VERSO UN LESSICO DI VALENZA DEL LATINO EMPIRICAMENTE MOTIVATO

Berta González Saavedra, *Marco* Passarotti

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 • Valence, Latin, Treeban

1. Introduzione

Nonostante una secolare tradizione lessicografica, la lingua latina manca ancora di risorse lessicali di tipo computazionale aggiornate allo stato dell'arte. Ciò è strettamente connesso alla limitata disponibilità di corpora testuali latini annotati linguisticamente, sulla cui base empirica possano essere costruite nuove risorse lessicali. Tuttavia, una serie di progetti mirati allo sviluppo di treebank a dipendenze per il latino è stata avviata nel corso dell'ultimo decennio. Egualmente, ha preso avvio la realizzazione di risorse lessicali fondamentali come *Latin WordNet* (Minozzi, 2010).

Questo articolo presenta Latin Vallex, un lessico di valenza per il latino costruito in connessione con l'annotazione semantico-pragmatica di due treebank latine comprensive di testi di epoche e generi diversi.

L'articolo è organizzato nel modo seguente. La sezione 2 riporta lo stato dell'arte relativo ai lessici di valenza, concentrandosi particolarmente su quelli per la lingua latina. La sezione 3 presenta *Latin Vallex*, descrivendone la struttura delle entrate lessicali e l'interrogazione dei dati. La sezione 4 conclude il lavoro e abbozza i prossimi sviluppi della risorsa.

2. Valenza e lessici

La nozione di valenza è generalmente intesa come il numero di complementi obbligatori richiesti da una parola: essi sono usualmente nominati 'argomenti', mentre i complementi non obbligatori sono detti 'aggiunte', o 'satelliti'. Il precursore dell'idea moderna di valenza è

considerato essere Lucien Tesnière in virtù del suo testo *Eléments de syntaxe structurale*, pubblicato postumo nel 1959 (Tesnière, 1959).

Approcciare la semantica lessicale attraverso il concetto di valenza è una pratica diffusa nella ricerca linguistica. Esso è, infatti, presente in diversi contesti teorici, tra i quali spicca la *Frame Semantics* di Charles Fillmore (1982), in base alla quale il significato di alcune parole può essere compreso appieno solo conoscendo gli elementi che fanno parte del *frame* evocato dalla parola stessa. Alcuni di questi elementi sono obbligatori (*core frame element*), mentre altri sono occasionali (*not core frame element*)[1].

In base al numero dei propri argomenti, le parole sono considerate essere zerovalenziali (ad esempio, *piovere*), monovalenziali (*camminare*), bivalenziali (*mangiare*), trivalenziali (*dare*), tetravalenziali (*spostare*) etc. Le posizioni argomentali possono venire arricchite con ruoli semantici, ossia etichette che specificano la relazione semantica che intercorre tra l'argomento e la parola di cui quest'ultimo è complemento (obbligatorio). Ad esempio, nel caso del verbo trivalenziale *dare* i ruoli semantici dei tre argomenti sono rispettivamente Agente (1), Paziente (2) e Destinatario (3): "(1) dà (2) a (3)".

I criteri a supporto della distinzione tra argomenti e aggiunte non sono rigidamente definiti e unanimemente accettati (al punto che la distinzione stessa è talvolta messa in dubbio); ciò richiede che essi vengano esplicitati ogni volta che la nozione di valenza è utilizzata e applicata a dati empirici (si veda 3.1).

Sono oggi disponibili per molte lingue descrizioni del lessico fondate sulla valenza. Esse sono rappresentate da risorse lessicali che recitano un ruolo importante nel trattamento automatico del linguaggio (TAL) grazie alla loro ampia applicabilità in ambiti come il *semantic role labeling*, la *word sense disambiguation*, l'acquisizione di restrizioni preferenziali e la realizzazione di treebank (Urešová, 2004).

Così come altre risorse linguistiche, anche i lessici di valenza possono essere costruiti in modalità *intuition-based* o *corpus-driven*, in base all'importanza del ruolo giocato dall'intuizione umana e dall'evidenza empirica estratta da corpora testuali nel corso della realizzazione della risorsa. Ad esempio, lessici come *PropBank* (Kingsbury & Palmer, 2002), *FrameNet* (Ruppenhofer et al., 2006) e *PDT-Vallex* (Hajič et al., 2003) sono stati inizialmente sviluppati in modalità *intuition-based* per venire successivamente controllati e perfezionati attraverso il confronto con dati estratti da corpora. Esempi di lessici acquisiti automaticamente dall'evidenza testuale sono, invece, *VALEX* (Korhonen et al., 2006) e *LexShem* (Messiant et al., 2008).

Mentre sono numerosi i lessici di valenza compilati per le lingue moderne, molto lavoro resta da fare per la realizzazione di risorse simili per le lingue antiche e, in particolare, per il latino e il greco. In merito al latino, Happ riporta una lista di verbi associati alle proprie valenze (Happ, 1976: 480-565). Bamman & Crane (2008), invece, descrivono un *dynamic lexicon* estratto automaticamente dalla Perseus Digital Library utilizzando la Latin Dependency Treebank (Bamman & Crane, 2006) come evidenza su cui addestrare un PoS tagger e un parser a dipendenze. Questa risorsa associa a ciascuna entrata lessicale informazione qualitativa e quantitativa in merito ai suoi modelli di sottocategorizzazione e alle sue restrizioni preferenziali. Infine, *IT-VaLex* (McGillivray & Passarotti, 2009) è un lessico di sottocategorizzazione le cui entrate (verbali) sono indotte automaticamente a partire dal livello di annotazione sintattica della *Index Thomisticus* Treebank (Passarotti, 2011). La medesima struttura di *IT-VaLex* è ripresa da un lessico sviluppato sulla base della Latin Dependency Treebank e descritto da McGillivray (2013: 31-60).

[1] Un'ampia descrizione del lessico basata sulla *Frame Semantics* è fornita dalla risorsa lessicale *FrameNet* sviluppata alla University of California (Berkeley, USA) dal gruppo di ricerca che faceva capo a Charles Fillmore: https://framenet.icsi.berkeley.edu/fndrupal/.

