A Layered Annotation Workflow for Semitic Epigraphy

Tal Bernstein[1], **Shai Gordin**[2,3], **Letizia Cerqueglini**[1]

[1] Tel Aviv University, [2] Ariel University

[1] The Department of Hebrew Language and Semitic Linguistics, Tel Aviv University, Tel Aviv 69978, Israel

[2] Department of Land of Israel and Archaeology, Ariel University, Ariel 40700, Israel

[3] Digital Humanities and Social Sciences Hub, Open University of Israel, Ra'anana, Israel

talbernstein@mail.tau.ac.il, shaigo@ariel.ac.il, cerqueglini@tauex.tau.ac.il

Abstract

This paper presents a layered annotation workflow for the historical linguistic study of Semitic epigraphic texts. Using a curated Phoenician corpus primarily based on *Kanaanäische und Aramäische Inschriften* (KAI), the system models inscriptions as multi-layered objects that encode graphemic, morphosyntactic, phonological, semantic, and contextual information as independently queryable layers. Annotation is embedded in a structured editorial workflow supporting peer review, expert validation, version tracking, and the representation of variant readings and uncertainty. A case study demonstrates how recurring formulaic constructions can be modeled as morphosyntactic configurations retrievable across inscriptions. Although the current corpus is limited in scope, the data model is language-agnostic, designed for extension to other Semitic epigraphic traditions.

Keywords: Layered annotation, Semitic epigraphy, historical linguistics, collaborative annotation, linguistic corpora, Phoenician inscriptions

1. Introduction

1.1 Background and Motivation

Over the past two decades, research on ancient languages has increasingly relied on digital corpora and language technologies. In Semitic epigraphy, this has led to the digitization of inscriptions and the development of online repositories, such as DASI (Digital Archive for the Study of Pre-Islamic Arabian Inscriptions) and OCIANA (Online Corpus of the Inscriptions of Ancient North Arabia), which provide large-scale access to epigraphic data with rich metadata (Avanzini, Prioletta & Rossi 2014; Al-Jallad 2015).

Epigraphic research nevertheless faces challenges such as fragmentary preservation, orthographic variation, and diachronic and diatopic diversity. Whereas many digital resources offer high-quality editions and metadata, fewer support fine-grained linguistic annotation within collaborative editorial workflows. These limitations are particularly evident in the Phoenician and Punic corpus, whose wide geographical and chronological spread calls for tools capable of representing both structural regularities and local variation in a systematic and reusable way.

1.2 Research Problem

How should the linguistic complexity of Semitic epigraphic texts be represented in a digital environment, to support historical linguistics research while remaining philologically responsible? Classical scholarship has long identified recurring formulaic structures and morphosyntactic patterns in Phoenician and Punic inscriptions, as well as in related Northwest Semitic languages such as Moabite, Ammonite, and Ekronite, often interpreted as a continuum of closely related Canaanite varieties (Koller 2013; Sanders 2008). However, such analyses are typically expressed in descriptive form and remain difficult to reproduce or compare systematically across corpora.

From a computational perspective, the challenge goes beyond digitizing inscriptions to designing an annotation model and workflow that encodes multiple layers of linguistic interpretation without flattening variation or obscuring editorial uncertainty. This includes supporting and preserving alternative readings, and enabling structured comparison across inscriptions distributed through time and space. It also calls for distinguishing between a *diplomatic representation*, which reflects the inscription as written, with line breaks, orthographic variation, damage, and restorations, and the *normalized representation*, which reflects linguistic interpretation, a standardized orthography, morphological segmentation, and lexical identification. Addressing such a problem requires an infrastructure in which linguistic annotation is tightly integrated with collaborative editorial processes.

1.3 Contribution of This Work

This paper presents a layered linguistic

annotation (LLA) workflow and digital infrastructure for Semitic epigraphy. It introduces a multi-layer annotation model grounded in Semitic linguistic principles, embeds it in a structured editorial workflow with peer review and expert approval, and demonstrates its analytical potential through a case study of Phoenician formulaic morphosyntax. Rather than aiming at full automation, such a workflow supports expert-driven scholarship through structured encoding and comparison of linguistic interpretations, and is extensible to other Semitic epigraphic traditions with comparable characteristics.

1.4 Structure of the Paper

The paper is structured as follows: Section 2 reviews related work. Section 3 describes the corpus. Section 4 presents the annotation model and workflow. Section 5 outlines the implementation. Section 6 presents a linguistic case study. Section 7 discusses limitations and future directions, followed by a brief conclusion.

2. Related Work

2.1 Annotation Frameworks for Historical and Ancient Languages

Digital corpora of historical languages increasingly rely on structured annotation frameworks to support linguistic analysis and computational reuse. TEI-based models and domain-specific standards such as EpiDoc alongside graph-oriented frameworks like LAF/GrAF, enable multi-level encoding of textual, linguistic, and contextual information (Bodard & Stoyanov 2016; Ide & Suderman 2014). Yet annotation practices remain largely project-specific, shaped by individual research goals and source materials rather than shared conventions (De Santis & Rossi 2018).

In parallel to XML and graph-based annotation frameworks, recent work has explored RDF and Linked Data approaches for representing LLA and interlinked textual resources. The Open Annotation Data Model provides a general framework for representing annotations as web resources, while models such as POWLA demonstrate how multi-layer linguistic corpora can be represented in RDF and Linked Data approaches for linguistic annotation (Sanderson et al. 2017; Chiarcos 2012; Chiarcos et al. 2012). These approaches emphasize interoperability, dataset linking, and flexible representation of annotation layers.

This tension is particularly evident in epigraphic corpora. In inscriptions such as the Phoenician Ahiram sarcophagus or Eshmunazar texts, fragmentary preservations, restored letters and competing interpretations[1] require concurrent representation of diplomatic readings, normalized forms and morphosyntactic analysis. Although existing standards allow such layered encoding, they do not inherently provide an integrated editorial workflow in which variant readings, certainty levels and linguistic interpretation remain dynamically interlinked and collaboratively validated. The challenge, therefore, is not only structural representation, but the coordination of LLA with transparent scholarly decision-making.

