts linguistically. Cryer argued that there is not enough linguistic variation in Biblical Hebrew to conclude that the Bible developed over a long period of time. The language is simply too homogeneous (Cryer, 1994).

Another critique from the 90s comes from Philip Davies (1995), who argued that the whole Hebrew Bible was a post exilic composition and that CBH and LBH co-existed in the post-exilic period.

The most comprehensive critique of linguistic dating was given by Young, Rezetko and Ehrensvärd (2 volumes, 2008). In 2 volumes, they discuss the principles and methods of linguistic dating. The authors come to the conclusion that it is not possible to date biblical texts using language alone. They acknowledge that one can distinguish between CBH and LBH, but in their opinion these are two styles that co-existed before and after the exile (Young, Rezetko and Ehrensvärd, 2008, volume 2, chapter 2). In later works they opt for an integrated approach (e.g. Rezetko and Young, 2014).

3.3 Redaction criticism

Redaction critical scholarship on Jeremiah owes a lot to the work of Bernhard Duhm (1901) and Sigmund Mowinckel (1914). Duhm distinguished two different categories of prose in the Book of Jeremiah: biographical and nonbiographical. The biographical prose parts appear in chapters 26–45, while the non-biographical parts appear throughout the book. According to Duhm, the non-biographical sections often draw heavily from other biblical texts and were added by later editors (Wilson, 1999, 414).

Mowinckel, by contrast, divided the Book of Jeremiah into three main sources. He assigned the label “A” to the poetic oracles in chapters 1–25, which he saw as the original core of the book. He labelled the biographical sections and the rhetorical prose passage—which he saw as linked with Deuteronomistic literature—“B” and “C” respectively (Wilson, 1999, 414). Mowinckel’s work was very influential, and much subsequent work by other researchers was devoted to investigating how C was related to A and to literature outside of the book, especially Deuteronomy. The date of the different sources also became a source of debate.

4 State of the art: Large Language Models

4.1 LLMs

Recently, Large Langue Models (LLMs), like GPT-4 (OpenAI, 2023) and LLaMA-3 (Llama team, 2024) have set benchmarks in various NLP tasks, including translation, summarization and conversation. Importantly, LLMs do not require hand coded features and thus reduce the risk of replicating traditional biblical scholarship on the segmentation and dating of biblical texts.

These models are able to achieve such a high level of performance by ingesting huge quantities of training data—usually billions or even trillions of words. Biblical Hebrew, however, is a low-resource language comprising approximately 262,934 words. Therefore, it is important to find a solution to the problem of the lack of data.

In recent years, developing LLMs for low-resource languages has become an active field of research. One solution is to reduce the number of parameters in the model. Wdowiak, for example, successfully built a language model for Sicilian using only 266,514 words by reducing BERT’s 12-layer architecture to just a single a single layer (Wdowiak 2021). The size of the Sicilian corpus used in this study is similar to that of the Hebrew Bible. Other studies opt for alternative solutions for low-resource languages (e.g. Alam et al., 2024;

Cahyawijaya et al., 2024; Nag et al., 2024 and Nguyen et al., 2023).

Complex models like LLMs are often called black-box models, because it is difficult to get an impression of how they make predictions. For the present project it is important that the models are not only capable of segmenting and classifying strands within the Book of Jeremiah, but also that they do this in an explainable way. The research results should be meaningful from the perspective of linguists. Explainability of LLMs is an emerging and active field of research, and there are various ways in which one can attempt to create transparency (e.g. Sundararajan, 2017. For a survey: Zhao et al., 2023).

4.2 Initial experiments

We have done some initial experiments to test whether it is possible to distinguish between different linguistic phases of Biblical Hebrew using Machine Learning.

Wilson-Wright finetuned a RoBERTa Base model with an adapted architecture using the verses of the Hebrew Bible as inputs (Wilson-Wright, "BERiT"). The model features a single attention block with four attention heads, smaller embedding and feedforward dimensions (256 and 1024), a smaller max input length (128), and an aggressive dropout rate (.5) at both the attention and feedforward layers. For further details, see the respective HuggingFace model cards for the architecture, parameters and training data for both BERiT and COHeN. Wilson-Wright then trained a linear classifier on top of the language model using labelled data drawn from CBH and LBH text. (Wilson-Wright, "COHeN"). The classifier also included data from two other hypothetical stages of Biblical Hebrew, Archaic Biblical Hebrew (ABH) and Transitional Biblical Hebrew (TBH). ABH is thought to precede CBH, while TBH represents the transitional phase between CBH and LBH. The classifier model achieved 73.4% accuracy on the validation dataset. The application of an explainability framework in the form of integrated gradients revealed that the classifier had independently learned at least one feature that scholars have argued distinguishes CBH from LBH, namely the occasional spelling of the personal name David as דָּוִיד in LBH (vs. דָּוִד everywhere else).