3. Latin Vallex

3.1 La struttura di Latin Vallex

Latin Vallex (LV) è stato realizzato congiuntamente allo sviluppo del livello di annotazione semantico-pragmatica di due treebank latine a dipendenze: la *Index Thomisticus* Treebank (IT-TB), che include testi di Tommaso d'Aquino, e la Latin Dependency Treebank (LDT), che riporta estratti di testi (sia in prosa che in poesia) dell'età classica. Ciascuna parola valenziale che occorre nella porzione delle due treebank annotata a livello semantico-pragmatico è associata a una *frame entry* in LV.

La struttura di LV richiama quella del lessico di valenza per la lingua ceca *PDT-Vallex*, prodotto nel contesto teorico della *Functional Generative Description* (FGD; Sgall et al., 1986). La FGD è la teoria linguistica che motiva lo stile di annotazione semantico-pragmatica anche delle treebank latine, corrispondente al cosiddetto livello 'tectogrammaticale' della *Prague Dependency Treebank* del ceco (PDT). Questo livello viene realizzato a partire da uno precedente di tipo sintattico, chiamato 'analitico', e include l'annotazione dei ruoli semantici e la risoluzione delle ellissi e delle anafore/catafore. Il *dialogue test* di Panevová (1974-1975) e i criteri descritti da Mikulová *et alii* (2005: 100-102, 116-162) sono utilizzati per distinguere tra argomenti e aggiunte. Sia il livello analitico che quello tectogrammaticale sono graficamente rappresentati attraverso alberi a dipendenze.

A livello macroscopico, il lessico è diviso in entrate lessicali. Un'entrata lessicale consiste in una sequenza di *frame entry*, ciascuna delle quali tende a corrispondere a uno dei sensi della parola in questione. La *frame entry* contiene una descrizione del *valency frame* e dei suoi attributi (*frame attribute*). Un *valency frame* è una sequenza di posizioni argomentali (*frame slot*), ciascuna delle quali rappresenta un complemento della parola ed è associata ad alcuni suoi tratti morfologici di superficie (con alcune deviazioni: si veda 3.2). Gli attributi corrispondono a nomi di ruoli semantici (chiamati *functor* in FGD) utilizzati per esprimere i tipi di relazioni semantiche che intercorrono tra la parola e i suoi complementi.

La struttura di un'entrata di LV può essere riassunta come segue:

Nome dell'entrata lessicale (lemma) – PoS

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

I ruoli semantici riportati nelle *frame entry* di LV sono quelli per gli argomenti (chiamati *inner participant*), che, in base alla FGD, corrispondono ai complementi a cui sono assegnati i seguenti *functor*: Agente (ACT[or]), Paziente (PAT[ient]), Destinatario (ADDR[essee]), Risultato (EFF[ect]) e Origine (ORIG[o]). Anche alcune aggiunte (*free modification*) possono rientrare nelle *frame entry*, venendo registrate come posizioni opzionali. L'insieme dei valori per i ruoli semantici utilizzato nella IT-TB e nella LDT è il medesimo descritto nel manuale di annotazione tectogrammaticale della PDT (Mikulová et al., 2005).

La sola differenza esistente tra LV e *PDT-Vallex* è conseguenza del fatto che il cosiddetto *argument shifting* non viene applicato nell'annotazione tectogrammaticale della IT-TB e della LDT. L'*argument shifting* (Mikulová et al., 2005: 103-105) è un criterio utilizzato per determinare il tipo di argomento e, dunque, assegnare il *functor* più confacente alla posizione argomentale in questione: esso stabilisce che al primo argomento va sempre assegnato il *functor* ACT, mentre il secondo argomento riceve sempre il *functor* PAT. Tutti gli altri *functor* argomentali (ADDR, EFF e ORIG) diventano (*to shift*) ACT e PAT nel caso in cui occorrano come primo o secondo argomento[2].

Ad esempio, in base all'*argument shifting*, se un verbo ha un argomento di tipo ORIG ma non include un PAT nel proprio *frame* argomentale, la posizione del PAT (seconda posizione del *frame*) viene occupata dall'argomento ORIG, a cui viene assegnato il *functor* PAT. Ciò si riflette nei dati di *PDT-Vallex*, in cui infatti non occorrono *frame entry* costituite da due posizioni i cui attributi siano, ad esempio, ACT e ORIG, in quanto all'argomento con attributo ORIG sarebbe assegnato il *functor* PAT proprio in virtù dell'*argument shifting*. Invece, una situazione del genere può accadere in LV, come risultato dell'annotazione tectogrammaticale della IT-TB e della LDT. Ad esempio, l'entrata del verbo *resulto* ("risultare") in LV include una *frame entry* con due posizioni, i cui attributi sono rispettivamente ACT e ORIG.

Un'occorrenza testuale di questa *frame entry* di *resulto* è nella seguente frase tratta dalla IT-TB (*Summa contra Gentiles*, libro 1, capitolo 27, numero 4):

(1) "ex unione formae et materiae resultat aliquid compositum" ("dall'unione della forma e della materia risulta qualcosa composto").

Nella frase (1), gli argomenti per il verbo *resultat* sono *aliquid* ("qualcosa") ed *ex unione* ("dall'unione"). Alla parola *aliquid* viene assegnato il *functor* ACT, mentre *ex unione* ha *functor* ORIG. Se l'*argument shifting* fosse stato applicato, *ex unione* avrebbe ricevuto functor PAT[3].

Come anticipato, oltre ai *functor* per gli argomenti, anche alcuni *functor* per le aggiunte possono essere presenti nelle *frame entry*. Questi *functor* sono registrati come opzionali. I *functor* opzionali più frequenti nelle *frame entry* sono quelli di tipo spaziale e direzionale, che sono per lo più impiegati per i verbi di movimento (Mikulová et al., 2005: 503-514). Ad esempio, la *frame entry* prototipica del verbo *venio* ("venire") include tre posizioni, i cui *functor* sono rispettivamente ACT, DIR1 (Direzione-Da) e DIR3 (Direzione-A).

[2] Il livello di annotazione tectogrammaticale della IT-TB e della LDT non include l'*argument shifting*, in quanto nelle treebank latine questo livello è inteso come meno direttamente connesso alla struttura sintattica rispetto alla PDT e più orientato all'analisi semantica.

