Vademecum

The aim of this corpus

Ancient Italy in the first millennium B.C.E. presents a unique trove of linguistic information. Rarely, if anywhere, in the ancient world is such a diversity of languages preserved in so small a region, and their study is of great potential interest to Indo-Europeanists, historical linguists and typologists alike. Unfortunately, however, the accessibility of this data is currently limited: most of the languages of ancient Italy are documented only through printed corpora, and these sometimes lack any kind of linguistic analysis. And where online corpora do exist, they are either incomplete, or lack the level of annotation required for anything more than relatively superficial linguistic research.

This project aims to fill this lacuna by providing a *linguistically oriented and publically available digital research corpus for the languages of Ancient Italy*. As such, it aims to provide the full epigraphic corpus of each of the languages within its purview, along with high-resolution linguistic information for every token. In its present version, the corpus is (almost) complete for the Sabellic, Messapic and Venetic languages, as well as Latin epigraphy before 100 A.D. The current scope of this corpus can therefore informally regarded as the "indigenous" Indo-European languages of Italy.

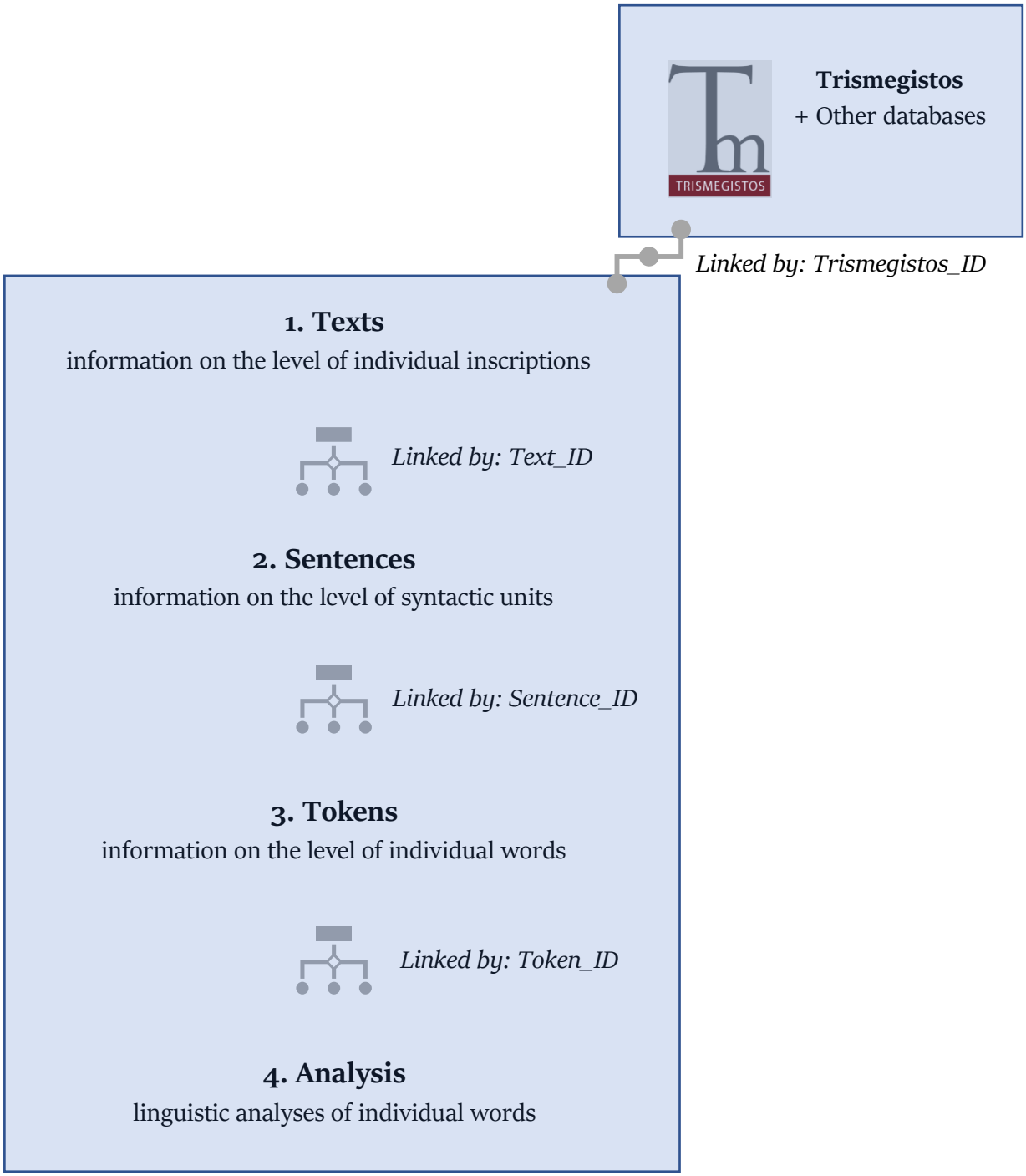
This corpus was created in the context of a research project on language contact in Ancient Italy and is therefore strongly tailored to the needs of linguistic research (whether synchronic or diachronic in nature), with a particular focus on the *intercomparability* of these ancient Indo-European languages. This means that there is less focus on the archaeological and epigraphical context of these texts. However, the Trismegistos link provided for each inscription can be used to track bibliography and metadata, and link the texts in this corpus to other projects, such as EDCS or EDR.

