Latin Vallex. A Treebank-based Semantic Valency Lexicon for Latin

Marco Passarotti, Berta González Saavedra, Christophe Onambele
CIRCSE Research Centre
Università Cattolica del Sacro Cuore
Largo Gemelli, 1 – 20123 Milan, Italy
E-mail: {marco.passarotti, berta.gonzalezsaavedra, christophe.onambele}@unicatt.it

Abstract

Despite a centuries-long tradition in lexicography, Latin lacks state-of-the-art computational lexical resources. This situation is strictly related to the still quite limited amount of linguistically annotated textual data for Latin, which can help the building of new lexical resources by supporting them with empirical evidence. However, projects for creating new language resources for Latin have been launched over the last decade to fill this gap. In this paper, we present *Latin Vallex*, a valency lexicon for Latin built in mutual connection with the semantic and pragmatic annotation of two Latin treebanks featuring texts of different eras. On the one hand, such a connection between the empirical evidence provided by the treebanks and the lexicon allows to enhance each frame entry in the lexicon with its frequency in real data. On the other hand, each valency-capable word in the treebanks is linked to a frame entry in the lexicon.

Keywords: valency, Latin, lexicography

1 Introduction

Despite a centuries-long tradition in lexicography, Latin lacks state-of-the-art computational lexical resources. This situation is strictly related to the still quite limited amount of linguistically annotated textual data for Latin, which can help the building of new lexical resources by supporting them with empirical evidence. However, projects for creating dependency treebanks for Latin have been launched over the last decade, as well as for creating fundamental lexical resources, like the (still very small) Latin WordNet (Minozzi, 2010).

In this paper, we present *Latin Vallex*, a new lexical resource for Latin consisting in a valency lexicon built in conjunction with the semantic and pragmatic annotation of two Latin treebanks featuring texts of different eras.

The paper is organized as follows. Section 2 presents the related work on valency lexica. Section 3 describes *Latin Vallex*, by detailing its structure and explaining the way it is built and how to query the data. Section 4 includes the conclusion and sketches the future work.

2 Related Work. Valency and Lexica

The notion of valency is generally defined as the number of obligatory complements required by a word: these are usually named ‘arguments’, while the non-obligatory ones are referred to as ‘adjuncts’. Viewing lexical semantics through the notion of valency is a widespread approach in linguistics. This is strictly related to the basic assumption of frame semantics (Fillmore, 1982), according to which the meaning of some words can be fully understood only by knowing the frame elements that are evoked by that word.

Valency lexica for several languages are today available. These lexica play an important role in NLP thanks to their large applicability in tasks like semantic role labeling, word sense disambiguation, automatic verb classification, selectional preference acquisition and also treebanking (Urešová, 2004).

Like other language resources, also valency lexica can be built in intuition-based or in corpus-driven fashion according to the role played by human intuition and by the empirical evidence extracted from textual corpora. For instance, lexica like PropBank (Kingsbury & Palmer, 2002), FrameNet (Ruppenhofer et al., 2006) and PDT-Vallex (Hajič et al., 2003) were first created in intuition-based fashion and then checked and refined by using data taken from corpora. Examples of valency lexica automatically acquired from annotated corpora are VALEX (Korhonen et al., 2006) and LexShem (Messiant et al., 2008). These lexica reflect the evidence provided by data, with very little human intervention.

While several valency lexica have been compiled for modern languages, much work in this area still remains to be done for classical languages. Regarding Latin, Happ reports a list of Latin verbs along with their valencies (Happ, 1976: 480-565). Bamman & Crane (2008) describe a “dynamic lexicon” automatically extracted from the Perseus Digital Library by using the Latin Dependency Treebank (Bamman & Crane, 2006) as a training set. This lexicon displays qualitative and quantitative information on subcategorization patterns and selectional preferences for each word of the corpus. IT-VaLex (McGillivray & Passarotti, 2009) is a corpus-driven subcategorization lexicon whose (verbal) entries are automatically induced from the syntactic layer of annotation of the *Index Thomisticus* Treebank (Passarotti, 2011). The same structure of IT-VaLex is resembled by a lexicon created from the Latin Dependency Treebank and described by McGillivray (2013: 31-60).

3 Latin Vallex

3.1 The Structure of *Latin Vallex*

Latin Vallex (LV) was developed while performing the semantically annotated subset of two Latin dependency treebanks, namely the *Index Thomisticus* Treebank (IT-TB), which includes works of Thomas Aquinas, and the Latin Dependency Treebank (LDT), which features works of different authors of the Classical era. Each valency-capable word occurring in the semantically annotated portion of the two treebanks is assigned one frame entry in LV.