[3] In base a una interpretazione più agentiva o più risultativa del soggetto sintattico del verbo *resulto*, alle sue posizioni argomentali potrebbero venire assegnati rispettivamente i *functor* ACT ed EFF invece che ACT e ORIG. Nell'esempio discusso, *aliquid* riceverebbe *functor* EFF, mentre (*ex*) *unione* sarebbe ACT: *aliquid* sarebbe così inteso come il risultato dell'azione rappresentata dal nome deverbale *unio*. Seppur semanticamente motivata, questa interpretazione scarta, tuttavia, dallo stile di annotazione tectogrammaticale della PDT e ad essa è dunque preferita quella che fa uso dei *functor* ACT e ORIG.

Un altro esempio è l'entrata del verbo *termino* che, in base alla struttura delle entrate lessicali di LV, include due *frame entry*, corrispondenti a due diversi sensi della parola: (a) "marcare il confine di qualcosa" e (b) "limitare qualcosa a qualcos'altro". La *frame entry* per il primo senso è costituita da un *valency frame* con due posizioni, la prima delle quali è rappresentata da un nome al nominativo (n1) mentre la seconda è un nome all'accusativo (n4). Gli attributi di queste due posizioni sono rispettivamente ACT e PAT. La *frame entry* per il secondo senso, invece, consiste in un *valency frame* con tre posizioni: un nome al nominativo (n1), un nome all'accusativo (n4) e un sintagma preposizionale governato dalla preposizione *in* (*in*+n4). I *functor* per questi tre argomenti sono rispettivamente ACT, PAT e DIR3.

L'entrata lessicale di *termino* in LV appare, dunque, come segue.

termino – V

- Frame Entry 1 ("marcare il confine di qualcosa"):
  - Valency Frame:
    - Frame slot 1: n1
    - Frame slot 2: n4
  - Frame Attribute:
    - Functor 1: ACT
    - Functor 2: PAT
- Frame Entry 2 ("limitare qualcosa a qualcos'altro"):
  - Valency Frame:
    - Frame slot 1: n1
    - Frame slot 2: n4
    - Frame slot 3: in+n4
  - Frame Attribute:
    - Functor 1: ACT
    - Functor 2: PAT
    - Functor 3: DIR3

I tratti morfologici (PoS e caso) riportati nei frame slot risultano dal confronto con l'evidenza testuale fornita dalle due treebank latine sulla cui base LV è costruito.

3.2 La realizzazione di Latin Vallex

Ciascuna parola valenziale che gli annotatori incontrano nel corso dell'annotazione tectogrammaticale della IT-TB e della LDT viene collegata a una *frame entry* di LV[4]. Le parole valenziali possono essere verbi (*do* - "dare"), aggettivi (*contrarius* - "contrario"), nomi (*descriptio* - "descrizione") e avverbi (*similiter* - "similmente")[5].

Al momento, LV include 1.373 entrate lessicali e 3.406 *frame entry*: nello specifico, 1.049 verbi (2.903 *frame*), 236 nomi (394 *frame*), 86 aggettivi (106 *frame*) e 2 avverbi (3 *frame*). Queste entrate risultano dall'annotazione tectogrammaticale delle prime 2.000 frasi della *Summa contra Gentiles* di Tommaso d'Aquino (IT-TB), dell'intera *De Catilinae coniuratione* di Sallustio (701 frasi) e di 100 frasi tratte dal *De bello gallico* di Cesare e dalle *Orationes in Catilinam* di Cicerone (LDT)[6].

Dal momento che la IT-TB e la LDT non sono bilanciate in modo tale da essere sufficientemente rappresentative della lingua latina (o di una sua specifica variante), nel corso della realizzazione di LV alle entrate lessicali realizzate in modalità *corpus-driven* è stata accostata una serie di entrate prodotte per intuizione. Nello specifico, per fini di rappresentatività della risorsa, LV include le entrate di tutte le parole valenziali presenti tra le 1.000 più frequenti della lingua latina riportate da Delatte *et alii* (1981). Benché la maggior parte di queste parole siano già presenti in LV, in quanto risultanti dal lavoro di annotazione tectogrammaticale, 163 di esse non sono state ancora trovate nei testi annotati. Dunque, le entrate per queste parole sono state inserite in LV sulla base dell'intuizione del lessicografo (e non sono, dunque, collegate ad alcuna occorrenza nelle treebank), che assegna ad esse quella che è considerata essere la loro *frame entry* prototipica, registrando solo i *functor* per le posizioni argomentali e non anche i tratti morfologici, indisponibili in quanto risultanti dal confronto con l'evidenza fornita dalle treebank. Benché la maggior parte delle entrate *intuition-based* di LV riceva una sola *frame entry* (prototipica), ci sono casi in cui più di una *frame entry* viene loro assegnata: in totale, il numero di *frame entry* per queste 163 entrate è 168. Nessuna di queste *frame entry* è connessa ad alcuna occorrenza testuale nelle treebank fino al momento in cui gli annotatori non ne incontrino la prima nel corso del lavoro di annotazione: a quel punto, la *frame entry* viene modificata in base all'evidenza empirica, aggiungendo i tratti morfologici.

La figura 1 mostra una porzione dell'albero tectogrammaticale della seguente frase tratta dalla IT-TB (*Summa contra Gentiles*, libro 1, capitolo 5, numero 2):

[4] Essendo le treebank e le *frame entry* di LV biunivocamente collegate (tranne che per le 163 entrate prodotte in modalità *intuition-based* descritte più avanti nella sezione), le modifiche eventualmente apportate a una risorsa vengono automaticamente applicate anche all'altra. Una serie di query scritte in linguaggio SQL (*Structured Query Language*) consentono altresì di indurre alcune *frame entry* di LV a partire dall'annotazione tectogrammaticale delle treebank. Tuttavia, questo metodo di creazione del lessico di valenza non è esaustivo in quanto ancora manchevole nell'estrazione dell'informazione da strutture orizzontali, come ad esempio le costruzioni coordinate. Le query SQL sono al momento utilizzate per fini di verifica della qualità di LV, confrontando il loro risultato con quanto prodotto dagli annotatori.