All files are made available here as .csv files (utf-16 encoded). The data can be analysed via Python or R, or opened with spreadsheet software such as LibreOffice or Excel. For the benefit of those users who prefer to use spreadsheet software, the final .csv file contains all the corpus data in a single file.

Please note that this corpus as currently published is a work in progress and should be used with caution. Although exhaustivity and accuracy are the ultimate aim, this has not yet been achieved. The text in red type in this vademecum attempts to inform the user of the current state of various parts of the corpus.

Structure and organisation

ACELAI is organised on four levels. Each level stands in a one-to-many relationship with the immediately subordinate level, as visualised below.



1. Texts

This table contains information pertaining to individual inscriptions as a whole. The sources for this information include Trismegistos, EDCS, as well as MLV, Studi Etruschi and other secondary literature.

1.1 Links

Text_ID	The ID of the inscription within this corpus. This number can be used to link this table to other tables (⬇ sentences.csv and trismegistos.csv).
Name	The informal name of the inscription. For instance, values of this field include *Fibula Praenestina* and *Iguvine Tables*.

1.2. Language information

Language	The main language in which the inscription is written. Values in the corpus are *Latin, Oscan, Umbrian, Old Sabellic, Messapic* and *Venetic*.
Language_Family	The higher-level classification of the main language of the inscription. For instance, Latin is classified as *Indo-European::Italic::Latino-Sabellic*. This classification is maximalist and includes speculative levels (such as placing Messapic under "Hellenic"), to allow for maximal search possibilities.
Language_Variety	This is generally equivalent to "Language", but may specify a variety on a lower taxonomic level. For instance, in line with recent research, this corpus classes Faliscan as a variety of Latin.
Script	The alphabet or script in which the inscription is written. This is relevant mostly for Sabellic, which was written in various scripts.

1.3. Chronological information

Date_after	The earliest possible date of this inscription.
Date_before	The latest possible date of this inscription.

1.4. Provenance

Provenance	The location (town or city) in which this inscription was found.

	In some cases inscriptions are known to have been written in a different location to where they were found. In those cases, provenance provides the location where the text was *written*.
Latitude	The latitude of the location provided under “provenance”.
Longitude	The longitude of the location provided under “provenance”.

1.5. Additional information

Finite_verb	*[Automatically generated field]* This field returns the value “TRUE” if the inscription in question contains at least one finite verb form, “FALSE” otherwise. The purpose of this field is to search specifically for inscriptions that constitute full sentences (as opposed to inscriptions containing only names, for instance).
Analysable_token	*[Automatically generated field]* Similar to the above, but broader, this field returns “TRUE” if the inscription contains at least one token with a linguistic analysis. This allows the user to sift out inscriptions which are wholly unintelligible or, for instance, contain only single letters.
Text_length	*[Automatically generated field]* The number of individual tokens this text links to. This is essentially a count of the number of words (and clitics) that the inscription contains.

