# Feature List for the Typological Encoding of Corpus Languages

## 1. Introduction

This document describes the LitTEL dataset and the structural features it contains. It serves both as a set of guidelines for the annotation of new languages in LitTEL, and as a repository for the evidence base underlying the existing LitTEL dataset.

The LitTEL database stands in a tradition of cross-linguistic typological projects, well-known from published databases such as WALS or Grambank. As implied by its name, however, the 'Feature List for the Typological Encoding of Corpus Languages' is distinguished from this previous work by using a feature list which is specifically tailored to the strengths of ancient corpus languages. These are often 'little' languages in documentary terms, whose attestation may be quantitatively and qualitatively limited in important ways.

These 'little' languages are not always amenable to the kind of typological description that is employed by existing work. Thus, for instance, the existence of an inclusive - exclusive distinction in the first person plural, while relatively easily elicited for spoken languages, is impossible to infer for the majority of ancient corpus languages, as first person plurals tend to be absent from certain epigraphic genres. Similar issues hold for many other semantic, phonological, prosodic and structural features included in WALS and Grambank. Conversely, there are functions (such as 'indirect object') and structural phenomena (such as word order) which do play to the strengths of even the most minimally attested corpus languages, and the aim of LitTEL is to provide a dataset which can capture the full linguistic diversity of the ancient world.

There is evidence that the typological distribution of features in the ancient world differed in significant ways from the distribution of these features today. By operationalising the evidence from the 'little' languages of the ancient world. LitTEL will make it possible to map these distributions, as well as their changes through time, at a higher resolution than ever before.

This document will first discuss some general organisational and annotational principles (section 2) before giving an overview of the features (section 3) and the languages (section 4) in the LitTEL database.

# 2. The Annotation of LitTEL Features

## *2.1. Comparative concepts*

A typological database must accurately describe not one, but many different languages. In order to do so, it must make use of comparative concepts, 'that is, concepts specifically designed for the purpose of comparison that are independent of [language-specific] descriptive categories' (Haspelmath 2010). In other words, the basis of comparison in the LitTEL features will always appeal to general conceptual and formal notions, rather than assuming that terms like 'adjective' or 'object' have cross-linguistically consistent instantiations.

This means, for instance, that when the term 'direct object' is used in this document, it refers to 'the more patient-like argument of a ditransitive verb', a comparative concept which can be assessed in any language. In many Indo-European languages, these arguments are coded in an 'accusative' case: however, that is not in itself a criterion that can be cross-linguistically operationalised.

## *2.2. Tailoring Features to Little Corpora*

LitTEL aims to ask typological questions of little corpora. Doing so without bias requires a fine-tuned methodology.

Suppose one were to annotate Venetic, for instance, for the WALS features. Many features (like the inclusive first person pronoun mentioned in section 1) would plainly be impossible to annotate due to a lack of data. This is, in itself, not a major problem: we can still theoretically compare the 'little' Venetic dataset with a 'big' language like Latin by eliminating the features for which Venetic attests no data.

For other features, however - the number of nominal cases, for instance - the annotator could certainly enter a value. Five are in evidence, but who is to say that Venetic did not possess more cases that are simply not attested? This problem is more pernicious, as it means the annotation of Venetic will never be properly comparable to the annotation of, say, Latin (where we have enough data to be sure that no cases are missing). At what point can a researcher conclude s/he has enough data to decide that s/he has a complete inventory of cases, or that case marking is fully obligatory, or that the position of case markers is not syntactically free?

Such decisions are always arbitrary, and will remain an intractable source of observation bias in a dataset comprising both little and big corpus languages. Unlike the WALS features, the LitTEL features are defined to avoid this issue of observation bias as much as possible.

Thus, LitTEL avoids asking 'are there any...' questions. Likewise, LitTEL features do not make reference to statistical concepts such as 'obligatoriness', nor do they presuppose the attestation of variation, or concepts which require variation to establish (such as degrees of syntactic freedom). Instead, LitTEL features first specify a construction (or multiple constructions), and then - if the construction is attested - they define two possible, mutually exclusive, properties of this construction. A construction is either attested or it is not. Thus, little languages may be more likely to contain NA values (see section 2.5), where relevant

constructions are simply not attested, but they are not more likely to contain a specific linguistically relevant value (0 or 1).

These desiderata are schematised below with examples.

*Table 2. LitTEL Feature Design*

| **Potential LitTEL feature** | **Status** | **Comment** |
|---|---|---|
| How many cases are attested? | ✘ | The answer is likely to be correct for well attested languages, while figures for poorly attested language may be incomplete in unknowable ways. |
| Is there an accusative case? | ✘ | In a poorly attested language, an accusative case may exist but simply not be attested, which is much less likely to be the case for a well attested language. |
| Is accusative marking obligatory? | ✘ | 'Obligatory' is a statistical concept which requires large datasets to reasonably establish, and may give misleading results when applied to small corpora. |
| Can the accusative marker be separated from its head noun? | ✘ | Smaller corpora will attest less variation and are thus less likely to provide evidence for relevant syntactic variation than larger corpora. |
| In overt accusative constructions with an attribute, does that attribute agree? | ✓ | If at least one relevant construction is attested, it either agrees or it does not. This is true for both poorly attested and well attested languages. |

Note that in some cases, features appeal to concepts such as lexical open-endedness or productivity. Although strictly speaking these concepts contradict the above desiderata, it is assumed that a linguist will have a sense of where these criteria apply even on limited evidence. For instance, a linguist will be aware that a locative attested only for placenames does not constitute good evidence for the existence of a locative case which could apply productively to the entire lexicon. Such a common-sense judgement can arguably be made based even on only a single construction.

