Guidelines for PaVeDa annotation

Version 1.1 – 31 January 31, 2024

Contributors should provide information about themselves and about the language they are working on (see file Languages_and_contributors on Dropbox, folder New Database PaVeDa.

The template Excel file has five sheets: 1) Meanings and microroles; 2) Basic usages; 3) Alternations; 4) Examples; 5) Cognates.

Please do not alter the order of the sheets and the order of the columns, as importing scripts are designed based on this template.

In each sheet, mandatory columns are highlighted in yellow.
If you do not want to add any information in the optional columns, just leave the cells empty.
If you need to add the same piece of information in multiple cells (e.g., a coding frame such as “1-nom V.subj[1]” for one-place verbs), please copy and paste it.

1. Meanings and microroles

In the ValPaL database microroles are associated with verb meanings, and thus they are not language-specific. However, the number-microrole mapping is language specific.

In the sheet 1) “Meanings and microroles”, Columns A, B, and C are mandatory.

Insert the verb meanings (in capital letters) in column A.
Insert all coding frames (both basic or derived ones) in column B.

You need to specify in column C the coding frame type of column B. Column C has two possible values, B/A:
B = basic
A = alternation
“B” indicates that the corresponding coding frame in Column B is basic. This means that in this row you are adding the microroles of the basic coding frame (basic usage of the verb).
“A” indicates that the corresponding coding frame in Column B is associated to an alternation. This means that in this row you are adding the microroles of the derived coding frame for a specific alternation.

The microrole-number mapping should not be modified from basic to derived coding frames. The reasons to add microroles for alternations, too, are that (i) microroles coding sets can be alternation specific; (ii) derived coding frames may require additional microroles.

In further columns, 1-n numbers indicate to which argument microrole names, coding sets, and argument types refer to. Microrole names, coding sets, and argument types should be added for each argument of the verb.

mr = microrole name
Insert the microrole of each argument: first argument (1-mr), second argument (2-mr), third argument (3-mr), etc. E.g., for the meaning ASK FOR: asker, requested thing, askee.
The order of microroles in this sheet should mirror the order of arguments indicated in the basic and derived coding frames (see Sections 2-3) and is thus language-specific.

cs = coding set
Insert the coding set of each argument: first argument coding set (1-cs), second argument coding set (2-cs), third argument coding set (3-cs), etc. Note that this information is language-specific. Consider, for instance, the 1-cs for the meaning ASK FOR (cf. cell E2): "1-nom V.subj[1]" indicates the case of 1-mr (1-nom) and the fact that it is indexed on the verb (V.subj[1]).

at = argument type
Insert the argument type of each argument: first argument (1-at), second argument (2-at), third argument (3-at), etc., by choosing among the following set of argument types, updated by the contributors in respect of the ValPaL list available in ValPaL:

A = transitive verb subject
P = direct object
S = intransitive verb subject
X = other types of arguments, including third arguments of ditransitive verbs
I = instrument
L = locative (including goal)

If you want to add new meanings, you should compile a file with all the information required (see file new-meanings-and-microroles on Dropbox folder New Database PaVeDa). Before adding new meanings please check whether in the ValPaL database there are some compatible meanings (also among the list of additional meanings)
E.g., BE A HUNTER → additional meaning: HUNT FOR
SHAVE (a body part/person) → additional meaning: SHAVE (self)

<u>Microroles assignment</u>

For verb meanings that are already in the ValPaL database, please stick to ValPaL microrole names. For new meanings, please use the following procedure:
(i) look at similar verb meanings in the ValPaL database: e.g., for the group of emotion verbs added for Latin (so-called "impersonal" verbs, e.g., BE ASHAMED *pudet*), other experiential verb meanings already in the ValPaL database have been considered (e.g., BE HUNGRY, BE SAD).
(ii) check other resources: e.g., PropBank and Framenet (UVI: https://uvi.colorado.edu).

