IJCoL

Italian Journal of Computational Linguistics

Rivista Italiana di Linguistica Computazionale

Volume 2, Number 2
december 2016

Special Issue:
Digital Humanities and Computational Linguistics

Guest Editors:
John Nerbonne, Sara Tonelli

aAccademia university press

editors in chief

Roberto Basili
Università degli Studi di Roma Tor Vergata
Simonetta Montemagni
Istituto di Linguistica Computazionale "Antonio Zampolli" - CNR

advisory board

Giuseppe Attardi
Università degli Studi di Pisa (Italy)
Nicoletta Calzolari
Istituto di Linguistica Computazionale "Antonio Zampolli" - CNR (Italy)
Nick Campbell
Trinity College Dublin (Ireland)
Piero Cosi
Istituto di Scienze e Tecnologie della Cognizione - CNR (Italy)
Giacomo Ferrari
Università degli Studi del Piemonte Orientale (Italy)
Eduard Hovy
Carnegie Mellon University (USA)
Paola Merlo
Université de Genève (Switzerland)
John Nerbonne
University of Groningen (The Netherlands)
Joakim Nivre
Uppsala University (Sweden)
Maria Teresa Pazienza
Università degli Studi di Roma Tor Vergata (Italy)
Hinrich Schütze
University of Munich (Germany)
Marc Steedman
University of Edinburgh (United Kingdom)
Oliviero Stock
Fondazione Bruno Kessler, Trento (Italy)
Jun-ichi Tsujii
Artificial Intelligence Research Center, Tokyo (Japan)

editorial board

Cristina Bosco
Università degli Studi di Torino (Italy)
Franco Cutugno
Università degli Studi di Napoli (Italy)
Felice Dell'Orletta
Istituto di Linguistica Computazionale "Antonio Zampolli" - CNR (Italy)
Rodolfo Delmonte
Università degli Studi di Venezia (Italy)
Marcello Federico
Fondazione Bruno Kessler, Trento (Italy)
Alessandro Lenci
Università degli Studi di Pisa (Italy)
Bernardo Magnini
Fondazione Bruno Kessler, Trento (Italy)
Johanna Monti
Università degli Studi di Sassari (Italy)
Alessandro Moschitti
Università degli Studi di Trento (Italy)
Roberto Navigli
Università degli Studi di Roma "La Sapienza" (Italy)
Malvina Nissim
University of Groningen (The Netherlands)
Roberto Pieraccini
Jibo, Inc., Redwood City, CA, and Boston, MA (USA)
Vito Pirrelli
Istituto di Linguistica Computazionale "Antonio Zampolli" - CNR (Italy)
Giorgio Satta
Università degli Studi di Padova (Italy)
Gianni Semeraro
Università degli Studi di Bari (Italy)
Carlo Strapparava
Fondazione Bruno Kessler, Trento (Italy)
Fabio Tamburini
Università degli Studi di Bologna (Italy)
Paola Velardi
Università degli Studi di Roma "La Sapienza" (Italy)
Guido Vetere
Centro Studi Avanzati IBM Italia (Italy)
Fabio Massimo Zanzotto
Università degli Studi di Roma Tor Vergata (Italy)

editorial office
Danilo Croce
Università degli Studi di Roma Tor Vergata
Sara Goggi
Istituto di Linguistica Computazionale "Antonio Zampolli" - CNR
Manuela Speranza
Fondazione Bruno Kessler, Trento

Registrazione presso il Tribunale di Trento n. 14/16 del 6 luglio 2016

Rivista Semestrale dell'Associazione Italiana di Linguistica Computazionale (AILC)
© 2016 Associazione Italiana di Linguistica Computazionale (AILC)

direttore responsabile
Michele Arnese

Pubblicazione resa disponibile
nei termini della licenza Creative Commons
Attribuzione – Non commerciale – Non opere derivate 4.0

isbn 978-88-99982-26-3

Accademia University Press
via Carlo Alberto 55
I-10123 Torino
info@aAccademia.it
www.aAccademia.it/IJCoL_2_2

Accademia University Press è un marchio registrato di proprietà
di LEXIS Compagnia Editoriale in Torino srl

IJCoL

Volume 2, Number 2
december 2016

Special Issue:
Digital Humanities and Computational Linguistics

Guest Editors:
John Nerbonne, Sara Tonelli

CONTENTS

On Singles, Couples and Extended Families. Measuring Overlapping between *Latin Vallex* and Latin WordNet

Gian Paolo Clemente*
Università Cattolica del Sacro Cuore, Milano

Marco C. Passarotti**
Università Cattolica del Sacro Cuore, Milano

Different lexical resources may pursue different views on lexical meaning. However, all of them deal with lexical items as common basic components, which are described according to criteria that may vary from one resource to another. In this paper, we present a method for measuring the degree of similarity between a valency-based lexical resource and a WordNet. This is motivated by both theoretical and practical reasons. As for the former, we wonder if there are lexical classes that "impose" themselves regardless of the fact that they are explicitly recorded as such in source lexical resources. As for the latter, our work wants to contribute to the research task dealing with merging lexical resources. In order to apply and evaluate our method, we propose a normalized coefficient of overlapping that measures the overlapping rate between a valency lexicon and a WordNet. In particular, in the context of the exploitation of the linguistic resources for ancient languages built over the last decade, we compute and evaluate the overlapping between a selection of homogeneous lexical subsets extracted from two lexical resources for Latin.

1. Introduction

Viewing lexical semantics through predicate-argument structure strictly relates with 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. The notion of *semantic frame* subsumes that of *valency* ((Ágel and Fischer 2010); (Tesnière 1959)), which is defined as the number of obligatory complements required by a word. These obligatory complements are usually named *arguments*, while the non-obligatory ones are referred to as *adjuncts*. Although different parts of speech (PoS) can be valency-capable, scholars have mainly focused on verbs, so that the notion of valency tends to coincide with that of verbal valency.

There is large use of the notion of valency in lexical resources. The degree of semantic granularity of the set of semantic roles assigned to arguments is one of the aspects that mostly distinguishes valency-based lexical resources like PropBank (Palmer, Gildea, and Kingsbury 2009), VerbNet (Kipper 2005) and FrameNet (Baker, Fillmore, and Lowe 1998) one from the other. In this respect, PropBank is semantically more coarse-grained than both VerbNet and FrameNet, as it labels arguments according to syntactic subcategorization instead of assigning them semantic roles.

* Department of Mathematics, Finance and Econometrics - Via Necchi 9, 20123 Milan, Italy.
E-mail: gianpaolo.clemente@unicatt.it
** CIRCSE Research Centre - Largo Gemelli 1, 20123 Milan, Italy.
E-mail: marco.passarotti@unicatt.it

© 2016 Associazione Italiana di Linguistica Computazionale

Instead, both syntactic subcategorization and semantic roles do not play any role in WordNet (Miller 1995), which pursues a different view on lexical meaning based on the idea of synonymy in the broad sense . Words are included into *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"[1].

Despite their differences, the views on lexical meaning pursued by valency-based lexical resources and WordNet are not incompatible. In this paper, we present a method for measuring the degree of similarity between such resources by proposing a normalized coefficient of overlapping (OVL). In particular, we apply the OVL to two lexical resources for Latin, namely the valency lexicon *Latin Vallex* and the Latin WordNet. The motivation of the work presented in the paper is twofold.