### *2.3. Binary feature values*

All features in the LitTEL feature set are binary. This means that they have two possible answers, coded as 0 (zero) and 1 (one). For instance, the first LitTEL feature (section 3.1) describes the word order of coordinative morphs, and has the following two values:

*Table 1. Coordinative Constructions*

| | |
|---|---|
| **0** | The coordinator is placed medially between the two coordinands (template A *coord* B). |
| **1** | The coordinator is placed after the second coordinand (template A B *coord*). |

The previous section drew attention to the issue of observation bias. LitTEL addresses this problem by avoiding the use of yes-no questions to obtain binary feature values, and instead describing both values in their own terms. Of course, this means that values must be formulated in such a way as to exclude logically possible third options in a typologically informed manner.

For instance, a logically possible third alternative for coordination - in addition to the templates A *coord* B (zero) and A B *coord* (one) shown above - is *coord* A B, with the coordinative morpheme preposed before the first coordinand. However, there do not seem to be any natural human languages which instantiate this construction. Consequently, existing typological knowledge informs the binary value set.

Existing typological knowledge also informs the coding of the two binary values. The value instantiated by 0 (zero) is the value which, based on currently available information, is more common in the languages spoken today. For the example given above, medial coordination appears to be significantly more common in the languages of the world than postpositional coordination (Stassen 2003): consequently, medial coordination is 0 (zero) while postpositional coordination is 1 (one).

### *2.4. Annotating competing constructions*

Human language is characterised by variation. Consequently, it is likely that in many cases languages - particularly 'big' corpus languages - will attest constructions answering to both values (0 and 1) of a given feature. In such cases, a decision is made using the following criteria:

1. Is there evidence that one construction is more frequent than the other? If so, this value takes precedence when annotating the language in question. In the best case, this quantitative criterion is invoked based on corpus data.
2. Is there evidence that one construction is more pragmatically neutral than the other? If so, this value takes precedence when annotating the language in question. For instance, some languages use word order productively to change the salience of particular constituents: in such cases, the most 'neutral' word order is the most relevant for cross-linguistic comparison.
3. Is there evidence that one construction is syntactically or semantically more free than the other? If so, this value takes precedence when annotating the language in question. A construction which only coordinates semantically related coordinands, for instance (such as 'sun and moon', 'father and mother') is dispreferred over a construction which can coordinate any constituent, regardless of its semantic properties.

In many cases, definitions will be formulated in such a way as to help the annotator discriminate between competing constructions. In particular, many definitions make reference to open-ended lexical classes. This helps to clarify that pronouns, for instance, which often have marginal constructions of their own, are less relevant to annotating values than lexical nouns.

LitTEL deliberately does not employ 'both' as an annotational value. This is because the extent of attested variation is strongly dependent on the degree and type of a language's documentation and would consequently risk introducing observation bias to the dataset. Where two constructions are attested and no

choice can be made through any criterion, the annotation is NA ('not applicable'), as described in the following section.

### *2.5. Annotating values as NA*

In addition to 0 (zero) and 1 (one), a third value NA ('not applicable') is available for annotational purposes. These are values which any operations in the LitTEL software will ignore. Consequently, this is a value which essentially eliminates the language in question from further consideration and should be used sparingly. NA fields essentially exist to keep track of datapoints an annotator could not find a satisfactory value for.

NA may be used in the following cases only:

- ➔ When the attestation of a language is insufficient to establish the correct value of a feature. Although LitTEL aims to be maximally suited for the description of poorly attested languages, it will still occasionally be the case that simply no relevant constructions are attested. Note that, all other things being equal, a single attested construction is considered sufficient to enter a value.
- ➔ Where the language employs a device that differs from, or is incompatible with, both alternatives given. In general, the feature values should be described in a way that precludes this possibility, but exceptions are conceivable. For instance, a language may use only zero-marked coordination, in which case the position of the coordinative morph is undefined. In such a case, NA is the only meaningful value.
- ➔ Where both alternatives exist but are either so rare or so well-balanced that the criteria in the previous section (2.3) cannot distinguish between them.

# 3. The LitTEL Features

## *3.1. Overview*

Unproblematic features which are well-defined and can be annotated for all or most languages:

- ➔ **Pos_Coord**: The position of the coordinative morph (our poster boy feature)
- ➔ **Ord_Obj**: Common and typically well-described word order features
- ➔ **Ord_Gen**: Common and typically well-described word order features
- ➔ **Ord_Adj**: Common and typically well-described word order features
- ➔ **Rank_Temp**: Balanced or deranked verb in temporal subordinate clauses
- ➔ **Aff_Nom**: Position of nominal affixes
- ➔ **Aff_Vrb**: Position of verbal affixes
- ➔ **Agr_Gend**: Gender agreement on attributes
- ➔ **Exp_Case**: Case-number co-exponence

Maybe problematic because typologically significant third options exist (e.g. no marking):

- ➔ **Loc_Obj**: Locus of object marking (head or dependent)?
- ➔ **Loc_Gen**: Locus of genitive marking (head or dependent)?
- ➔ **Align_Nom**: Does nominal alignment follow the accusative or ergative pattern?

