Linguistic Analysis in the *Inscriptions of Israel/Palestine* Project

https://www.inscriptionsisraelpalestine.org/

Christine Roughan (New York University), Christopher B. Zeichmann (Toronto Metropolitan University), Michael L. Satlow (Brown University)

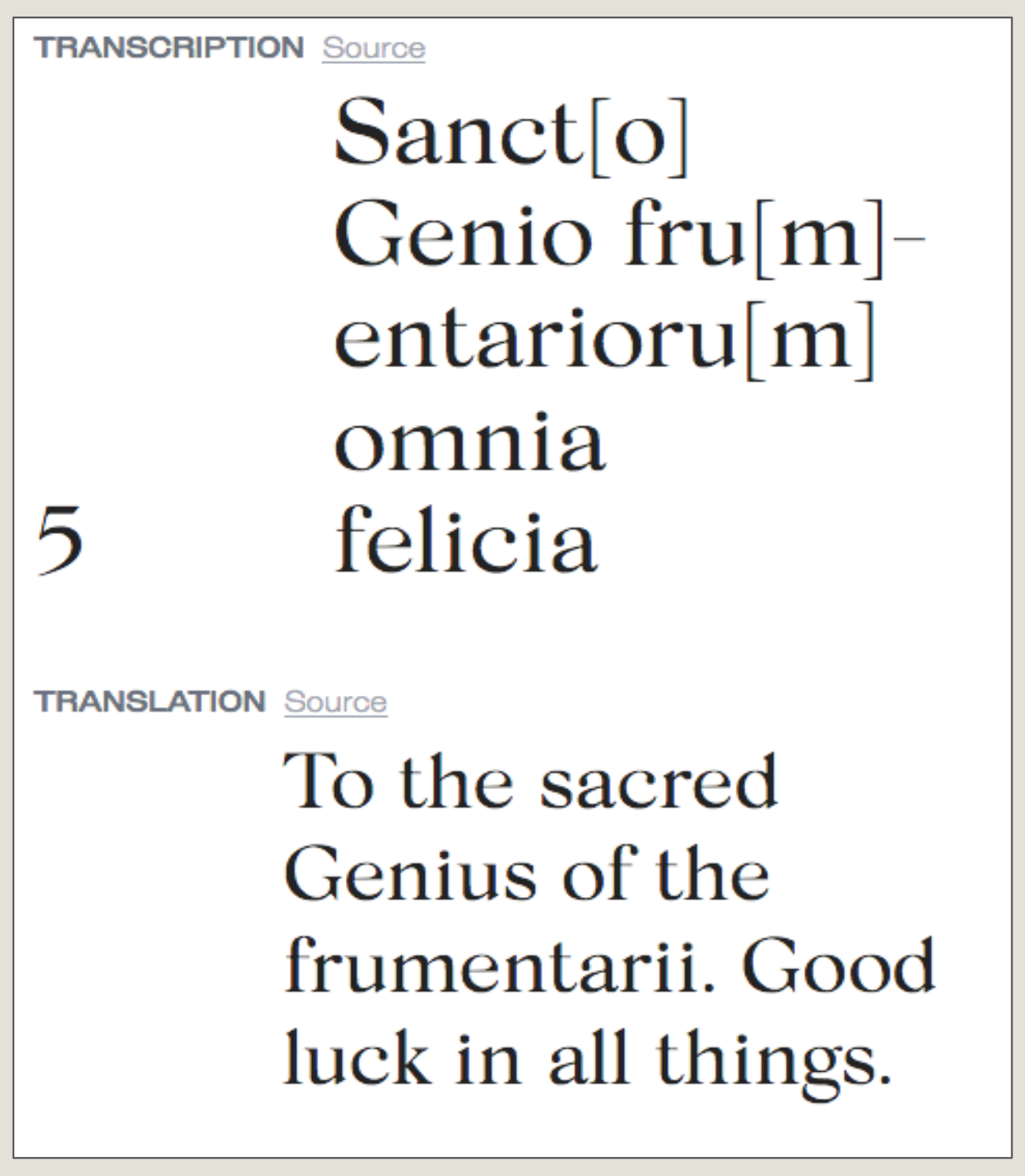
Introduction

The Inscriptions of Israel/Palestine (IIP) project presents a digital corpus of inscriptions from Israel and Palestine dating between the sixth century BCE and the seventh century CE. As of April 2023, this corpus includes 5,282 inscriptions encoded in EpiDoc-compliant XML, an example of which can be seen in the XML excerpt to the right. Four languages comprise the bulk of the corpus: 2,941 of these inscriptions contain Greek, 1,739 contain Aramaic, 457 contain Hebrew, and 262 contain Latin. Other languages (Phoenician, Classical Armenian, Syriac, Arabic, and Georgian) are represented in smaller amounts.