Take as an example the meaning ASK FOR. In the ValPaL database this meaning is associated with the following general microroles: asker, askee, requested thing, ask for causer. Consider now the Latin corresponding verb *petō* and see how its arguments are mapped with specific microroles, coding sets, and argument types. The verb ASK FOR *petō* has three arguments. The first argument is associated with the microrole “asker” (1-mr = asker). The “asker" argument is encoded with the nominative case and it is indexed on the verb (1-cs = 1-nom V.subj[1]). The “asker" argument is a transitive subject (1-at = A). The second argument of *petō* is associated with the microrole “requested thing”. The “requested thing'' argument is marked with the accusative case (2-cs = 2-acc). The “requested thing" argument is a direct object (2-at = P). The third argument is associated with the microrole “askee” (3-mr = askee). The “askee” argument is encoded with a prepositional phrase with *a*/*ab* and the ablative case (3-cs = a/ab 3-abl). The “askee” argument is the addressee of the event of asking, so its argument type is labeled with the “X” (3-at = X).

2. Basic usages

In the sheet 2) “Basic usages”, Columns A, B, C, D, E are mandatory, whereas Columns F and G are optional.

Insert the verb meanings in column A “Meaning”.

Column B “ValPaL” asks you to specify whether the verb meaning is stored in the ValPaL database. It has 3 possible values: yes, additional, no.

- You should use “yes” if the meaning in column A is in the ValPaL core meaning list.
- You should use “additional” if the meaning in column A is already in the ValPaL but does not belong to the ValPaL core meaning list.
- You should put “no” if the meaning in column A has been added by you, so it is a new meaning.

For each verb meaning insert the verbal counterpart (the lemma) in column C “Verb”.

In column D “Coding frame” insert the basic coding frame for each verb.
The coding frame should follow the ValPaL conventions. For example, look at the following coding frames:
(i) Lat. 1-nom V.act.subj[1] 2-acc

- 1 and 2 indicate the first and second arguments respectively
- -nom and -acc indicate case marking
- V stands for ‘verb’ and indicates that the verb is inflected in the unmarked voice, that, in the case of Latin, is the active voice.
- .subj[n] indicates the argument, if any, that is indexed on the verb
- word order is not specified, as the symbol “>” is not employed.

For languages with flexible and/or free word order you can place the verb in the middle.

(ii) It. 1 > V.subj[1] > 2 (con+3)

- Italian has no inflectional cases; thus, only numbers are used for subject and direct object, whereas “con+3” indicates that the third argument is introduced by the preposition “con”

- here, word order is specified by “>”, as it is relevant to encode grammatical relations in Italian
- Round brackets indicate that a certain argument is optional.

Some motion verbs allow for more than one prepositional phrase (PP) as locative/direction argument. How can we signal this information in the basic coding frame? We encourage to select the PP that is semantically neuter (and/or the most frequent) for the basic coding frame and listing the other PPs in the notes.

E.g. meaning GO Latin verb *eo* coding frame: 1-nom Vact.subj[1] ad 2-acc

ad = Latin preposition denoting direction vs. *in* = Latin preposition denoting direction but implying also containment

The contributor selected the PP: *ad* + acc for the basic coding frame and we listed *in* + Acc as one of the possible direction arguments of GO *eo* in the notes of the basic coding frame.

Column E “S/C” asks you to provide minimal information concerning the form of the verbal lemma. It has two possible values:
- S = simplex
- C = complex

You should mark with “S” verbs that are simple forms. You should mark with “C” verb meanings that are instantiated by multi-word expressions (e.g., Eng. BE AFRAID *be afraid*, It. BLINK *sbattere le palpebre*), as well as compound verbs with preverbs/prefixes, such as Latin ASSASSINATE *interficiō* (*inter+faciō*).

In columns F and G, you can optionally add comments and notes regarding the verbal lemma (cf. Column F “notes-lemma”) and the basic coding frame (cf. column G “notes-cf).

3. Alternations

Sheet 3) has 12 columns. Columns A-B, E-H, and J are mandatory. Columns C, D, I, K, and L are optional.
Each alternation has a dedicated row in the sheet. In every row, repeat the verb meaning (column A “Meaning”) and the corresponding verb (column B “Verb”). If a verb does not feature any alternations, please repeat the meaning and the verb form in column A and B and put “na” (= not attested) in every other cell of the row, except for the cell in column J (see detail later).

In column C “Comparative concept” you should add a comparative concept for the language specific alternation in the row. The tagset of comparative concepts is still to be finalized, so this information is not mandatory at this stage.