Another experiment was done by Naaijer (2020, 149–176). He trained an LSTM-based sequence classifier that distinguishes between CBH and LBH to find out whether the language of the biblical books of Jonah and Ruth shares more characteristics with CBH or LBH. Instead of training the model with the raw Hebrew text, clauses were represented as sequences of parts of speech or phrase functions. Models were trained for narrative and quoted speech. In general, Naaijer found that the language of Jonah and Ruth shares more characteristics with CBH than with LBH. This is an interesting result, but it is somewhat unsatisfying because LSTM models are a black box.

5 How to move forward

The main data source for this research is the ETCBC dataset of the Hebrew Bible (e.g. Roorda 2018). The first step will be to figure out how best to train an LLM for Biblical Hebrew using the available data. Questions we will consider include: What is the best architecture, what is the best representation of the Hebrew text (vocalized or unvocalized), and how should the text be tokenized? Also relevant is whether it is possible to use transfer learning by training the model on texts in related languages.

After training a masked language model for Biblical Hebrew, we will finetune the model to be a text classifier with the goal of segmenting and classifying parts of the Book of Jeremiah. Here, it is very important that explainability is one of the key ingredients of the research process.

6 Conclusions

There are many interpretations of when and how the Book of Jeremiah was composed and edited. With the newest developments in the field of Natural Language Processing we think it is possible to take groundbreaking new steps in combining redaction criticism and linguistic analysis of the Book of Jeremiah.

Acknowledgments

We thank the Swiss National Science Foundation for funding this research under grant number 10001381.

References

Firoj Alam, Shammur Absar Chowdhury, Sabri Boughorbel and Maram Hasanain. 2024. LLMs for Low Resource Languages in Multilingual, Multimodal and Dialectal Settings. In *Proceedings* *of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*. 27–33. https://aclanthology.org/2024.eacl-tutorials.5.

Samuel Cahyawijaya, Holy Lovenia and Pascale Fung. 2024. LLMs Are Few-Shot In-Context Low-Resource Language Learners. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. 405–433. https://aclanthology.org/2024.naacl-long.24.

Robert Chazan, William W. Hallo, and Lawrence H. Schiffman. 1999. *Ki Baruch Hu, Ancient Near Eastern, Biblical, and Judaic Studies in Honor of Baruch A. Levine*. Eisenbrauns, Winona Lake.

Carly L. Crouch. 2021. The Historical Contexts of the Books of Jeremiah. In Stulman and Silver, 2021, chapter 1.

Frederick H. Cryer. 1994. The Problem of Dating Biblical Hebrew and the Hebrew of Daniel. In Knud Jeppesen, Kirsten Nielsen and Bent Rosendal (eds.). *In the Last Days: On Jewish and Christian Apocalyptic and Its Period*. Aarhus University Press, Aarhus, 185–198.

Philip R. Davies. 1995. *In Search of "Ancient Israel"*, Sheffield Academic, Sheffield, 2nd edition.

Bernhard Duhm. 1901. *Das Buch Jeremia*. Mohr, Tübingen and Leipzig.

Steven E. Fassberg. 2016. What is Late Biblical Hebrew, *Zeitschrift für die Alttestamentliche Wissenschaft*, 128(1). 1–15.

Georg Fischer. 2005. *Jeremia 1-25; Jeremia 26-52*. 2 Volumes. Herders Theologischer Kommentar, Freiburg.

William L. Holladay. 1986. *Jeremiah 1. Commentary on the book of Jeremiah. Chapters 1–25*. Hermeneia, A Critical and Historical Commentary on the Bible, Fortress Press, Philadelphia.

Aaron Hornkohl. 2014. *Ancient Hebrew Periodization and the Language of the Book of Jeremiah, The Case for a Sixth-Century Date of Composition*, Brill, Leiden.