Features that perhaps should be added or split up into multiple features:

- ➔ **Agr_Acc**: A split-up of case agreement into individual cases
- ➔ **Agr_Dat**: A split-up of case agreement into individual cases
- ➔ **Agr_Loc**: A split-up of case agreement into individual cases
- ➔ **Rel_Clause**: Relativisation strategies (relative pronoun or other)
- ➔ **Prohib_Clause**: Do prohibitives differ morphologically from imperatives?
- ➔ **Overt_Acc**: Is the direct object overtly marked through any morph (on either head or dependent)?

Features that should probably be removed:

- ➔ **Synth_Acc**: Syntactic freedom is a poor criterion for little languages
- ➔ **Synth_Dat**: Syntactic freedom is a poor criterion for little languages
- ➔ **Synth_Loc**: Syntactic freedom is a poor criterion for little languages

Ideas that could be threshed out further:

- ➔ Something about flexivity?
- ➔ Something about ablaut?
- ➔ Something about reduplication?

### *3.2. Coordinative Morph Position (Pos_Coord)*

**The relevant construction**: This feature looks for constructions involving neutral coordinative relations. In such constructions, a morph, referred to as the 'coordinator', specifies a 'horizontal' syntactic relationship between two or more constituents (referred to as the 'coordinands'). This relationship is semantically neutral (English 'and'), and does not disjoin ('or') or contrast ('but') the coordinands.

It is sometimes difficult to be sure that a construction is a true instance of neutral coordination. Commonly, the morphs in question will have polysemies involving adverbial functions such as 'also' or 'with'. The principal criterion for distinguishing a true coordinative construction is whether the construction functions as a single constituent (e.g. in terms of triggering plural agreement on its verb).

If attestation allows, this feature should be annotated on the basis of bisyndetic constructions, involving exactly two nominal coordinands without attributes (such as 'mother and father').

**The feature values**: The values of this feature query the ordering of the coordinator with regard to the coordinands, specifically whether it precedes or follows the second coordinand. As noted previously (section 2.2), there do not appear to be any languages which employ the logically conceivable third option, where the coordinator precedes both coordinands (*coord* A B).

*Feature Values for Pos_Coord*

| | |
|---|---|
| **0** | The coordinator is placed medially between the two coordinands (template A *coord* B). |
| **1** | The coordinator is placed after the second coordinand (template A B *coord*). |

**Example 1:** English (value = 0)

In English, the construction 'A and B' is the most common way to express neutral coordination, and the morph is always placed medially. That is, 'mother and father', never †'mother father and'.

**Example 2**: Latin (value = 1 evolving towards 0)

Classical Latin has both medial ('mater et pater') and postposed ('mater paterque') coordinative morphs. However, the medial morph is somewhat more common and has more syntactic and semantic freedom. Increasingly, there is a tendency to restrict '-que' to coordinands which are semantically closely related (as in 'terra marique', for instance). Consequently, Classical Latin (from about -200) is annotated as 0.

The earliest epigraphic Latin, conversely, preferred '-que' in all contexts and is annotated as 1 (up to about -300). It is difficult to establish the dominant strategy in the intervening period (the third century BCE).

**Example 3:** Sumerian (value = 1)

Sumerian uses a comitative marker ('with') as a coordinator and postposes it after the second coordinand. Since this triggers plural agreement on verbs, it should be analysed as a relevant coordinative construction.

### *3.3. Word Order Features (Ord_Obj, Ord_Gen, Ord_Adj)*

**The relevant constructions**: This group of related features looks at the ordering of constituents relative to their syntactic head. It focuses on three types of constituents, referred to as 'direct object' (2-DOO), 'genitive attribute' (2-GAO) and 'adjectival attribute' (2-AAO).

These constructions are defined as follows:

- ➔ 'Direct objects' are identified in constructions involving bivalent verbs, that is, verbs which take two arguments. In prototypical transitive constructions, one of these arguments is more 'agent-like' while the other is more 'patient-like'. This more 'patient-like' argument will be referred to as the 'direct object'.
- ➔ 'Genitives' are identified in constructions where a noun stands in an attributive relationship with another noun phrase. Prototypically, 'possessive' relations (A of B) are considered prototypical genitive constructions, but in many languages genitives cover a wide range of different semantic relationships. These will tend to be translatable by English 'of'.
- ➔ 'Adjectives' are identified in constructions in which an attribute describes a quality or property of the head noun it is syntactically dependent on. In many - but by no means all - languages such attributes form a distinct word class of their own (adjectives). However, in line with the criteria offered by WALS, the concept of 'adjective' is here defined semantically, not formally, so the existence of such a word class is not relevant to the identification of relevant constructions for this feature. Prototypical examples of relevant constructions in English would be 'a good man', 'a big house' or 'a tasty meal'.

**The feature values**: For all three of the constructions defined above, the values of the features in question query the position of the subordinate constituent with regard to the constituent which governs it.

In many languages, word order can be relatively free and subject to much variation. Often, this variation is pragmatically marked. Consequently, close attention should be paid to the criteria in section 2.3 when annotating languages for these features.

Note that for the direct object and genitives, the 0 (zero) value is preceding their heads, while for adjectives the 0 (zero) value is following their heads. This is due to the typological frequency of the different word orders and should obviously be taken into account while annotating.

*Feature Values for Ord_Obj*

| | |
|---|---|
| **0** | The direct object tends to precede its predicate (template OV). |
| **1** | The direct object tends to follow its predicate (template VO). |

*Feature Values for Ord_Gen*

| | |
|---|---|
| **0** | The genitive attribute tends to precede its head (template GN). |
| **1** | The genitive attribute tends to follow its head (template NG). |

*Feature Values for Ord_Adj*

| | |
|---|---|
| **0** | The adjectival attribute tends to follow its head (template NA). |
| **1** | The adjectival attribute tends to precede its head (template AN). |

## *3.?. Degree of Case Relation Synthesis (Synth_Acc, Synth_Dat, Synth_Loc)* **RETIRE?**

**The relevant constructions**: This feature looks for constructions involving case relations. Case relations are defined as semantic and syntactic relations obtaining between a noun and its head. Such relations are expressed in many languages through adpositional elements or morphological case markers.