Column D “Alternation class” has four possible values:
R = valency re-arranging
A = valency augmenting (cf. Malchukov’s 2015: 96 ‘increasing’)

D = valency decreasing (both demoting strategies, such as passive, and deleting strategies, such as object omission)
I = argument identifying (reflexive and reciprocal alternations)

Insert in column E “Language-specific alternation”, the label of the language-specific alternation. This label can specify formal details regarding the realization of the alternation in a specific language and/or conform to the terminological conventions of the grammatical tradition of the project language.
E.g., Latin passive alternation

Following precedent literature, the contributor chose the label “*-r passive*” in order to point out that passive forms in the *infectum* stem have a characteristic affix in *-r.*

In column F, add the description of the language-specific alternation.
E.g., Latin *-r* passive

In Latin the passive alternation is marked by continuants of the PIE mediopassive or ‘middle’ which in Latin, in some cases, is to be interpreted as a passive. F
or the *infectum* stem, synthetic forms with a characteristic affix in *-r* are used. There are no special *perfectum* stem forms for the passive. Instead, periphrastic forms consisting of the passive perfect participle and forms of the verb *sum* ‘to be’ are used (Pinkster 2015: 51). In Latin A can be expressed with an adjunct PP with a/ab + ablative.

Column G “type” has two possible values: coded and uncoded. You should indicate here if the alternation in the row is coded or uncoded. Coded alternations are overtly marked by an affix, a clitic or an auxiliary (e.g, English passive form, Latin passive *-r* form). Uncoded alternations are not marked on the verb in this way (e.g., object omission in Italian and the dative alternation in English).

Insert in column H “Derived coding frame” the derived coding frame for each alternation. The microrole-number mapping should not be altered with respect to that of the basic coding frame. For example, look at the following coding frame:
Lat. 2-nom (a/ab 1-abl) passV’.subj[2]
In addition to what indicated in Section 2 for the basic coding frames, please note the changes concerning the verb form:

- “pass” indicates that the verb is in the passive form
- V’ (with the single quote mark) indicates that the alternating verb form has undergone some changes with respect to that in the basic coding frame.

<u>Tips for the annotation of the coding frames of frequent alternations</u>

- Reflexive alternation

COVER Gothic verb *(ga)-huljan*
reflexive alternation derived coding frame: 1=2-nom Vact.subj[1=2] 1=2-acc-sik
We advise to put 1=2 to signal the reflexive alternation.

- Reciprocal alternation

SEARCH FOR Ancient Greek verb *zētéō*
reciprocal alternation derived coding frame: 1/2-nom V.act.subj[1] 1/2-allḗlōn-acc
We advise to put 1/2 to signal the reciprocal alternation.

- Cognate object

PLAY Ancient Greek verb *paízō*
Cognate object alternation derived coding frame: 1-nom V.act.subj[1] 2-acc-cognate
We advise to put the word “cognate” in the derived coding frame to signal that an argument of the verb is a cognate object.

Warning: Identical coding frames cannot be associated with different alternations. Please avoid using ambiguous coding frames. Try to find a way that allows disambiguating the coding frames. For instance, in the derived coding frame of the Latin -r passive alternation, the contributor added the passive agent within brackets:
2-nom (a/ab 1-abl) passV'.subj[2]
This allows differentiating the derived coding frame of the -r passive alternation from the derived coding frame of the anticausative alternation with the mediopassive -r form: 2.nom passV’subj[2].

In column I you can add the derived verb lemma for coded alternations.

Columns J-K refer to the frequency of an alternation in your corpus data.
Column J has four possible values
R = regularly
M = marginally
N = never
D = no data
Choose “N” if an alternation is not attested (see verbs featuring no alternations)
Choose “D” if your corpus data regarding that specific alternation are not conclusive. For instance you can opt for D when an alternation is attested in lexicographic resources, previous literature or other corpora but it is not found or is under-represented in your reference corpus.

In column K “Frequency”, you can insert the exact frequency of each alternation in your corpus data. Please insert a number between 0 and 100 (percentage format) for the frequency. Percentages of each alternation are calculated out of the total number of occurrences of the verb.

In column L, you can add comments regarding the language-specific alternation. For instance, you can add information concerning:

a) lexical aspect, root or other morphological restrictions on the lemma that alternates
b) grammatical aspect tense restrictions of a specific alternation
c) semantic (e.g., animacy) or Part-Of-Speech (POS) restrictions of arguments of the alternations (limited to those that are relevant to define classes).