The structure of the lexicon resembles that of the valency lexicon for Czech *PDT-Vallex* in the theoretical context of Functional Generative Description (FGD; Sgall et al., 1986). FGD is the framework that also guides the style of the semantic and pragmatic layer of annotation of the two Latin treebanks, which corresponds to the so-called "tectogrammatical" layer of the Prague Dependency Treebank for Czech (PDT). This is built on a surface syntactic layer called "analytical" and it includes semantic role labelling, information structure and ellipsis/anaphora resolution. The *Dialogue Test* by Panevová (1974-1975) and the criteria reported in Mikulová et alii (2005: 100-102, 116-162) are used to distinguish arguments from adjuncts.

On the topmost level, the lexicon is divided into word entries. A word entry consists of a non-empty sequence of frame entries relevant for the lemma in question, where each different frame entry usually corresponds to one of the lemma's senses. Each frame entry contains a description of the valency frame itself and of the frame attributes. A valency frame is a sequence of frame slots. Each frame slot represents one complement of the given lemma. The surface morphological features of the frame slots are recorded, with some deviation (see 3.2). Attributes are semantic roles (called 'functors' in FGD) used to express types of relations between lemmas and their complements.

The structure of an entry of LV can be resumed as follows:

Name of the Word Entry (lemma) – PoS

- Frame Entry 1:
  - Valency Frame:
    - Frame slot 1
    - Frame slot *n*
  - Frame Attributes:
    - Functor 1
    - Functor *n*
- Frame Entry *n*:…

The semantic roles reported in the frame entries of LV are those for arguments ('inner participants'), which according to FGD are those complements that are assigned the following functors: Actor (ACT), Patient (PAT), Addressee (ADDR), Effect (EFF) and Origo (ORIG). Even some adjuncts ('free modifications') enter the frame entries and are recorded as optional slots. The set of functors is the one provided in the guidelines for tectogrammatical annotation of the PDT (Mikulová et al., 2005).

The only difference between LV and PDT-Vallex results from the fact that the so-called *argument shifting* is not applied in tectogrammatical annotation of the IT-TB and LDT. Argument shifting (Mikulová et al., 2005: 103-105) is a criterion used for determining the type of argument in question and, thus, assigning functors. Basically, it states that the first argument is always the Actor (functor ACT) and the second argument is always the Patient (functor PAT). All the other functors for arguments beside ACT and PAT (ADDR, EFF and ORIG) shift to ACT and PAT.

For instance, if a verb has an Origo-like argument but not a Patient, the Patient position is taken up by the Origo-like argument, i.e. Origo shifts to the position of Patient. This is reflected in PDT-Vallex, which features no frame entry provided with two slots whose frame attributes are, for instance, ACT and ORIG (because ORIG would be replaced by PAT by argument shifting).

Instead, this can happen in LV, as resulting from the tectogrammatical annotation of the IT-TB and LDT. For instance, the entry for the verb *resulto* ("to result") features a frame entry with two frame slots whose attributes are ACT and ORIG.

One occurrence of this frame entry for *resulto* is in the following sentence from the IT-TB (*Summa contra Gentiles*, book 1, chapter 27, number 4): "ex unione formae et materiae resultat aliquid compositum" ("from the union of form and matter it results some kind of composition (literaly: "something composed")"). In this sentence, the arguments for the verb *resultat* are *aliquid* ("something") and *ex unione* ("from the union"). The word *aliquid* is assigned functor ACT, while *ex unione* is assigned ORIG. If argument shifting had applied, *ex unione* would have been assigned functor PAT.

Beside the functors for arguments, also some functors for adjuncts can occur in frame entries. Such functors are marked as optional and do not undergo argument shifting. The most frequent functors for adjuncts appearing in LV are the locative and directional ones, which are mostly used in the frame entries for verbs of movement (Mikulová et al., 2005: 503-514). For instance, the prototypical frame entry for the verb *venio* ("to come") features three slots, whose functors are ACT, DIR1 (Direction-From) and DIR3 (Direction-To.)

Another example is the entry for the verb *termino* which, according to the structure of word entries in LV, includes two frame entries corresponding to two different senses of the word: (a) "to mark the boundaries of something" and (b) "to limit something to something else".

The frame entry for the first sense has a valency frame with two frame slots, the first of which is represented by a nominative noun (n1) and the second by an accusative noun (n4). The frame attributes report the functors, which are Actor (ACT) for the first frame slot and Patient (PAT) for the second.

The frame entry for the second sense features a valency frame with three frame slots: a nominative noun (n1), an accusative noun (n4) and a prepositional phrase introduced by the preposition *in* ("in"), which governs an accusative noun (*in*+n4). The functors for these three frame slots are Actor (ACT), Patient (PAT) and Direction-To (DIR3) respectively.

The entry for *termino* in LV looks as follows:

termino – V

- Frame Entry 1 (“to mark the boundaries of something”):
  - Valency Frame:
    - Frame slot 1: n1
    - Frame slot 2: n4
  - Frame Attributes:
    - Functor 1: ACT
    - Functor 2: PAT
- Frame Entry 2 (“to limit something to something else”):
  - Valency Frame:
    - Frame slot 1: n1
    - Frame slot 2: n4
    - Frame slot 3: *in*+n4
  - Frame Attributes:
    - Functor 1: ACT
    - Functor 2: PAT
    - Functor 3: DIR3

The morphological information reported in frame slots results from the confrontation with the textual evidence provided by the two Latin treebanks LV is built on.

