Quantitative methods for identifying systematic polysemy classes

Laure Vieu, Elisabetta Jezek, Tim van de Cruys

Quantitative methods for identifying systematic polysemy classes

Laure Vieu
IRIT-CNRS, Toulouse &
LOA-ISTC-CNR, Trento
vieu@irit.fr

Elisabetta Jezek
Università di Pavia
jezek@unipv.it

Tim Van de Cruys
IRIT-CNRS, Toulouse
tim.vandecruys@irit.fr

***Abstract*—In this paper we report the results of four experiments conducted to extract lists of nouns that exhibit inherent polysemy from corpus data following semiautomatic and automatic procedures. We compare the methods used and the results obtained. We argue that quantitative methods can be used to distinguish different classes of polysemous nouns in the language on the basis of the variability of copredication contexts.**

I. Motivation and goals

In this paper we examine nouns which exhibit systematic (or regular) polysemy, i.e. general sense alternations of the kind `animal-food`, where the same relation holds between the meanings for a series of lexical items in a language (such as *chicken*, *rabbit*, *codfish*, *lamb*, etc.) and is not particular to a single one (cf. [1] and subsequent work). The goal is twofold: to acquire nouns belonging to different polysemy alternations from corpora and group them in different classes based on the underlying nature of their systematic polysemy, which we assume not to be univocal. Specifically, we are interested to tell apart nouns which exhibit systematic polysemy because of their regular ability to, in a single occurrence, convey multiple aspects and thus denote entities having a complex type (for instance, `physicalobject-informationobject`, e.g., 'book', `event-food`, e.g., 'lunch') from nouns usually presenting a single aspect in an occurrence and whose systematic polysemy is more likely due to (more or less lexicalized) coercion effects triggered by the linguistic and/or the pragmatic context (`animal-food`, e.g., 'chicken', `container-containee`, e.g., 'bottle').

Previous work has shown that copredication,[1] the usual test employed in the literature to distinguish the first kind of nouns (variously called "complex or dotted type nouns" [3], "nouns with facets" [4], "dual aspect nouns" [2] "inherently polysemous nouns" [5]) from other kinds of systematically polysemous nouns, including "selectional polysemy" [5] and "pseudo-dots" [3] such as `animal-food` or `container-containee`, is not sufficient because copredication is also possible, albeit less frequent, with expressions which exhibit polysemy due to coercion effects. This is the case of the noun *sandwich* in such contexts as *Sam grabbed and finished the sandwich in one minute*, in which *sandwich* is predicated both as a physical object and as the event of eating it. In an earlier work [6], we proposed that variability of pair of predicates in copredication contexts is the key to distinguish inherently polysemous nouns from nouns subject to coercion. According to this hypothesis, high variability of pair of predicates in copredication contexts is evidence of inherently polysemous nouns, while low variability points to nouns subject to coercion effects. Our work has also shown that the bottleneck of a quantitative methodology meant to distinguish different classes of polysemous nouns is the identification of predicates selecting for the different aspects of the nouns with high precision. Particularly, manual selection has proved to be very time consuming.

In this paper, we report the result of experiments we run to evaluate the whole methodology developed in [6], and to test the possibility to expand it in a way to automatize the selection of predicates exploiting distributional methods. Specifically, Section II-A outlines the methodology we adopted in [6]. Section II-B introduces the distributional method for selecting predicates we used for two experiments. Related work is discussed in III. Section IV presents the two main experiments and an evaluation procedure used to compare them with two baselines, the manual experiment of [6] and another one based on Lexit, a lexical resource for Italian. Section V discusses the results, and Section VI draws conclusions and offers hints for further work.

