Guidelines for Building the Latin Valency Lexicon VALLEX

Marco Passarotti
Berta González Saavedra
The *Index Thomisticus* Treebank Project
CIRCSE Research Centre
Università Cattolica del Sacro Cuore, Milan Italy

** Work in Progress **
November, 2015

This document describes the treatment of some specific issues in building the Latin valency lexicon VALLEX.

Table of Contents

Morphological Tag Assignment and Tagset

Each frame element in the lexicon is assigned one tag for case and one for PoS (except for the cases reported below).

Newly Added Nodes:

A)

If a frame element is represented by a newly added node with t-lemma #Gen, it is not assigned any morphological tag. See, for instance: SlaT-005.SCG*LB1.CP--++1.N.-2.1-1.5-1-n6.

B)

If a frame element is represented by a newly added node with t-lemma #PersPron, it is assigned the following morphological tags: (.uANY-POSSIBLE-CASE). This is in order to distinguish between an impersonal construction (case (A), with t-lemma #Gen) and a personal one, with the subject not represented by any lexical item in the sentence. See, for instance: a#a-005.SCG*LB1.CP--++1.N.-2.13-3.15-7W2.

C)

If a frame element is a newly added node that is a copy of an analytic node, it is assigned the morphological tags as if the newly added node was present at analytical level. See for instance the valency frame of *tango* at 005.SCG*LB1.CP--++1.N.-7.3-4.6-5 (*sapientia* is the analytical newly added node).

The tag for the case is not assigned if a frame element is assigned one of the following PoS:

- interjection (i)
- uninflected adverb (d)
- subclauses (c & v)
- conjunction (j)
- a verbal inflected form of verb (v; i.e. not a gerundive, gerund, participle form)
- direct speech (s)
- the (f) value, if the frame element is an infinitive. In all the other cases (participle, gerunds, and gerundives), the tag for the case must be assigned.

The tagset used is the same of the PDT, except for the following modifications:

- PoS "c": modified from "direct subclause" -> "subclause"
- Pos "f": participles, gerunds, and gerundives are added

PoS:

- a: adjective
- d: adverb
- i: interjection
- n: noun
- j: subordinating conjunction
- u: possessive form of a noun or pronoun
- f: infinitive clause (i.e. Accusativus cum Infinitivo, or infinitive with nominal function) OR nominal inflection of verbs (IT tagset: tag "2", first position): participles, gerunds, and gerundives
- s: direct speech
- c: content clause (a subordinate clause beginning with a relative, or indefinite pronoun/adverb)
- v: subordinate clause, with any kind of conjunction: j[.v]

Case:
- 1: nominative
- 2: genitive
- 3: dative
- 4: accusative
- 5: vocative
- 6: ablative
- 7: adverbial

Formal Notation of Valency Frames

In the valency frame, each functor is followed by round brackets. Apart from the case of newly added argument-nodes (where round brackets are left empty), the round brackets include as many "forms" of the functor as are those occurring in the sentence. Each form is opened by a dot, which is followed by (a) the PoS tag and the Case tag, or (b) only the PoS tag (see above for more details). The order tag-sensitive (always the PoS tag first).
For instance: ACT(.n1) PAT(.n4) -> ACT=noun-nomin. PAT=noun-acc.

In case of functors represented by more than one lexical item (i.e. when paratactic constructions are at work), all the different forms of the lexical items are reported in the valency frame. The forms are separated by a ;.
For instance: ACT(.n1;.f) -> ACT=noun-nomin. & infinitive clause
NB: only the different forms are reported in the valency frame, not their number, i.e. if a functor is represented by more than one lexical item and all of them have the same PoS and case, only one form is reported in the valency frame.

Every time a prepositional clause is concerned, the preposition must be reported before the tags. The tags must be reported between square brackets [ ]. See the following format:
NAME_OF_PREPOSITION[.TAG(S)]
See for instance the valency frame of *pertineo* at 005.SCG*LB1.CP--++1.N.-2.13-3.15-7: "ad quam (PAT) pertinet eius finis (ACT)". In this case, the lexical frame is ACT(.n1) PAT(ad[.u4]).