2.2 Digital Epigraphic Corpora and Tools

Digital epigraphic corpora have substantially improved the preservation, accessibility, and dissemination of inscribed materials by combining textual editions with structured metadata and visual documentation. Such platforms enable browsing, cataloguing, and basic textual search, and play an important role in safeguarding endangered cultural heritage (Vitale et al. 2021). However, their development is often constrained by funding cycles, software obsolescence, and long-term maintenance challenges, motivating infrastructure initiatives focused on persistent identifiers and archival solutions (Derntl et al. 2023). These efforts address sustainability and access, but many epigraphic corpora remain oriented toward documentation rather than toward fine-grained linguistic annotation or iterative scholarly interpretation.

2.3 Semitic and Northwest Semitic Digital Resources

Numerous Semitic corpora, both epigraphic and manuscript-based, have been digitized through libraries- and academic institutions-led projects. Codex and literary traditions in Arabic, Aramaic, Ge'ez and Hebrew are represented in major initiatives such as e.g., Arabic Collections Online (ACO; NYU Libraries), Corpus Coranicum

[1] The final two words of the Ahiram sarcophagus inscription (KAI 1) are read by Donner & Röllig (2002) as לפן גבל (LPN GBL, "before Byblos"), whereas Avishur (1979), unresolveably reads לפפ שבל (LPP ŠBL), noting that "this phrase is difficult and there are no comments worth mentioning". Others have suggested an alternative לעד שרל (LʿD ŠRL), reflecting ongoing uncertainty at both graphemic, lexical and semantic levels.

(Berlin-Brandenburg Academy of Sciences and Humanities), the Ktiv project of the National Library of Israel, and syri.ac: An Annotated Bibliography of Syriac Resources Online. Cuneiform texts have also been made available through large digital platforms such as the Cuneiform Digital Library Initiative (CDLI), the Open Richly Annotated Cuneiform Corpus (ORACC), the Ebla Digital Archives (EbDA), and the Amarna Letters project in Akkadian in the Eastern Mediterranean World (AEMW/Amarna, Lauinger & Yoder 2014). These resources are typically organized around specific languages, regions, scripts, or textual traditions, reflecting domain-specific priorities. Even though coverage varies, most corpora emphasize cataloguing and access rather than fine-grained linguistic annotation.

2.4 Positioning the Present Workflow

The proposed workflow builds on existing annotation and epigraphic frameworks while addressing their limitations for linguistic analysis. It adopts a modular, multi-layered model tailored to Semitic inscriptions, enabling the representation of graphemic, morphosyntactic, semantic, and contextual information as interrelated yet independently queryable layers. Unlike many corpora that prioritize static editions, this workflow integrates linguistic annotation with a structured editorial process, including peer review, expert validation, and transparent versioning. Annotations retain variant readings and degrees of certainty, preserving interpretive flexibility and philological rigor. The resulting data support structured querying and export to standard formats (e.g., EpiDoc, LAF), enabling interoperability and comparative synchronic and diachronic research.

3. Corpus Overview

3.1 Linguistic Scope

The initial phase of the platform's corpus centers on Phoenician inscriptions from the Levantine "Motherland" (Byblos, Sidon, Tyre, and surrounding regions), drawing from the canonical selection in *Kanaanäische und Aramäische Inschriften* (KAI, Donner & Röllig 2002). Though selective, KAI serves as a backbone for Northwest Semitic epigraphy, offering historically and linguistically significant texts. The platform demonstrates the annotation and editorial workflow through diverse literary types, including royal inscriptions, temple dedications, funerary stelae, tariffs, and bilingual texts[2] on stone, metal, ostraca, and pottery.

3.2 Epigraphic Scope and Sources

Each inscription in the platform is encoded as an ordered array of lines, segmented into tokens and subdivided into graphemic segments, representing clitics, stems, suffixes, and inflectional morphemes. Linguistic annotations are stored in separate queryable layers for lemma, root, part of speech, morphology, phonology, and context-sensitive glosses. Transcriptions follow published editions and support both diplomatic and normalized readings (see Section 1.2), in accordance with established editorial practices[3].

3.3 Digital Preparation and Metadata Model

Metadata fields include inscription name and variants, language, dialect, script, provenance and current locations (with GIS coordinates), archaeological dating, discovery date and legacy sigla (e.g., KAI, CIS). Bibliographic references are linked to online copies, when available, and visual media is normally derived from Wikimedia Commons files. The easily customizable model supports object classification details and physical dimensions, following best practices from Trismegistos (Depauw & Gheldof 2014) and DASI projects.

The platform uses a standard Latin-based transliteration system alongside full Unicode display of the original script, as well as auto-transliteration into square Hebrew script to support orthographic comparison. In line with KAI, the system applies a consistent set of epigraphic sigla to mark restorations, damage, and other textual features. These can be presented in an optional layer of graphemic annotations to preserve the clarity of the main text while supporting detailed scholarly analysis.

3.4 Corpus Growth and Future Extensions

The current corpus, built from an initial set of about twenty inscriptions, functions as a testing ground for the full editorial workflow: editors annotate texts from established scholarly

[2] See Gzella (2011) for linguistic and historical overview.

[3] Cf. Sahle (2016) that defines and characterizes Scholarly Digital Editions (SDEs) as the critical representations of historical documents, distinct from digitized text or digital libraries.

editions, submit their work for peer review, for the inscription data to receive a final approval from a subject specialist before being published and added to the platform. The system is designed for modular growth and for expansion across languages, since its annotation model is grounded in core Semitic morphological concepts such as root, stem, and state, among others. This foundation enables cross-linguistic querying across Canaanite and the wider Semitic family as the project grows. The corpus and its editorial tools were conceived not as a static repository, but as a collaborative research environment built for comparative linguistic work.

4. Layered Annotation Workflow

4.1 Conceptual Architecture

The annotation framework is based on the principle that Semitic inscriptions require multiple interdependent interpretive layers. Unlike monolithic transcription models, the platform adopts a modular architecture in which distinct annotation layers interact. This architecture reflects both the non-linear nature of epigraphic interpretation and the scholarly need to document variation, uncertainty, and contextual dependency at multiple levels.

The system draws conceptual inspiration from multi-layer annotation standards in computational linguistics (e.g., LAF/GrAF; Ide & Romary 2004), but tuned to the specificities of Semitic historical texts.

