

Somewhere between Valency Frames and Synsets. Comparing Latin *Vallex* and Latin WordNet

Marco Passarotti, Berta González Saavedra, Christophe Onambélé Manga

CIRCSE Research Centre

Università Cattolica del Sacro Cuore

Largo Gemelli, 1 – 20123 Milan, Italy

{marco.passarotti, berta.gonzalezsaavedra,
christophe.onambele}@unicatt.it

Abstract

English. Following a comparison of the different views on lexical meaning conveyed by the Latin WordNet and by a treebank-based valency lexicon for Latin, the paper evaluates the degree of overlapping between a number of homogeneous lexical subsets extracted from the two resources.

Italiano. *Alla luce di un confronto tra gli approcci al significato lessicale realizzati dal WordNet latino e da un lessico di valenza prodotto sulla base di due treebank latine, l'articolo descrive la valutazione del grado di sovrapposizione tra alcuni sottoinsiemi lessicali omogenei estratti dalle due risorse lessicali.*

1 Introduction

Several lexical resources are today available for many languages, ranging from dictionaries to wordnets, ontologies, valency lexica and others. Although such resources deal with the same basic constituents, i.e. lexical entries, these are organized according to different criteria, corresponding to different views on lexicon and, in particular, on lexical meaning.

On the one hand, a widespread approach to lexical meaning comes from 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 those words. Following such an assumption, there is a large use of the concept of *valency* and of labels for semantic roles in lexical resources. The degree of semantic granularity of the set of semantic roles used is what mostly distinguishes resources like PropBank, VerbNet and FrameNet one from the other.

On the other hand, a lexical resource largely used in both theoretical and computational linguistics is WordNet, which is centred on the idea of synonymy in the broad sense. Words are included in *synsets*, which are sets of words “that are interchangeable in some context without changing the truth value of the proposition in which they are embedded” (from the glossary of WordNet: <http://wordnet.princeton.edu>).

Despite their differences, these two views are not incompatible. Over the last decade, several attempts at linking different lexical resources together have been launched. One of the best known projects is Semlink, which makes use of a set of mappings to link PropBank, VerbNet, FrameNet and WordNet (Palmer, 2009). Pazienza et alii (2006) study the semantics of verb relations by mixing WordNet, VerbNet and PropBank. Shi and Mihalcea (2005) integrate FrameNet, VerbNet and WordNet into one knowledge-base for semantic parsing purposes.

Regarding the relations between valency lexica and wordnets, Hlavácková (2007) describes the merging of the Czech WordNet (CWN) with the database of verb valency frames for Czech VerbaLex, whose lexical entries are related to each other according to the CWN synsets. Hajič et alii (2004) use CWN while performing the lexico-semantic annotation of the Prague Dependency Treebank for Czech (PDT), which is in turn exploited to improve the quality and the coverage of CWN. In order to pick out the semantic constraints of the verbal arguments in the Polish WordNet (PolNet), the valency structure of verbs is used as a property of verbal synsets, because it is “one of the formal indices of the meaning (it is so that all members of a given synset share the valency structure)” (Vetulani and Kochanowski, 2014, page 402).

Despite a centuries-long tradition in lexicography, the development of state-of-the-art computational lexical resources for Latin is still in its infancy. However, some fundamental resources were built over the last decade. Among them are a WordNet and a treebank-based valency lexicon. In this paper, we present the first steps towards a comparison of these two resources, by evaluating the degree of overlapping of a number of their lexical subsets.

2 The Lexical Resources

2.1 The Latin Valency Lexicon *Vallex*

The Latin valency lexicon *Vallex* (LV; González Saavedra and Passarotti, forthcoming) was developed while performing the semantic annotation of two Latin treebanks, namely the *Index Thomisticus* Treebank, which includes works of Thomas Aquinas (Passarotti, 2014), and the Latin Dependency Treebank, which features works of different authors of the Classical era (Bamman and Crane, 2006). All valency-capable lemmas occurring in the semantically annotated portion of the two treebanks are assigned one lexical entry and one valency frame in LV.