3.2 Building *Latin Vallex*

All valency-capable words annotators get through while performing the tectogrammatical annotation of the IT-TB and LDT are assigned a frame entry in LV. These can be verbs (*do* - “to give”), adjectives (*contrarius* - “contrary”), nouns (*description* - “description”) and adverbs (*similiter* - “similarly”).

Presently, LV includes 1,373 lexical entries and 3,406 frame entries: 1,049 verbs (2,903 frames), 236 nouns (394 frames), 86 adjectives (106 frames), 2 adverbs (3 frames). These result from the tectogrammatical annotation of 2,000 sentences from the *Summa contra Gentiles* of Thomas Aquinas (IT-TB), of the full text of Sallust’s *Bellum Catilinae* (701 sentences) and of 100 selected sentences from Caesar’s *De bello gallico* and Cicero’s *In Catilinam* respectively (LDT).

Since the IT-TB and the LDT are not balanced to be representative of Latin (or of a variety of it), we enhanced the corpus-driven building of LV with a number of intuition-based entries. In particular, we wanted LV to include the lexical entries for all the valency-capable words occurring among the first 1,000 most frequent words of Latin reported by Delatte et alii (1981). Although most of such words are already present in LV as resulting from the annotation of the treebanks, 163 of them were not found yet in the texts. Thus, we built in intuition-based fashion those that we consider to be the prototypical frame entries for these words, by filling only the frame attributes and not also the frame slots, which have to be assigned by confrontation with textual evidence. Most of the intuition-based lexical entries of LV are assigned one prototypical frame entry, the total number of frame entries for the 163 intuition-based entries being 168. No frame entry of an intuition-based built entry of LV is linked to any occurrence in the treebanks until annotators get through its first occurrence and the frame entry is modified accordingly.

Figure 1 shows the tectogrammatical subtree of an excerpt from the IT-TB (*Summa contra Gentiles*, book 1, chapter 5, number 2): “[…] christianae religioni […], quae singulariter bona spiritualia et aeterna promittit” (“[…] to the Christian religion […], which in a unique way promises spiritual and eternal goods”).

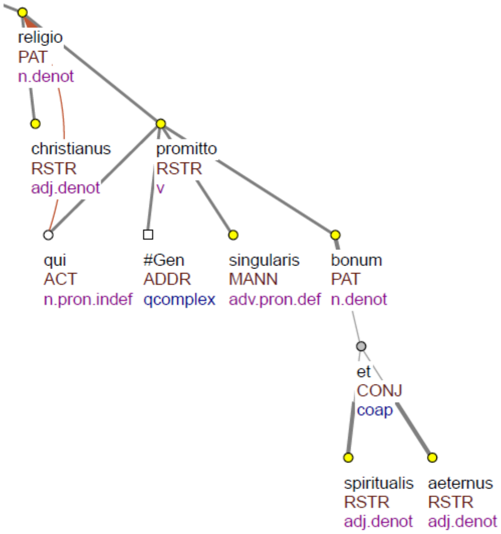

Figure 1: A tectogrammatical tree[1]

While building the tectogrammatical subtree shown in figure 1, annotators get through an occurrence of the valency-capable verb *promitto* (“to promise”) and they build (or assign it) the relevant frame entry in the lexicon.

The valency frame of the frame entry for this occurrence of *promitto* includes three frame slots. The first is a pronoun at the nominative case (*quae*; node *qui*): The second is a noun in the accusative case (*bona*; node *bonum*). The third frame slot results from a resolved ellipsis of an argument that cannot be retrieved contextually and, thus, it is considered to be a “generic” argument (marked #Gen); since this argument has no surface realization, it is not assigned any PoS and morphological feature in the frame slot. The frame attributes for these three slots are the following: Actor (ACT), Patient (PAT) and Addressee (ADDR).

Beside the three nodes that enter the frame entry, in the tree of figure 1 the node for *promitto* governs also a fourth node, which corresponds to the word *singulariter* in the sentence (node *singularis*) and it is assigned functor MANN (Manner). This node does not correspond to any frame slot in the frame entry of *promitto* because MANN is a functor assigned to free modifications not reported in frame entries.

[1] In the default visualization of tectogrammatical trees, forms are replaced by lemmas. For instance, *qui* is the lemma for the form *quae*.

More than 60% of the frame entries of LV feature a valency frame provided with two slots ('valency-2 frame entries'). For most of them, the frame attributes are Actor and Patient. The second most frequent kind of valency frame in the lexicon (around 20% of total) is that provided with three slots ('valency-3 frame entries'). The frame attributes for valency-3 frame entries present a more diverse configuration than those for valency-2 ones.