- ➔ 'Accusatives' are case relators expressing the 'direct object', or the more 'patient-like' argument, in a prototypical transitive sentence (as defined as in the previous section).
- ➔ 'Datives' are case relators expressing the more 'beneficiary-like' argument in prototypical ditransitive sentences.
- ➔ 'Locatives' are case relators which represent a general productive device for expressing static spatial location. These will tend to correspond to English 'in'. Where no spatial constructions are attested, static location in time is considered an acceptable alternative.

**The feature values**: The values of these features query the degree of synthesis of the morph expressing the case relation in question. Since the notion of 'word' is not cross-linguistically defined, synthesis is, in reality, a cline rather than a dichotomy. The value definitions below represent this idea using Haspelmath's (2018) compound term 'anasynthesis', with analytic and synthetic representing two ends of an 'anasynthetic cline'.

A case relator will be considered more on the 'synthetic' end of the anasynthetic cline if it is obligatorily adjacent to the syntactic head of the noun phrase it qualifies. Conversely, if the case relator can be placed next to an attribute instead, or if it can be separated from the noun phrase by other constituents, it is considered to fall on the 'analytic' side of the anasynthetic cline.

This criterion is purely formal, as in ancient corpus language prosody is typically not available as a diagnostic (although there are exceptions). At the same time, it is important not to conflate synthesis with affix ordering. In many languages preposed morphs are less grammaticalised than suffixed morphs, but this is not relevant for the annotation of this feature (see example 3 below). Consequently, this analysis must focus solely on the degree of syntactic freedom of the case relator.

Of course, 'syntactic freedom' is itself a complicated concept in poorly attested corpora, as it presupposes the existence of variation or sufficient attestation to be able to exclude it. For this reason, the following additional criterion applies: if the case relator stands in a clear paradigmatic relationship with other case relators, their properties may be generalised across the board. For instance, the Latin locative '-i' may be relatively rare in some subcorpora of the language, but it stands in a strong paradigmatic relationship with other suffixal case endings with clearly synthetic properties.

*Possible Feature Values for Synth_Acc*

| | |
|---|---|
| **0** | The accusative case relation is either zero-marked, or stands on the analytic side of the anasynthetic cline. |
| **1** | The accusative case relation stands on the synthetic side of the anasynthetic cline. |

*Possible Feature Values for Synth_Dat*

| | |
|---|---|
| **0** | The dative case relation is either zero-marked, or stands on the analytic side of the anasynthetic cline. |
| **1** | The dative case relation stands on the synthetic side of the anasynthetic cline. |

*Possible Feature Values for Synth_Loc*

| | |
|---|---|
| **0** | The locative case relation is either zero-marked, or stands on the analytic side of the anasynthetic cline. |
| **1** | The locative case relation stands on the synthetic side of the anasynthetic cline. |

**Example 1**: English (value = 0, 0, 0)

English zero-marks the accusative case relation and expresses others using adpositions such as ‘to’ or ‘in’. In all cases, elements may intervene between the adposition and the head noun. For instance, in the phrase ‘in the house’, a determiner ‘the’ intervenes between the case relator ‘in’ and the head of the noun phrase ‘house’.

**Example 2**: Latin (value = 0, 0, 1)

Latin uses integrated morphological markers for the accusative and dative functions, which are obligatorily adjacent to the head noun (as well as being marked redundantly on any inflecting attributes).

Latin has two strategies for expressing the locative case relation. The older strategy is comparable to the other morphological cases and is clearly synthetic, but is restricted in all documented periods of Latin to a small and unproductive set of nouns (e.g. ‘domi’, ‘humi’, ‘ruri’ as well as some place names). Consequently, this is analysed as a marginal strategy. The more productive construction in Latin is the use of a preposition ‘in’, which is less tightly integrated (‘in magna urbe’, with an intervening attribute).

**Example 3**: Messapic (value = 0, 0, 1)

Messapic is comparable to Latin in its values and argumentation, but serves as an instructive example of how these features should be annotated for poorly attested languages.

### *3.4. Ranking of Subordinate Clauses (Rank_Temp)*

**The relevant construction**: This feature looks for adverbial clauses which specify that an event occurred in some temporal relationship with the event of the main clause. Often, these will correspond to English subordinate clauses introduced by 'when' or 'while'. The semantic nature of the temporal relationship can vary significantly and languages are likely to have various competing devices (for which see section 2.3).

For a construction to be relevant, it must be used to indicate such a temporal relationship in a productive and open-ended manner. Essentially, it should be capable of expressing temporal clauses corresponding to any possible simple clause: for any *sentence*, the construction should be able to express 'when or while *sentence*, something else happened'.

In many languages, however, such a construction may still involve a nominalised or otherwise non-finite verb form. In other words, the presence of a finite verb form is not an operative criterion.

**The feature values**: This feature queries the verb form used in temporal clauses. Following the terminology used by WALS, the verb in question may be 'balanced' or 'deranked'.

If the verb is balanced, then the same verb form used in the temporal clause can also stand in a main clause, without any alteration to its form or the loss or addition of any markers. In other words, it should be possible to convert the temporal clause into a main clause describing the same event without changing the verb and its accessories (see examples below).