[5] L'annotazione tectogrammaticale viene prodotta da due annotatori, che lavorano in fasi diverse. Il primo costruisce l'albero tectogrammaticale di una frase sulla base dell'output di una serie di script di conversione automatica dal livello analitico a quello tectogrammaticale (González Saavedra & Passarotti, 2014). Il secondo annotatore verifica il lavoro del primo. I casi di disaccordo tra le scelte dei due annotatori vengono raccolti e discussi regolarmente, risultando in periodiche revisioni delle regole di annotazione descritte in Mikulová *et alii* (2005).

[6] Le frasi tratte dai testi di Cesare e di Cicerone sono le prime 100 disponibili di essi nella LDT.

(2) "[…] christianae religioni […], quae singulariter bona spiritualia et aeterna promittit" ("[…] alla religione cristiana […], che unicamente promette beni spirituali ed eterni").

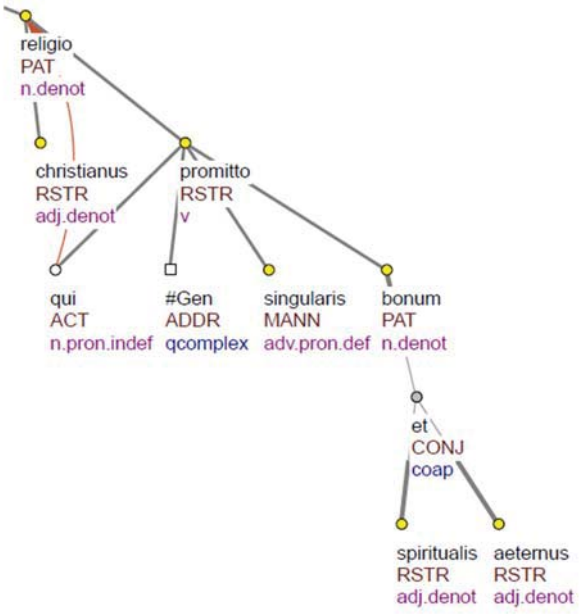

Figura 1: Un albero tectogrammaticale[7]

Nel corso della costruzione della porzione di albero tectogrammaticale riportato nella figura 1, gli annotatori incontrano una occorrenza della parola valenziale *promitto* ("promettere") e la associano alla corrispondente *frame entry* in LV o costruiscono la *frame entry* ex novo, qualora essa non sia già presente nella risorsa.

La *frame entry* per questa occorrenza di *promitto* include tre posizioni argomentali. La prima è rappresentata da un pronome al caso nominativo (*quae*; nodo *qui*). La seconda è un nome all'accusativo (*bona*; nodo *bonum*). La terza posizione risulta dalla risoluzione dell'ellissi di un argomento che non può essere identificato contestualmente ed è, quindi, considerato un argomento "generico" (#Gen): dal momento che questo argomento non ha una realizzazione testuale, ad esso non è associato alcun tratto morfologico (PoS e caso) nel *frame slot*. I *functor* per i tre *slot* sono rispettivamente i seguenti: ACT, PAT e ADDR.

Oltre ai tre nodi che entrano a far parte della *frame entry*, nella porzione di albero riportata nella figura 1, *promitto* governa anche un quarto nodo, che corrisponde alla parola *singulariter* nella frase (nodo *singularis*) e riceve *functor* MANN[er] (Modo). Questo nodo non rientra nella *frame entry* di *promitto* perché il *functor* MANN è assegnato ad aggiunte non riportate nelle entrate lessicali di LV.

Più del 60% delle *frame entry* di LV presentano un *valency frame* con due posizioni argomentali. Nella maggior parte di questi *valency frame*, i ruoli semantici sono ACT e PAT. Il

[7] Nella usuale visualizzazione degli alberi tectogrammaticali, le forme di parola sono sostituite dal corrispondente lemma. Ad esempio, *qui* è il lemma della forma *quae*. Per le specifiche sui *functor* che occorrono negli alberi riportati nelle Figure 1 e 5, si rimanda a Mikulová *et alii* (2005).

secondo più frequente tipo di *valency frame* in LV (circa il 20% del totale dei *frame*) include tre argomenti. Diversamente dalle *frame entry* bivalenziali, quelle trivalenziali presentano una configurazione piuttosto varia dei ruoli semantici.

La figura 2 rappresenta le *frame entry* trivalenziali di LV attraverso un network indotto automaticamente dalla risorsa[8]. I nodi colorati in rosso sono quelli per i *functor*, mentre quelli in bianco sono per le *frame entry*, il cui nome è formato dal lemma dell'entrata lessicale a cui appartengono e una lettera assegnata alla specifica *frame entry*. Ad esempio, il nodo nominato *amo*-a corrisponde alla *frame entry* 'a' del lemma *amo* ("amare"). Nel network, un nodo per un *functor* è connesso a un nodo per una *frame entry* attraverso un ramo se quel *functor* è presente in almeno una posizione argomentale della *frame entry*.

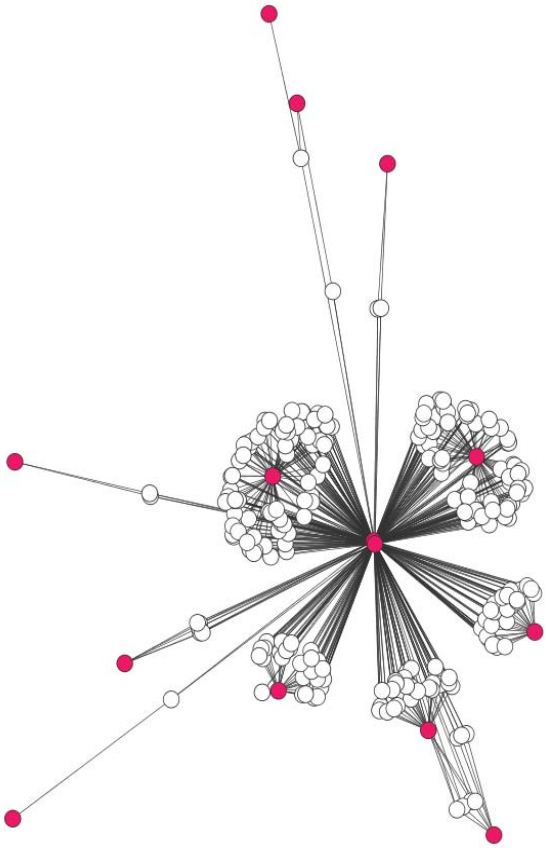

Figura 2: Network delle *frame entry* trivalenziali di *Latin Vallex*

Il centro della figura 2 è occupato da due nodi rossi, che corrispondono ai *functor* ACT e PAT: la maggior parte dei nodi per le *frame entry* sono connessi ad essi. Ciò significa che la maggior parte delle *frame entry* trivalenziali in LV hanno un ACT e un PAT tra i propri *functor*.