The structure of the lexicon resembles that of the valency lexicon for Czech *PDT-Vallex* (Hajič et al., 2003). On the topmost level, the lexicon is

divided into lexical entries. Each entry consists of a sequence of frame entries relevant for the lemma in question. A frame entry contains a sequence of frame slots, each corresponding to one argument of the given lemma. Each frame slot is assigned a semantic role. The set of semantic roles is the same used for the semantic annotation of the PDT (Mikulová et al., 2005). Since the development of the lexicon is directly related to textual annotation, the surface form of the semantic roles run across during the annotation is recorded as well.

Presently, LV includes 983 lexical entries and 2,062 frames: 760 verbs (1,728 frames), 161 nouns (263 frames), 60 adjectives (68 frames), and 2 adverbs (3 frames).

2.2 The Latin WordNet

The Latin WordNet (LWN; Minozzi, 2010) was built in the context of the MultiWordNet project (Pianta et al., 2002), whose aim was to build a number of semantic networks for specific languages aligned with the synsets of Princeton WordNet (PWN). The language-specific synsets are built by importing the semantic relation among the synsets for English provided by PWN.

At the moment, LWN includes 8,973 synsets and 9,124 lemmas (4,777 nouns; 2,609 verbs; 1,259 adjectives; 479 adverbs).

3 Comparing the Lexical Resources

3.1 Method

To provide a basic understanding of the differences and similarities between the views on lexical meaning pursued by LV and LWN, we evaluate the degree of overlapping between some lexical subsets extracted from the two resources.

The lexical subsets of LWN that we use are the synsets, while for LV they are groups of words (lemmas) that share the same properties of arguments at frame entry level. For each subset extracted from LV we calculate its degree of overlapping with the synsets of LWN. The maximum overlapping holds when all the words belonging to the same LV subset do occur in the

same synset(s) of LWN. Conversely, the minimum overlapping holds when no word of an LV subset shares the same synset with any of the other words of the same LV subset, i.e. when all the words of a LV subset are “single” words. Our starting point, thus, are the LV subsets, whose contents are compared with the synsets of LWN.

We use two metrics to evaluate the results: (a) the number of single words, and that of couples, triples ... n -tuples of words in the LV subset that share the same LWN synset(s) (if the same word occurs in more n -tuples, the n -tuple with the higher value of n is considered); (b) the number of words in the LV subset that share the same LWN synset with n words of the same group (“connection degree”).

Loosely speaking, a good overlapping degree between an LV subset and the LWN synsets is given by (a) a low percentage of singles, (b) a high number of couples and n -tuples (this being as more meaningful as the value of n is higher) and (c) a high number of words with high connection degree.

3.2 Selecting the *Vallex* Subsets

LV subsets include words that share the same properties of arguments at frame entry level. We use frame entries instead of lexical entries because the frame is the level of the lexical entry that is mostly bound to meaning, a frame entry usually corresponding to one of the word’s senses. We focus on verbal entries only, as verbs are the most valency-capable words and the best represented PoS in LV.

Three selection criteria for LV subsets are at work: the quality of the arguments (i.e. their semantic role), their quantity and their surface form. For reasons of space, the groups discussed here are only a small selection of those that we built. In particular, we focus on a number of “semantically rich” frame entries, which feature such semantic roles (and some morphological features of them) that are expected to select verbs with a substantial degree of common semantic properties. According to these criteria, we selected the following LV subsets:

- (A) frame entries with three arguments: (a) an Actor, a Patient and a Direction-To (A_P_TO) and (b) an Actor, a Patient and an Addressee (A_P_AD);
- (B) the subsets in (A) are further specified by the following: an Actor, a Patient and a Direction-To expressed by a prepositional phrase headed by the preposition (a) *ad* (A_P_TO-ad) and (b) *in* (A_P_TO-in); (c) an Actor, a Patient and an Addressee expressed by a noun in dative (A_P_AD-dat);
- (C) frame entries featuring a Patient expressed (a) by a verbal phrase (P-VP) and (b) by a conjunction phrase (P-VP-conj), the latter being a subset of the former.