First, given that a valency lexicon and a synonymy-based lexical resource look at lexical items from two different theoretical perspectives, namely one standing between syntax and semantics (valency) and the other being closer to referential semantics (synonymy), we have a theoretical interest in understanding what these have in common: are there lexical classes that "impose" themselves regardless of the fact that they are explicitly recorded as such in source lexical resources? what is common to words if we imagine them to be somewhere between valency frames and synsets? In this respect, there are aspects that are still partially known. For instance, we are aware of the fact that not all resultative change of state verbs are synonyms, but we do not know which of them are synonyms (and if there are), and what criteria rule this.

Second, evaluating the degree of overlapping between a valency-based lexical resource and a WordNet is among the first steps towards merging such resources and exploiting at best the different information they provide on lexical items. Actually, so far most works aimed at merging and/or aligning lexical resources have basically focussed on collecting meta-linguistic information about words carried by different lexical resources. Our approach is different. We do not collect into one lexical description features about words taken from various lexical resources. Instead, we find which classes of words automatically result from "comparing" homogeneous subsets of lexical items extracted from the resources. Our assumption is that, before merging/aligning resources, we must find what they have in common, just to exploit better what they have not: and lexical resources share the very object they deal with, i.e. words. In this respect, the fact that the two lexical resources used in this work carry different kind of information (i.e. they describe words from different perspectives) is an added value and not a drawback. Looking for what such different perspectives on the same objects have in common will help with alignment just because it allows us to connect lexical resources not on one, flat level (i.e. by just summing them up) but on a multi-level scale, ranging from general word classes (common to all merged resources and automatically induced from them, i.e. not theoretically imposed) to specific word classes (proper of single resources).

We apply our language-independent coefficient of overlapping to two lexical resources for Latin, because we believe that times are mature enough to move towards the next step for language resources for Classical languages. Indeed, over the last decade several research projects have focussed on building fundamental textual and lexical resources for such languages, with the aim of moving them out of their under-resourced status. Now we have to switch from building new resources to exploiting the available ones by first comparing and merging their contents, in order to make them

1 Taken from the glossary of WordNet: `http://wordnet.princeton.edu`.

collaborate fruitfully for different purposes, ranging from NLP to information extraction and theoretical linguistics. In particular, for what concerns lexical resources, both Latin and Ancient Greek show a centuries long tradition in lexicography, which nowadays can be enhanced by enriching lexical analysis through computational resources that are built according to the same criteria used for other (living) languages.

The paper is organized as follows: Section 2 surveys the previous work, Section 3 introduces the lexical resources used in the experiments, Section 4 describes the criteria for comparing the resources and presents the results, Section 5 details the coefficient of overlapping, which is in turn evaluated in Section 6. Section 7 concludes the paper and sketches the future work.

2. Previous Work

Much work has been done on the integration of lexical resources[2]. The main theoretical framework used in this context is the so called *linking theory* (Tenny and Pustejovsky 2000), which describes how verbal arguments are linked to the positions for syntactic subject and object(s). In particular, two approaches are mostly applied while dealing with the interaction between syntax and semantics: *mono-stratal* approaches advocate a direct detection of semantic roles from syntax, while *multi-stratal* ones make use of an intermediate grammatical level to ease the transition. The former seems to be trickier to apply ((Moschitti 2004), (Pradhan et al. 2005)). Instead, the latter is largely used in lexical resources, among which are most of those mentioned in Section 1.

In this respect, the SemLink project partially merges PropBank, VerbNet, FrameNet and WordNet by combining their information through a set of mappings (Palmer 2009). (Pazienza, Pennacchiotti, and Zanzotto 2006) study the semantics of verbal relations by mixing the lexical relations made available by WordNet, the classes of VerbNet and the Penn Treebank to connect relational verbal semantics with surface syntactic realizations. Such connection is established through the use of PropBank as a link between the lexical resources and the textual evidence provided by the treebank. In order to build a knowledge base for robust semantic parsing purposes, (Shi and Mihalcea 2005) use mappings between VerbNet classes and FrameNet frames on one side, and between selectional restrictions for roles in VerbNet and semantic classes of WordNet on the other. (Giuglea and Moschitti 2004) use VerbNet as a link between PropBank syntactic arguments and FrameNet semantic arguments to improve the accuracy of a system for semantic role labeling.

Predicate models exploiting the relations between valency lexica and WordNets have also been built. (Vetulani and Kochanowski 2014) use the valency structure of verbs as a property of verbal synsets to detect the semantic constraints of verbal arguments in the Polish WordNet (PolNet). (Hlaváčková 2007) merges a database of verbal valency frames for Czech with the Czech WordNet (CWN) in order to create classes enhanced with semantic roles for verbal arguments. The CWN is also used by (Hajič et al. 2004) both to perform the lexico-semantic annotation of the Prague Dependency Treebank (PDT) and to improve the quality and coverage of the CWN.

2 Among others, see (Burchardt, Erk, and Frank 2005), (Crouch and King 2005), (Johansson and Nugues 2007) , (De Cao et al. 2008), (Pennacchiotti et al. 2008), (Pianta and Tonelli 2009), (Laparra, Rigau, and Cuadros 2010), (Necsulescu et al. 2011), (Gurevych et al. 2012) and (López de Lacalle, Laparra, and Rigau 2014).

Yet, to our knowledge, no specific investigation has been made so far on measuring the overlapping between WordNet-like resources and valency-based ones[3].

3. Lexical Resources

The valency lexicon *Latin Vallex* (LV; (Passarotti, González Saveedra, and Onambele 2016)) was developed while performing the semantic annotation of two Latin treebanks, namely the *Index Thomisticus* Treebank, which includes works of Thomas Aquinas, written in Medieval Latin (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. Each argument is assigned a semantic role. The surface form of the semantic roles run across during treebank annotation (in terms of PoS and case) is recorded as well. The set of semantic roles is the same used for the semantic annotation of the PDT (Mikulová and others 2005). The *Dialogue Test* by (Panevová 1974) and (Panevová 1975) and the criteria reported in ((Mikulová and others 2005) pages 100-102, 116-162) are used to distinguish arguments from adjuncts. Presently, LV includes 983 lexical entries and 2,062 frames.

The Latin WordNet (LWN; (Minozzi 2010)) was developed in the context of the MultiWordNet project (Pianta, Bentivogli, and Girardi 2002), the aim of which was to build semantic networks for specific languages aligned with the synsets of Princeton WordNet. At the moment, LWN includes 9,124 lemmas and 8,973 synsets.

4. Comparing Lexical Resources

To understand the differences and similarities between the views on lexical meaning pursued by LV and LWN, we evaluate the degree of overlapping between a selection of homogeneous lexical subsets extracted from the two resources.

Synsets are the lexical subsets of LWN that we use, while for LV they are groups of words (lemmas) that share the same argumental properties 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 senses of the word. We focus on verbal entries only, as verbs are the most valency-capable words and the best represented PoS in LV (759 out of the 983 entries of LV are verbs).

4.1 *Latin Vallex* Subsets

Three selectional criteria for LV subsets are at work (not necessarly all at the same time): (a) the quality of the arguments (i.e. their semantic role), (b) their number (quantity) and (c) their surface form.

3 A first sketch of the idea behind the work presented in this paper is reported by (Passarotti, González Saveedra, and Onambele 2015).

Following these criteria, we extracted 28 subsets from LV. Table 1 shows a small selection of them. The first line of Table 1 concerns the LV subset whose members are provided with (at least) the following three arguments (quantity = 3): ACT[or], PAT[ient] and ADDR[essee]. The surface form for the Addressee is represented by a noun phrase (NP) with the noun in the dative case. Both the second and the third lines concern LV subsets that include at least one argument (quantity = 1): this is a Patient expressed respectively by a noun phrase (with the noun in the dative case) and by a verbal phrase headed by a subordinating conjunction (VP(sconj)).