When a prepositional clause features two or more lexical items (with different forms), all of them must be reported in the valency frame, according to the following format:
FUNCTOR(PREP[.TAG(S)];PREP[.TAG(S)])
NB: the PREP must be repeated as many times as are the different forms of the lexical items involved.
Example: PAT(ad[.n4];ad[.u4]) -> a Patient expressed by a PP headed by "ad" with two (depending) forms: n4 (noun, accusative) and u4 (pronoun, accusative).
NB: in principle, this is valid for prepositional clauses only, and not also for conjunction clauses, because the latter involve verbal clauses only, whose tag is always "v" (thus, no more than one value is at work).

Every time a subclause introduced by a subordinative conjunction is concerned (postag .j or .v), the subordinative conjunction must be reported before the tags (this is not the case for AcI constructions,

which are not introduced by a subordinative conjunction and are just assigned the PoS tag .f). The tags must be reported between square brackets [ ]. See the following format:
NAME_OF_CONJUNCTION[.TAG(S)].
For instance, the valency frame of *dico* in a sentence like "puella dicit quod…" is is ACT(.n1) PAT(quod[.v]).
A subordinate clause, with any kind of conjunction is reported as follows: j[.v]

The question mark preceding the functor specification indicates optionality; if the question mark is not present, the modification is obligatory.

Every functor (including functors of non-arguments and alternations) must occur in the record no more than once.

A valency frame lists the valency modifications in the following order: ACT, CPHR, DPHR, PAT, ADDR, ORIG, EFF, BEN, LOC, DIR1, DIR2, DIR3, TWHEN, TFRWH, TTILL, TOWH, TSIN, TFHL, MANN, MEANS, ACMP, EXT, INTT, MAT, APP, CRIT, REG.

Frame elements are either optional or obligatory. The record of an optional element is preceded by a question mark.

Valency of Nouns. Basic Principles

The valency frames of derived nouns depend on the valency frames of their base verbs. The valency frames of non-derived nouns are considered independently (of any other valency frames).

In its event use, a noun usually has all the arguments and obligatory adjuncts of the base verb; towards the fully substantivized (lexicalized) meanings the verbal arguments become less and the nominal ones more frequent (predominant).

The role absorption may be defined as follows: the semantics of a certain argument (or possibly an adjunct) which is part of the valency frame of the base verb is incorporated in the meaning of the derived noun. As a result, the derived noun does not have the incorporated (absorbed) argument (adjunct) in its valency frame.
The argument shifting principle does not apply in these cases, so that the parallel between the valency frames of the noun and its corresponding verb stayed preserved. For example, if the Actor role is absorbed, it does not mean that the Patient is to be moved to its position (if the base verb has a Patient); the Actor is simply missing from the valency frame of the noun.
See, for instance, *artifex*: PAT(.u2) (005.SCG*LB1.CP--++1.N.-2.25-2.27-1)
Another example: *motor*: PAT(.u2) (005.SCG*LB1.CP--++1.N.-4.1-1.3-2)

Verbal and event nouns referring to events or states in principle share the valency frames with their base verbs.
See, for instance, *consideratio*: ACT(.u2) PAT() (005.SCG*LB1.CP--++1.N.-3.8-1.11-1)
Event and non-event use must be separated. While the event use requires the valency frame parallel to the one of the corresponding verb, the

result/affected object use is assigned a reduced valency frame (i.e. reduced in comparison to the valency frame of the verb):

- Event-use: see, for instance, *studium*: ACT() PAT(.n2) (005.SCG*LB1.CP--++2.N.-1.1-1.3-3)
- Non-event-use: see, for instance, *officium*: ACT(.n2) (005.SCG*LB1.CP--++1.N.--.1-1.1-4).
  Non-event-use: see, for instance, *scientia*: PAT(.n2) (005.SCG*LB1.CP--++1.N.-5.1-1.2-4)

See: http://ufal.mff.cuni.cz/pdt2.0/doc/manuals/en/t-layer/html/ch06s02s03.html#valence3.2.4.2