3.3 Results and Discussion

For each LV subset, table 1 shows the number of its members (column “W[ords]”), the percentage of them occurring in LWN (“Cover[age]”), the number of single words (“Singles”) and their percentage (“S%”). For instance, there are five singles in the A_P_TO subset: *conduco* (to drive), *educō,-ere* (to lead out), *immo* (to dip), *instigo* (to incite), *termino* (to limit)¹.

Table 1 shows also the number of couples (“Couples”) and n -tuples (“ n ples”) for each LV subset. For instance, the A_P_AD subset features one sextuple, i.e. six members of this subset share the same LWN synset: *doceo* (to teach), *exhibeo* (to present), *offero,-erre* (to offer), *ostendo*, (to show), *praebeo* (to offer), *praesto* (to offer).

Finally, table 1 shows the connection degree values (columns “1” to “10”). For instance, there are two words (*do* - to give - and *offero,-erre* - to

¹ Given the difference in size between LV and LWN, we considered only those subsets having a coverage ≥ 0.7 . The English translations provided here report the sense of the Latin word in the frame concerned. For instance, *termino* has two main senses, which are conveyed by two different frames: A_P (to mark the boundaries of) and A_P_TO (to limit).

offer -) in the A_P_AD subset that share the same synset at least once with nine different words belonging to the same synset. The nine words sharing the same synset with *do* are: *attribuo* (to assign), *dedo* (to consign), *largior* (to donate), *mitto* (to send), *offero,-erre* (to offer), *perhibeo* (to present), *praebeo* (to offer), *refero* (to report), *tribuo* (to assign).

According to the results, we can organize the LV subsets into three groups by overlapping degree. Low overlapping is shown by the subsets that include the Direction-To argument, A_P_TO-in and A_P_TO-ad presenting lower overlapping with LWN than A_P_TO. The two subsets featuring a Patient expressed by a VP (P-VP and P-VP-conj) show medium overlapping. The highest overlapping degree holds when the subsets with the Addressee argument are concerned (A_P_AD and A_P_AD-dat).

The results show a correspondence between the level of granularity of the semantic roles of the LV frame entries and the overlapping degree. Since Actor and Patient are quite semantically poor labels and they are common to all A_P_TO and A_P_AD subsets, it is the more fine-grained (and more strictly selecting) meaning of the Addressee than that of the Direction-To argument to make the A_P_AD subsets more overlapping with LWN than the A_P_TO ones.

Another aspect that biases the overlapping degree between LV and LWN are some morphological features of the semantic roles. A Patient expressed by a verbal phrase performs a quite strict selection of the verbs that can have

such a construction. These verbs must be able to subcategorize a Patient that is an event or a state expressed by a verb. Several of them tend to be verbs of perception and cognition, like for instance *verba dicendi*, *putandi* and *sentiendi*. In fact, one of the quadruples of the P-VP subset includes four *verba putandi*: *cogito* (to think), *credo* (to believe), *opinor* (to suppose), *suspicio* (to suspect).

4 Conclusion and Future Work

Although a valency lexicon like LV accounts for the different senses that one word may have by assigning it different frame entries, these are not as much semantically defined as the LWN synsets are. However, there is a certain degree of correspondence between these two resources: the more/less fine-grained a frame-based LV subset is, the higher/lower its overlapping with the LWN synsets. For instance, LV includes 1,060 frame entries of verbs formed by an Actor and a Patient: such a subset is both too large and semantically coarse-grained to allow for a sufficient overlapping with the LWN synsets. For this reason, while evaluating the overlapping degree between LV and LWN, we have first focussed on a number of “semantically rich” LV subsets. The evaluation metrics that we used are still very simple. The values of the *n*-tuples must be weighted at evaluation stage (one sextuple is “heavier” than one triple) and the lexical subsets of LWN must be extended beyond synonymy, by exploiting also other relations between words, like hyperonymy and hyponymy.