Conversely, if the verb is deranked, then it

> "may lack some or all of the categorial distinctions relevant to verbs in the language (such as tense, aspect, mood or person agreement distinctions), or display special markers not used in independent clauses, e.g. special tense, aspect, mood or person markers, nominalizers, case markers or adpositions." (Cristofaro 2013)

*Possible Feature Values for Rank_Temp*

| | |
|---|---|
| **0** | The verb form in temporal clauses is balanced. |
| **1** | The verb form in temporal clauses is deranked. |

**Example 1**: English (value = 0)

English has both balanced and deranked constructions, and the criteria in section 2.3 need to be applied.

For instance, in 'when we arrived at our destination, we got off the train', the subordinate clause can be converted into a main clause ('we arrived at our destination') without any change to the verb.

Conversely, in 'having arrived at our destination, we got off the train', this is not possible: a main clause would need to be 'we arrived at our destination' or something along those lines.

The previous construction is more common in English?

**Example 2**: Latin (value = 1)

Latin, too, has balanced as well as deranked constructions.

### *3.5. Affix Ordering (Aff_Nom, Aff_Vrb)*

**The relevant constructions**: These features look for constructions involving nominal and verbal affixes, respectively. Although all affixes are relevant, it preferentially focuses on those productive affixes which are maximally synthetic, where it is possible to make such a distinction within the attested corpora.

**The feature values**: These features essentially relate to the ordering of affixes, examining whether they tend to be prefixed or suffixed. This involves taking an inventory of the most grammaticalised affixal paradigms in the language and establishing whether more or less than 50% of the morphs in question are suffixed.

*Possible Feature Values for Aff_Nom*

| | |
|---|---|
| **0** | Most nominal affixes are suffixed. |
| **1** | Most nominal affixes are prefixed. |

*Possible Feature Values for Aff_Vrb*

| | |
|---|---|
| **0** | Most verbal affixes are suffixed. |
| **1** | Most verbal affixes are prefixed. |

**Example 1**: English (value = 0, 0)

**Example 2**: Latin (value = 0, 0)

**Example 3**: Hebrew (value = 0, 0)

### *3.6: Locus of Marking (Loc_Obj, Loc_Gen)*

**The relevant construction**: This feature looks for constructions involving 1) a direct object and 2) a genitive, respectively (defined as for previous features).

**The feature values**: If a language has any overt marking for the syntactic relationship involved in the construction (direct object and genitive, respectively), this marking may be on the syntactic head or on the syntactic dependent, or both. The feature values query this variation in terms of head-marking.

That is, if the construction has any kind of head-marking (with or without additional dependent-marking) it is annotated as 'head-marking', while if it does not it is annotated as 'other'.

Be aware that, for reasons of the differing typological frequencies of the two constructions, these options correspond to different binary feature values (as shown below).

*Possible Feature Values for Loc_Obj*

| | |
|---|---|
| **0** | The direct object is either head-marked, or both head-marked and dependent-marked |
| **1** | The direct object is either only dependent-marked or not marked at all |

*Possible Feature Values for Loc_Gen*

| | |
|---|---|
| **0** | The genitive attribute is either only dependent-marked or not marked at all |
| **1** | The genitive attribute is either head-marked, or both head-marked and dependent-marked |

### *3.7: Nominal Agreement (Agr_Case, Agr_Gend)*

**The relevant construction**: This feature focuses on marking in the nominal system along the structural dimensions of 1) case relations and 2) gender or noun classification.

Case relations are defined as semantic and syntactic relations obtaining between a noun and its head. Such relations are expressed in many languages through adpositional elements or morphological case markers. The following relations are relevant here:

- ➔ 'Accusatives' are case relators expressing the 'direct object', or the more 'patient-like' argument, in a prototypical transitive sentence (as defined as in the previous section).
- ➔ 'Datives' are case relators expressing the more 'beneficiary-like' argument in prototypical ditransitive sentences.
- ➔ 'Locatives' are case relators which represent a general productive device for expressing static spatial location. These will tend to correspond to English 'in'. Where no spatial constructions are attested, static location in time is considered an acceptable alternative.

**The feature values**: If an overt case relation morph is redundantly expressed on an open-ended class of attributes as well as on their head nouns, the construction in question is considered an instance of case agreement. Conversely, if no case relation morph is redundantly expressed on attributes, then case agreement is absent.

Comparably, if the form of an open-ended class of attributes changes depending on the gender (or noun classification) of its head noun, then the language in question is characterised by gender agreement.

As always, marginal constructions (for instance, involving only determiners) are precluded by the specification that agreement should involve an open-ended lexical class.

In some languages where agreement exists it is obligatory, in others it is optional. To be annotated as 1 (one), agreement does not have to be absolutely obligatory, but it should at least be the preferred option (in terms of the criteria in section 2.3) with the class of attributes for which agreement is attested.