2. Sentences

This table contains information pertaining to individual syntactic units (sentences) within a given inscription. Frequently, an inscription contains only a single syntactic unit.

2.1. Links

Text_ID	The ID of the inscription to which this sentence pertains (allowing this table to be linked to ⬆ texts.csv).
Sentence_ID	A fixed and unique number referring to this specific sentence (allowing this table to be linked to ⬇ tokens.csv).
Sentence_position	A number which keeps track of the order of sentences within an inscription. This counter starts from 1 in every new inscription and is therefore not a unique ID.

2.2. Sentence

Sentence	This field contains the full text of the sentence in question. Since the focus of this corpus is linguistic, not epigraphic, the resolution here is low. The text provided is a trimmed-down version of the text provided by published corpora (or papers in which more recent inscriptions appear). This field is meant only for convenient reference and should on absolutely no level be considered as an alternative to consulting an epigraphic apparatus.
Section	In the rare cases where a text is long enough that a subdivision into larger units than sentences, this field keeps track of them. Currently this applies only to the Iguvine Tables (where this field keeps track of tables running from I to VIIb).

3. Tokens

This table contains information on individual tokens within a syntactic unit, usually words or enclitics.

3.1. Links

Sentence_ID	The ID of the sentence to which this token pertains (allowing this table to be linked to ⬆ sentences.csv).
Token_ID	A fixed and unique number referring to this specific token (allowing this table to be linked to ⬇ analysis.csv).
Token_position	A number which keeps track of the order of tokens within a sentence. This counter starts from 1 in every new sentence and is therefore not a unique ID.

3.2. Token

Token	The token itself.
Token_clean	[*Automatically generated field*] This field is intended for ease of searching. It reproduces the token stripped of all special or non-alphabetic characters, and in transliteration where a non-Latin symbol is used.

4.5 Syntax

Syntactic information is provided on the level of Token, not Analysis, because it concerns the way in which tokens relate to each other. Becausee it is tricky to incorporate multiple possible syntactic analyses into the design of a corpus, this has been avoided altogether.
This level of information has not yet been completed but should appear (at least in part) relatively soon.

Relation	The syntactic function of the token in its context, in accordance with the conventions of the Perseus Dependency Treebanks. For instance, a token might be PRED (predicate) or SBJ (subject). Many of these tags are intuitive.
Head	The Token_ID of the word immediately

4. Analysis

This table provides morphological information. This information is provided on a level subordinate to tokens, because in an obscure or fragmentary text it is possible for a given token to have multiple plausible analyses.

4.1. Links

Token_ID	The ID of the token to which this analysis pertains (allowing this table to be linked to ⬆ tokens.csv).
Analysis_ID	A fixed and unique number referring to this analysis.

4.2. Lemmatisation

Lemma	The lemma to which the current token belongs. Note that the lemma field does not disambiguate homonyms. Thus, searching *dico* will return forms of *dicare* and *dicere* without fear or favour. This avoids the need to resort to arbitrary and confusing disambiguations such as *dico_2*. If one wishes to find only forms of *dicere*, use this field in combination with "morphological type" to specify the conjugation. The lemma is the dictionary form in Latin. For other Italic languages it is undeclined root (so *deyk* in Sabellic), to avoid having to reconstruct unattested forms.
Lemma_simplex	The simplex form of a compound lemma. For instance, the "Lemma" entry for the token *perfecerit* is *perficio*, while the "Lemma_simplex" is simply *facio*.
Lemma_frequency	[*Automatically generated field*] The number of times this particular value occurs in the "Lemma" field. As noted above, be aware that "Lemma" does not discriminate between homonyms.
Language	The language this token is analysed as belonging to. This field may differ from the "main language" of the inscription as a whole (provided in ⬆ texts.csv), as inscriptions often contain tokens in multiple languages. In addition, the same token may have different possible analyses in different languages.

4.3. Morphology

Since the aim of this corpus is to facilitate comparison between different ancient languages, it strives for a maximally descriptive (and therefore linguistically accurate) terminology in its

morphological analyses. For ease of reference, the field POS_code converts the conventions employed here into standard POS tags.
Where a dimension of morphological description is not applicable for a given token, it is annotated simply as “N.A.”.