LV_Subset	W	Cover	Singles	S%	Couples	3ples	4ples	5ples	6ples	1	2	3	4	5	6	7	8	9	10
A_P_TO	24	0.833	5	0.25	16	1	-	-	-	8	4	2	1	-	-	-	-	-	-
A_P_AD	55	0.8	4	0.091	49	9	7	4	1	9	6	6	6	2	4	3	1	2	-
A_P_TO-ad	21	0.905	9	0.474	7	1	-	-	-	5	2	1	1	-	-	-	-	-	-
A_P_TO-in	17	1	9	0.529	4	-	-	-	-	6	1	-	-	-	-	-	-	-	-
A_P_AD-dat	35	0.8	5	0.178	31	5	8	2	-	3	2	6	6	3	3	-	1	-	-
P-VP	100	0.71	19	0.268	89	17	8	1	-	13	7	11	7	3	5	3	3	-	1
P-VP-conj	30	0.833	6	0.24	30	5	2	-	-	7	4	6	1	-	-	1	-	-	-

Table 1. Coverage, Singles and *n*-tuples, Connection Degree

References

- D. Bamman, D. and G. Crane. 2006. The design and use of a Latin dependency treebank. J. Nivre and J. Hajič (eds.), *Proceedings of the Fifth Workshop on Treebanks and Linguistic Theories (TLT2006)*, 67-78.
- C. Fillmore. 1982. Frame semantics. *Linguistics in the Morning Calm*. Hanshin Publishing Co., Seoul, 111-137.
- B. González Saavedra B. and M. Passarotti. Forthcoming. Verso un lessico di valenza del latino empiricamente motivato. *Atti del Workshop SLI "Dati empirici e risorse lessicali"*, La Valletta, Malta, 25 Settembre 2015.
- J. Hajič, M. Holub, M. Hučínová and M. Pavlík. 2004. Validating and Improving the Czech WordNet via Lexico-Semantic Annotation of the Prague Dependency Treebank. *Proceedings of the Workshop on "Building Lexical Resources from Semantically Annotated Corpora" at LREC 2004*, 25-30.
- J. Hajič, J. Panevová, Z. Urešová, A. Bémová, V. Kolárová and P. Pajas. 2003. PDT-VALLEX: Creating a large-coverage valency lexicon for treebank annotation. J. Nivre and E. Hinrichs (eds.), *Proceedings of the second workshop on treebanks and linguistic theories*, 57-68.
- D. Hlavácková. 2007. The Relations between Semantic Roles and Semantic Classes in VerbaLex. P. Sojka and A. Horák (eds.), *RASLAN 2007 Recent Advances in Slavonic Natural Language Processing*, 97.
- M. Mikulová et alii. 2005. Annotation on the tectogrammatical layer in the Prague Dependency Treebank. The Annotation Guidelines. Available at <http://ufal.mff.cuni.cz/pdt2.0/doc/manuals/en/t-layer/pdf/t-man-en.pdf>.
- S. Minozzi. 2010. The Latin WordNet project. In *Latin Linguistics Today*. Latin Linguistics Today. P. Anreiter and M. Kienpointner (eds.), *Akten des 15. Internationalen Kolloquiums zur Lateinischen Linguistik*, 4.-9. April 2009. Innsbrucker Beiträge zur Sprachwissenschaft, Innsbruck, 707-716.
- M. Palmer. 2009. Semlink: Combining English lexical resources. *Proceedings of the 5th. International Workshop on Generative Approaches to the Lexicon (GL2009)*, 9-15.
- M. Passarotti. 2014. From Syntax to Semantics. First Steps Towards Tectogrammatical Annotation of Latin. K. Zervanou and C. Vertan (eds.), *Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH) @ EACL 2014*. April 26, 2014. Gothenburg, Sweden, 100-109.
- M.T. Pazienza, M. Pennacchiotti and F.M. Zanzotto. 2006. Mixing wordnet, verbnet and propbank for studying verb relations. *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC-2006)*, 1372-1377.
- E. Pianta, L. Bentivogli and C. Girardi. 2002. MultiWordNet: developing an aligned multilingual database. *Proceedings of the first international conference on global WordNet*, Vol. 152, 55-63.
- L. Shi and R. Mihalcea. 2005. Putting pieces together: Combining FrameNet, VerbNet and WordNet for robust semantic parsing. *Computational linguistics and intelligent text processing*. Springer, Berlin Heidelberg, 100-111.
- Z. Vetulani and B. Kochanowski. 2014. "PolNet-Polish WordNet" project: PolNet 2.0-a short description of the release. *Proceedings of the Global Wordnet Conference*, 400-404.