There are 38,256 tokens in the IIP's current dataset, 32,933 of which have been tagged as words. Across the main languages of the project are 19,677 words in Greek, 9,115 in Aramaic, 2,091 in Hebrew, and 1,869 in Latin.

Inscription CAES0683

Photo: Zev Radovan

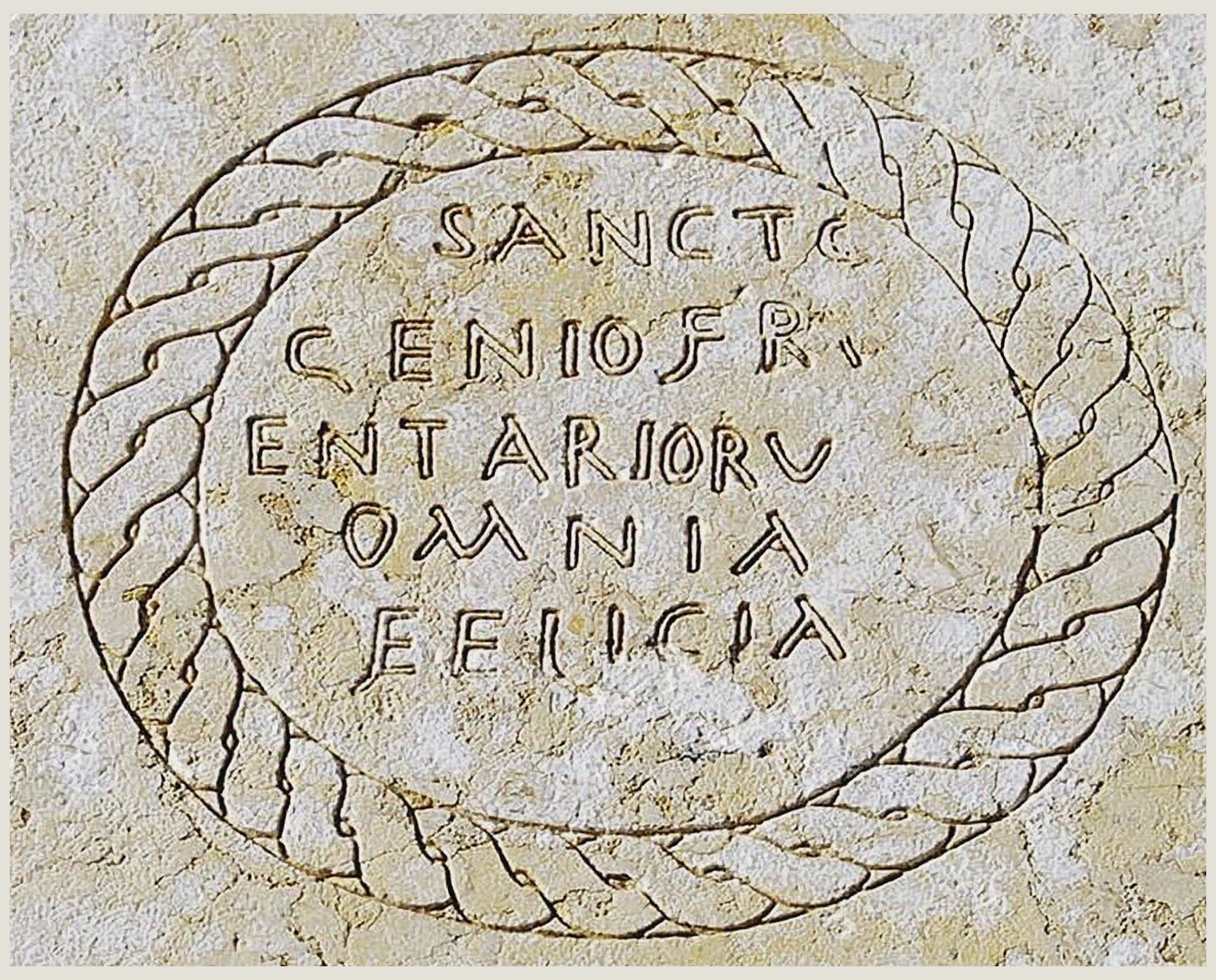


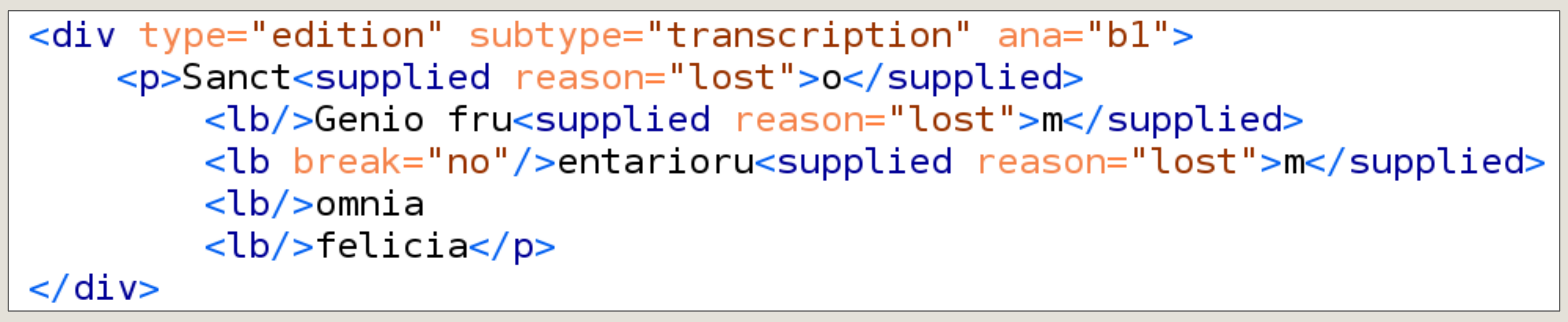
TRANSCRIPTION Source

Sanct[o]
Genio fru[m]-
entarioru[m]
omnia
5 felicia

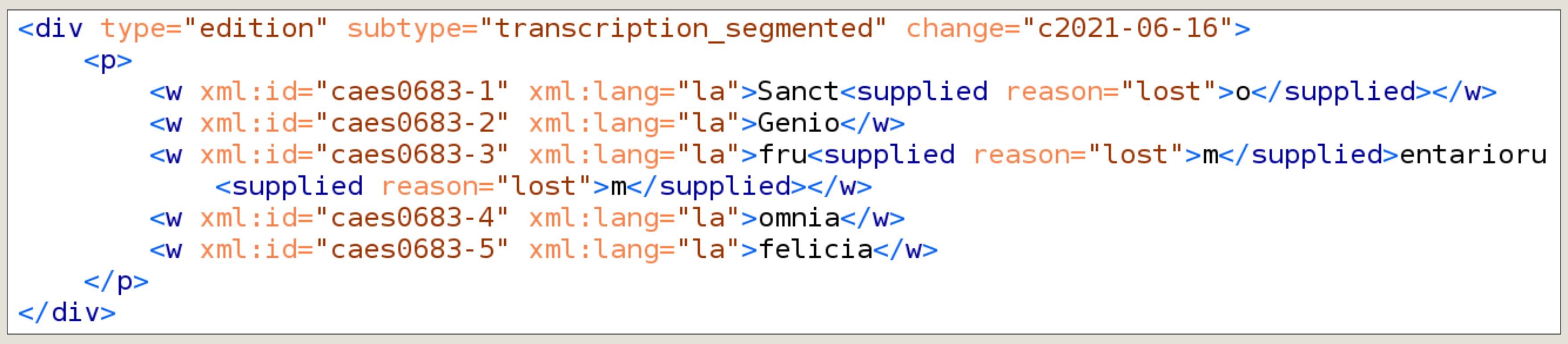
TRANSLATION Source

To the sacred
Genius of the
frumentarii. Good
luck in all things.