Figure 2 shows a network-like representation of valency-3 frame entries automatically induced from LV[2]. Red-colored nodes are for functors; white nodes are for frame entries, which are named from the word entry they belong to (the lemma) + a letter assigned to the single frame entry: for instance, the name *amo*-a stands for the frame entry 'a' of the word entry *amo* ("to love"). A functor node is connected to a frame entry node by an edge if that functor occurs in a frame slot of that frame entry.

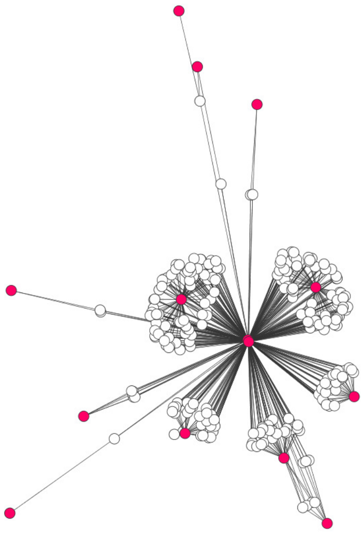

Figure 2: Network of valency-3 frame entries

In figure 2 there are two most central red nodes, which are those for ACT and PAT: most of the nodes for frame entries are linked to them. This means that most of the valency-3 frame entries feature an ACT and a PAT among their frame attributes.

There are five main clusters of nodes around the center of the network. From top left clockwise, they are those for ADDR (the biggest one), EFF, ORIG, DIR3 and LOC (locative). These are the most frequent functors assigned (as frame attributes) to the third frame slot of valency-3 frame entries (the first two being ACT and PAT). Thus, for instance, the nodes clustered around the node for ADDR are those for valency-3 frame entries whose frame attributes are ACT, PAT and ADDR (like *attribuo* - "to assign").

The most peripheral red nodes represent functors that are assigned few times to the third frame slot of valency-3 frame entries. For instance, the second red node from bottom left clockwise is that for ACMP (Accompaniment), which is linked to the nodes for the frame slots of the ACT-PAT-ACMP verbs *admisceo* ("to mix"), *conjungo* ("to conjoin") and *unio* ("to join").

The morphological information reported in frame slot does not reflect the surface form occurring in textual data for three kinds of constructions, namely: (a) passive clauses, (b) accusativus cum infinitivo and (c) ablative absolute. This is motivated by two main reasons. First, we want to keep the number of frame entries as limited as possible by collecting different surface forms into common frame entries. Second, LV is strictly related to the tectogrammatical layer of annotation of the base treebanks. Such layer of annotation represents the underlying syntax (also considered to be the literary meaning) of a sentence through a surface form independent pattern.

Although the surface form of the frame slots for these three constructions is not reported as it is in the LV entries for these constructions, it is not lost as it can always be retrieved from the morphological annotation of the treebanks. In more detail, the frame slots for words occurring in such constructions are built as follows.

3.2.1 Passive Clauses

Passive clauses are transformed into the corresponding active form before assigning (or building from scratch) a frame entry to their head verb.

For instance, see the following clause from the IT-TB (*Summa contra Gentiles*, book 1, chapter 1, number 2): "sapientes dicantur qui res recte ordinant" ("wise are said [to be] those who arrange things correctly"). In this clause, the head verb *dicantur* is in the passive form. The LV frame entry of the lemma *dico* ("to say") linked to this occurrence reflects the active form of the clause: "[they] say wise those who arrange things correctly". Thus, the frame entry for this occurrence of *dico* has a valency frame that includes three frame slots, and their corresponding frame attributes, as follows:

(1) a generic Actor (ACT);
(2) a Patient (PAT) expressed by a pronoun: "those (who arrange things correctly)";
(3) an Effect (EFF: functor assigned to obligatory predicative complements): "wise".

This solution allows to assign the same frame entry to the occurrence of *dico* in such a clause regardless of the fact that its surface form is active or passive.

3.2.2 Accusativus cum infinitivo

In Latin, accusativus cum infinitivo (AcI) is a construction formed by an infinitive verb whose subject in the accusative case.

The LV frame entry for an AcI corresponds to its counterpart construction with the finite form of the verb. In active constructions, the Actor of an AcI is assigned the nominative case in LV (instead of the accusative): In

[2] We used the free software *Cytoscape* for network creation and computing (Shannon et al., 2003; Saito et al., 2012). Nodes and edges are arranged according to the *Edge-weighted Spring Embedded* layout setting (Kohl et al., 2011).

passive constructions, the same happens after transforming the passive construction into the active.

For instance, see the following clause from the LDT (*Bellum Catilinae*, XX): “quis mortalium […] tolerare potest […] illos binas aut amplius domos continuare […]?” (“who in the world can endure that they should join together two houses or more?”). In this clause, the word *illos* (“they”) is a plural accusative pronoun playing the role of subject of the infinitive verb *continuare*.

The frame entry for this occurrence of the verb *continuo* (“to join together”) includes two slots:

(1) an Actor (ACT) expressed by a nominative pronoun: *illos* (accusative) → *illi* (nominative);

(2) a Patient (PAT) expressed by an accusative noun: *domos* (“houses”).