Intorno al centro del network sono visibili cinque gruppi principali di nodi. Ruotando in senso orario a partire dal gruppo in alto a sinistra, essi sono rispettivamente i gruppi per i *functor*

[8] Il network è stato creato con il software *Cytoscape* (Shannon et al., 2003; Saito et al., 2012). Nodi e relazioni sono organizzati in base al *setting* nominato *Edge-weighted Spring Embedded* (Kohl et al., 2011).

ADDR, (il più numeroso), EFF, ORIG, DIR3 e LOC[ative] (Locativo). Questi sono i *functor* assegnati più frequentemente alla terza posizione argomentale delle *frame entry* trivalenziali (le prime due posizioni essendo occupate da ACT e PAT rispettivamente). Dunque, ad esempio, i nodi raccolti attorno al nodo per il *functor* ADDR rappresentano quelle *frame entry* trivalenziali i cui ruoli semantici sono ACT, PAT e ADDR (come nel caso del verbo *attribuo* - "attribuire").

I nodi più periferici nel network rappresentano, invece, *functor* assegnati alla terza posizione argomentale con bassa frequenza. Ad esempio, il nodo per il *functor* ACMP (*Accompaniment* - Accompagnamento), posizionato in basso a sinistra nel network, è connesso a soli tre nodi, che corrispondono alle *frame entry* (del tipo ACT-PAT-ACMP) per i verbi *admisceo* ("mischiare"), *coniungo* ("congiungere") e *unio* ("unire").

I tratti morfologici riportati nei *frame slot* non riflettono i dati testuali nel caso di tre tipi di costruzioni: (a) proposizioni passive, (b) proposizioni infinitive e (c) ablativo assoluto. Ciò è motivato da due ragioni. La prima è legata alla necessità di non moltiplicare eccessivamente il numero di *frame entry*, raccogliendo più forme testuali in una *frame entry* comune. La seconda dipende dal fatto che LV è strettamente legato al livello di annotazione tectogrammaticale delle treebank: questo livello ha il fine di rappresentare la sintassi soggiacente di una frase (*underlying syntax* in FGD), altresì considerata il suo significato letterale, attraverso un modello il più possibile indipendente da quello della sintassi di superficie.

Benché i tratti morfologici delle posizioni argomentali di queste tre costruzioni non siano registrati nelle entrate di LV riflettendo in pieno il livello testuale, essi possono sempre essere estratti dal livello morfologico di annotazione delle treebank. Più precisamente, i *frame slot* per le occorrenze di parole valenziali nelle tre costruzioni menzionate sono costruiti come descritto nelle sezioni che seguono.

3.2.1 Proposizioni passive

Le proposizioni passive vengono trasformate nella corrispondente forma attiva prima di assegnare una *frame entry* al loro verbo principale (o costruirne una *frame entry* ex novo).

Si prenda ad esempio la seguente frase tratta dalla IT-TB (*Summa contra Gentiles*, libro 1, capitolo 1, numero 2):

(3) "sapientes dicantur qui res recte ordinant" ("[che] saggi siano detti coloro che ordinano le cose in modo retto").

Nella frase (3), il verbo principale (*dicantur*) è alla forma passiva. La *frame entry* di LV per il lemma *dico* ("dire") cui questa occorrenza della forma *dicantur* è connessa riflette la forma attiva della frase: "che [un soggetto generico] dica saggi coloro che ordinano le cose in modo retto". Dunque, la *frame entry* associata a questa occorrenza del lemma *dico* include un *valency frame* con tre posizioni argomentali cui sono associati i seguenti attributi:

(1) un ACT generico, ossia non espresso testualmente e non identificabile dal contesto;

(2) un PAT rappresentato da un pronome: "coloro (che ordinano le cose in modo retto)";

(3) un EFF, che è il *functor* assegnato ai complementi predicativi obbligatori: "saggio".

Questa soluzione consente di assegnare la medesima *frame entry* all'occorrenza del verbo *dico* in una frase come quella d'esempio indipendentemente dal fatto che essa sia espressa in forma attiva o passiva.

3.2.2 Proposizioni infinitve

In latino, una proposizione infinitiva (nota come "accusativo con l'infinito": AcI) è una costruzione formata da un verbo all'infinito il cui soggetto è flesso al caso accusativo.

La *frame entry* di LV per un AcI riflette l'equivalente costruzione con la forma verbale finita. Nelle costruzioni attive, l'ACT di un AcI riceve, dunque, il caso nominativo (invece dell'accusativo, che appare a livello testuale); ciò avviene anche in quelle passive, ma a fronte della previa trasformazione della costruzione da passiva ad attiva (secondo quanto detto in 3.2.1).

Ad esempio, si consideri la seguente frase tratta dalla LDT (*De Catilinae coniuratione*, XX):

(4) "quis mortalium […] tolerare potest […] illos binas aut amplius domos continuare […]?" ("chi tra i mortali può tollerare che loro costruiscano di seguito due o più palazzi?").

Nella frase (4), la parola *illos* ("loro") è un pronome all'accusativo plurale che recita il ruolo di soggetto della forma verbale all'infinito *continuare* ("assommare", "costruire di seguito").

La *frame entry* associata a questa occorrenza del verbo *continuo* include due posizioni argomentali:

(1) un ACT espresso da un pronome al nominativo: *illos* (accusativo) → *illi* (nominativo);

(2) un PAT rappresentato da un nome all'accusativo: *domos* ("palazzi").

In questo modo, la medesima *frame entry* viene assegnata all'occorrenza del verbo *continuo* indipendentemente dal fatto che essa compaia in un AcI o in una costruzione con il verbo in forma finita, quest'ultima essendo usualmente rappresentata da una proposizione introdotta da una congiunzione subordinativa, come ad esempio *quod* ("che").