4.2 Discrete Annotation Layer Tiers

- **Paleographic Layer:** Optional, showing relevant graphemic variation, uncertainty and visual distinctions (e.g., regional ductus).
- **Graphemic Layer:** Encodes the literal signs as segmented from the inscription, aligning each grapheme to its linear position and Unicode representation.
- **Morphosyntactic Layer:** Segments each token into its internal morphemes (e.g., prefixes, stems, suffixes), with annotations for lemma, root, and setting the main part of speech of that token, which then contextually offers additional features to set, such as gender, number, stem, and state.
- **Semantic and Pragmatic Layers:** Include translation equivalents, glosses, context-sensitive disambiguation, semantic field and entity type.
- **Contextual Layer:** Stores editorial notes, variant readings, references to published debates, and degrees of certainty.

Editorial interventions are modeled as span-based annotations over tokens. Tokens carry both diplomatic and normalized forms linked through token identifiers. In other words, lexical annotation, associated with normalized word entities, is linked to one or more tokens. Multiple tokens across lines may comprise a single lexical word, thus preserving the distinction between inscriptional form and linguistic interpretation. Such separation preserves epigraphic fidelity while enabling structured linguistic analysis and reproducibility. Because layers are modular and linked through unique identifiers, scholars can annotate and publish different aspects of an inscription without collapsing them into a single representation.

4.3 Data Flow and Inter-layer Relationships

The system follows a bottom-up workflow: editors enter graphemes, segment tokens, and then may add morphological, phonological and syntactic analysis. Layers remain interlinked but revisable; semantic reinterpretation may trigger morphological updates, all tracked through internal logs to ensure transparency. Prior to the expert-led annotation, the platform already provides limited semi-automated input assistance, with tokens split from transliterated or original script line-level entry. Editors can also copy annotations of previously set tokens through indexed lookups, but keeping their editorial control.

4.4 Collaborative Workflow Design

The annotation process is structured as a staged editorial workflow. Each inscription is assigned to an editor responsible for initial graphemic and linguistic encoding. Submitted annotations enter a peer review, where collaborators may comment, suggest revisions, and flag uncertainties, for the editor's consideration. Final approval is granted by a chief editor, after which the inscription becomes part of the published corpus. All stages are versioned and logged, ensuring traceability of editorial decisions.

4.5 Workflow Overview

Figure 1 shows the staged progression from assignment to publication. The dashed path shows an iterative peer feedback before delivery for approval, reflecting the non-linear nature of philological revision. The model emphasizes controlled state transitions rather than open editing, ensuring accountability while preserving collaborative input.

A typical workflow for an inscription includes the following stages:

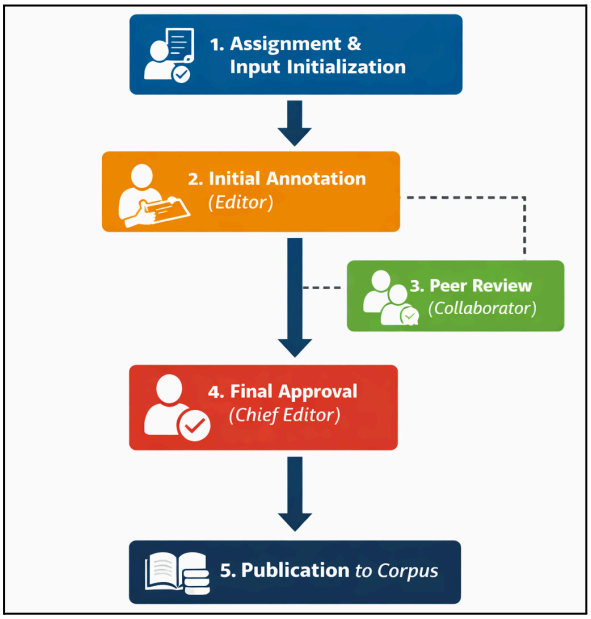

Figure 1. Collaborative Editorial Workflow for Layered Annotation of Semitic Inscriptions.

1. **Assignment & Input Initialization**: An editor is assigned an inscription (e.g., from KAI). The base text is initialized with line and token segmentation in both transliteration and original script Unicode, alongside a sourced transcription.
2. **Initial Annotation (Editor)**: The assigned editor verifies token segmentation and adds token and segment annotation, populating morphological, semantic, and contextual layers. Published editions, lexica, and relevant literature are consulted throughout. Optional commentary is attached at the grapheme, segment or token level, giving room to supplemental or even sometimes contradictionary interpretations as either token-linked linguistic notes or using the built-in Apparatus mechanism.
3. **Peer Review (Collaborator)**: A second editor reviews the annotations. Feedback includes disagreements about segmentation, alternate morphological readings, or suggested corrections in metadata or glossing. Comments and change suggestions are tracked in the platform and reviewed by the editor before final approval. Said disagreements are not automatically resolved but recorded as alternative editorial interpretations linked to relevant tokens or spans. The workflow offers multiple proposed readings or analyses to be reviewed by a chief editor, who makes the final decision, either preserving these proposals in the inscription's notes or version history or discarding them. This ensures transparency of editorial decisions and allows future revisions if new interpretations emerge.
4. **Final Approval (Chief Editor)**: A designated field expert performs a final pass, validating the quality and completeness of the annotation. Submissions, in whole or partially, can be approved as-is or sent back for revision. All editorial decisions are recorded in a transparent log.
5. **Publication**: Approved inscriptions may then be published into the corpus. All layers are stored as discrete, queryable structures. Users may browse, search, and later export annotations per layer or as full bundled records, in addition to the inscription's text and metadata.

Each step in the workflow is timestamped and version-controlled, supporting rollback. A draft lives alongside a public canonical version (should one already exist), until its editorial work is finalized, reviewed and approved, prior to being published for the first time or as a replacement to the existing copy. Researchers can reconstruct editorial decisions and preserve the academically-expected transparency in epigraphy.

The interface supports sequential multi-user annotation with locked editing states, visual differences tracking for chief editors, and source citation fields to anchor interpretations to specific scholarly sources.

The platform's export capabilities will include structured formats (e.g., TEI, LAF), ensuring compatibility with infrastructures like CLARIN or ELRC.

5. Implementation and Infrastructure

5.1 System Architecture

The platform is a browser-based annotation environment with a React frontend and Firebase backend, centered on Firestore, Google's cloud-hosted NoSQL database. Designed for collaborative editing, Firestore provides real-time synchronization, offline access, and fine-grained access control, enabling a responsive, multi-user workflow without custom server infrastructure. It maps naturally onto the hierarchical structure of epigraphic data, where inscriptions contain

metadata, segmented text, and apparatus entries within nested Firestore documents. This supports efficient editing and UI rendering without relying on complex relational joins. Its declarative security rules enable role-specific editorial permissions, essential for managing contributions across contributor, reviewer, and approver roles. Priority is thus given to operational simplicity and scholarly usability: versioned real-time updates are enforced via TypeScript schemas, reducing backend complexity. While Firestore has limits, e.g., no native joins or complex aggregations, these are mitigated through export pipelines planned for integration with SQL backends or TEI/XML-compatible formats for long-term interoperability and corpus-wide analysis (cf. Sahle 2016; Elliott et al. 2017-2026).