Additional remarks on alternations

- Always include alternations in which we have different case markings on the argument.

For example: in Turkish the forms *üzgün olmak* “be sad” and *üzülmek* “become sad” differ in the case marking of their arguments.

(i) Ben son. olay-lar-dan üzgün-üm.
I recent. event-PL-ABL sad-1sg
‘I am sad about the recent events.’

(ii) Ben son olay-lar-a çok üz-ül-dü-m.
I recent event-PL-DAT very sad-PASS-PAST-1sg
‘I am saddened by the recent events.’

- Try to maximize the number of alternations we include as far as they are marked morphologically rather than by an auxiliary.

  E.g., inchoative forms of stative predicates

  For example: in Turkish *aç olmak* “be hungry” > *acıkmak* “become hungry”. The inchoative verb has the same argument structure as its stative counterpart.

  (i) Ben aç-ım / karn-ım aç.
  I hungr1SG/tummy-1POSS hungry
  ‘I am hungry/My tummy is hungry.’

  (ii) Ben/karn-ım acık-tı-(m).
  I/tummy-1POSS become.hungry-PAST-AGR
  ‘I/my tummy got hungry.’

  Tentatively, use the term fientive to indicate instances like the examples above, in which we have an adjective denoting a state and a derived verb denoting the transition from a state to a new state.

- Metaphorical usages should not be included. But if you like, you can include them in a note.

4. Examples

Sheet 4) contains 7 columns. Columns A-D and I are mandatory, columns E-H and J are optional (details below).
Provide here examples for both basic and derived coding frames of each verb. Insert the verb meaning in column A and the coding frame (basic or derived) in column B.
In column C put “B” (basic) if the coding frame is basic; alternatively, put “A” (alternation) if the coding frame is derived.
In column D “primary text” insert the example exactly as it is in the language you are dealing with (raw text). In column E “analyzed text”, include information useful to understand the

glosses. E.g., for languages with writing systems other than the Latin one you can add the transliteration. You can signal here other phenomena, such as external sandhi (e.g., Eng. *don't* → *do not*; Lat. *deaeque* →*deae =que*).

Provide the glosses of the example in column F "Gloss".
Important: please remember that the number of blank spaces in column E "Analyzed text" should correspond to the number of blank spaces in column F "Glosses". Spaces are the means that scripts use to align text and glosses. If the example is very long, you can shorten it down to the argument structure that you are interested in and then keep the whole translation in brackets. Insert in column G "Translation", a translation of the example. If your reference corpus provides a suitable translation, feel free to use it. If the translation in the reference corpus is too free or inaccurate, please provide your own translation or a translation from a different source.
In column H "comment" you can add information regarding the example provided.
In column I "Source" indicates the source of the example (e.g., the locus or the reference to a section of the corpus you are using).
In column J you should provide the link to the example in the external resources you are using (corpus data). Links are not mandatory, but we encourage you to add them if you are using an open access corpus. If the link to the external corpus allows you to directly go to the exact example you want to use, you can just provide the link to the corpus, the primary text (column D) and the source (column I). However, we suggest filling out all the columns (including analyzed text, gloss and translation), as links are not stable.

5. Cognates

This sheet is optional. You can include here cognate verbs in different languages. Note that in the database different stages of the same language are considered different languages. Please insert, in the first place, the cognate(s) of the language(s) that is(are) chronologically and/or genealogically closer to the language you are working with. For example, for Latin, the contributor should initially insert cognates in Italian and in other Romance languages; later on, they can optionally add cognates in other IE languages as well.

Please provide the verb meaning in column A "Meaning" and the corresponding verb in column B "Verb".
In column C "Glottolog", insert the code of the language of the cognate verb (see the list of codes in Glottolog https://glottolog.org/). In column D "Meaning (rel col C)", add the meaning of the cognate verb of the language in column C: this meaning might be the same as the meaning of the verb in column A or might be different. In column E "Verb (rel col C)" insert the cognate verb in the language in column C.

E.g., for Lat. GIVE *do* we have a cognate verb It. GIVE *dare.* Here, the two cognate verbs have the same meaning. Instead,e.g. Engl. *teach* and Germ. *zeigen* are cognates with two different meanings, specifically, TEACH and SHOW (here, as a cognate, one could also add It. *dire* SAY).