Morphological_type	This field keeps track of inflectional declensions and conjugations, usually based on the final or characteristic sound of the stem. The terms used are descriptive, and based on the final or characteristic sound of the stem. The Latin “first declension” thus appears as “A-stem”. This terminology is less arbitrary than the traditional terms and facilitates intercomparison between the languages involved.
Part_of_speech	The part of speech of the token analysed (noun, verb, etc.).
Stem	This field keeps track of basic stem categories which host morphological inflection. These include the present and perfect stems of Latin verbs, but also, for instance, the degrees of comparison of adjectives.
Tense	Tense includes any category denoting temporal distinctions. For the Latin present stem this would include simple (e.g. “amo”), future (“amabo”) and past (“amabam”). Note that the perfect indicative is also analysed as the “simple” (unmarked) tense form, where as “past” is reserved for the pluperfect.
Mood	The mood of the token analysed. In line with some traditional grammatical analysis, it has been found convenient to group nominal forms of verbs (participle, gerundive, infinitive, supine) in this category as well: although these are not modal categories, they are in complementary distribution with modal exponents.
Diathesis	The grammatical voice or diathesis of the token analysed.
Person	The grammatical person of the token analysed.
Number	The grammatical number of the token analysed.
Gender	The nominal class of the token analysed. In the Italic languages this is usually (but not always) equivalent to the traditional category of “gender”, although note Etruscan, for instance, where an animate is frequently opposed to an inanimate class.
Case	The grammatical case of the token analysed.

Category	This is a wastebasket field for any categorial distinction not captured by the above. It assigns alphabetic indices (A, B, C…). Currently this is applicable only to the Latin imperatives “ama” and “amato”, which are labelled imperative A and imperative B in the corpus. However, the distinction between the Etruscan genitives would also be annotated in this field.
POS_code	To facilitate intercomparability, this

4.4. Semantic information

Meaning	A short English translation of the token in question.
Meaning_category	This organises a number of lemmata into very broad semantic categories. The corpus currently contains: • PROPER (for proper names e.g. *Manius, Zeus, Roman*) • RELATION (e.g. *daughter, son-in-law, mother*) • HUMAN (e.g. *slave, praetor, flute-player*) • BODY (e.g. *head, hand, liver*) • ANIMAL (e.g. *horse, sheep, lion*) • VICTUALS (e.g. *water, wine, meal*) • NUMBER (e.g. *three, twice, fourth*) • POSSESSIVE (e.g. *your, his, my*) • BUILDING (e.g. *house, statue, bridge*) • MATERIAL (e.g. *gold, silver, wood*) • COLOUR (e.g. *white, black, red*) • TIME (e.g. *day, year, hour*) • GEOGRAPHY (e.g. *sea, river, field*) • POLITICS (e.g. *vote, assembly, plebs*) • LAW (e.g. *fine, judge, punishment)* • ECONOMY (e.g. *money, debt, sell*) • RELIGION (e.g. *goddess, sacrifice, pray*)
Meaning_subcategory	This is currently used only to distinguish different categories of proper name, and contains the values PERSONAL, TOPONYM and THEONYM.
Classical_Latin_equivalent	This field contains a close cognate or semantic equivalent of the current lemma in orthographically standardised classical Latin. This field is more consistent and regularised than the “Translation” field above, and therefore facilitates intercomparison between different languages. For instance, one may wish to find all occurrences of a first person pronoun in Italic languages. One can do so simply by searching “ego” in this field.

	This field also standardises Old Latin forms to classical Latin. Thus the Old Latin verb *to, tare* “to steal”, for instance, becomes *furor* in this field. This field is still under construction, but is mostly complete except for Sabellic lemmata occuring fewer than four times.
TAM_analysis	Information on the tense, aspect and modality of verb forms. This column contains the data used for Pitts (2020) and is limited to finite verb forms in Sabellic.

Links

This table links to texts.csv and serves to link the corpus to extensive metadata and bibliographical information hosted by Trismegistos.

Text_ID	The ID of the text in this corpus (links to ⬆ texts.csv)
Trismegistos_ID	The ID of this text in the Trismegistos project.