In this way, the same frame entry is assigned to a textual occurrence of *continuo* either if it occurs in a AcI or in a finite construction (the latter usually introduced by the subordinating conjunction *quod* - “that”).

3.2.3 Ablative Absolute

In Latin, ablative absolute is a grammatical construction where a noun and (typically) a participle form a phrase that is disjoint from the grammar of the rest of the sentence; both the noun and participle are inflected in the ablative case.

In LV frame entries, ablative absolute constructions are treated like passive clauses and AcI are. In frame slots, the subject noun is assigned the nominative case (and the ACT functor) for active ablative absolute constructions (present participle); instead, for passive constructions (perfect participle), first the participle is turned into active and then the subject noun is assigned the accusative case (and the PAT functor).

For instance, see the following clause from the IT-TB (*Summa contra Gentiles*, book 1, chapter 43, number 10): “[…] qualibet quantitate finite data […]” (“being given any finite quantity”). The word *data* is a form of the perfect participle of the verb *do* (“to give”). Thus, the ablative construction is turned into active (“having [a generic subject] given any finite quantity”) and the subject noun of the participle (*quantitate*) is assigned the accusative case in the frame entry.

The frame entry for this occurrence of the verb *do* includes three slots:

(1) a generic Actor (ACT);

(2) a Patient (PAT) expressed by an accusative noun: *quantitate* (ablative) → *quantitatem* (accusative);

(3) a generic Addressee (ADDR).

3.3 Querying *Latin Vallex*

Latin Vallex and the treebanks can be freely downloaded from the website of the IT-TB (http://itreebank.marginalia.it/view/download.php). Both the lexicon and the treebanks can be queried through an implementation of the PML-TQ search engine (Prague Markup Language – Tree Query) (Štěpánek & Pajas, 2010; http://itreebank.marginalia.it/view/resources.php).

Figure 3 shows a PML-TQ query in graphical form. The query searches for those word entries of LV (node $n0) that feature a frame entry ($n1) that is provided with at least three slots, which are assigned the following frame attributes: ADDR ($n4), PAT ($n2) and ACT ($n3). Furthermore, the query states that the Addresse slot must be a word inflected at the dative case (case = “3”).

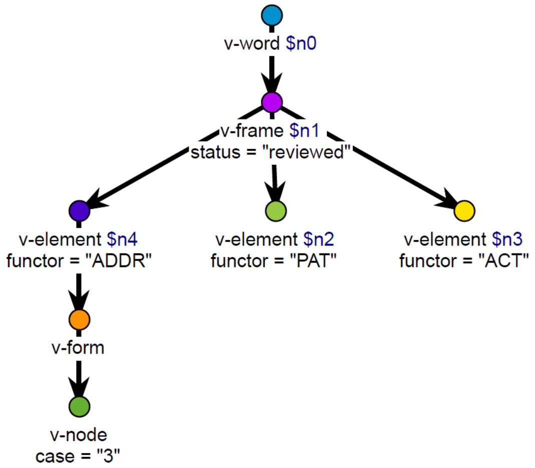

Figure 3: A graphical query on *Latin Vallex*

Figure 4 shows one of the outputs of the query pictured in figure 3. In particular, it reports one frame entry for the verb *confero* (“to confer”).

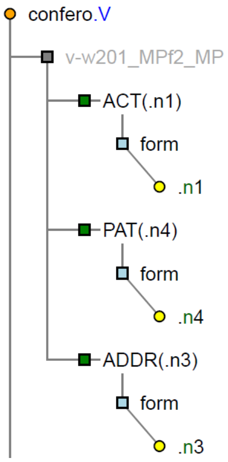

Figure 4: One frame entry for the verb *confero*

Following the query, this frame entry includes three slots: Actor, Patient and Addressee. The frame slots are further specified: the Actor is a nominative noun (n1), the Patient is an accusative noun (n4) and the Addressee is a dative noun (n3).

One can move from a specific frame entry in the lexicon to its occurrences in the treebanks by running a query like the following[3]:

[3] This is possible if the frame entry in question is not one built in intuition-based fashion, in which case the “form” sub-node in the frame entry is assigned the value “typical”. Instead, the

```
t-node $t := [val_frame.rf v-frame $v
:= [ id = "v-w201_MPf2_MP"]]
```

This query searches in the tectogrammatical layer of annotation of the treebanks for all the nodes (`t-node $t`) that are assigned a valency frame reference (`val_frame.rf`) linking to the frame entry in LV provided with `id` equal to "`v-w201_MPf2_MP`", i.e. the frame entry shown in figure 4. Figure 5 presents one of the outputs of this query.