5.2 Data Model and File Formats

The platform uses a structured JSON data model reflecting the distinction between physical inscriptional units and linguistic analysis. An inscription comprises metadata, an ordered sequence of lines, and a lexical layer linking tokens to abstract word-level objects[4]. Lines consist of tokens representing physical units of the inscription while tokens store graphemic content, diplomatic and normalized forms, and optional morphological segments. Linguistic analysis is modeled separately: word objects linked via token reference IDs store lemma, root, part of speech, morphology, phonology gloss, and related features. Editorial interventions and variant readings are represented through span-based annotations at either the line or inscription (apparatus) level .

The schema is enforced through TypeScript interfaces to ensure structural consistency and machine readability. While conceptually aligned with standards such as EpiDoc, the model is implemented natively in JSON, with planned export to TEI-compliant formats. A representative JSON example of the inscription data model is provided in Appendix A. Detailed sample annotated data will be made available upon publication.

5.3 Storage, Versioning, and Export

Firestore functions as the platform's primary storage layer, supporting structured queries and cloud-based scalability. Each modification generates a timestamped audit trail at both token and document levels, enabling editors to track annotation history and restore prior states.

The platform is planned to support data export in standard scholarly formats, including TEI-XML for EpiDoc compatibility, TSV or CSV for linguistic analysis, and JSON-LD for semantic web integration. While current exports are user-initiated, the architecture is designed for automated format generation through cloud functions as the corpus scales.

5.4 Interoperability and Integration Potential

Although designed for human-led annotation, the modular architecture facilitates integration with computational tools and external research infrastructures. Planned extensions include interoperability with NLP pipelines for lemmatization, pattern detection, and grapheme-level classification.

The platform is being developed to integrate with research infrastructures such as CLARIN (Váradi et al. 2008) and other archival standards. In addition to TEI and LAF/GrAF export, future development will include RDF serialization compatible with Linked Data frameworks such as the Open Annotation Model and POWLA, allowing annotated inscriptions and linguistic data to be linked with external repositories and research datasets. The system is also designed with compatibility in mind for cultural heritage and museum data standards such as CIDOC-CRM (Crofts, Doerr, & Gill 2003) and Linked Open Data frameworks (Bizer, Heath, & Berners-Lee 2009), enabling connections between inscriptions, places, persons and related historical datasets.

6. A Linguistic Case Study Enabled by Layered Annotation

6.1 Research Question and Linguistic Background

Formulaic language is a hallmark of Phoenician and other Northwest Semitic inscriptions, especially in royal and votive contexts. Across the corpus, recurring motifs such as dedication formulae, genealogical openings, and blessing or curse sequences tend to follow stable morphosyntactic patterns while exhibiting significant variation in orthography, morphology, and vocabulary. Traditionally, these patterns have been discussed qualitatively and compared through philological judgment.

[4] This distinction is required as inscriptional word boundaries may not align with physical line boundaries (KAI 10:L.10↔11, L.11↔12)

This case study examines whether recurring dedication formulae can be operationalized as structured morphosyntactic configurations that are computationally retrievable across inscriptions. Rather than simply identifying new linguistic patterns, the platform shows how familiar constructions can be modeled as structured units through LLA, clarifying their internal architecture, their points of variation, and their reproducibility for further research.

6.2 Data Selection and Annotation Basis

Drawing on a subset of the Phoenician corpus encoded in the platform, the focus is given to five royal and two dedicatory inscriptions (e.g. the Ahiram Sarcophagus, Eliba῾l Osorkon Bust), dated between the 10th and the 3rd centuries BCE, featuring particularly prominent formulaic language.

All selected inscriptions are tokenized and segmented into graphemic and morphological units. Linguistic annotations include lemma, root, part of speech, and relevant morphological features, alongside both diplomatic and normalized representations. This annotation provides a consistent basis for identifying recurring constructions while maintaining alignment with published editions.

6.3 Representing Formulaic Morphosyntax

As a concrete example, the case study focuses on dedication and self-presentation clauses built around verbs of making, dedicating, or establishing, typically followed by an agent (often a royal or donor proper name), an object, and a recipient deity introduced by a prepositional marker. While surface realization varies across inscriptions, the underlying morphosyntactic structure is stable.

Within the LLA model, each component of such a clause is encoded explicitly: the verbal predicate is identified by lemma and morphological features; proper names are tagged as nominal tokens with onomastic and grammatical properties; prepositions and pronominal elements are segmented at the morpheme level; and divine names are annotated as distinct lexical entities. Because token order and segmentation are preserved, equivalent constructions can be compared across inscriptions despite orthographic or morphological variation.

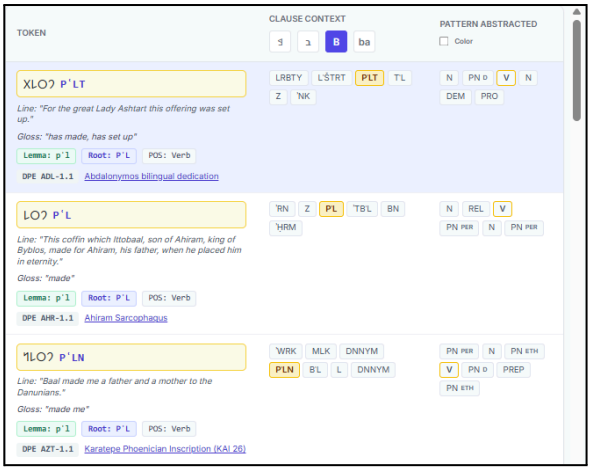

Figure 2: Retrieval of verbal forms

The query in Figure 2 filters tokens by morphological features (POS=**verb**, root=√**P῾L**) and retrieves their linked clause-level annotations. Such structured retrievals expose recurrent predicate-agent-recipient sets across inscriptions. Since these relations are encoded explicitly in the data model, equivalent morphosyntactic templates can be compared independently of surface variation. Formulaic structures can therefore be modeled as configurations of linguistic roles rather than linear text sequences. This representation makes formulaic morphosyntax comparable across inscriptions while maintaining philological interpretation.