II. Methodology

A. Manual method

As referenced in Section I, we previously conducted a corpus-based study to assess the possibility to empirically distinguish between complex type nouns and nouns subject to coercion through the analysis of copredication contexts [6]. We here take up the same global semi-automatic method, whose concrete goal is, for a given complex type, to compute the variability of copredication contexts of interest for each candidate noun in order to rank them. This variability is measured by the ratio of copredication contexts of interest over all copredication contexts for that noun in a corpus. Decreasing ratios are supposed to rank nouns from most likely being of the complex type at stake to most likely being of some other type but subject to coercion. We use the SketchEngine (henceforth SE, [7]) tagged Italian corpus ItTenTen10 (2,5 Gigawords) and its tools. The complex or dot type chosen for the previous study and here is `informationobject•physicalobject` (or `info•phys`) of which 'book' is taken to be the prototype in

[1] Copredication can be formally defined as a "grammatical construction in which two predicates jointly apply to the same argument" [2]. We focus here on copredication contexts in which the two predicates select for disjoint types. An example is *They burned the controversial books*, where the predicate *burned* selects for the `physicalobject` aspect (or sense) of the argument *books* while *controversial* selects for the `informationobject` aspect.

the literature, and the copredication pattern used, [V [Det N Adj]], exploits verbs and adjectives as predicates.

1) Manual predicate extraction: The copredication contexts of interest are those based on a verb and an adjective that each select for a different type. The first step of the method is therefore to pick four lists of predicates: transitive verbs selecting for `informationobject` (`info`) or `physicalobject` (`phys`) as objects and adjectives postmodifying nouns of either type. The starting point in [6] was 10 seed nouns[2] considered as good candidates of the complex type. Having gathered the most frequent 200 verbs and adjectives in the collocational profiles (*WordSketches*) of each of these seed nouns, 2-by-2 intersections and then union were performed, yielding 427 verbs and 388 adjectives. These predicates were manually doubly classified into `phys` and `info`, avoiding those too polysemic, generic, or subject to metaphorical uses. 65 `phys` and 53 `info` verbs, 18 `phys` and 127 `info` adjectives were so collected.

2) Computing the copredication context variability: We adopt the method developed in [6]. For each noun N to be tested,[3] all occurrences of the [V [Det N Adj]] pattern with V and Adj free, are automatically extracted from the corpus. These hits are grouped by pairs ⟨V, Adj⟩, that is, "copredication contexts" for this noun. Among these, the contexts of interest combine selected predicates from the four lists, either ⟨V`phys`, Adj`info`⟩ or ⟨V`info`, Adj`phys`⟩. The ratio of relevant contexts among all contexts is an indicator of the variability of `info`•`phys` copredication contexts for each noun, and this variability a sign of the conventionalisation of the ability of the lemma to jointly denote both `phys` and `info` referents. The hit ratio yields a different order than the context ratio, since a single relevant context may have a large incidence.

In [6], on the basis of a manual annotation of 200 (0,8%) hits for the noun *libro* (book), the recall had been estimated at 6%. Precision had also been estimated for *libro*: 118 (86%) extracted copredication hits were relevant cases. This brief quantitative evaluation was completed by a qualitative evaluation of the ranking, intuitions regarding the inherently polysemous character of the nouns being corroborated by the results. In spite of a limited evaluation, the previous study supported the conclusion that an experimental method to separate nouns of complex types from nouns subject to coercion appeared possible. The extension of this earlier work was also restricted by the criticality of the manual predicate selection. The task is difficult because there are almost no monosemous predicates, and one cannot rely only on highly specialized unfrequent predicates since the relevant copredications are sparse. It is also very time consuming and the process cannot be easily extended to other complex type nouns. This is why in the present paper we investigate proposals for expanding the method by substituting this critical phase with a predicate selection procedure as automatic as possible.

[2] *articolo, diario, documento, etichetta, fumetto, giornale, lettera, libro, racconto, romanzo* ('article', 'diary', 'document', 'label', 'comic', 'newspaper', 'letter', 'book', 'short novel', 'novel')

[3] [6] proposed a method to extract candidate nouns, which we ignore here.

B. Exploiting distributional semantics to select predicates

In this section we describe how we exploit a latent semantic distributional model in order to semi-automaticaly extract predicates from corpus. The key idea is that the latent dimensions that come out of our model hint at particular co-predication contexts, such as `phys`, `info` or both.

1) Non-negative matrix factorization: Our latent model uses a factorization technique called non-negative matrix factorization (NMF) [8] in order to find latent dimensions. The key idea is that a non-negative matrix **A** is factorized into two other non-negative matrices, **W** and **H**

$$\mathbf{A}_{i\times j} \approx \mathbf{W}_{i\times k}\mathbf{H}_{k\times j} \tag{1}$$

where k is much smaller than i, j so that both instances and features are expressed in terms of a few components. Non-negative matrix factorization enforces the constraint that all three matrices must be non-negative, so all elements must be greater than or equal to zero.