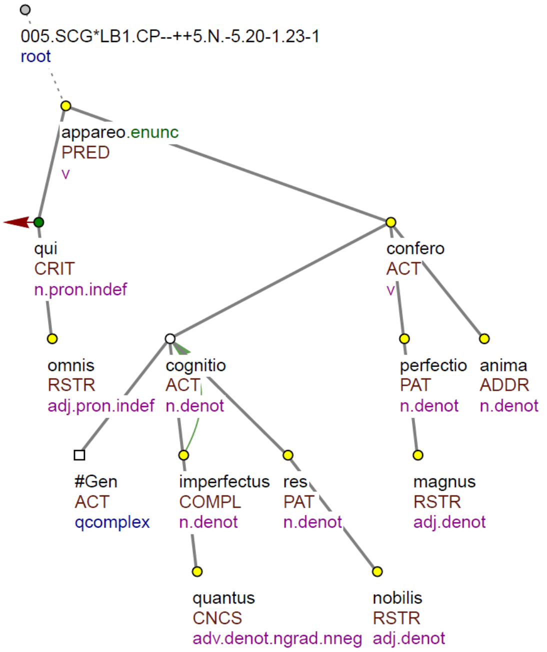

Figure 5: An occurrence of a frame entry

Figure 5 shows the tectogrammatical tree for the following sentence from the IT-TB (*Summa contra Gentiles*, book 1, chapter 5, number 5): "ex quibus omnibus apparet quod de rebus nobilissimis quantumcumque imperfecta cognitio maximam perfectionem animae confert" ("according to all these [things], it turns out that the cognition of the choicest things, though imperfect, confers maximum perfection to soul").

In this tree, the occurrence of the verb *confero* (*confert*) comes with an Actor represented by a nominative noun (*cognitio*), an accusative noun which is assigned functor Patient (*perfectionem*; node *perfectio*) and a dative noun playing the role of Addressee (*animae*; node *anima*).

Following the example reported in 3.2.2, we present a query that connects a frame entry of LV with its occurrences in the treebanks that feature a specific surface form. In 3.2.2 we discussed an occurrence of the AcI construction whose head verb is *continuo*. The entry for *continuo* in LV can be retrieved by running the following query, which searches for the `v-word` (word entry) whose attribute `lemma` has the value "`continuo`":

```
v-word [lemma = "continuo" ]
```

Figure 6 shows the LV entry for *continuo*. The id for the only frame entry for *continuo* informs that this is the entry number 508 of LV (w508) and that this is the first frame entry (f1) for *continuo*. The frame entry includes an Actor (ACT) expressed by a nominative pronoun (u1) and a Patient (PAT) expressed by an accusative noun (n4).

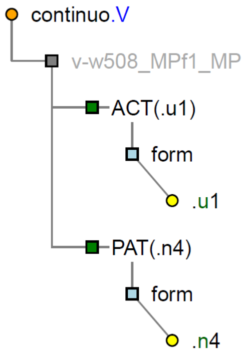

Figure 6. The LV entry for the verb *continuo*

Among the occurrences in the treebanks of the frame entry of *continuo* reported in figure 6, the following query searches for those where *continuo* heads an AcI construction.

```
  t-node $n0 :=
  [ val_frame.rf v-frame $n3 :=
    [ id = "v-w508_MPf1_MP" ],
    a/lex.rf a-node $n1 :=
    [  (m/tag ~ "^3..[HQ]" or m/tag ~
"^v...n"),  a-node $n2 :=
       [   afun  =  "Sb",   (m/tag  ~
"......[DM]" or m/tag ~ ".......a") ] ]
];
```

Figure 7 shows the same query in graphical form.

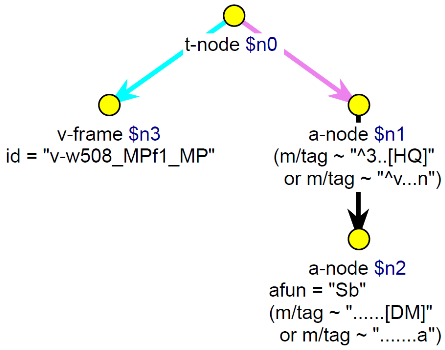

Figure 7. A graphical query on the treebanks

This query searches for a node of a tectogrammatical tree in the treebanks (`t-node $n0`) whose frame entry in

corpus-driven entries of LV are linked to all their occurrences in the tectogrammatical tree structures of the IT-TB and LDT.

LV has id equal to "v-w508_MPf1_MP" (v-frame $n3). The t-node $n0 corresponds to a node in the analytical layer of annotation of the treebanks (a-node $n1) whose morphological tags (m/tag) are those for infinitive verbs. The a-node $n1 heads another analytical node (a-node $n2), which is assigned the syntactic label for subjects (afun = "Sb") and it is a word inflected at the accusative case[4].

Figure 8 shows one of the outputs resulting from the query.