6.4 Implications for Comparative Semitic Studies

Although initialized with a limited dataset, the model of structured fields is grounded in shared Semitic morphological principles: root-based lexemes, affixal morphology, state distinctions and syntactic roles. This encoding allows comparison of equivalent constructions across the entire Northwest Semitic corpora and beyond, transcending both time, space, language and dialect boundaries, as shown in Figure 3.

In this respect, not only qualitative comparison is supported, but also systematic investigation of formulaic morphosyntax, lexical distributions and structural variation across related varieties. Bilingual inscriptions further provide controlled environments for aligning normalized representations across linguistic systems. LLA functions as an analytical substrate for comparative Semitic linguistics, enabling reproducible cross-corpus research.

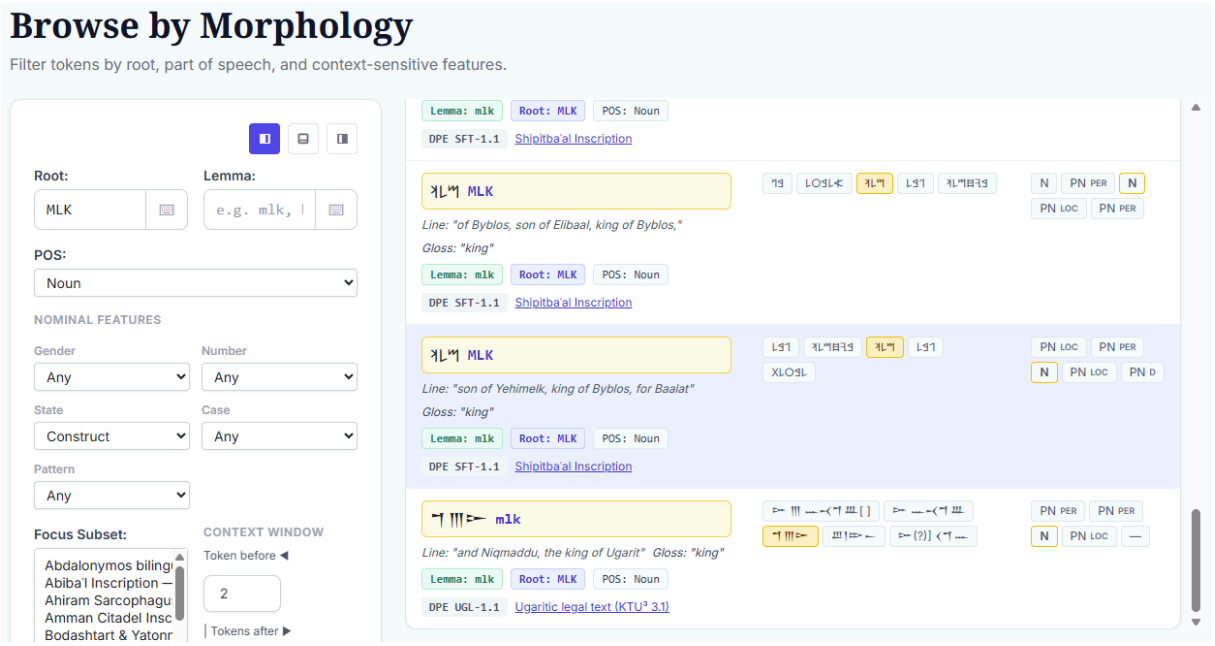

Figure 3. Cross-inscription morphological lookup

7. Limitations and Future Work

7.1 Current Limitations

The platform and corpus are currently in an early stage of development, and several limitations must be acknowledged. The corpus is intentionally restricted to a curated subset of Phoenician inscriptions, with only limited coverage of other Canaanite and Northwest Semitic languages. While sufficient for demonstrating the annotation architecture and supporting focused case studies, this scope does not yet allow large-scale quantitative analysis or broad typological generalizations.

Although inscriptions are tokenized and morphologically segmented, annotation depth and completeness vary across texts, reflecting the iterative and collaborative nature of the editorial workflow. More advanced analytical operations, such as pattern-based multi-token queries or large-scale aggregation, are not yet available at the interface level, despite being supported conceptually by the data model.

Finally, the system currently prioritizes human-led annotation and editorial control. Automated linguistic support, such as morphological pre-annotation or pattern suggestion based on recurring structures, remains under development and is not yet fully integrated into the workflow.

7.2 Planned Development

Corpus coverage and usability are expanded as a main focus. This includes adding further Phoenician, Punic, and other Northwest Semitic inscriptions while maintaining editorial consistency using the review and approval workflow. Interface improvements will support more effective browsing and comparison of annotated structures, including expressive queries over linguistic features and token sequences.

From a technical perspective, future work includes export to standardized formats such as TEI/EpiDoc, ensuring interoperability and long-term preservation. Semi-automated tools, including AI-assisted draft annotation or suggestion mechanisms, are envisaged as optional preliminary editorial aids, to support routine tasks while adhering to expert oversight and final scholarly authority.

While the annotated corpus will be released in stages, with editorial considerations in mind, the platform code is planned to be released as an open-source project, following rigorous stability and multiple user roles operability tests. Separating the annotation infrastructure from the corpus data will allow other projects to reuse the platform for different epigraphic or historical linguistic corpora and contribute to further development of the annotation workflow.

7.3 Long-term Research Roadmap

In the longer term, the project aims to establish a generalizable annotation workflow applicable to a wider range of Semitic language corpora, extending beyond Phoenician to other Canaanite

and Northwest Semitic traditions[5], as well as closely related epigraphic corpora. By maintaining a layered, language-agnostic representation grounded in shared morphological principles, the infrastructure is designed to support comparative research across time, region, and genre.

Future directions include closer integration with analytical environments for corpus and historical linguistics, and the exploration of computational approaches to modeling formulaic language at scale. Throughout these developments, computational methods are intended to support, rather than replace, philological expertise, ensuring that scholarly responsibility and editorial control remain central to the workflow.

8. Conclusion

This paper has presented a layered annotation workflow and digital infrastructure for the computational representation of Semitic epigraphic texts. Fine-grained LLA combined with a structured editorial workflow aligns philological practice with reproducible computational research. Inscriptions are structured, versioned research objects whose analytical layers can be encoded, reviewed, and compared systematically.