Using the minimization of the Kullback-Leibler divergence as an objective function, we want to find the matrices **W** and **H** for which the divergence between **A** and **WH** (the multiplication of **W** and **H**) is the smallest. This factorization is carried out through the application of two update rules, iteratively updating both matrix **W** and **H** in an alternating fashion (see [8] for details).

2) Combining different copredication contexts: Using an extension of non-negative matrix factorization [9], it is possible to jointly induce latent factors for three different modes (nouns, verbs, and adjectives) that appear within our investigated copredication pattern [V [Det N Adj]]. As input to the algorithm, two matrices are constructed that capture the pairwise co-occurrence frequencies for the different modes. The first matrix contains co-occurrence frequencies of nouns cross-classified by verbs, and the second matrix contains co-occurrence frequencies of nouns cross-classified by adjectives that appear within the designated pattern. NMF is then applied to the two matrices, and the separate factorizations are interleaved (i.e. matrix **W**, which contains the nouns by latent dimensions, is shared between both factorizations). A graphical representation of the interleaved factorization algorithm is given in figure 1. The numbered arrows indicate the sequence of the updates.

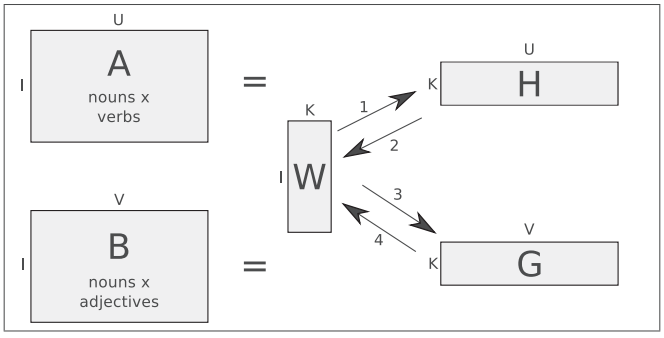

Fig. 1. A graphical representation of the interleaved NMF

When the factorization is finished, the three different modes (nouns, verbs, and adjectives) are all represented according

to a limited number of latent dimensions. Our hypothesis is that certain dimensions are tied up with certain basic types (such as phys or info), and complex types are constituted by a number of those basic dimensions; by exploiting these dimensions, we are able to perform a semi-automatic extraction of those complex types.

III. Related work

A. Extracting systematically polysemous nouns

Previous attempts have been made to semi-automatically acquire nouns considered as complex type nouns from lexical resources and corpora, see for example the Corelex database [10], and the experiments reported in [11] and [12]. The difference with the present work and [6] on which it builds is that in those contributions no distinction is drawn between the two types of systematic polysemy we aim at assessing with our method. The notion of systematic (or regular) polysemy as a whole and the sub-case of inherent polysemy are conflated, either for practical purposes or from disregard of the theoretical distinction. Moreover, systematic association of a certain noun with two aspects is not assessed in a copredication setting but in different syntactic contexts, thus not exploiting the characterization of inherent polysemy. The extracted lists of "complex type" nouns actually contain inherently polysemous nouns as well as pseudo-dots and other systematically polysemous nouns. As a result, the question whether nouns exhibiting systematic polysemy may not constitute a homogeneous class and may crucially differ in their underlying semantic representation is not assessed empirically.

B. Extracting selectional preferences of predicates

Our method is related to previous work on selectional preference acquisition. Most closely related is the work of [13], who present a clustering method that is able to extract selectional preferences for complex types by using the notion of contextualized similarity. By automatically clustering predicates that appear within a particular context and manually tagging the resulting clusters, they are able to acquire basic selectional preferences for separate aspects of complex type nouns. Within the broader context of selectional preference acquisition, our work is related to [14], who propose an Expectation-Maximization (EM) clustering algorithm for selectional preference acquisition based on a probabilistic latent variable model, and [15], who presents a model for multi-way selectional preference induction, making use of a latent factorization model for three-way co-occurrences.