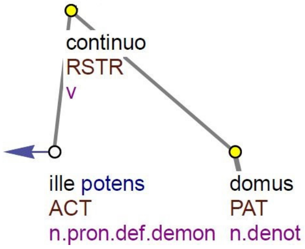

Figure 8. A subtree resulting from a query

Figure 8 shows the tectogrammatical subtree for the clause "illos […] domos continuare" (see 3.2.2). The node for the word *continuare* (*continuo* in the subtree) heads an Actor (ACT) whose surface form is an accusative pronoun (*illos*; node *ille*) and a Patient (PAT) expressed by an accusative noun (*domos*; node *domus*)[5].

Although the AcI construction is not reflected in the frame entries of LV, the treebanks' occurrences in AcI constructions of the LV word entries can always be retrieved.

4 Conclusion and Future Work

We presented *Latin Vallex*, a valency lexicon for Latin built in mutual connection with the semantic and pragmatic annotation of two Latin treebanks covering texts of different eras. On the one hand, such a connection between the empirical evidence provided by texts and the description of the lexicon allows to enhance each corpus-driven frame entry of LV with the frequency of its utterances in the treebanks. On the other hand, each valency-capable word in the treebanks is linked to a frame entry in LV.

In order to balance the representativity of LV, we enhanced the corpus-driven entries with a number of intuition-based ones. The relation between these two strategies represents one of the trickiest issues in building LV. Indeed, if a fully corpus-driven lexicon has the pro of both being empirically motivated and featuring a mutual relation with textual evidence, one con of such an approach is that texts could be not representative enough, possibly resulting in leaving out prototypical valency frames just because they do not occur in the reference texts.

Conversely, a fully intuition-based approach to building the lexicon risks to report just those frame entries that annotators believe are the most relevant ones for a specific word entry. But this is mostly based on annotators' knowledge of language, which is always something dangerous to deal with when an ancient language is concerned and no native speakers are available.

Thus, a steady confrontation with the evidence provided by more and more texts is needed, in order both to enlarge the coverage provided by the corpus-driven approach and to evaluate the quality of the contents of LV built in intuition-based fashion.

As mentioned in the Introduction, a valency lexicon has several applications in the area of NLP. In this respect, LV is part of a group of lexical resources for Latin that includes also the morphological analyzer LEMLAT (Passarotti, 2004), the syntactic-based subcategorization lexicon IT-VaLex and the Latin WordNet. In the near future, we must integrate all these resources in order to exploit at best the different kind of lexical information they provide in support of both NLP and theoretical linguistics purposes (Passarotti et al., 2015).

References

Bamman, D. and Crane, G. (2006). The design and use of a Latin dependency treebank. In J. Nivre and J. Hajič (Eds.), *Proceedings of the Fifth Workshop on Treebanks and Linguistic Theories (TLT2006)*. Prague, Czech Republic: ÚFAL, pp. 67--78.

Delatte, L., Evrard, E., Govaerts, S. and Denooz, J. (1981). *Dictionnaire fréquentiel et Index inverse de la langue latine*. Université de Liège: Laboratoire d'analyse statistique des langues anciennes.

Fillmore, C. (1982). *Frame semantics. Linguistics in the Morning Calm*. Seoul: Hanshin Publishing Co., pp. 111--137.

Hajič, J., Panevová, J., Urešová, Z., Bémová, A., Kolárová-Rezníckova, V. and Pajas, P. (2003). PDT-VALLEX: Creating a Large Coverage Valency Lexicon for Treebank Annotation. In J. Nivre and E. Hinrichs (Eds.), *TLT 2003 – Proceedings of the Second Workshop on Treebanks and Linguistic Theories*. Volume 9 of *Mathematical Modelling in Physics, Engineering and Cognitive Sciences*, Växjö, Sweden: Växjö University Press, pp. 57--68.

[4] Since the IT-TB and the LDT have different morphological tagsets, the query searches for two different sequences of tags both for infinitive verbs and for accusative words. In both cases, the first sequence (i.e. the one preceding the operator or in the query) makes use of the IT-TB tagset, while the second is built according to the LDT one. See http://itreebank.marginalia.it/view/documentation.php for the full documentation on the morphological tagsets of the IT-TB and LDT. Both the treebanks have recently been made available in the *Universal Dependencies* repository (http://universaldependencies.org/; McDonald et al. 2013) with a common tagset following the Google Universal PoS tagset (Petrov et al., 2012).

[5] As mentioned above, anaphora resolution is performed in the tectogrammatical layer. In the subtree of figure 8, the node for the pronoun *ille* links to that for the lemma *potens* ("powerful"), which occurs in the previous sentence of the text. This link is graphically represented by the arrow pointing left out from the node of *ille*. Such an annotation informs that these *illos* correspond to the previously mentioned "powerful people".

Happ, H. (1976). *Grundfragen einer Dependenz-Grammatik des Lateinischen*. Goettingen, Germany: Vandenhoeck & Ruprecht.

Kingsbury, P. and Palmer, P. (2002). From Treebank to Propbank. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)*. Las Palmas - Gran Canaria, Spain: ELRA, pp. 1989--1993.