Using the focused corpus and a concrete linguistic case study, the paper has shown how well-established epigraphic and morphosyntactic phenomena are explicit, queryable, and comparable across texts. The layered data model enables the representation of graphemic, morphological, semantic, and contextual information. In other words, synchronic and diachronic analysis isn't flattening the interpretive complexity inherent to epigraphic sources.

Although the system is currently limited in scope and remains under active development, its modular design and language-agnostic principles make it extensible to other Semitic epigraphic traditions and related corpora. The workflow demonstrates how computational tools can augment, rather than replace, expert-driven scholarship. By foregrounding annotation as an interpretive and collaborative process, this work contributes a sustainable model for future digital research on historical and ancient languages.

[5] I.e., the epigraphic and written traditions that are linguistically, structurally, and methodologically adjacent to Northwest Semitic corpora, and therefore compatible with the same LLA principles.

9. Required Statements

9.1 Ethics Statement

This research is based on the study of historical and ancient epigraphic texts that are part of the public cultural heritage. The corpus consists exclusively of inscriptions that have been previously published in scholarly editions and does not involve human subjects, personal data, or sensitive information. The annotation workflow is designed to support transparent scholarly interpretation, including the explicit representation of uncertainty, alternative readings, and editorial responsibility. No ethical risks associated with data collection, annotation, or dissemination have been identified.

9.2 Data and Code Availability

The corpus and annotation workflow described in this paper are under active development. The annotated data are derived from published epigraphic sources and are encoded within a custom digital infrastructure designed for collaborative scholarly annotation. At the time of submission, the full dataset and source code are not yet publicly released due to ongoing editorial work and review processes. The authors intend to make selected components of the corpus, along with documentation of the annotation model and workflow, available in future releases, in accordance with licensing constraints and best practices for the sharing of language resources. The platform infrastructure is designed to be released independently of the corpus data, allowing reuse of the annotation environment by other projects.

9.3 Acknowledgements

The authors would like to thank colleagues and mentors for scholarly guidance and discussion that contributed to the development of this project. Particular thanks are due to academic collaborators for their feedback on the conceptual design, linguistic scope, and methodological framing of the workflow. Any remaining errors or limitations are the sole responsibility of the authors.

10. Bibliographical References

Al-Jallad, A. (2015). An Outline of the Grammar of the Safaitic Inscriptions. Brill.

Avanzini, A., Prioletta, A., & Rossi, I. (2014). The Digital Archive for the Study of Pre-Islamic Arabian Inscriptions: an ERC project. https://doi.org/10.13131/UNIPI.ARABIANTICA.0000000002

Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked data: The story so far. International Journal on Semantic Web and Information Systems, 5(3), 1–22.

Bodard, G., & Stoyanova, S. (2016). Epigraphers and encoders: Strategies for teaching and learning digital epigraphy. In G. Bodard & M. Romanello (Eds.), Digital classics outside the echo-chamber: Teaching, knowledge exchange & public engagement (pp. 51–68). Ubiquity Press. https://doi.org/10.5334/bat.d

Chiarcos, C. (2012). POWLA: Modeling linguistic corpora in OWL/DL. In C. Chiarcos, S. Nordhoff, & S. Hellmann (Eds.), Linked data in linguistics: Representing and connecting language data and language metadata (pp. 225–246). Springer. https://doi.org/10.1007/978-3-642-30284-8_22

Chiarcos, C., Nordhoff, S., & Hellmann, S. (Eds.). (2012). Linked data in linguistics: Representing and connecting language data and language metadata. Springer. https://doi.org/10.1007/978-3-642-28249-2

Crofts, N., Doerr, M., & Gill, T. (2003). The CIDOC Conceptual Reference Model: A standard for communicating cultural contents. Cultivate Interactive, 9.

De Santis, A., Rossi, I. 2018. Crossing Experiences in Digital Epigraphy. From Practice to Discipline. Warsaw-Berlin: De Gruyter.

Depauw, M., & Gheldof, T. (2014). Trismegistos: An interdisciplinary portal of the ancient world. In S. Berti & M. Costa (Eds.), Ancient Worlds in Digital Culture (pp. 293–303). Brill. https://www.trismegistos.org/

Derntl, M., Gietz, P., Helling, P. 2023. FORGE 2023 - Anything Goes?! Forschungsdaten in den Geisteswissenschaften - kritisch betrachtet. Tübingen, 4.-6. Oktober 2023. Zenodo. https://10.5281/zenodo.8341605.

Di Filippo, F. 2018. “Sinleqiunnini: Designing an Annotated Text Collection for Logo-Syllabic Writing Systems”. In De Santis, Annamaria and Irene Rossi (eds.) Crossing Experiences in Digital Epigraphy (pp. 49–64). Warsaw-Berlin: De Gruyter.

Donner, H. and Röllig, W. Kanaanäische und Aramäische Inschriften, 5th ed. (Wiesbaden: Harrassowitz, 2002).

Elliott, T., Bodard, G., Mylonas, E., Stoyanova, S., Tupman, C., Vanderbilt, S., et al. (2017–2026). EpiDoc guidelines: Ancient documents in TEI XML (Version 9). https://epidoc.stoa.org/gl/latest/

Gzella, H. (2011). Phoenician. In H. Gzella (Ed.), Languages from the World of the Bible (pp. 55-75). Berlin: De Gruyter.

Ide, N., & Romary, L. (2004). A registry of standard data categories for linguistic annotation. In Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC’04) (pp. 135-138). European Language Resources Association.

Ide, N., & Suderman, K. (2014). The linguistic annotation framework: A standard for annotation interchange and merging. Language Resources and Evaluation, 48(3), 395–418. https://doi.org/10.1007/s10579-014-9268-1

Koller, A. 2013. “Ancient Hebrew מעצד and עצד in the Gezer Calendar”. Journal of Near Eastern Studies 72: 179-193.

Pierazzo, E. (2016). Textual scholarship and text encoding. In S. Schreibman, R. Siemens, & J. Unsworth (Eds.), A new companion to digital humanities (pp. 307-321). Wiley-Blackwell.

Renan, E. (1881-1962). Corpus Inscriptionum Semiticarum (CIS). Historical multi-volume collection of Semitic inscriptions.