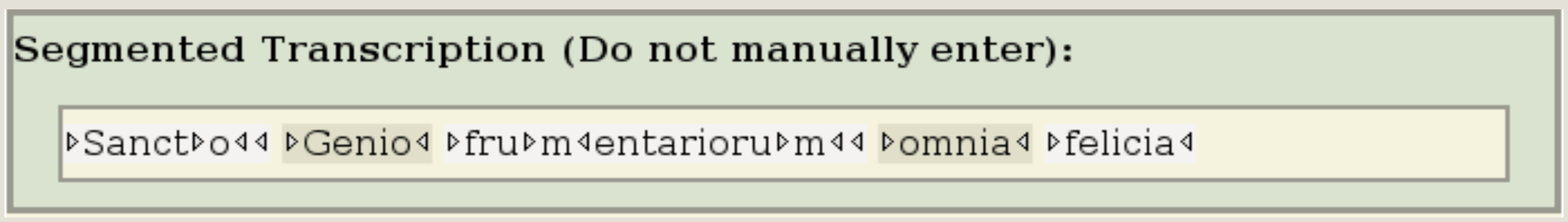


```
<div type="edition" subtype="transcription" ana="b1">
    <p>Sanct<supplied reason="lost">o</supplied>
        <lb/>Genio fru<supplied reason="lost">m</supplied>
        <lb break="no"/>entarioru<supplied reason="lost">m</supplied>
        <lb/>omnia
        <lb/>felicia</p>
</div>
```



```
<div type="edition" subtype="transcription_segmented" change="c2021-06-16">
    <p>
        <w xml:id="caes0683-1" xml:lang="la">Sanct<supplied reason="lost">o</supplied></w>
        <w xml:id="caes0683-2" xml:lang="la">Genio</w>
        <w xml:id="caes0683-3" xml:lang="la">fru<supplied reason="lost">m</supplied>entarioru
            <supplied reason="lost">m</supplied></w>
        <w xml:id="caes0683-4" xml:lang="la">omnia</w>
        <w xml:id="caes0683-5" xml:lang="la">felicia</w>
    </p>
</div>
```

Most of the EpiDoc XML tags from the transcription are preserved in the segmented data output for the new <div>.



Segmented Transcription (Do not manually enter):

▹Sanct▹o◃◃ ▹Genio◃ ▹fru▹m◃entarioru▹m◃◃ ▹omnia◃ ▹felicia◃

The IIP project's Author mode template in Oxygen XML will also display the segmented data.

Segmentation

After the IIP project's encoders transcribe an inscription in XML, an automated process handles word segmentation. This produces a new <div> element with words tagged as <w>, numbers as <num>, and unclear tokens as <orig>. Each token receives its own identifier, and both this and the language of the token is indicated in the XML (see the example on the left).

As of 2023, a Python script handles the word segmentation workflow.

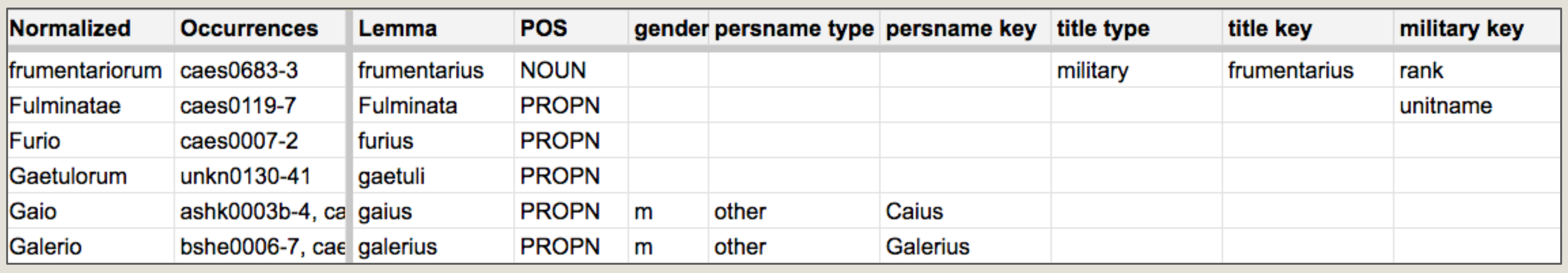
Linguistic Analysis

The word segmentation process additionally outputs every token into a CSV file; this file is then passed forward to the linguistic analysis workflows of the project.