*Possible Feature Values for Agr_Acc*

| | |
|---|---|
| **0** | No open-ended class of attributes expresses agreement for the accusative case relation. |
| **1** | An open-ended class of attributes typically or obligatorily expresses agreement for the accusative case relation. |

*Possible Feature Values for Agr_Dat*

| | |
|---|---|
| **0** | No open-ended class of attributes expresses agreement for the dative case relation. |
| **1** | An open-ended class of attributes typically or obligatorily expresses agreement for the dative case relation. |

*Possible Feature Values for Agr_Loc*

| | |
|---|---|
| **0** | No open-ended class of attributes expresses agreement for the locative case relation. |
| **1** | An open-ended class of attributes typically or obligatorily expresses agreement for the locative case relation. |

*Possible Feature Values for Agr_Gend*

| | |
|---|---|
| **0** | No open-ended class of attributes expresses agreement for gender. |
| **1** | An open-ended class of attributes typically or obligatorily expresses agreement for gender. |

**Example 1**: English (value = 0)

Case relations in English are mostly expressed through prepositions, which are not repeated with attributes. Thus, for instance, 'in water' can take an attribute and become 'in deep water', but never †'in deep in water'. Comparably, the genitive marker '-s' has scope over the entire noun phrase and is consequently not repeated for attributes ('the tall man's son' but never †'the tall's man's son').

English is thus a clear non-agreeing language: it has no case relations which show even optional agreement.

**Example 2**: Latin (value = 1)

In Latin, agreement is extremely strict for morphological case on inflected attributes. Only a marginal (and closed) set of indeclinable adjectives (e.g. 'frugi') fails to mark such agreement. In this regard, Latin stands on the opposite end of the scale to English.

Note that, in addition to these agreeing morphological cases, Latin also uses non-agreeing prepositions to express case relations. This has no relevance for the annotation.

**Example 3**: Tocharian (value = 1)

Tocharian has inherited cases which mark agreement obligatorily, as well as postpositional case relators where agreement is optional but not obligatory. Since agreement is preferred or obligatory for at least a subset of case relations, Tocharian is annotated as agreeing.

### *3.8. Overt Expression of Case (Overt_Acc)*

**The relevant construction**:

**The feature values**: Is the accusative case relation signalled through an overt morph, either on the direct object itself or on its verbal head, rather than zero-marked?

*Possible Feature Values for Overt_Acc*

| | |
|---|---|
| **0** | The accusative is not overtly marked |
| **1** | The accusative case relation is overtly marked. |

### *3.9. Coexponence of Case and Number (Coex_Numb)*

**The relevant constructions**: This construction looks for a case relation morph which is attested in both the singular and the plural. Most languages have grammatical morphs for many different case relations. A single case relation should be selected to evaluate this feature.

The morph chosen to evaluate this feature should be selected in line with the following criteria:

- ➔ By preference, select the most 'core' or 'grammatical' case relation available: so, for instance, prefer the accusative over the dative, and the dative over the locative.
- ➔ Select only a case relation that is expressed overtly in both its singular and its plural slot. In other words, do not select zero-marked cases.
- ➔ By preference, select forms from the noun class or gender which is more morphologically basic in the language in question. If this cannot be established from the language itself, make reference to its closest genealogical relative where it can be established. If no such relative is available, choose at random (but remark on this in the comment field).
- ➔ By preference, select forms from the most common inflectional class.

If no two parallel (singular and plural) forms are attested for any case relation in the language, the annotation should be NA.

**The feature values**: This feature queries whether the same marker expresses both case and number. If it is possible to distinguish a morph referring to the case relation from a morph referring to the number relation, annotate 0. If this is not possible, annotate 1.

*Possible Feature Values for Coex_Numb*

| | |
|---|---|
| **0** | Case and number are not coexpressed by the selected morph |
| **1** | Case and number are coexpressed by the selected morph |

**Example 1**: English (value = 0)

English zero-marks the nominative and accusative relations, so the dative was selected to evaluate this feature. The dative can be overtly marked by a preposition, which is invariant in the singular and plural and distinct from the non-contiguous plural marker ('to the house', 'to the houses').

**Example 2**: Latin (value = 1)

The Latin nominative has different forms in the singular and the plural, in all genders and inflectional classes. No distinct plural marker can be distinguished (cf. -us versus -i, -a versus -ae, -s versus -es).

**Example 3**: Proto-Semitic (value = 0)

### *3.10. Nominal Marking Alignment (Align_Nom)*

**The relevant constructions**: This feature rests on the comparison of a transitive construction (involving a verb with exactly two arguments) and an intransitive construction (involving a verb with a single argument). In these constructions:

- ➔ The argument of the intransitive construction is referred to as S.
- ➔ The more 'agent-like' argument of the transitive construction is referred to as A.
- ➔ The more 'patient-like' argument of the transitive construction is referred to as P.

**The feature values**: This feature queries the alignments in the morphological marking of the S, A and P arguments. Only dependent-marking is relevant for the purposes of this feature.

- ➔ If S and A are marked in the same way, but P is marked differently, we have accusative alignment.
- ➔ If S and P are marked in the same way, but A is marked differently, we have ergative alignment.
- ➔ The third logical possibility, where A and P are aligned against S, does not appear to occur.
- ➔ If S, A and P are all marked in the same way, or not marked at all, we have neutral alignment. Since this feature is annotated in terms of the presence or absence of ergativity, neutral alignment is grouped with accusative alignment.

In many ergative languages, ergativity is restricted to particular groups of referents (e.g. inanimates) or to particular verbal categories (e.g. aspects). By preference, therefore, this feature focuses on animate referents, and where evidence otherwise exists for multiple alignments the criteria in section 2.4 should be followed.

*Possible Feature Values for Align_Nom*

| | |
|---|---|
| **0** | Dependent-marked arguments are typically characterised by accusative or neutral alignment. |
| **1** | Dependent-marked arguments are typically characterised by ergative alignment. |

### *3.11. Imperatives and Prohibitions (Prohib_Verb)*

**The relevant constructions**: This feature rests on the comparison of two constructions:

- ➔ The first, henceforth the 'imperative' construction, is a construction in which a party other than the speaker is instructed or requested to perform the described action.
- ➔ The second, henceforth the 'prohibitive' construction, is a construction in which a party other than the speaker is forbidden from performing the described action ('preventative'), or is enjoined to desist from performing the described action if it is already ongoing. It serves as the negative equivalent of the 'imperative' construction.