Sahle, P. (2016). What is a scholarly digital edition? In M. J. Driscoll & E. Pierazzo (Eds.), Digital scholarly editing: Theories and practices (pp. 19-39). Open Book Publishers. https://doi.org/10.11647/OBP.0095

Sanders, S. 2008. “Writing and Early Iron Age Israel: Before National Scripts, Beyond Nations and States”. In Tappy, Ron and P. Kyle McCarter (eds.) Literate Culture and Tenth-Century Canaan: The Tel Zayit Abecedary in Context. Winona Lake, IN (pp. 100-102).

Sanderson, R., Ciccarese, P., & Van de Sompel, H. (2017). Web annotation data model. World Wide Web Consortium (W3C). https://www.w3.org/TR/annotation-model/

TEI Consortium. (2022). Guidelines for Electronic Text Encoding and Interchange (Version P5, 4.5.0). https://tei-c.org/release/doc/tei-p5-doc/en/html/

Váradi, T., Wittenburg, P., Krauwer, S., Wynne,

M., Koskenniemi, K., & others. (2008). CLARIN: Common language resources and technology infrastructure. In Proceedings of LREC 2008. ACL Anthology. https://aclanthology.org/L08-1008/

Vitale, V., de Soto, P., Simon, R., Barker, E., Isaksen, L. and Kahn, R. 2021. Pelagios - Connecting Histories of Place. Part I: Methods and Tools. International Journal of Humanities and Arts Computing, 15/1-2: 5–32.

10.1. Digital Corpora and Online Language Resources

ACO Arabic Collections Online. (n.d.). *Arabic Collections Online*. Digital Library Technology Services. https://dlib.nyu.edu/aco/

AICC AI Cuneiform Corpus. https://aicuneiform.com/

Beta maṣāḥəft: Manuscripts of Ethiopia and Eritrea. www.betamasaheft.uni-hamburg.de

CASPO Corpus of Akkadian Shuila Prayers Online. http://www.shuilas.org/

CDL Cambridge Digital Library. Ethiopian Manuscripts. https://cudl.lib.cam.ac.uk/collections/ethiopianmanuscripts/1

CDLI Cuneiform Digital Library Initiative. (n.d.). *Home*. https://cdli.earth/

Corpus Coranicum https://www.arabic-latin-corpus.philosophie.uni-wuerzburg.de/

CSAI Corpus of South Arabian Inscriptions. https://dasi.cnr.it/index.php?id=42&prjId=1&corId=0&collId=0

DANES Online Resources for Digital Ancient Near Eastern Studies. https://opendanes.org/nav/DANES-resources.html

Digital Archive for the Study of Pre-Islamic Arabian Inscriptions (DASI). (n.d.). *DASI*. https://dasi.cnr.it/

DCGAS Digital Corpus for Graeco-Arabic Studies. https://www.graeco-arabic-studies.org/home.html

Ebla Digital Archives (EbDA). (n.d.). *Ebla Digital Archives*. http://ebda.cnr.it/

eHammurabi https://ehammurabi.org/

EPIDAT The Database of Jewish Epigraphy. https://www.re3data.org/repository/r3d100012712

National Library of Israel. (2024). Ktiv: The International Collection of Digitized Hebrew Manuscripts. https://www.nli.org.il/en/discover/manuscripts/hebrew-manuscripts

Lauinger, J. and Yoder, T. 2014. AEMW/ Amarna. https://oracc.museum.upenn.edu/aemw/amarna/index.html

OCIANA. (n.d.). *Online Corpus of the Inscriptions of Ancient North Arabia*. https://dasi.cnr.it/index.php?id=42&prjId=4

Open Richly Annotated Cuneiform Corpus (ORACC). (n.d.). *ORACC*. https://oracc.museum.upenn.edu/

PDLIM Princeton Digital Library of Islamic Manuscripts. https://dpul.princeton.edu/islamicmss

PFDP Polonsky Foundation Digitization Project. http://bav.bodleian.ox.ac.uk/hebrew-manuscripts

QAC The Quranic Arabic Corpus. https://corpus.quran.com/

syri.ac. (n.d.). *An Annotated Bibliography of Syriac Resources Online*. https://syri.ac/digimss

Appendix A. Inscription Data Model

A.1 Overview

This appendix presents a representative JSON structure for an inscription document in the annotation platform. The data model separates inscription-level metadata, line-based diplomatic transcription, token- and grapheme-level epigraphic annotation, and word-level linguistic annotation. In the implementation, each inscription is stored as a single document containing nested arrays for lines, translations, editorial spans, and tokens, together with a separate `wordsById` object for lexical entries that may link to one or more tokens. This design preserves the diplomatic transcription while allowing linguistic, phonological, and editorial annotations to be queried independently.

A.2 Top-level Inscription Document Structure

A representative inscription document contains metadata, line objects, word-level records, and workflow state:

```
{
  "metadata": {
    "iid": "DPE-AHR-1.1",
    "ctg_id": "DPE-AHR-1",
    "legacy_file_id": "KAI_1",
    "language": "Phoenician",
    "script": "Phoenician Alphabet",
    "date": "10th Century BCE",
    "date_year_sort": -950,
    "period": "Iron Age IIA",
    "provenance": {
      "place": "Byblos, Lebanon",
           "coordinates": { "lat": 34.1236,
"long": 35.6513 }
    },
    "location": {
      "place": "National Museum of Beirut"
    },
    "legacy_refs": [
       { "type": "KAI", "number": "1", "ref":
"KAI 1" }
    ],
    "sigla": [
      { "label": "KAI 1" }
    ],
    "bibliography": [
       { "citation": "Donner, H. & Röllig, W.
(1962). KAI, vol. 1, no. 1." }
    ],
    "workflowStatus": "published",
    "visibility": "public"
  },
  "lines": [...],
  "wordsById": {...},
  "peerReviewState": "completed"
}
```

The `metadata` object stores bibliographic, chronological, geographic, and administrative information. The inscription text itself is represented in `lines`, while `wordsById` stores the separate lexical layer.