3.2.3 Ablativo assoluto

L'ablativo assoluto è una costruzione in cui un nome e un participio formano una proposizione che è disgiunta rispetto alla struttura sintattica del resto della frase in cui occorre. Sia il nome che il participio sono flessi al caso ablativo; il nome ha il ruolo di soggetto del participio.

Nell'assegnare le *frame entry* di LV, gli ablativi assoluti sono trattati nello stesso modo delle proposizioni passive e degli AcI. Nella *frame entry*, il nome che funge da soggetto riceve *functor* ACT e caso nominativo nel caso di un ablativo assoluto alla forma attiva (ossia con il participio al tempo presente). Invece, in caso di ablativo assoluto passivo (participio perfetto), prima il participio è trasformato nella forma attiva e, poi, il nome che aveva originariamente il ruolo di soggetto riceve *functor* PAT e caso accusativo[9].

[9] Rispetto all'attribuzione del *functor* PAT al nome in ablativo con ruolo di soggetto possono esserci eccezioni, come nel caso del verbo *doceo* ("insegnare"), dove il *functor* non è PAT ma ADDR in conseguenza della sua costruzione con il doppio accusativo.

Ad esempio, si veda la seguente frase tratta dalla IT-TB (*Summa contra Gentiles*, libro 1, capitolo 43, numero 10):

(5) "[…] qualibet quantitate finita data […]" ("essendo stata data una qualsiasi quantità finita").

La parola *data* è un participio perfetto del verbo *do* ("dare"). Essendo passivo, l'ablativo assoluto viene innanzitutto reso alla forma attiva ("avendo [un soggetto generico] dato una qualsiasi quantità finita"); quindi, il nome che fa da soggetto del participio (*quantitate*) riceve il caso accusativo nella *frame entry*.

La *frame entry* associata a questa occorrenza di *do* include tre posizioni argomentali:

(1) un ACT generico;

(2) un PAT espresso da un nome all'accusativo: *quantitate* (ablativo) → *quantitatem* (accusativo);

un ADDR generico.

3.3 Interrogare Latin Vallex

Latin Vallex ed entrambe le treebank del latino sono liberamente disponibili presso il sito Internet della IT-TB (http://itreebank.marginalia.it/view/download.php) [10] e accessibili attraverso un'implementazione del motore di ricerca PML-TQ (*Prague Markup Language – Tree Query*) (Štěpánek & Pajas, 2010; http://itreebank.marginalia.it/view/resources.php).

Le query vanno dapprima scritte in un apposito box secondo il formato del linguaggio PML-TQ; successivamente, la corrispondente forma grafica può essere visualizzata in un grafo ad albero. Ad esempio, la seguente query cerca tutte le *frame entry* (`v-frame`) che abbiano (`child`) una posizione argomentale (`v-element`) rappresentata da un Destinatario (`functor = "ADDR"`):

```
v-frame [ child v-element [ functor = "ADDR" ] ]
```

La figura 3 mostra la forma grafica di una query disegnata secondo il linguaggio PML-TQ. La query cerca quelle entrate lessicali di LV (nodo $n0) che includano una *frame entry* ($n1) con almeno tre posizioni argomentali, cui siano associati rispettivamente i seguenti *functor*: ADDR ($n4), PAT ($n2) e ACT ($n3). Inoltre, la query impone che l'argomento con *functor* ADDR sia rappresentato da una parola flessa al caso dativo (*case* = "3").

[10] Della LDT è accessibile la porzione annotata a livello tectogrammaticale e la corrispondente controparte di livello analitico. L'intera LDT è disponibile presso la Perseus Digital Library: https://perseusdl.github.io/treebank_data/.

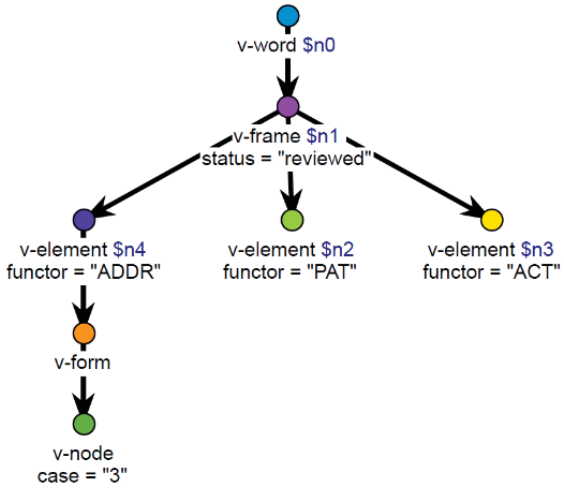

Figura 3: Una query grafica di PML-TQ su *Latin Vallex*

La figura 4 presenta uno degli output prodotti dalla query sopra descritta. In particolare, riporta una *frame entry* del verbo *confero* (“conferire”).

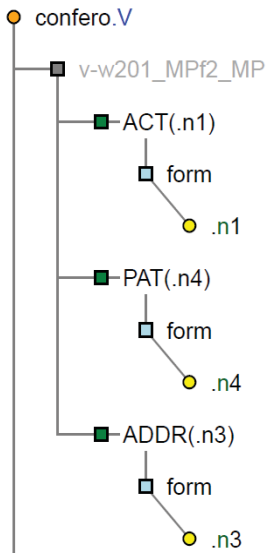

Figura 4: Una *frame entry* del verbo *confero*

In accordo con quanto stabilito nella query, questa *frame entry* include tre posizioni argomentali: un Agente, un Paziente e un Destinatario. Le posizioni argomentali sono ulteriormente specificate dalla PoS e dal caso (nodi “form”): l’Agente è un nome al nominativo (n1), il Paziente è un nome all’accusativo (n4) e il Destinatario è un nome al dativo (n3).

Data la connessione biunivoca tra il lessico e i dati delle treebank, è possibile spostarsi da una specifica *frame entry* di LV alle sue occorrenze nei dati testuali attraverso una query come la seguente[11]:

t-node $t := [val_frame.rf v-frame $v := [id = “v-w201_MPf2_MP”]]

Questa query cerca nel livello tectogrammaticale di annotazione delle treebank quei nodi (`t-node $t`) cui sia associato un identificativo del *valency frame* (`val_frame.rf`) che connetta il nodo tectogrammaticale con la *frame entry* di LV che ha id `“v-w201_MPf2_MP”`, ossia la *frame entry* riportata nella figura 4. La figura 5 presenta uno degli output di questa query.