Table 1
Three selected LV subsets

Semantic Roles	Quantity	Surface Form
ACT-PAT-ADDR	3	ACT-PAT-ADDR_NP(dat)
PAT	1	PAT_NP(dat)
PAT	1	PAT_VP(sconj)

The 28 LV subsets are detailed in the following[4].

(1) ACMP: verbs with at least one argument that is assigned semantic role ACMP (Accompaniment). Example: *admisceo* ("to mix with").
(2) ACT-DIR1-DIR3: verbs with at least three arguments, whose semantic roles are ACT[or], DIR1 (Direction-From) and DIR3 (Direction-To). Example: *eo* ("to go").
(3) ACT-DIR3: verbs with at least two arguments, whose semantic roles are ACT[or] and DIR3 (Direction-To). Example: *advenio* ("to come to").
(4) ACT-ORIG: verbs with at least two arguments, whose semantic roles are ACT[or] and ORIG[o]. Example: *abstineo* ("to keep off").
(5) ACT-PAT-ADDR: verbs with at least three arguments, whose semantic roles are ACT[or], PAT[ient] and ADDR[essee]. Example: *do* ("to give").
(6) ACT-PAT-ADDR_NP(dat): verbs with at least three arguments, whose semantic roles are ACT[or], PAT[ient] and ADDR[essee], the latter being expressed by a noun phrase (with the noun in the dative case). Example: *do* ("to give").
(7) ACT-PAT-DIR3: verbs with at least three arguments, whose semantic roles are ACT[or], PAT[ient] and DIR3 (Direction-To). Example: *extendo* ("to extend").
(8) ACT-PAT-DIR3_PP(ad): verbs with at least three arguments, whose semantic roles are ACT[or], PAT[ient] and DIR3 (Direction-To), the latter being expressed by a prepositional phrase headed by the preposition *ad* ("to"). Example: *termino* ("to limit [something to something else]").
(9) ACT-PAT-DIR3_PP(in): verbs with at least three arguments, whose semantic roles are ACT[or], PAT[ient] and DIR3 (Direction-To), the latter being expressed by a prepositional phrase headed by the preposition *in* ("in', "into', "to"). Example: *addo* ("to add").
(10) ACT-PAT-EFF: verbs with at least three arguments, whose semantic roles are ACT[or], PAT[ient] and EFF[ect] (i.e. the semantic role assigned to predicative complements). Example: *censeo* ("to estimate").

4 The same verb can belong to different subsets.

(11) ACT-PAT-ORIG: verbs with at least three arguments, whose semantic roles are ACT[or], PAT[ient] and ORIG[o]. Example: *capio* (“to take”).
(12) ACT-PAT-ORIG_PP(ab): verbs with at least three arguments, whose semantic roles are ACT[or], PAT[ient] and ORIG[o], the latter being expressed by a prepositional phrase headed by the preposition *ab* (“by”, “from’). Example: *accipio* (“to receive”).
(13) ACT-PAT-ORIG_PP(ex): verbs with at least three arguments, whose semantic roles are ACT[or], PAT[ient] and ORIG[o], the latter being expressed by a prepositional phrase headed by the preposition *ex* (“by”, “from’). Example: *colligo* (“to obtain by begging”).
(14) ADDR: verbs with at least one argument that is assigned semantic role ADDR[essee]. Example: *confero* (“to confer”).
(15) ADDR_NP(dat): verbs with at least one argument that is assigned semantic role ADDR[essee] expressed by a noun phrase (with the noun in the dative case). Example: *nuntio* (“to announce”).
(16) ADDR_PP(ad): verbs with at least one argument that is assigned semantic role ADDR[essee] expressed by a prepositional phrase headed by the preposition *ad* (“to”). Example: *dico* (“to say”).
(17) DIR1: verbs with at least one argument that is assigned semantic role DIR1 (Direction-From). Example: *venio* (“to come”).
(18) DIR3: verbs with at least one argument that is assigned semantic role DIR3 (Direction-To). Example: *redeo* (“to go back”).
(19) DIR3_PP(ad): verbs with at least one argument that is assigned semantic role DIR3 (Direction-To) expressed by a prepositional phrase headed by the preposition *ad* (“to”). Example: *eo* (“to go”).
(20) DIR3_PP(in): verbs with at least one argument that is assigned semantic role DIR3 (Direction-To) expressed by a prepositional phrase headed by the preposition *in* (“in’, “into’, “to”’). Example: *adduco* (“to lead to”).
(21) Four_Roles: verbs whose arguments are assigned at least four different semantic roles in their frame entries. Example: *moveo* (“to move”).
(22) ORIG: verbs with at least one argument that is assigned semantic role ORIG[o]. Example: *assumo* (“to receive”).
(23) ORIG_PP(ab): verbs with at least one argument that is assigned semantic role ORIG[o] expressed by a prepositional phrase headed by the preposition *ab* (“by”, “from’). Example: *acquiro* (“to acquire”).
(24) ORIG_PP(ab/ex): verbs with at least one argument that is assigned semantic role ORIG[o] expressed by a prepositional phrase headed by the preposition *ab* or *ex* (“by’, “from”’). Example: *accipio* (“to receive”).
(25) PAT_NP(dat): verbs with at least one argument that is assigned semantic role PAT[ient] expressed by a noun phrase (with the noun in the dative case). Example: *consentio* (“to agree”).
(26) PAT_VP: verbs with at least one argument that is assigned semantic role PAT[ient] expressed by a verbal phrase. Example: *dico* (“to say”).
(27) PAT_VP(sconj): verbs with at least one argument that is assigned semantic role PAT[ient] expressed by a verbal phrase headed by a subordinating conjunction. Example: *ostendo* (“to show”).
(28) Three_Roles: verbs whose arguments are assigned at least three different semantic roles in their frame entries. Example: *facio* (“to make”).

4.2 Evaluation Metrics

We use the following three metrics to evaluate how much overlapping a LV lexical subset and a LWN synset are:

(a) **Coverage** is the number of words in a LV subset that are also in LWN. Given the difference in size between LV and LWN, we considered only subsets with a coverage $\geq$0.6, i.e. those in which more than half of the words are included also in LWN. If the coverage for a LV subset is under this threshold, the subset is left out.
(b) **Singles, couples, triplets...*n*-tuplets** (called *co-occurrences*) refer to the number of words in a LV subset that share the same LWN synset(s)[5]. Singles are words of a LV subset that do not share the same LWN synset with any of the other words of that subset. Couples, triplets and n-tuplets are groups of 2, 3 and n words of a LV subset that share the same LWN synset. For each LV subset we calculate the number of singles, couples, triplets,... and n-tuplets.
(c) **Connection Degree** is the number of words in a LV subset that share the same LWN synset(s) with a word x of the same subset. The connection degree for x is calculated as follows:
   (i) extract all the n-tuplets for x in a LV subset;
   (ii) list the distinct words that occur in the n-tuplets for x;
   (iii) the number of items in the list is the connection degree for x.

4.3 Results

We applied the evaluation metrics described in 4.2 to the 28 subsets extracted from LV. Results are reported in Table 2.

For each LV subset, we calculate the number of its members (column "W[ords]"), the number ("N[umber]") and the percentage ("C[overage] R[atio]") of those occurring also in LWN, the number of singles ("S[ingle]s") and that of couples ("C[ouple]s"), triplets ("3s"), quadruplets ("4s"), quintuplets ("5s") and sextuplets ("6s")[6]. The column "MD" (Maximum Degree) reports the maximum value of connection degree observed in the LV subset. "AvD" (Average Degree) is the average connection degree of the LV subset.