Avi Hurvitz. 1974. The Date of the Prose-Tale of Job Linguistically Reconsidered. *The Harvard Theological Review*, 67(1). 17–34.

Avi Hurvitz. 2014. *A Concise Lexicon of Late Biblical Hebrew, Linguistic Innovations in the Writings of the Second Temple Period*. Brill, Leiden.

Jan Joosten. 2016. Diachronic Linguistics and the Date of the Pentateuch. In Jan C. Gertz, Bernard M. Levinson, Dalit Rom-Shiloni and Konrad Schmid (eds.). *The Formation of the Pentateuch*. Mohr Siebeck, Tübingen.

Llama team. 2024. The Llama 3 Herd of Models. https://ai.meta.com/research/publications/the-llama-3-herd-of-models.

Mark Leuchter. 2021. The Historical Jeremiah. In Stulman and Silver, 2021, chapter 4.

Sigmund Mowinckel. 1914. *Zur Komposition des Buches Jeremia*. Dybwad, Kristiania.

Martijn Naaijer. 2020. *Clause Structure Variation in Biblical Hebrew: A Quantitative Approach.* PhD thesis, Vrije Universiteit Amsterdam. https://research.vu.nl/en/publications/clause-structure-variation-in-biblical-hebrew-a-quantitative-appr.

Arijit Nag, Soumen Chakrabarti, Animesh Mukherjee and Niloy Ganguly. 2024. Efficient Continual Pre-training of LLMs for Low-resource Languages. https://arxiv.org/abs/2412.10244.

Xuan-Phi Nguyen, Sharifah Mahani Aljunied, Shafiq Joty and Lidong Bing. 2023. Democratizing LLMs for Low-Resource Languages by Leveraging their English Dominant Abilities with Linguistically-Diverse Prompts. https://arxiv.org/abs/2306.11372.

OpenAI. 2023. Gpt-4 technical report. https://arxiv.org/abs/2303.08774v6.

Shalom Paul. 2012. Signs of Late Biblical Hebrew in Isaiah 40–66. In Cynthia Miller-Naudé and Ziony Zevit (eds.). *Diachrony in Biblical Hebrew.* Eisenbrauns, Winona Lake.

Gary A. Rendsburg. 2012. Late Biblical Hebrew in the Book of Haggai. In Rebecca Hasselbach and Naama Pat-El (eds.). *Language and Nature. Papers Presented to John Huehnergard on the Occasion of his 60th Birthday*. Studies in Ancient Oriental Civilization. Number 67. The Oriental Institute of the University of Chicago, Chicago.

Robert Rezetko and Ian Young. 2014. Historical Linguistics & Biblical Hebrew, Steps Toward an Integrated Approach. SBL Press, Atlanta.

Dirk Roorda. 2018. Coding the Hebrew Bible. In Research Data Journal for the Humanities and Social Sciences, Volume 3 Issue 1. 27–41. https://doi.org/10.1163/24523666-01000011.

Hermann-Josef Stipp. 2021. Two Ancient Editions of the Book of Jeremiah. In Stulman and Silver, 2021, 93–113.

Mukund Sundararajan, M., Ankur Taly, and Qiqi Yan. 2017. Axiomatic Attribution for Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning*. https://dl.acm.org/doi/10.5555/3305890.3306024.

Louis Stulman and Edward Silver. 2021. *The Oxford Handbook of Jeremiah*, Oxford University Press.

Eryk Wdowiak. 2021. Sicilian Translator: A Recipe for Low-Resource NMT. https://arxiv.org/abs/2110.01938.

Robert R. Wilson. 1999. Poetry and Prose in the Book of Jeremiah. In Chazan, Hallo and Schiffman (1999). 413–428.

Aren M. Wilson-Wright. “BERiT.” https://huggingface.co/gngpostalsrvc/BERiT.

Aren M. Wilson-Wright. “COHeN.” https://huggingface.co/gngpostalsrvc/COHeN.

Ian Young. 1993. *Diversity in Pre-Exilic Hebrew*. Mohr, Tübingen.

Ian Young., Robert Rezetko and Martin Ehrensvärd. 2008. *Linguistic Dating of Biblical Texts*, 2 Volumes, Equinox Publishing, London.

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin and Mengnan Du. 2023. Explainability for Large Language Models: A Survey. https://arxiv.org/abs/2309.01029.