Lexical Analysis

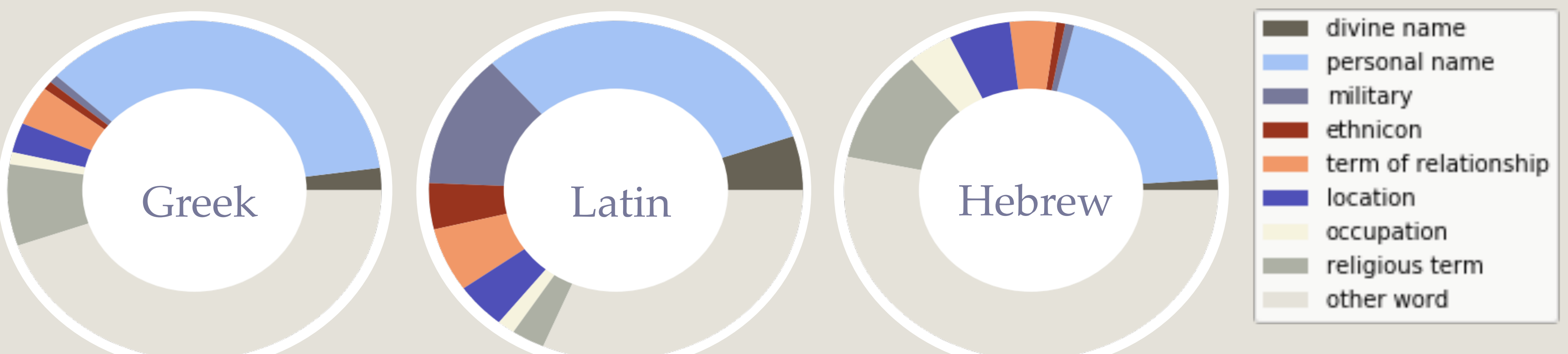
Lemmatizing and morphological parsing are one part of these workflows. The IIP project uses a combination of manual and automated approaches and has collaborated with LiLa: Linking Latin and DICTA to produce an initial batch of lexical data. In addition, the project is currently exploring automated lemmatizers and parsers (such as those offered in the Classical Language Toolkit for Python). The already extant parsed data will allow for evaluations of how effective such automated tools are for work with the IIP project's corpus.

A sample selection of the lexical data and tagged features produced from the segmented words in Latin inscriptions.



Normalized	Occurrences	Lemma	POS	gender	persname type	persname key	title type	title key	military key
frumentariorum	caes0683-3	frumentarius	NOUN				military	frumentarius	rank
Fulminatae	caes0119-7	Fulminata	PROPN						unitname
Furio	caes0007-2	furius	PROPN						
Gaetulorum	unkn0130-41	gaetuli	PROPN						
Gaio	ashk0003b-4, ca	gaius	PROPN	m	other	Caius			
Galerio	bshe0006-7, cae	galerius	PROPN	m	other	Galerius			