By preference, select the second person singular if attested in both constructions, and select the 'preventative' variant if multiple semantic nuances of the prohibitive exist. Plural or third person commands are also acceptable. If no two comparable constructions are attested, annotate NA.

**The feature values**: This feature queries the distinction between imperative and prohibitive constructions. Although it makes sense that prohibitives will tend to employ some marker of negation, many languages additionally distinguish the two constructions through further overt means, for instance by their verbal morphology (e.g. employing a different mood).

*Possible Feature Values for Prohib_Verb*

| | |
|---|---|
| **0** | Prohibitives are distinguished from imperatives only by negative morphology. |
| **1** | Prohibitives are distinguished from imperatives by some overt means beyond or in addition to negative morphology. |

**Example 1**: English (value = 0)

In English, 'jump!' and 'do not jump!' use the same imperative form, distinguished only by negative morphology. Note that the auxiliary 'do' is required by the syntax of the negative particle and does not qualify as special prohibitive morphology.

**Example 2**: Latin (value = 1)

In Latin, the usual way of expressing prohibitives is 'ne' followed by the subjunctive, contrasting with the use of the imperative mood in the simple imperative.

### *3.12. Relative Clauses (Rel_Clause)*

**The relevant constructions**: A relative clause is a subordinate clause which serves as an attribute to a head noun. As per the definition given by WALS, "a relative clause is a clause narrowing the potential reference of a referring expression by restricting the reference to those referents of which a particular proposition is true".

Note that these constructions do not have to involve finite verbal morphology (a concept which cannot be defined cross-linguistically). This means that, in many languages, participial attributes with arguments can count as relative clauses. They do, however, need to be open-ended and productive.

**The feature values**: Relativisation strategies can be distinguished along several dimensions. These features query three of these dimensions.

- ➔ Is the head itself inside the relative clause? That is, does it syntactically or positionally belong to the relative clause rather than to the main clause?
- ➔ Is the head referenced in the relative clause at all? That is, does some form of pronominal or non-pronominal element represent the head's function in the relative clause, or is there simply a gap?

The examples below may help to clarify these questions.

*Possible Feature Values for Rel_Head*

| | |
|---|---|
| **0** | The head of the relative clause is not inside the relative clause as such, although it may be represented there by a (relative) pronoun or particle. |
| **1** | The head of a relative clause is inside the relative clause or belongs to it syntactically or positionally. |

*Possible Feature Values for Rel_Gap*

| | |
|---|---|
| **0** | The head of a relative clause is not represented in the relative clause. |
| **1** | The head of a relative clause is either inside the relative clause, or is represented there by a relative pronoun, a non-relative pronoun or a particle. |

**Example 1**: English (value = 0, 1)

English relative clauses involve a pronoun 'who', 'which' or 'that' which represents the head in the clause. For instance, 'I saw the man who jumped' contains a relative clause 'who jumped'. The head of the clause is outside the clause (serving as the direct object of 'I saw') but it is represented inside the clause by a pronoun 'who'.

**Example 2**: Latin (value = 0, 1)

Classical Latin relative clauses work in much the same way as their English equivalents.

In archaic and particularly legal Latin, an alternative correlative construction is attested: ‘qui homo… is…’. Here, the head is inside the relative clause, and the main clause contains a resumptive ‘is’ (‘he’). If this were the dominant strategy in Latin, it would be annotated as 1 for both features.

**Other possible features:**

Flex_Cats: some parameter of flexivity? (evidence for more than one form determined lexically)

Presence of features such as: ablaut, reduplication

# 4. The LitTEL Dataset

## *4.1. LitTEL as a CLLD StructureDataset*

LitTEL is structured as a CLLD StructureDataset. It consists of the following principal files:

- ➔ **languages.csv** contains information about individual languages.
- ➔ **values.csv** contains the actual datapoints, with references.
- ➔ **parameters.csv** contains a list of the features and their IDs.
- ➔ **codes.csv** contains a list of the possible (binary) feature values.

## *4.2. Values*

The actual data which forms the LitTEL database is stored in a ValueTable, which should, as per the CLDF guidelines, contain at least the following columns: ID, Language_ID, Feature_ID and Value.

The values can be consulted in this Google spreadsheet. The spreadsheet contains the following columns:

- ➔ **ID**: A numeric ID which keeps track of each of the datapoints.
- ➔ **Feature_ID**: The ID of the relevant feature.
- ➔ **Language_ID**: The Glottolog ID of the language in question, allowing the data to be linked to the languages.csv file, described in the previous section.
- ➔ **Value**: The value of the specified feature in the specified language at the specified date. Must in all cases be one of 0 (zero), 1 (one) or NA (not applicable). No other values should be entered. Note that 'NA' datapoints are eliminated before any software operations and thus essentially exist only for bookkeeping purposes.
- ➔ **Date**: The date at which the value given is correct. If no chronological information is offered, enter no date, and the software will default to assuming the value is correct for the entire documentation of the language. Dates CE are given as positive numbers, while dates BCE are negative numbers (e.g. -100). Note that 0 serves as 1 BCE, which has ramifications for exact dates: for instance, Caesar was assassinated in 44 BCE, which corresponds to -43.
- ➔ **Source**: An academic citation (or other source where relevant) for the value given.
- ➔ **Note**: An observation, clarification, justification or summary of the evidence relevant to the value given.
- ➔ **Code_ID**: An automatically generated link to the code table.