For instance, the LV subset ACMP includes only singles (3). Instead, the ACT-PAT-ADDR subset features one sextuplet, 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" and *praesto* "to offer". Absolute values must be interpreted carefully while evaluating the degree of overlapping of a LV subset; for instance, the subset named Three_roles shows the highest values for all the evaluation metrics, but this is biased by the fact that it is the largest subset among those reported in Table 2 ($N = 113$). We will face this issue while building the OVL (see Section 5).

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-tuplets;

5 If the same co-occurrence appears in more than one LWN synset, it counts as one. The sequence of words in a co-occurrence is not meaningful.
6 In the LV subsets that we extracted, sextuplets are the longest n-tuplets that we found.

Table 2
Coverage, Singles, Couples, ..., n-tuplets, Connection Degree

LV_Subset	W	N	CR	Ss	Cs	3s	4s	5s	6s	MD	AvD
(1) ACMP	4	3	75.00%	3	0	0	0	0	0	0	0
(2) ACT-DIR1-DIR3	4	4	100.00%	4	0	0	0	0	0	0	0
(3) ACT-DIR3	26	21	80.77%	8	7	0	0	0	0	2	0.67
(4) ACT-ORIG	10	6	60.00%	4	1	0	0	0	0	1	0.33
(5) ACT-PAT-ADDR	54	43	79.63%	4	28	7	2	4	1	26	5.34
(6) ACT-PAT-ADDR_NP(dat)	35	28	80.00%	5	16	5	4	2	0	16	5.36
(7) ACT-PAT-DIR3	24	20	83.33%	5	11	1	0	0	0	4	1.40
(8) ACT-PAT-DIR3_PP(ad)	21	18	85.71%	9	6	1	0	0	0	4	1.00
(9) ACT-PAT-DIR3_PP(in)	17	16	94.12%	9	4	0	0	0	0	2	0.50
(10) ACT-PAT-EFF	33	27	81.82%	7	14	6	1	1	1	17	4.67
(11) ACT-PAT-ORIG	17	14	82.35%	7	7	0	0	0	0	3	1.00
(12) ACT-PAT-ORIG_PP(ab)	10	6	60.00%	6	0	0	0	0	0	0	0
(13) ACT-PAT-ORIG_PP(ex)	13	9	69.23%	3	5	0	0	0	0	3	1.11
(14) ADDR	59	47	79.66%	6	30	8	2	5	1	26	5.40
(15) ADDR_NP(dat)	35	28	80.00%	5	16	5	4	2	0	16	5.29
(16) ADDR_PP(ad)	8	8	100.00%	6	1	0	0	0	0	1	0.25
(17) DIR1	19	17	89.47%	10	5	1	0	0	0	5	0.94
(18) DIR3	51	42	82.35%	10	22	2	1	0	0	5	1.62
(19) DIR3_PP(ad)	21	18	85.71%	9	6	1	0	0	0	4	1.00
(20) DIR3_PP(in)	18	17	94.44%	7	7	0	0	0	0	3	0.82
(21) Four_Roles	17	17	100.00%	10	6	0	0	0	0	3	0.71
(22) ORIG	27	20	74.07%	9	9	0	0	0	0	3	0.90
(23) ORIG_PP(ab)	11	7	63.64%	5	1	0	0	0	0	1	0.29
(24) ORIG_PP(ab/ex)	22	15	68.18%	7	6	0	0	0	0	3	0.80
(25) PAT_NP(dat)	19	12	63.16%	5	3	1	1	0	0	7	2.00
(26) PAT_VP	100	70	70.00%	18	38	15	7	1	0	20	3.86
(27) PAT_VP(sconj)	30	24	80.00%	6	15	4	1	0	0	11	2.75
(28) Three_Roles	143	113	79.02%	19	68	31	11	6	2	30	5.61

(c) a high number of words with high connection degree.

However, proposing a formal overlapping measure needs more than this. Two main aspects must be considered. First, one resource (LV) is significantly smaller than the other (LWN). Second, the number of couples and n-tuplets is more meaningful as the value of n is higher. For instance, a sextuplet is "heavier" than a triplet. Thus, the value of n in the n-tuplets must be taken into account at evaluation stage by a weighting function able to consider that some n-tuplets count more than others.

5. Overlapping Measure

Although in statistical analysis it is more common to compare two distributions by looking at their characteristics such as mean or median, an overlapping measure has been also proposed to quantify the absolute or relative degree of overlapping of two distributions[7].

To our knowledge, such an overlapping measure has not been applied yet in merging lexical resources. In the specific case of our work, such an application is quite peculiar, as we do not deal with two probability distributions, but we have to measure the degree of overlapping of two groups of items (namely, LV subsets and LWN synsets).

7 See (Jaccard 1901), (Weitzman 1970), (Inman and Bradley Jr. 1989), (Tan, Steinbach, and Kumar 2005) and (Goldberg, Hayvanovych, and Magdon-Ismail 2010).

For this purpose, we developed an ad-hoc overlapping measure (OVL). Before entering the formal aspects of this measure, we briefly introduce some basic notation.

Given a generic subset LV_h of LV, the coverage of LV_h (i.e. the number N of members of LV_h occurring also in LWN) is $N_h = |LV_h \cap LWN|$[8].

For LV_h we can measure the type and the number of n-tuplets (i.e. how many singles, couples...n-tuplets are observed in LV_h) and the connection degree for each item j (with $j = 1, 2, ..., N_h$) belonging to LV_h.

Our aim is to include the effect of these two evaluation metrics into a single measure able to summarize in one value the degree of overlapping for each LV subset. Another desideratum is the possibility of decomposing the measure to emphasize the contribution made by each of its components.

In the following two subsections, we start to build the measure by weighting separately the effect of co-occurrences and connection degree.

5.1 Weighting Co-occurrences

We assume that a semantic relation of synonymy holds between the words that share the same LWN synset. We define two words belonging to the same LWN subset as co-occurrent in that LWN and we make use of this co-occurrence while building the OVL.

To evaluate the contribution to overlapping made by each item j in terms of co-occurrence, we propose a weighting measure represented by the coefficient c_j defined in Equation (1).

Definition 1 (Co-occurrence based overlapping weight)

Let $Ns_h = N_h - s_h$ where s_h refers to the number of singles in the subset LV_h. For each item j ($j = 1, .., Ns_h$), we define the following coefficient:

$$c_j = \sum_{t=2}^{T} \frac{t \cdot s_{j,t}}{M_j} \tag{1}$$

where the vector $\mathbf{s_j} = [s_{j,2}, ..., s_{j,T}]$ shows how many couples $s_{j,2}$, triplets $s_{j,3}$ and so on are observed for the item j . T is the longest n-tuplet observed.

M_j represents the maximum value that the numerator of Equation (1) can assume. It allows to normalize c_j between 0 and 1. If we consider all potential combinations of items, we have $M_j = \sum_{t=2}^{N_h} t \cdot \binom{N_h - 1}{t-1}, \forall j$. However, in order to take into account also the different distributions that one single item has in LWN, we limit the evaluation of M_j to the number of synsets where the item j is present. It is noteworthy that c_j is not defined for singles, because in that case the vector $\mathbf{s_j}$ is empty.

In order to explain the proposed methodology, we present an example run on the LV subset ACT-PAT-ADDR. As shown in Table 2, 43 out of the 54 items of this LV subset occur also in LWN ($N_h = 43$) and 4 of them are singles ($Ns_h = 39$). For each of this 39 items, we apply Equation (1).