Passive clauses

The valency lexicon entry of a verb used in passive form corresponds to its active use.
For instance: "sapientes dicantur qui res recte ordinant" (SlaT-005.SCG*LB1.CP--++1.N.-2.1-1.5-1-n11). Here, *dicantur* is passive. The correponding lexical entry of *dico* in the valency lexicon reflects the active use of the verb, as if the clause was "[they: #Gen, ACT()] dicunt sapientes [EFF(.n4)] (eos) qui res recte ordinant [PAT(.c)]".
Another example (005.SCG*LB1.CP--++1.N.-2.16-1.20-1): "[…] circa quam medicinalis [PAT] versatur #Gen [ACT]". Valency frame: ACT() PAT(.n4)
Another example: "qui [PAT] intenditur a primo motore [ACT]" (005.SCG*LB1.CP--++1.N.-4.1-1.3-2). Valency frame of *intendo*: ACT(.n1) PAT(.u4). The valency frame is assigned as if the sentence was "quem primus motor intendit".

EXCEPTION: *videor*
In order to represent in the valency lexicon the semantic shift of *videor* from *video*, in the case of *videor* with the meaning "to seem" the passive-to-active conversion is not performed while compiling the valency frame, but the cases occurring in the passive construction are kept as they appear in the analytical layer.
Thus, the ACT is assigned the tag for dative (.3), while the nominal forms of the PAT and of the EFF respectively are assigned the tag for the nominative (.1).
Examples:
"hoc videtur dicendum": ACT() PAT(.u1) EFF(.f1)
"mihi videtur hoc bonum": ACT(.u3) PAT(.u1) EFF(.a1)
"mihi videtur hoc bonum esse": ACT(.u3) PAT(.u1) EFF(.f)

Accusativus cum infinitivo

The valency lexicon entry does not reflect the AcI construction.
All the arguments entering the valency frame are assigned the morphological tags as in the finite construction.
This assignment is done to maintain the original structures of the verbs in the lexicon entries, and in this way two different valency-frames are not created with the same semantic structure, according only to syntax.

Active Infinitives
The ACT of an active infinitive is not assigned the accusative case, but the nominative (like all the arguments related to the ACT, e.g. PNom->PAT).

Example: "[…] unam (ACT) esse gubernativam (PAT_M) et quasi principem (PAT_M) [...]" (005.SCG*LB1.CP--++1.N.-2.13-3.15-7). In this case, the valency frame of the verb *sum* is ACT(.u1) PAT(.a1), as if the clause was "una est gubernativa et quasi princeps".

Passive Infinitives

A) first, apply the rule about "Passive clauses" (i.e.: turn the clause to the active voice)

B) then, apply the above rule about active infinitives.

Example: "[…] regulam ex fine sumi necesse est" (005.SCG*LB1.CP--++1.N.-2.8-1.10-4).

A) "#Gen (ACT) sumere regulam (PAT)"

B) the PAT is assigned (.n4), because it is in the accusative case in the active voice, and the ACT is assigned () because it is an added node #Gen. If the ACT in the passive voice were expressed by an ablative with *ab* (*ab aliqua re*), it would be considered the ACT of the active voice and then it would be assigned (.PoS1) according to the rule of the "passive clauses".

Nominal infinitives

Consistently with what has been explained at the preceding point, the nominal infinitives are considered as finite clauses where the nodes of reconstructed arguments are added. Those nodes are assigned the morphological tags, according with the rules expressed in "Newly added nodes".

Ablative Absolute

The entry in the valency lexicon does not reflect the ablative absolute construction.

So, in these cases the ablative case (.6) is not assigned to the PAT (past participle head), or to the ACT (present participle head) depending on the participle, but the corresponding case according to the (active use of the) predicate is given.

If the head verb has another argument in its valency frame, this is a newly added node (#Gen, or #PersPron) in most cases.

Example of past participle head: "[…] [#Gen] carne [PAT] induta […]" (005.SCG*LB1.CP--++1.N.-4.10-1.12-5). In this case, the valency frame of the verb *induo* is ACT() PAT(.n4).

Example of present participle head: "natura manente": here the valency frame of the verb *maneo* is ACT(.n1).

PoS-Tag of Elided Nouns, Nominalized Adjectives and n.quant. Words in Valency Frames Notation

In the valency frame of a valency capable headword, the PoS-Tag of these types of words is assigned as follows:

A) Elided Nouns

Elided nouns heading an attribute (expressed by an agreeing adjective, a noun in genitive, or a relative clause) are reconstructed. In the valency frame of their (valency-capable) headword they are assigned PoS-Tag .n

For instance: 005.SCG*LB1.CP--++7.N.-6.15-7.17-5 (SCG_2, 20)

unde nec demonstrationis vim habent , sed vel sunt rationes probabiles vel sophisticae

A new node (with t-lemma *ratio*) is added heading the node *sophisticus*. In the valency frame of *sum*, it is assigned PoS-Tag .n.