## *4.3. Languages*

LitTEL requires information about individual languages over and beyond the information supplied by Glottolog (involving a LanguageTable).

The languages can be consulted in this Google spreadsheet. The spreadsheet contains the following columns:

- **ID**: This field contains the Glottolog ID for the language in question. Proto-languages are assigned the code of the family they are ancestral to.
- **ID_status**: When Glottolog lacks an ID for a language or dialect, a new code is created by taking the first four letters of the name and adding a number from 1200 (e.g. Proto-Tyrrhenian becomes tyrr1200). To signal the creation of a new ID, ID_status is set to 0. For languages which have a Glottolog ID, ID_status is 1. Temporary field.
- **Name**: The English name of the language.
- **Alternative_name**: Alternative names for the language used in the (English) literature.
- **Family**: The top-level family to which the language belongs. Isolates and languages of unclear genealogical affiliation are classed as 'unclassified'.
- **Phylum**: The intermediate family to which the language belongs. These include taxons such as 'Italic' or 'Semitic'. If a language cannot be categorised into one of the known phyla of a family, 'unclassified' may be used on this level too.
- **Earliest**: The (if necessary approximate) date at which the language in question is first attested. This date may refer to the creation date of a physical bearer, but alternatively to the composition date of a work which survives only on physical copies that were made later.
- **Floruit**: The date around which the largest number of texts in the language in question were written. This date is used as a reference date for values when there is no evidence for diachronic change in a language. If no floruit is given, the average of the earliest and latest dates will be used.
- **Latest**: The date at which a language goes extinct in the written record. Where none is given, the software will assume the language persisted until the end of the period covered by LitTEL (1000 CE).
- **Issue**: A Boolean field stating whether or not a language had issue (linear descendants). For instance, Latin has issue in the Romance languages (1 or TRUE), while Etruscan goes extinct with no descendants (0 or FALSE).
- **Attestation**: This field states the nature of a language's attestation. This may be 'reconstructed' for languages that are not attested at all, 'epigraphic' for languages, 'literary' for languages with literary texts.
- **Status**: This field disambiguates languages from dialects.
- **Latitude**: The latitude of a point central to the oldest known heartland of the language in question. This field, together with the following field, serves to visualise data on a map, as well as providing the parameters for any other geographical operations on the data.
- **Longitude**: The longitude of a point central to the oldest known heartland of the language in question.
- **Parent**: The Glottolog ID of the immediate parent of a language (if known). For dialects, this field should contain the ID of the language they are subsumed under (e.g. Latin for Faliscan).
- **Source**: The default or principal source used for the annotation of a language.
- **Relevance**: This field specifies which languages should be used for the analysis (eliminating rows kept only for bookkeeping purposes, e.g. because the language is invalid or non-existent).
- **Note**: Any further information on the language.

### *4.3. Features and codes*

The LitTEL features, discussed in great length in this document, are summarised in their own spreadsheet. As per the CLDF guidelines, this involves a ParameterTable (specifying the comparative concepts involved) and a CodeTable (specifying the categorial values that each feature allows).

The ParameterTable can be consulted here and contains the following columns:

- **ID**: An alphabetic code containing two abbreviated words an a hyphen (e.g. Pos_Coord). These codes are designed to be maximally human-readable for ease of annotation.
- **Name**: An English interpretation of the feature ID.
- **Description**: A longer description of the feature.

The CodeTable can be consulted here and contains the following columns:

- **ID**: An automatically generated ID combining the Parameter_ID and the binary value.
- **Parameter_ID**: A reference to the parameter to which the code belongs.
- **Name**: A short name for the value.
- **Description**: A longer description of the value.
- **Value**: The code (0 or 1) of the value.
- **Typology**: The source of evidence for setting 0 to a certain value (viz. typological work indicating that a particular value is more common).

### *4.4. Scripts*

The LitTEL dataset can be operationalised with the following Python scripts.

- **main.py**: initialises LitTEL.
- **imports.py**: imports the required files.
- **summary.py**: creates some summary statistics based on the LitTEL data.
- **timeslice.py**: uses the LitTEL data to create a TimeSlice object, representing the distribution of a value for a feature at any given date.
- **timelineage.py**: uses the LitTEL data to create a TimeLineage object, representing the distribution of a value for a feature over the history of a single lineage.
- **changelog.py**: creates multiple TimeLineages to generate a ChangeLog of all unique diachronic changes to a specific feature in a specific direction.
- **geography.py**: visualises a TimeSlice on a map.
- **mds.py**: performs MDS on multiple TimeSlices.
- **correlation.py**: compares ChangeLogs to track the correlation between value changes.

# 5. Colophon

**General editors**:

Reuben Pitts (reuben.pitts@kuleuven.be)

Maxime Maleux (maxime.maleux@kuleuven.be)

**Annotators**

Reuben Pitts (reuben.pitts@kuleuven.be)

Maxime Maleux (maxime.maleux@kuleuven.be)

Jobstudenten...?

**Special thanks to**

Toon Van Hal

Freek Van de Velde

**Link to share (read-only)**

https://docs.google.com/document/d/1mVsx3DZ5uEtzaE-zht0_toEb0enlRFNy4XLprubvugo/edit?usp=sharing

https://www.britishmuseum.org/collection/object/W_SOC-93