8 It goes without saying that coverage is the only evaluation metrics affected by those items that belong to LV_h but are not in LWN, as n-tuplets and connection degree are not measurable for such items.

For instance, to evaluate the contribution to overlapping made by the word *exhibeo* "to present", we first calculate that this word appears in 1 couple, 1 quadruplet, 1 quintuplet and 1 sextuplet. Thus, we have $c_j = \frac{1 \cdot 2 + 1 \cdot 4 + 1 \cdot 5 + 1 \cdot 6}{M_j} = \frac{17}{M_j}$.

Furthermore, *exhibeo* occurs in 4 synsets of LWN. Potentially, (a) it could share the same LWN synset with all the other items of its LV subset (namely, ACT-PAT-ADDR), thus leading to observe one 43-tuplet, and (b) it could share the other three synsets with 42 different words, thus resulting in three 42-tuplets. We can then compute $M_j = 43 + 3 \cdot 42 = 169$ and $c_j = \frac{17}{M_j} = 10.06\%$.

This result means that *exhibeo* covers 10.06% of its potential maximum value of co-occurence.

5.2 Weighting Connection Degree

Like for n-tuplets, we consider the connection degree for each item j, through the coefficient w_j defined in Equation (2). Also in this case, we skip singles, because their connection degree is equal to 0.

Definition 2 (Connection degree based overlapping weight)

For each item j ($j = 1, .., Ns_h$), we define the following coefficient:

$$w_j = \frac{d_j}{N_h - 1} \tag{2}$$

as the ratio of the connection degree of the item j (d_j) to the maximum observable degree ($N_h - 1$). In other words, we have $d_j = N_h - 1$ when the item is connected to any other item in the same LV subset. Also this coefficient is normalized in the range $(0, 1)$.

Using again the same example above, we can apply Equation (2) to the word *exhibeo*. Since this word has connection degree equal to 13 (i.e. it shares the same LWN synset(s) with 13 other words in the LV subset it belongs to), we have $w_j = \frac{13}{42} = 30.95\%$.

This result means that *exhibeo* is connected to 30.95% of the items in its LV subset.

5.3 Measuring Overlapping Degree

Definition 3 (OVL measure)

Given a generic subset LV_h of LV, we define the normalized overlapping coefficient between $LV_h \cap LWN$ and LWN as

$$OVL = \frac{1}{2} \sum_{j=1}^{Ns_h} \left(\frac{c_j + w_j}{N_h} \right) \tag{3}$$

With (3), we ensure that singles do not make any positive contribution to the overlapping degree, as they affect only the denominator of Equation (3). This means that, if only singles were observed, the OVL would be equal to 0.

Equation (3) can be easily rewritten as

$$OVL = \frac{1}{2}\left(\sum_{j=1}^{Ns_h}\frac{c_j}{N_h} + \sum_{j=1}^{Ns_h}\frac{w_j}{N_h}\right) = \frac{(\bar{c}+\bar{w})}{2} \quad (4)$$

In this way, we can both join and keep separated at the same time the single contribution to overlapping made by co-occurrences ($\bar{c}$) and connection degree ($\bar{w}$) respectively, because the OVL is calculated as the average of the single values of them.

The terms $\bar{c}$ and $\bar{w}$ are obtained by averaging the coefficients c_j and w_j calculated at the level of each single item. Furthermore, Equation (3) depends on the term $OVL_j = \frac{(c_j+w_j)}{2}$ which quantifies the contribution made by one item j to the overlapping degree. The total OVL for the subset $LV_h \cap LWN$ is the average of the OVL_j for each item belonging to that subset.

To explain in detail how the OVL is calculated, let's consider again the word *exhibeo*. The total contribution made by *exhibeo* to the overlapping degree is $OVL_j = \frac{(c_j+w_j)}{2} = \frac{(10.06\%+30.95\%)}{2} = 20.51\%$.

Once this computation is applied to all the other items of the LV subset which *exhibeo* belongs to (namely, ACT-PAT-ADDR), the average of the single contributions of the items gives the OVL rate for this subset.

6. Evaluation and Discussion

To evaluate the OVL, we measured[9] the degree of overlapping between the 28 subsets that we extracted from LV and the synsets of LWN by computing Equation (3).

Figure 1 provides an overview of the results by plotting the OVL rate and the coverage ratio for all the subsets. The closer to the upper right corner a subset is, the better it performs. As expected, higher connection degree and/or higher number of co-occurrences lead to higher OVL rate. See, for instance, subset *6* (ACT-PAT-ADDR_NP(dat)), which shows a quite high average connection degree (5.35) and a significant presence of co-occurrences (see Table 2), thus resulting in high OVL rate (14.58%)[10].

However, rather than computing the overlapping degree in absolute terms, we measure it in relative terms by taking into account also the size of the subsets (N_h) and the number of synsets in which each item of a subset occurs. In Figure 1, this is well represented by subset *28* (Three_Roles). Like for subset *6*, also in this case we have high average connection degree (5.61) and a significant number of co-occurrences (see Table 2). But the OVL rate is very low (3.53%). Indeed, if subset *28* is very similar to *6* in absolute terms, this does not hold true in relative terms. Each item of *28* is connected on average with roughly 5 other items of *28* (connection degree = 5.61), but it could be potentially connected with 112 other items of *28* ($N_h = 113$). Likewise, each item of *6* is connected on average with roughly 5 other items of *6* (connection degree = 5.35), but it could be potentially connected with other 27 items ($N_h = 28$). Following Equation

9 It is noteworthy that the computation of Equation (3) is quite a light task. To give an idea of the computational time needed, we run the procedure on an Intel Core 2 Duo Processor E7500, 2.93 GHz - 4GB RAM for all the analysed subsets, getting the estimated overlapping coefficient in about 35 seconds.The computational complexity is higher as the size of subsets increases.

10 OVL rates for all LV subsets are detailed in decreasing order in Table 3.

(3), this makes the OVL rate for *6* much higher than for *28*. Things are similar if co-occurences instead of connection degree are considered.

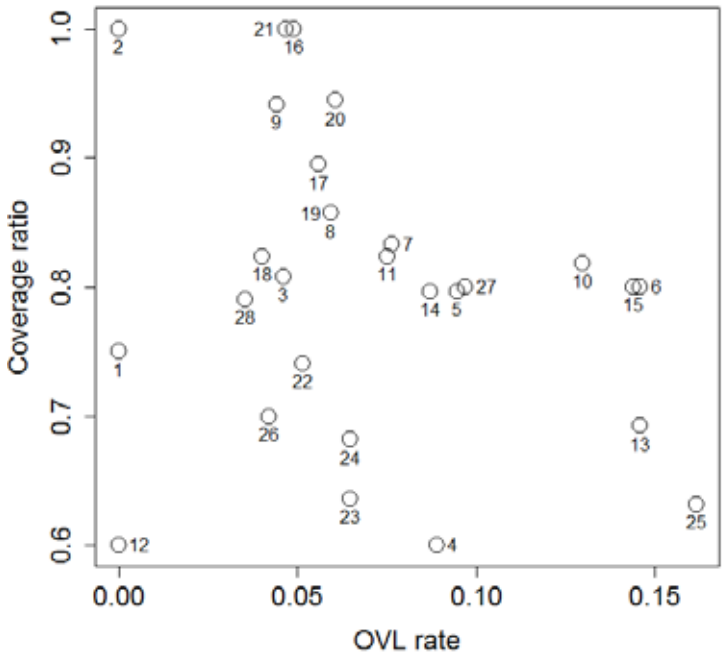

Figure 1
OVL rate and coverage ratio for each subset of LV

Figure 2 shows the specific contribution to the OVL made by co-occurrences and connection degree. The coordinates of each point in the plot are obtained by considering the co-occurrences coefficient $\bar{c}$ and the connection degree coefficient $\bar{w}$. The OVL rate for the 28 LV subsets in Figure 1 is the simple average of these two coordinates.