A.3 Line and Token Structure

The diplomatic transcription is stored as an ordered array of `lines`. Each line contains `tokens` and may also contain `translations` and editorial spans.

```
{
  "id": "L1",
  "line_number": 1,
  "translations": [
    {
      "id": "tr-L1-en",
      "language": "en",
       "text": "The coffin which Ittobaal, son
of Ahiram, king of Byblos, made",
      "source": "Gibson 1982"
    }
  ],
  "editorialSpans": [
    {
      "id": "es-1",
      "kind": "supplied",
      "cert": "high",
       "start": { "tokenId": "T1.3", "pos": 0
},
      "end": { "tokenId": "T1.3", "pos": 3 },
      "sourceEdition": "KAI"
    }
  ],
  "tokens": [
    {
      "id": "T1.1",
      "form": "ʾrn",
      "script_form": "𐤀𐤓𐤍",
      "normalized_form": "ʾarōn",
      "wordId": "W1.1",
      "lemma": "ʾrn",
      "root": "ʾry",
      "pos": "Noun",
      "gloss": "coffin",
      "morphology": {
        "gender": "Masculine",
        "number": "Singular",
        "state": "Absolute"
      }
    }
  ]
}
```

This structure keeps the physical sequence of the inscription at line and token level, while allowing each token to carry normalized, lexical, and morphological information. `editorialSpans` encode line-level editorial phenomena such as supplied text, gaps, or corrections.

A.4 Grapheme and Segment Annotation

Token-internal annotation can be represented at grapheme and segment level. Grapheme objects capture character-level epigraphic information such as attestation, restoration, damage, and certainty, while segment objects encode morpheme-level analysis aligned to graphemes.

```
{
  "id": "T1.1",
  "form": "ʾrn",
  "graphemes": [
    {
      "id": "T1.1:g:1",
      "n": 1,
      "base": "ʾ",
      "presence": "attested",
      "certainty": "secure",
      "damage": "none",
           "anchor": { "kind": "tokenSpan",
"start": 0, "end": 1 }
    },
    {
      "id": "T1.1:g:3",
      "n": 3,
      "base": "n",
      "presence": "restored",
      "certainty": "uncertain",
      "damage": "none",
           "anchor": { "kind": "tokenSpan",
"start": 2, "end": 3 }
    }
  ],
  "segments": [
    {
      "id": "T1.1:m:1",
      "n": 1,
      "kind": "stem",
      "form": "ʾrn",
      "lemma": "ʾrn",
      "gloss": "coffin",
      "morph": "N.ms.abs",
      "anchor": {
        "kind": "graphemes",
                "graphemeIds": ["T1.1:g:1",
"T1.1:g:2", "T1.1:g:3"]
      }
    }
  ]
}
```

This layer allows epigraphic and morphological annotation to remain linked without collapsing them into a single representation.

A.5 Token-Word Linking

The model distinguishes between inscriptional tokens and word-level lexical records. Tokens preserve the diplomatic sequence of the inscription and may contain local annotation, while `wordsById` stores a separate lexical object that can link to one or more tokens through `tokenRefs`. This is particularly useful when an interpreted lexical word spans multiple inscriptional units, or when clitics and compound constructions must be represented without losing the original inscriptional segmentation.

A representative word-level object is shown below:

```
{
  "W1.1": {
    "id": "W1.1",
    "tokenRefs": [
        { "lineId": "L1", "tokenId": "T1.1",
"order": 0 }
    ],
    "leaderTokenId": "T1.1",
    "leaderLineId": "L1",
    "displayPolicy": "leaderOnly",
    "lemma": "ʾrn",
    "root": "ʾry",
    "pos": "Noun",
    "gloss": "coffin",
    "morphology": {
      "gender": "Masculine",
      "number": "Singular",
      "state": "Absolute"
    }
  }
}
```

In this structure, `tokenRefs` records the inscriptional token or tokens associated with a lexical entry, while `leaderTokenId` and `leaderLineId` identify the default display anchor. This separation makes it possible to preserve inscriptional form and editorial segmentation at token level while storing normalized lexical analysis in a reusable word-level layer.

A.6 Translation and Metadata Structure

Translations are stored per line rather than only at inscription level, which allows multiple aligned renderings and source attributions.

```
{
  "translations": [
    {
      "id": "tr-L1-en",
      "language": "en",
       "text": "The coffin which Ittobaal, son
of Ahiram, king of Byblos, made",
      "source": "Gibson 1982"
    }
  ]
}
```

Metadata fields are stored in a separate inscription-level object and include identifiers, chronology, provenance, present location, bibliography, sigla, and legacy references. A compact example is shown below:

```
{
  "metadata": {
    "iid": "DPE-AHR-1.1",
    "ctg_id": "DPE-AHR-1",
    "legacy_file_id": "KAI_1",
    "language": "Phoenician",
    "script": "Phoenician Alphabet",
    "date": "10th Century BCE",
    "period": "Iron Age IIA",
    "provenance": {
      "place": "Byblos, Lebanon",
           "coordinates": { "lat": 34.1236,
"long": 35.6513 }
    },
    "location": {
      "place": "National Museum of Beirut"
    },
    "legacy_refs": [
       { "type": "KAI", "number": "1", "ref":
"KAI 1" }
    ],
    "sigla": [
      { "label": "KAI 1" }
    ]
  }
}
```

In the implementation, the metadata layer may additionally include multilingual names, images, bibliographic records, visibility settings, and workflow state.

A.7 Main Layers in the Inscription Document

The main components of the data model are as follows:

- `metadata`: per document (root), bibliographic, geographic, chronological, and administrative information
- `lines[]`: per document, the physical line divisions of the inscription:
  - `translations[]`: per-line translations, optionally with source attribution
  - `editorialSpans[]`: per-line, editorial ranges such as supplied, unclear, or corrected text
  - `tokens[]`: per-line, inscriptional token units with local annotation:
    - `graphemes[]`: per token, grapheme-level segmentation and physical condition flags
    - `segments[]`: per token, morpheme-level segmentation aligned to graphemes
    - `phonology`: per token, phonological transcription and feature bundles
- `wordsById`: per document (root), word-level lexical objects linked to (spanning) one or more tokens, for clitics, compound constructions, etc.

A representative set of key enumerated values includes:

- presence: `attested`, `restored`, `expanded`, `corrected`, `missing`
- certainty: `secure`, `uncertain`, `highly_uncertain`
- damage: `none`, `damaged`, `illegible`
- pos: `Noun`, `Verb`, `ProperNoun`, `Adjective`, `Adverb`, `Preposition`, `Conjunction`, `Demonstrative`, `DefiniteArticle`, `Numeral`, `Suffix`, `Particle`, `Pronoun`
- state: `Absolute`, `Construct`, `Emphatic`
- editorialSpan.kind: `supplied`, `gap`, `unclear`, `add`, `del`, `corr`, `choice`, `abbrExpan`