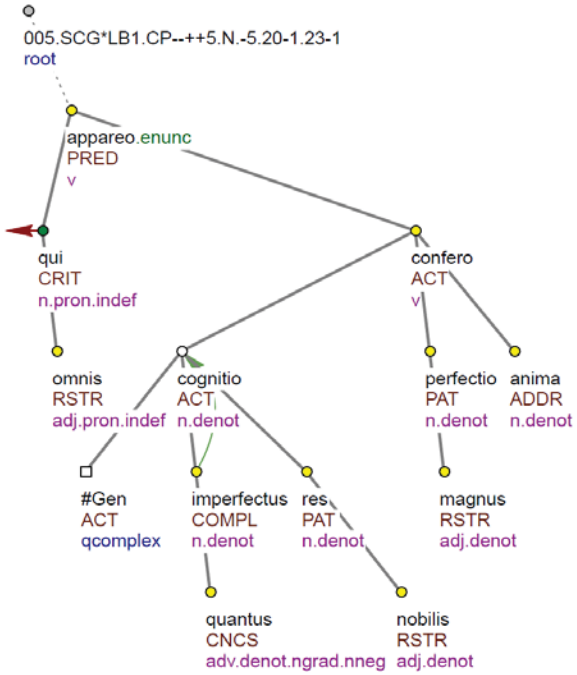

Figura 5: Una occorrenza testuale di una *frame entry*

La figura 5 mostra l’albero tectogrammaticale della seguente frase della IT-TB (*Summa contra Gentiles*, libro 1, capitolo 5, numero 5):

(6) “ex quibus omnibus apparet quod de rebus nobilissimis quantumcumque imperfecta cognitio maximam perfectionem animae confert” (“in base a tutte le qual cose, risulta che la conoscenza delle cose più nobili, per quanto imperfetta, conferisce la massima perfezione all’anima”).

[11] Ciò è possibile a condizione che la *frame entry* in questione non rientri tra quelle costruite in modalità *intuition-based*, nel qual caso i nodi “form” della *frame entry* ricevono il valore “typical” e la *frame entry* non è connessa ad alcuna occorrenza nelle treebank.

In questo albero, l'occorrenza del verbo *confero* (*confert*) governa un Agente rappresentato da un nome al nominativo (*cognitio*), un nome all'accusativo come Paziente (*perfectionem*; nodo *perfectio*) e un nome al dativo con il ruolo di Destinatario (*animae*; nodo *anima*).

Richiamando l'esempio discusso in 3.2.2, la prossima query collega una *frame entry* di LV con sue occorrenze testuali che presentano specifiche proprietà morfologiche. L'esempio in questione concerne una occorrenza di un AcI la cui testa verbale è una forma del verbo *continuo*. L'entrata lessicale di *continuo* in LV può essere richiamata attraverso la seguente query, che cerca la `v-word` (entrata lessicale) il cui attributo `lemma` ha valore `"continuo"`:

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

La figura 6 mostra l'entrata di *continuo*. L'identificativo dell'unica *frame entry* associata a *continuo* informa che questa è l'entrata numero 508 di LV (w508) e che quella riportata è la sua prima (e unica) *frame entry* (f1). La *frame entry* include un Agente espresso da un pronome al nominativo (u1) e un Paziente rappresentato da un nome all'accusativo (n4).

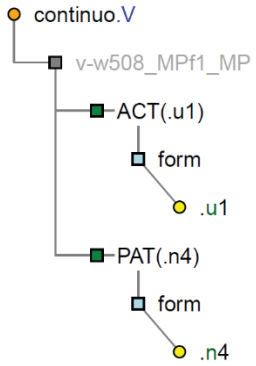

Figura 6. L'entrata lessicale del verbo *continuo*

Tra le occorrenze testuali della *frame entry* di *continuo* riportata nella figura 6, la query seguente cerca quelle dove *continuo* governa una costruzione AcI.

```
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") ] ] ];
```

La figura 7 presenta la medesima query in formato grafico.

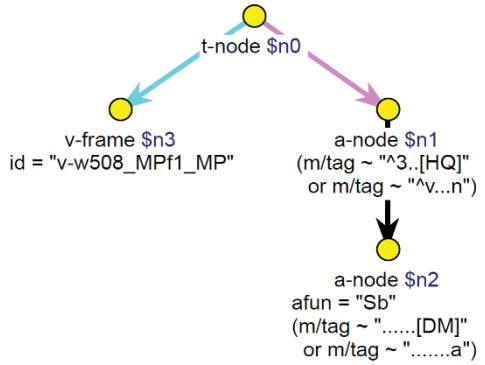

Figura 7. Una query sulle treebank in formato grafico

Questa query cerca nelle treebank i nodi tectogrammaticali (`t-node $n0`) la cui *frame entry* in LV abbia `id` uguale a `"v-w508_MPf1_MP"` (`v-frame $n3`). Il t-node $n0 corrisponde a un nodo nel livello analitico di annotazione delle treebank (`a-node $n1`) i cui codici morfologici (`m/tag`) sono quelli per le forme verbali al modo infinito. L'a-node $n1 governa un altro nodo analitico (`a-node $n2`), che ha funzione sintattica di soggetto (`afun = "Sb"`) ed è una parola flessa al caso accusativo[12].

La figura 8 mostra uno degli output di questa query.