We observe a strong linear dependency between the two coefficients (i.e. the higher/lower $\bar{c}$ is, the higher/lower $\bar{w}$ is), because the connection degree of an item is related to the number and size of its co-occurrences. However, Figure 2 confirms that dependency is not "full" (i.e. a linear correlation coefficient equal to 1) and that both measures contribute positively to the OVL.

As explained in 5.3, the total contribution of each lemma OVL_j in a subset is derived by averaging its co-occurrences (c_j) and connection degree (w_j) coefficients. Then, the overall OVL for the subset is just the average of the OVL_j for each item belonging to that subset. Once applied to each lemma in a subset, these two coefficients can be used in order to identify groups of lemmas showing similar OVL-driven behaviour in a subset. For instance, Figure 3 plots the contribution of each lemma of the ACT-PAT-ADDR_NP(dat) subset according to co-occurrences and connection degree coefficients.

Moving from the upper right to the lower left corner, the following areas, which correspond to as many groups of lemmas, can be identified in Figure 3:

- verbs meaning “to give” (*do*) and “to offer”’ (*praebeo*);

Table 3
OVL rates for LV subsets

LV_Subset	OVL
(25) PAT_NP(dat)	16.15%
(13) ACT-PAT-ORIG_PP(ex)	14.6%
(6) ACT-PAT-ADDR_NP(dat)	14.58%
(15) ADDR_NP(dat)	14.4%
(10) ACT-PAT-EFF	12.99%
(27) PAT_VP(sconj)	9.69%
(5) ACT-PAT-ADDR	9.49%
(4) ACT-ORIG	8.89%
(14) ADDR	8.7%
(7) ACT-PAT-DIR3	7.64%
(11) ACT-PAT-ORIG	7.51%
(23) ORIG_PP(ab)	6.46%
(24) ORIG_PP(ab/ex)	6.46%
(20) DIR3_PP(in)	6.07%
(8) ACT-PAT-DIR3_PP(ad)	5.95%
(19) DIR3_PP(ad)	5.95%
(17) DIR1	5.6%
(22) ORIG	5.15%
(16) ADDR_PP(ad)	4.91%
(21) Four_Roles	4.67%
(3) ACT-DIR3	4.62%
(9) ACT-PAT-DIR3_PP(in)	4.41%
(26) PAT_VP	4.2%
(18) DIR3	4%
(28) Three_Roles	3.53%
(1) ACMP	0%
(2) ACT-DIR1-DIR3	0%
(12) ACT-PAT-ORIG_PP(ab)	0%

- verbs meaning “to show” (*exhibeo*, *perhibeo*, *ostendo*, the latter standing in the middle between this group and the previous one), “to assign” (*attribuo*, *tribuo*) and “to dispense” (*largior*, *praesto*). The verb *propono* (“to propose”) is also in this group[11];

11 The original meaning of the verb *propono* in Classical Latin is “to put forth” and, later, “to display”. The use of *propono* with meaning “to propose” is recorded in LV (thus falling into the subset ACT-PAT-ADDR_NP(dat)) because it is largely attested in the texts of Thomas Aquinas included in the *Index Thomisticus* Treebank, which is one of the Latin treebanks used for building LV. One example of the use of *propono* with this meaning is in the sentence “Quod [that] veritas [truth-PAT] divinorum [of divine issues] ... hominibus [to men-ADDR] credenda [to be believed] proponitur [is proposed]” (“That the truth about God ... is proposed to men for belief”, *Summa contra Gentiles*, 1-4). As the sentence is impersonal, the Actor is not expressed explicitly; its ellipsis is resolved by the semantic annotation of the Latin treebanks that feed LV by adding a new node in the tree for the sentence with the feature “Generic", because it is a resolved ellipsis for an argument whose referent cannot be contextually retrieved. The same holds also for the verb *praesto*, whose meaning in Classical Latin is “to stand out”/“to excel”, while several of its occurrences in Thomas Aquinas mean “to give”, “to provide”, “to dispense”. See for instance the following sentence: “Huic [to this] autem [but] errori [error] quatuor [four] sunt [are] quae

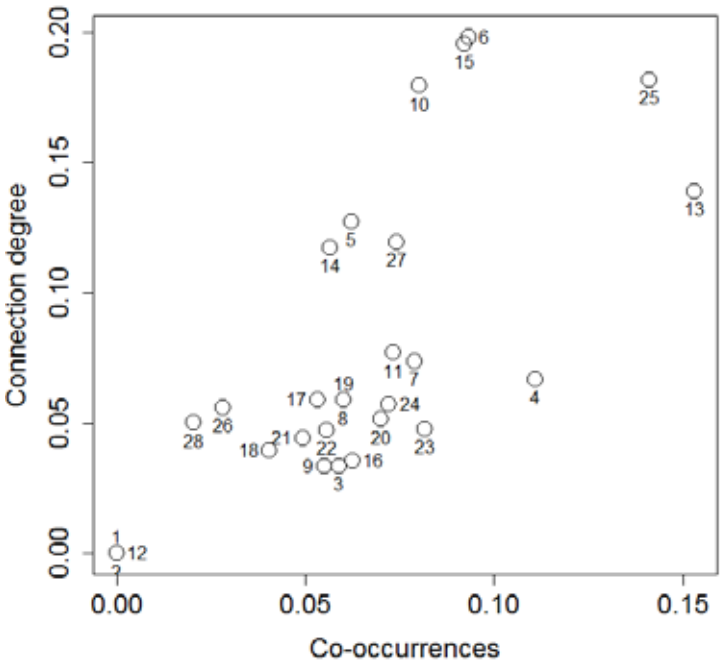

Figure 2
Co-occurrences ($\bar{c}$) and connection degree ($\bar{w}$) coefficients for all LV subsets

- verbs with the prefix *ad-* and meaning "to add" (*addo, adhibeo*) and "to adapt"/"to apply" (*adapto, applico*; also *apto*, with the same meaning but without the prefix *ad-*); verbs with the prefix *in-* and meaning "to put into"/"to introduce" (*immitto, indo, insero*);
- verbs meaning "to say"/"to speak" (*dico, dissero*) and "to confer" (*confero*[12]);
- remnants: verbs with different meanings showing low co-occurrences and connection degree coefficients: *appropinquo* ("to come near"), *debeo* ("to owe"), *persuadeo* ("to convince"), *promitto* ("to promise"), *respondeo* ("to answer").

By looking at the distribution of meanings of lemmas in Figure 3, one can claim that groups of verbs of the same subset resulting from higher co-occurrences and connection degree coefficients (upper right corner) are semantically tighter than those showing lower rates for these coefficients (lower left corner). This is confirmed by Figure 4,

[those that] videntur [seem] praestitisse [to have dispensed] fomentum [fomentation]" ("Four factors seem to have contributed to the rise of this error", *Summa contra Gentiles*, 1-26).

12 The meaning of *confero* in Classical Latin is "to bring together". Thomas Aquinas uses *confero* with the same meaning of the corresponding Italian verb *conferire* ("to confer"). See an example in this sentence: "De [of] rebus [things] nobilissimis [most noble] ... cognitio [knowledge] maximam [greatest] perfectionem [perfection] animae [to soul] confert [confers]" ("The knowledge about the most noble realities confers the greatest perfection to the soul", *Summa contra Gentiles*, 1-5).