B) Nominalized Adjectives

Adjectives functioning as nouns are assigned PoS-Tag .n in the valency frame of their (valency-capable) headword.

For instance: 005.SCG*LB1.CP--++8.N.-3.6-2.8-3 (SCG_2, 32)

qui enim pie infinita prosequitur

Infinitus is assigned PoS-Tag .n in the valency frame of *prosequor*.

C) n.quant. Words

Like the words that are assigned sempos n.pron, also the words that are assigned sempos n.quant are considered pronouns. So they take are assigned PoS-Tag .u in the valency frame of their (valency-capable) headword in the valency lexicon.

For example: 005.SCG*LB1.CP-1++3.N.18.2-5.4-5 (SCG_3, 128)

et unum illorum invenitur sine altero , probabile est quod alterum absque illo inveniri possit

Unus (sempos: n.quant.) is assigned PoS-Tag .u in the valency frame of *invenio*.

Disambiguation in Naming the Lemma

When two lemmas are homograph of each other and they have the same part of speech, in the valency lexicon they are disambiguated by adding a 'note' in their name after the symbol ^.
This note for disambiguation can be chosen freely, but it is usually one of the following:

- the infinitive of the verb, in case of homographic verbs of different conjugation. Example: *volo^velle* vs. *volo^volare*;

educo^educere vs. *educo^educare*

- the lemma of the base-lemma from which the lemma in question is derived. Example: *accido^cado* vs. *accido^caedo*

- a word providing a semantic description of the lemma in question. Example: *liber^volumen* vs. *liber^filius*

In order to keep the link between the entry in the valency lexicon and its occurrences in the tectogrammatical layer of the treebank, the note for disambiguation must be added also in the name of the lemma in all its occurrences in TGTS. In particular, it must be added in the "t_lemma" box.

NB
Some of these addings are already present in the lemmatization available in ATS (strictly speaking, in the morphological layer), some not.

The former do not require any change in the name of the lemma both in the valency lexicon and in the "t_lemma" box.

The notes for disambiguation already available in ATS are inherited from the original lemmatization of the IT. They are those that are needed to disambiguate those homographic lemmas that cannot be distingushed by their morphological features (like, for instance, the conjugation).

Example: the two lemmas *accido* cannot be distinguished by their morphological tags (they share the same conjugation). Thus, in the IT they are disambiguated by adding "^cado" or "caedo" after the name of the lemma. The same holds also for "liber".

Instead, the notes for disambiguation that are not present in ATS but must be added both in TGTS ("t_lemma" box) and in the valency lexicon are those that are needed to disambiguate those homographic lemmas that can be distingushed by their morphological features (like, for instance, the conjugation).

Example: the two lemmas *volo* are not disambiguated in their name in ATS, since they can be disambiguated by their morphological tags: one *volo* belongs to the first conjugation, the other to the

regurlarly irregular one.

Since the morphological tags are not available in the valency lexicon, a note for disambiguating those homographic lemmas taht have the same part of speech is needed.

Valency of adjectives ending in *-bilis*

Deverbal adjectives ending in *-bilis*:

- are assigned one argument node ACT if they are derived from a two-argument verb
- are assigned two argument nodes (ACT and PAT, respectively) if they are derived from a three-argument verb

The additional argument (present in the valency frame of the base-verb) is the head-noun of the adjective in the dependency tree.

(Fictional) examples:

A) *mirabilis* < *miro, -are* (two-argument verb)

- *res mirabilis*: in the TGTS *mirabilis* heads a newly added node with functor ACT and t_lemma #Gen
- *res mirabilis ab omnibus*: in the TGTS *mirabilis* heads the node of *omnis* which is assigned functor ACT

B) *assimilabilis* < *assimilo, -are* (three-argument verb)

- *res assimilabilis*: in the TGTS *assimilabilis* heads two newly added nodes: (a) one with functor ACT and t_lemma #Gen; (b) the other with functor PAT and t_lemma #Gen
- *res assimilabilis alicui*: in the TGTS *assimilabilis* heads two nodes: (a) one with functor ACT and t_lemma #Gen; (b) the other is the node of *aliquis* which is assigned the functor PAT

Intuition-based Lexical Entries

In order to make the valency lexicon as much representative as possible, we decided that it must include at least the lexical entries for the first 500 most frequent verbs reported in L. Delatte *et al.* (1981), *Dictionaire fréquentiel et index invers de la langue latine*.