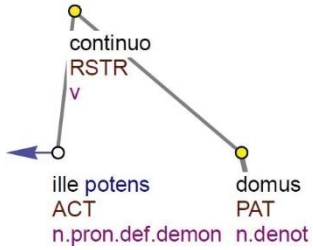

Figura 8. Una porzione di albero risultante da una query

La figura 8 presenta la porzione di un albero tectogrammaticale corrispondente alla proposizione "illos […] domos continuare" (si veda la frase (4) in 3.2.2). Il nodo per la parola *continuare* (*continuo* nella porzione di albero in figura) governa un Agente rappresentato da un

[12] Dal momento che la IT-TB e la LDT usano *tagset* morfologici diversi, la query cerca due diverse sequenze di codici sia per i verbi all'infinito che per le parole accusativo. In entrambi i casi, la prima sequenza di codici nella query (ossia quella che precede l'operatore or) usa il *tagset* della IT-TB, mentre la seconda è costruita in base a quello della LDT. La documentazione relativa ai *tagset* della IT-TB e della LDT è disponibile a questa URL: http://itreebank.marginalia.it/view/documentation.php. Entrambe le treebank sono state recentemente rese disponibili nell'ambito del progetto *Universal Dependencies* (http://universaldependencies.org/; McDonald et al. 2013) con codici morfologici comuni in accordo con il *Google Universal PoS tagset* (Petrov et al., 2012).

pronome all'accusativo (*illos*; nodo *ille*) e un Paziente espresso da un nome all'accusativo (*domos*; nodo *domus*)[13].

Dunque, benché la costruzione AcI non sia riportata come tale nelle *frame entry* di LV, le occorrenze testuali in forma di AcI delle entrate lessicali di LV possono sempre essere recuperate.

4. Conclusioni

In questo articolo abbiamo presentato *Latin Vallex*, un lessico di valenza per il latino costruito in connessione con l'annotazione semantico-pragmatica di due treebank latine che includono testi di epoche e generi diversi. Da un lato, questa connessione tra l'evidenza testuale e la descrizione lessicale consente di assegnare a ciascuna *frame entry* di LV (costruita in modalità *corpus-driven*) la frequenza delle sue occorrenze nelle treebank. Dall'altro, ogni parola valenziale che occorre nei testi delle treebank è connessa a una *frame entry* in LV.

Al fine di rendere LV sufficientemente rappresentativo del lessico latino, abbiamo inserito in esso anche un certo numero di entrate costruite in modalità *intuition-based*. La relazione tra le due strategie di realizzazione delle entrate lessicali è uno degli aspetti più delicati nello sviluppo di LV. Infatti, se un lessico completamente *corpus-driven* ha il pregio di essere empiricamente motivato in virtù di una mutua relazione con l'evidenza testuale che lo supporta, uno svantaggio di questo approccio consiste nel fatto che i testi possono non essere sufficientemente rappresentativi della lingua oggetto, essendo possibile che *frame* valenziali prototipici non compaiano nel lessico semplicemente perché essi non occorrono nei testi usati come base empirica dello stesso.

D'altra parte, un lessico realizzato in modalità totalmente *intuition-based* presenta il rischio di includere solo quei *frame* valenziali che il lessicografo crede essere i più prototipici per una specifica parola. Ciò è particolarmente problematico nel momento in cui si ha a che fare con una lingua antica e, dunque, non si hanno a disposizione parlanti nativi.

Dunque, è necessario un confronto serrato con l'evidenza fornita da un numero sempre maggiore di testi, sia per aumentare la copertura lessicale realizzata in modalità *corpus-driven* che per valutare la qualità dei contenuti di LV costruiti sulla base dell'intuizione del lessicografo.

Come menzionato nell'Introduzione, un lessico di valenza può avere numerose applicazioni nell'area del TAL. In questo senso, LV rientra in un gruppo di risorse lessicali per il latino di cui fanno parte anche l'analizzatore morfologico *LEMLAT* (Passarotti, 2004), il lessico di sottocategorizzazione sintattica *IT-VaLex* e *Latin WordNet*. La nostra speranza è di poter integrare tutte queste risorse al fine di sfruttare al meglio i diversi tipi di informazione lessicale che esse portano a supporto sia del TAL che di ricerche di tipo linguistico teorico.

BILIOGRAFIA

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 Treebank 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.

[13] Come detto, il livello tectogrammaticale di annotazione include anche la risoluzione delle anafore/catafore. Nella figura 8, il pronome *ille* rimanda al lemma *potens* ("[le persone] potenti"), che occorre nella precedente frase del testo. Questa connessione è graficamente rappresentata dalla freccia che punta a sinistra rispetto al nodo di *ille*.

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

González Saavedra, B. and Passarotti, M. (2014). Challenges in Enhancing the *Index Thomisticus* Treebank with Semantic and Pragmatic Annotation. In V. Enrich, E. Hinrichs, D. De Kok, P. Osenova and A. Prepiórkowski (Eds.), *Proceedings of the Thirteenth International Workshop on Treebanks and Linguistic Theories (TLT-13)*. Tübingen, Germany: Department of Linguistics, University of Tübingen, Germany, pp. 265--270.

Hajič, J., Panevová, J., Urešová, Z., Bémová, A., Kolárová-Reznícková, 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 Treebank and Linguistic Theories*. Volume 9 of *Mathematical Modelling in Physics, Engineering and Cognitive Sciences*, Växjö, Sweden: Växjö University Press, pp. 57--68.

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.

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 Treebank in a Uniform Way. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC 2010)*. Valletta, Malta: ELRA, pp. 1828--1835.

Tesnière, L. (1959). *Éléments de syntaxe structural*. Paris, France: Editions Klincksieck.

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.

BERTA GONZÁLEZ SAAVEDRA • graduated in Classical Philology at the Universidad Complutense of Madrid (Spain). She has been working at the Index Thomisticus Treebank since 2015, collaborating with Marco Passarotti in the annotation of Latin texts.She also graduated in Italian Philology at the Universitat of València (Spain). Her research fields are Historical Linguistics, Semantics and Indo-European Linguistics.

E-MAIL • berta.gonzalezsaavedra@unicatt.it

MARCO PASSAROTTI • is researcher at Università Cattolica del Sacro Cuore (Milan) in the area of computational linguistics. A pupil of one of the pioneers of humanities computing, father Roberto Busa SJ, his main research interests deal with developing and disseminating language resources and NLP tools for Latin. Since 2006, he heads the "Index Thomisticus" Treebank project. In 2009, he founded the CIRCSE research centre on computational linguistics at Università Cattolica. Currently, he is Principal Investigator of a FIR-2013 funded project and Coordinator of a Marie Sklodowska-Curie Individual Fellowships. He organized and chaired several international scientific events. He co-chairs the series of workshops on 'Corpus-based Research in the Humanities' (CRH). He is author of one book and of about seventy papers published in scientific reviews and proceedings of national and international conferences.

E-MAIL • marco.passarotti@unicatt.it