Kohl, M., Wiese, S. and Warscheid, B. (2011). Cytoscape: software for visualization and analysis of biological networks. *Methods in Molecular Biology*, 696, pp. 291--303.

Korhonen A., Krymolowski, Y. and Briscoe, T. (2006). A Large Subcategorization Lexicon for Natural Language Processing Applications. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*. Genoa, Italy: ELRA, pp. 1015--1020.

McDonald, R.T., Nivre, J., Quirmbach-Brundage, Y, Goldberg, Y., Das, D., Ganchev, K., Hall, K., Petrov, S., Zang, H., Täckström, O., Bedini, C., Castelló, N.B. and Lee, J. (2013). Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*. Stroudsburg, PA, USA: ACL, pp. 92--97.

McGillivray, B. and Passarotti, M. (2009). The Development of the *Index Thomisticus* Treebank Valency Lexicon. In *Proceedings of LaTeCH-SHELT&R Workshop 2009*. Athens, Greece: ACL, pp. 43--50.

McGillivray, B. (2013). *Methods in Latin Computational Linguistics*. Leiden: Brill.

Messiant, C., Korhonen, A. and Poibeau, T. (2008). LexSchem: A Large Subcategorization Lexicon for French Verbs. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*. Marrakech, Morocco: ELRA, pp. 533--538.

Mikulová, M. et alii. (2005). *Annotation on the tectogrammatical layer in the Prague Dependency Treebank. The Annotation Guidelines*. Prague, Czech Republic: ÚFAL.

Minozzi, S. (2010). The Latin WordNet project. In P. Anreiter and M. Kienpointner (Eds.), *Latin Linguistics Today. Latin Linguistics Today. Akten des 15. Internationalen Kolloquiums zur Lateinischen Linguistik*. Innsbruck, Austria: Innsbrucker Beiträge zur Sprachwissenschaft, pp. 707--716.

Panevová, J. (1974-1975). On Verbal Frames in Functional Generative Description. Part I, *Prague Bulletin of Mathematical Linguistics*, 22, pp. 3--40; Part II, *Prague Bulletin of Mathematical Linguistics*, 23, pp. 17--52.

Passarotti, M. (2004). Development and perspectives of the Latin morphological analyser LEMLAT. In A. Bozzi, L. Cignoni and J.L. Lebrave (Eds.), *Digital Technology and Philological Disciplines. Linguistica Computazionale*, XX-XXI, pp. 397--414.

Passarotti, M. (2011). Language Resources. The State of the Art of Latin and the *Index Thomisticus* Treebank Project. In M.S. Ortola (Ed.), *Corpus anciens et Bases de données, «ALIENTO. Échanges sapientiels en Méditerranée», N°2*. Nancy, France: Presses universitaires de Nancy, pp. 301--320.

Passarotti, M., González Saavedra, B. and Onambélé Manga, C. (2015). Somewhere between Valency Frames and Synsets. Comparing Latin Vallex and Latin WordNet. In C. Bosco, S. Tonelli, and F.M. Zanzotto (Eds.), *Proceedings of the Second Italian Conference on Computational Linguistics (CLiC-it 2015)*. Torino, Italy: Academia University Press, pp. 221--225.

Petrov, S., Das, D., and McDonald, R. (2012). A universal part-of-speech tagset. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012)*. Istanbul, Turkey: ELRA, pp. 2089--2096.

Ruppenhofer, J., Ellsworth, M., Petruck, M.R.L., Johnson, C.R. and Scheffczyk. J. (2006). *FrameNet II. Extendend Theory and Practice*. E-book available at http://framenet.icsi.berkeley.edu/index.php?option=com_wrapper&Itemid=126.

Sgall, P., Hajičová, E. and Panevová, J. (1986). *The Meaning of the Sentence in its Semantic and Pragmatic Aspects*. Dordrecht, NL: D. Reidel.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B. and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13(11), pp. 2498--504.

Štěpánek, J. and Pajas, P. (2010). Querying Diverse Treebanks in a Uniform Way. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC 2010)*. Valletta, Malta: ELRA, pp. 1828--1835.

Urešová, Z. (2004). *The Verbal Valency in the Prague Dependency Treebank from the Annotator's Point of View*. Bratislava, Slovakia: Jazykovedný ústav Ľ. Štúra, SAV.

Language Resources References

CIRCSE Research Centre. (2016). *Index Thomisticus* Treebank (tectogrammatical layer). Project FIR-2013 “Building and Integrating Advanced Language Resources for Latin”, distributed via http://itreebank.marginalia.it.

CIRCSE Research Centre. (2016). *Latin Vallex*. Project FIR-2013 “Building and Integrating Advanced Language Resources for Latin”, distributed via http://itreebank.marginalia.it.

CIRCSE Research Centre. (2016). *Latin Dependency Treebank* (tectogrammatical layer). Project FIR-2013 “Building and Integrating Advanced Language Resources for Latin”, distributed via http://itreebank.marginalia.it (source data distributed via the Perseus Digital Library at https://perseusdl.github.io/treebank_data/).