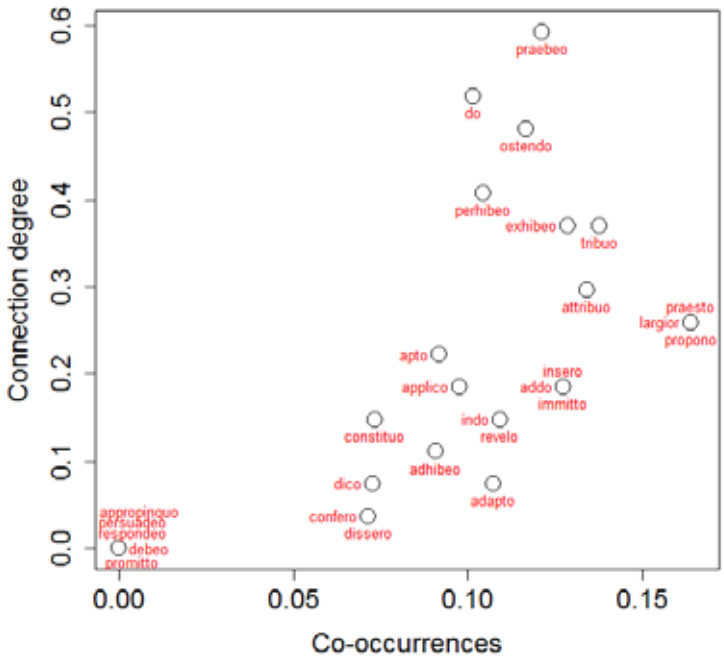

Figure 3
Co-occurrences (c_j) and connection degree (w_j) coefficients for each lemma of the ACT-PAT-ADDR_NP(dat) subset

which plots the contribution of each lemma of the ACT-PAT-EFF subset according to co-occurrences and connection degree coefficients.

The verbs in the upper right corner of Figure 4 share the common meaning "to estimate"/"to consider" when they are used with at least three arguments (Actor, Patient and Effect, respectively): *aestimo, arbitror, censeo, cogito, habeo*. Under this group in Figure 4 there is another cluster of verbs, which can be organised into two semantically tight subgroups of items featuring the same meaning: (a) "to look at"/"to see" (*conspicio*, *video*), (b) "to call"/"to say" (*appello*, *dico*, *voco* and also *enuntio*, which is placed slightly lower in the plot). Then, as much as one moves towards the lower left corner of Figure 4, the groups of lemmas become semantically less consistent.

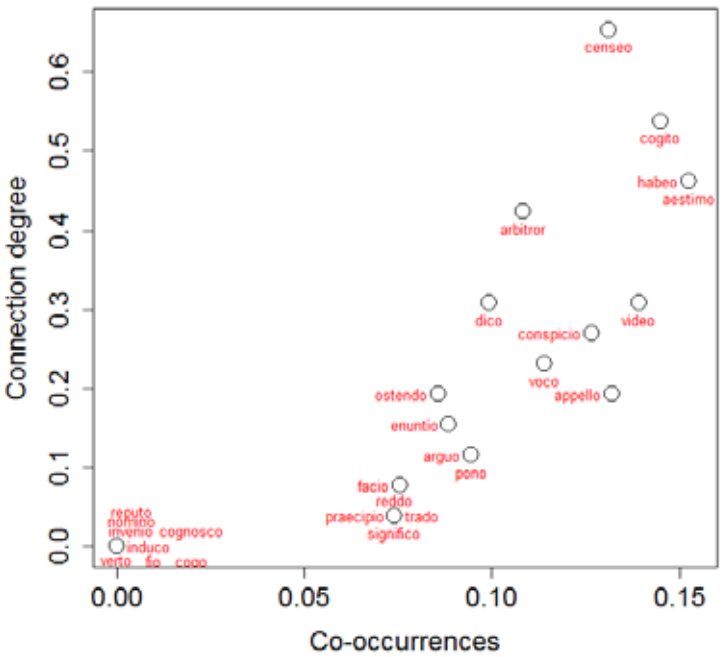

Figure 4
Co-occurrences (c_j) and connection degree (w_j) coefficients for each lemma of the ACT-PAT-EFF subset

7. Conclusion and Future Work

We presented a method for evaluating the degree of similarity between a valency lexicon and a WordNet, by proposing and applying a normalized coefficient of overlapping able to summarize in one value the overlapping rate holding between homogeneous lexical subsets extracted from two lexical resources for Latin (LV and LWN).

Our results show quite a diverse distribution of the overlapping rate of the subsets. This is due the fact that the more/less semantically fine-grained a LV subset is, the higher/lower its overlapping with the LWN synsets is. For instance, LV features 1,060 frame entries of verbs that include only 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. Thus, the selection criteria of the LV subsets play an essential role in evaluating the overlapping between LV and LWN. Indeed, while selecting the LV subsets, we did not consider those that are so broad to be uninformative, like for instance those including verbs with two arguments whose semantic roles are Actor and Patient respectively. Before including such subsets in our work, we should refine them by building consistent sub-subsets out of them (for instance, by using selectional restrictions available from treebanks). On the LWN side, we should extend the LWN subsets beyond synsets, by exploiting other available relations between words, like hyperonymy and hyponymy.

Since our coefficient of overlapping is meant to be language independent, we plan to apply it also to other (modern) languages, like Czech and English, which are provided both with WordNets and with valency lexica similar to LV. However, we are aware of the fact that dealing with resources for ancient languages raises a number of peculiar issues that are not common with modern languages. This is particularly true when semantic aspects of lexical items of ancient languages are concerned, as they can change (even heavily) over time and place. Indeed, since Latin shows a wide diachronic and diatopic span (around two millennia, all over Europe), a set of (merged) lexical resources for Latin must account for such variations. This desideratum is strictly connected to the new challenges opened for NLP by ancient languages and, more generally, by Digital Humanities, among which is the self adaptation of tools to the specific features shown by input texts in terms of time, place and genre.

Overall, evaluating the overlapping between valency lexica and WordNets is a fundamental step towards building more fine-grained lexical resources that result from merging the specific features provided by the already available ones. In this respect, our coefficient of overlapping enables to detect lexical classes resulting from different resources (in which such classes are not required to be explicitly recorded), the merging of which is supposed to make the whole greater than the sum of its parts. Furthermore, information taken from lexical resources of different kind can be used for hybridizing fully stochastic NLP methods in tasks like syntactic parsing, semantic role labeling and ellipsis resolution.

Acknowledgments

Special thanks should be given to Berta González Saavedra for building *Latin Vallex* and to Christophe Onambele for both providing helpful remarks and computing the results shown in Table 2.

References

Ágel, Vilmos and Klaus Fischer. 2010. *Dependency grammar and valency theory*. The Oxford Handbook of Linguistic Analysis. Oxford University Press.

Baker, Collin F., Charles J. Fillmore, and John B. Lowe. 1998. The berkeley framenet project. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1*, pages 86–90.

Bamman, David and Gregory Crane. 2006. The design and use of a latin dependency treebank. In *Proceedings of the Fifth Workshop on Treebanks and Linguistic Theories (TLT2006)*, pages 67–78.

Burchardt, Aljoscha, Katrin Erk, and Anette Frank. 2005. A wordnet detour to framenet. In *Proceedings of the GLDV 2005 workshop GermaNet II*, pages 408–421.

Crouch, Dick and Tracy H. King. 2005. Unifying lexical resources. In *Proceedings of the Interdisciplinary Workshop on the Identification and Representation of Verb Features and Verb Classes*, pages 32–37.