Those verbs that are included in these 500 but are not present in the data-driven valency lexicon are assigned a lexical entry built in intuition-based fashion, i.e. the lexical entry is not linked to any occurrence in the treebank, reflecting its formal properties.

To build the intuition-based lexical entries, we take inspiration from the following:

- our knowledge of Latin

- H. Happ (1976), *Grundfragen einer Dependenz-Grammatik des Lateinischen*,

- the contexts for the verb whose lexical entry must be built in the *Summa Contra Gentiles* of Thomas Aquinas available from the *Index Thomisticus*.

The intuition-based lexical entries include only information about the functors that take part in the valency frame(s) of the verb. No information about PoS and case is provided, since this depends on the formal properties of the single occurrences of the verb (and of its arguments) in the text. The tags for PoS and case (put between round parentheses after the name of the functor in the valency frame) are replaced with *, which stands for "typical".

Example: *iubeo* has one valency frame, which includes ACT and PAT. This valency frame looks as follows:

ACT(*) PAT(*)

In the TrEd visualization, this valency frame is shown like this:

ACT(typical)

PAT(typical)

There might be cases where more than one valency frame must be assigned, according to different senses of the verb. On example is the entry for *peto*, which has two valency frames:

ACT(*) PAT(*): when it means *to reach*

ACT(*) PAT(*) ORIG(*): when it means *to ask*

When the first occurrence of an intuition-based lexical entry is found in the data of the treebank, the following situations can happen:

- CONDITION: the lexical entry includes only one valency frame, which is the same of that found in the first occurrence in the treebank.

  ACTION: replace the * in parentheses with PoS and (if required) case tags reflecting the formal properties of the first occurrence found in the treebank

- CONDITION: the lexical entry includes more than one valency frame, one of which is the same of that found in the first occurrence in the treebank.

  ACTION: in this valency frame, replace the * in parentheses with PoS and (if required) case tags reflecting the formal properties of the first occurrence found in the treebank. Do not modify the other valency frame(s) (i.e. leave the *)

- CONDITION: the lexical entry includes one, or more than one valency frame, none of which is the same of that found in the first occurrence in the treebank.

  ACTION: add the new valency frame. Do not modify the other valency frame(s) (i.e. leave the *)

Notes about specific intuition-based lexical entries

MODAL VERBS

In tectogrammatical annotation modal verbs are supposed to collapse into the nodes of their dependent infinitives, thus not occurring in TGTSs.

However, modal verbs are assigned a lexical entry in the lexicon. This is because of the following: when an infinitive depends on two modal verbs (one heading the other: “possum velle amare”), in TGTSTs only the node for the second modal verb is collapsed into the node of the final infinitive, while the node for the first modal verb remains in the tree, thus requiring to have its entry in the valency lexicon.

The valency frames assigned to the modal verbs are:

- when the depending infinitive is assigned afun Sb in ATS (*oportet*, *licet*, etc.): ACT(*)

- when the depending infinitive is assigned afun OBJ in ATS (*possum*, *debeo*, etc.): ACT(*) PAT(*)

FIO

Although the lemmatization system for the IT-TB does not include the lemma *fio* (because the passive forms of *facio* are lemmatized under the lemma *facio*), we built the lexical entry for *fio* in the valency lexicon.

This entry won't be associated to any occurrence in the IT-TB, but it must be there for covering other lemmatization systems.

Fio is assigned the following intution-based valency frames:

- ACT(*) PAT(*): "something is done by somebody"

- ACT(*) PAT(*) EFF(*): "somebody is made something by somebody"

If a word has more than a spelling we follow the one used by the *Index Thomisticus*. If this word does not appear in the *Index Thomisticus* we follow the spelling used in the Forcellini's dictionary.