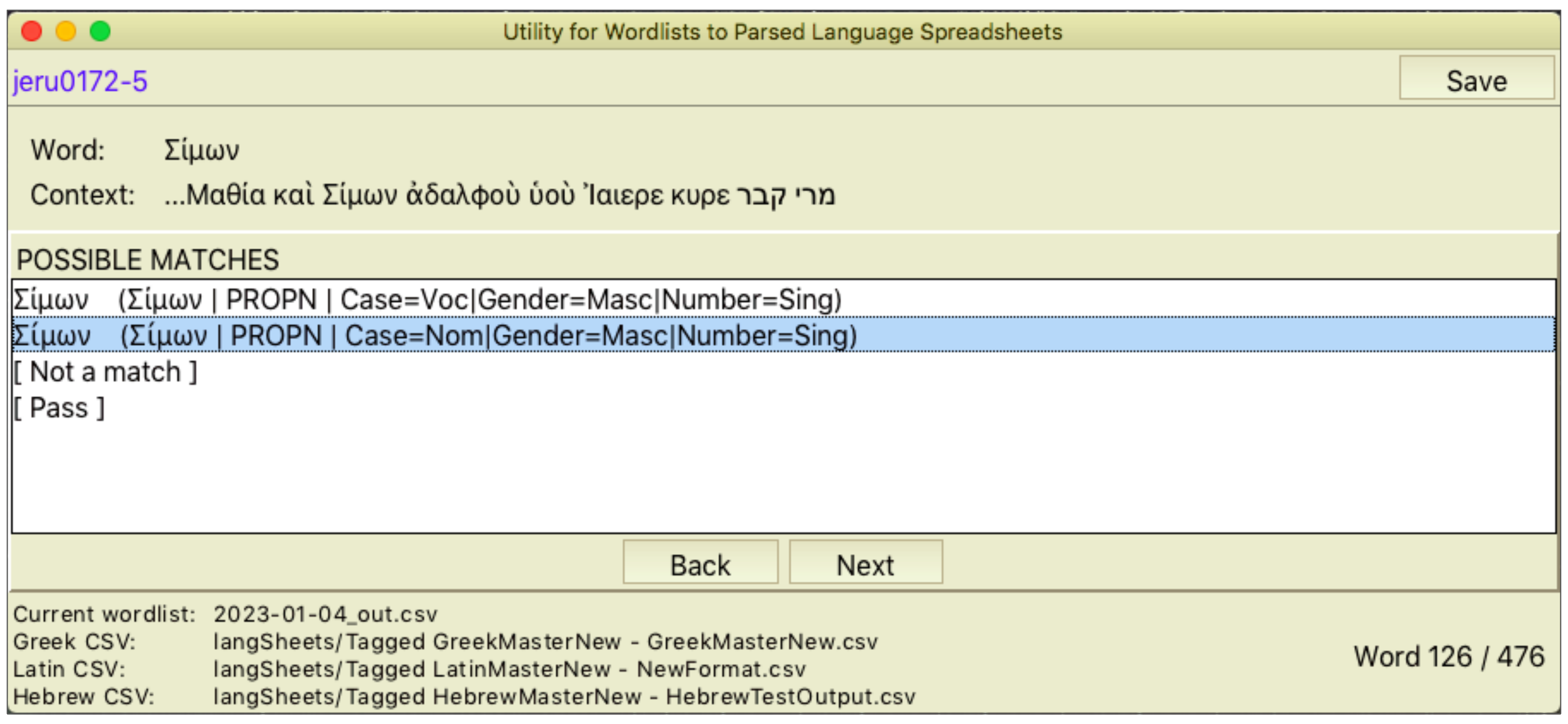
Proportions of Tagged Vocabulary



*Aramaic data is in progress.

Tagging Entities and Other Features of Interest

In addition to lexical data, the project tags words which refer to entities (such as names or locations) or other features of interest (such as military or religious terms). Automated tools can aid in recognizing named entities to some extent, but at present many of these and other features of interest are tagged manually. The screenshot above is from the Latin wordlist – the word "frumentariorum" which occurs in the inscription CAES0683 has been matched with its lemma and the relevant military rank and title. The pie charts show the range of features tagged by the project. Spelling of course varies across inscriptions and languages, so the project assigns each feature a key: this will allow researchers to search for it regardless of standardization. The IIP project has selected some features to support more complicated queries in the future, such as finding inscriptions that include a name with some specified familial relationship.



jeru0172-5

Word: Σίμων
Context: ...Μαθία καὶ Σίμων ἀδαλφοὺ ὑοὺ Ἰαιερε κυρε מרי קבר

POSSIBLE MATCHES
Σίμων (Σίμων | PROPN | Case=Voc|Gender=Masc|Number=Sing)
Σίμων (Σίμων | PROPN | Case=Nom|Gender=Masc|Number=Sing)
[Not a match]
[Pass]

Current wordlist: 2023-01-04_out.csv
Greek CSV: langSheets/Tagged GreekMasterNew - GreekMasterNew.csv
Latin CSV: langSheets/Tagged LatinMasterNew - NewFormat.csv
Hebrew CSV: langSheets/Tagged HebrewMasterNew - HebrewTestOutput.csv
Word 126 / 476

Python tkinter GUI utility to facilitate matching newly segmented words with the existing indices.

Updating and Validating the Data

This is a continually growing corpus. Present work includes the creation of tools to facilitate matching segmented words from new inscriptions to the extant parsed indices. One of these is a simple GUI utility, designed as a tool that will not require familiarity with Python or the command-line on the user's part.

The project is currently adding more robust validation processes to ensure that the segmented data and the parsed and tagged word indices accurately reflect the data from the original inscriptions' XML files. Similarly, the project is designing workflows to handle updates to the word indices when the XML files are changed (for example, if a correction needs to be made to the transcription). Since every word is given a unique identifier, they can be tracked and updated at later stages in the lexical analysis workflow.

ENCODE/Epigraphy.info VII | April 24-26, 2023