De Cao, Diedo, Danilo Croce, Marco Pennacchiotti, and Roberto Basili. 2008. Combining word sense and usage for modeling frame semantics. In *Proceedings of The Symposium on Semantics in Systems for Text Processing*, pages 85–101.

Fillmore, Charles J. 1982. *Frame semantics*. Linguistics in the Morning Calm. The Linguistic Society of Korea, (eds.).

Giuglea, Ana-Maria and Alessandro Moschitti. 2004. Knowledge discovering using framenet, verbnet and propbank. In *Proceedings of the Workshop on Ontology and Knowledge Discovering at the 15th European Conference on Machine Learning*, Pisa, Italy.

Goldberg, Mark, Mykola Hayvanovych, and Malik Magdon-Ismail. 2010. Measuring similarity between sets of overlapping clusters. In *Proceedings of the 2010 IEEE Second International Conference on Social Computing*, pages 303–308.

Gurevych, Iryna, Judith Eckle-Kohler, Silvana Hartmann, Michael Matuschek, Christian M. Meyer, and Christian Wirth. 2012. Uby - a large-scale unified lexical-semantic resource based on lmf. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 580–590.

Hajič, Jan, Martin Holub, Marie Hučínová, Martin Pavlík, Pavel Pecina, Pavel Straňák, and Pavel Martin Šidák. 2004. Validating and improving the czech wordnet via lexico-semantic annotation of the prague dependency treebank. In *Proceedings of the Workshop on "Building Lexical Resources from Semantically Annotated Corpora" at LREC 2004*, pages 25–30.

Hajič, Jan, Jarmila Panevová, Zdeňka Urešová, Alevtina Bémová, Veronika Kolářová, and Petr Pajas. 2003. Pdt-vallex: Creating a large-coverage valency lexicon for treebank annotation. In *Proceedings of the second workshop on treebanks and linguistic theories*, pages 57–68.

Hlaváčková, Dana. 2007. The relations between semantic roles and semantic classes in verbalex. In Petr Sojka and Aleš Horák, editors, *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2007*, pages 97–101.

Inman, Henry F. and Edwin L. Bradley Jr. 1989. The overlapping coefficient as a measure of agreement between probability distributions and point estimation of the overlap of two normal densities. *Communications in Statistics - Theory and Methods*, 18(10):3851–3874.

Jaccard, Paul. 1901. Etude comparative de la distribution florale dans une portion des alpes et des jura. *Bulletin de la Societe Vaudoise des Sciences Naturelles*, 37:547–579.

Johansson, Richard and Pierre Nugues. 2007. Using wordnet to extend framenet coverage. In *Proceedings of the Workshop on Building Frame-semantic Resources for Scandinavian and Baltic Languages*, pages 27–30, Department of Computer Science, Lund University.

Kipper, Karin. 2005. *VerbNet: A broad-coverage, comprehensive verb lexicon*. Ph.D. thesis, University of Pennsylvania, Leyden.

Laparra, Egoitz, German Rigau, and Montse Cuadros. 2010. Exploring the integration of wordnet and framenet. In *Proceedings of the 5th Global WordNet Conference (GWC'10)*, Mumbai, India.

López de Lacalle, Maddalen, Egoitz Laparra, and German Rigau. 2014. Predicate matrix: extending semlink through wordnet mappings. In *Proceedings of LREC'14*, pages 903–909.

Mikulová, Marie et al. 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*.

Miller, George A. 1995. Wordnet: A lexical database for english. In *Communications of the ACM 38*, pages 39–41.

Minozzi, Stefano. 2010. The latin wordnet project. In P. Anreiter and M. Kienpointner, editors, *Latin Linguistics Today. Akten des 15. Internationalem Kolloquiums zur Lateinischen Linguistik*, volume 137, pages 707–716.

Moschitti, Alessandro. 2004. A study on convolution kernels for shallow semantic parsing. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04)*, pages 335–342.

Necsulescu, Silvia, Núria Bel, Muntsa Padró, Montserrat Marimon, and Eva Revilla. 2011. Towards the automatic merging of language resources. In *Proceedings of the First International Workshop on Lexical Resources (WoLeR)*, pages 70–77.

Palmer, Martha. 2009. Semlink: Linking propbank, verbnet and framenet. In *Proceedings of the 5th International Conference on Generative Approaches to the Lexicon*, pages 9–15.

Palmer, Martha, Daniel Gildea, and Paul Kingsbury. 2009. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

Panevová, Jarmila. 1974. On verbal frames in functional generative description. *Prague Bulletin of Mathematical Linguistics*, 22:3–40.

Panevová, Jarmila. 1975. On verbal frames in functional generative description. *Prague Bulletin of Mathematical Linguistics*, 23:17–52.

Passarotti, Marco. 2014. From syntax to semantics. First steps towards tectogrammatical annotation of latin. In *Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*, pages 100–109.

Passarotti, Marco, Berta González Saveedra, and Christophe Onambele. 2015. Somewhere between valency frames and synsets. comparing latin vallex and latin wordnet. In *Proceedings of the Second Italian Conference on Computational Linguistics (CLiC-it 2015)*, pages 221–225, Trento, Italy.

Passarotti, Marco, Berta González Saveedra, and Christophe Onambele. 2016. Latin vallex. A treebank-based semantic valency lexicon for latin. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 2599–2606, Portorož, Slovenia.

Pazienza, Maria Teresa, Marco Pennacchiotti, and Fabio Massimo Zanzotto. 2006. Mixing wordnet, verbnet and propbank for studying verb relations. In *Proceedings of the Fifth*

International Conference on Language Resources and Evaluation (LREC-2006), pages 1372–1377.
Pennacchiotti, Marco, Diego De Cao, Roberto Basili, Danilo Croce, and Michael Roth. 2008. Automatic induction of framenet lexical units. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 457–465.
Pianta, Emanuele, Luisa Bentivogli, and Christian Girardi. 2002. Multiwordnet: developing an aligned multilingual database. In *Proceedings of the First International Conference on Global WordNet*, volume 152, pages 55–63.
Pianta, Emanuele and Sara Tonelli. 2009. A novel approach to mapping framenet lexical units to wordnet synsets. In *Proceedings of the Eighth International Conference on Computational Semantics*, pages 342–345.
Pradhan, Sameer, Kadri Hacioglu, Valerie Krugler, Wayne Ward, James H. Martin, and Daniel Jurafsky. 2005. Support vector learning for semantic argument classification. *Machine Learning*, 60(1):11–39.
Shi, Lei and Rada Mihalcea. 2005. Putting pieces together: Combining framenet, verbnet and wordnet for robust semantic parsing. In *Computational linguistics and intelligent text processing*. Springer-Verlag, Berlin, pages 100–111.
Tan, Pang-Ning, Michael Steinbach, and Vipin Kumar. 2005. *Introduction to Data Mining*. Addison-Wesley, Reading, MA.
Tenny, Carol and James Pustejovsky. 2000. A history of events in linguistic theory. In *Events as Grammatical Objects*. Stanford: Center for the Study of Language and Information, pages 3–38.
Tesnière, Lucien. 1959. *Éléments de syntaxe structurale*. Klincksieck, Paris.
Vetulani, Zygmunt and Bartlomiej Kochanowski. 2014. "polnet - polish wordnet" project: Polnet 2.0 - a short description of the release. In *Proceedings of the 7^{th} Global WordNet Conference*, pages 400–404.
Weitzman, Murray S. 1970. *Measure of the Overlap of Income Distribution of White and Negro Families in the United States*. U.S. Bureau of the Census.