IV. Experiments

The experiment described in [6] and taken up here as a baseline is called "*Manual*". As explained in Section II-A1, it is based on 4 manually selected lists of adjectives and verbs selecting for either the info or the phys aspect of nouns to be tested, containing from 18 (phys adjectives) to 126 (info adjectives) predicates. To test the possibility to bypass the costly manual selection of predicates, three more experiments have been run.

The two main ones rely on latent dimensions computed by non-negative matrix factorization, as explained in Section II-B. We extracted our co-occurrence frequencies from the freely available ItWaC corpus [16], using 1K verbs, 4K nouns and 2K adjectives that occur most frequenty in our corpus. We set our number of latent dimensions to $k = 100$. In the experiment called "*Dimensions-nouns*"/"*DimsN*", 15 dimensions most associated with the same 10 seed info•phys nouns as in *Manual* were examined. Fig. 2 shows two of these 15 dimensions, one considered as characterizing the info basic type, and the other the phys basic type. We manually selected between 2 and 5 dimensions among the 15 ones for each type of predicate and we gathered the first 20 predicates from each dimension. This yielded 4 lists containing from 37 (phys adjectives) to 91 (info adjectives) predicates. In the experiment called "*Dimensions-preds*"/"*DimsP*", we aimed at reducing even more the burden and possible arbitrariness of this manual selection of dimensions, although already considerably lower than for *Manual*. We gathered the first 20 predicates from the 5 first dimensions mostly associated with one of 4 seed lists of 10 predicates. These seed lists were manually extracted from the 4 predicates lists of *Manual*, aiming at generality, that is, avoiding predicates overly associated with info•phys nouns, albeit selecting for only one of the two aspects (e.g., keeping *spostare* (to move) but not *fotocopiare* (to photocopy) as seed phys verb). This yielded 4 balanced lists containing 85-93 predicates. The difference between these two dimension-based experiments, *DimsN* and *DimsP*, lies in where the manual intervention occurs: for picking the dimensions best corresponding to each type of predicate, or for a priori making up seed lists for the same types of predicates.

The last experiment, called "*Lexit*", was aimed to serve as a second baseline. It is a fully automatic method exploiting existing semantic resources. Namely, we utilized the distributional semantic profiles of verbs and adjectives in the Lexit resource (http://lexit.fileli.unipi.it/ [17]) extracted from the *La Repubblica* corpus and based on 24 noun semantic classes or "supersenses" in the Italian section of MultiWordNet [18]. We selected the 50 first verbs or adjectives most associated with nouns (in object position for the verbs, and in "mod-pre" position for the adjectives) of either of the two semantic classes "communication" (for info) and "artefact" (for phys).

For all four experiments, results were computed following the method described in II-A2 on the corpus ItTenTen10 from SE. To compare the results, we used an evaluation procedure exploiting an external resource. We evaluated precision and recall on the basis of the systematic polysemy relations in the resource Parole-Simple-CLIPS (henceforth PSC, [19]), in which 195 nouns have senses put into "PolysemySemioticArtefact-InformationObject" relation. Since copredication contexts are sparse, we retained only the 24 such nouns with high frequency (above 200K) in SE. These 24 nouns were complemented by 76 distractors randomly chosen among the 709 (699 manually cleaned) highly-frequent nouns in SE. As far as we know, PSC is the only resource for Italian encoding systematic polysemy. However, it doesn't distinguish inherent polysemy from other systematic polysemies. Indeed some nouns arguably aren't of the complex type phys•info: 6 out of 24 are derived from communication verbs, and their primary sense is eventive, combined with an info sense in some systematic polysemy, but not directly with any phys sense.[4] For computing preci-

[4]These 6 nouns are: *atto* (act), *contratto* (contract), *decreto* (decree), *rapporto* (report), *relazione* (report), *sentenza* (judgment). *Pubblicazione* (publication) also is a deverbal noun, but much closer to the target, since the result of a publishing event directly is an info•phys entity.

Verbs	Adjectives	Nouns
narrare (narrate)	*antico* (ancient)	*leggenda* (legend)
raccontare (tell)	*greco* (Greek)	*favola* (fable)
imparare (learn)	*volgare* (vulgar)	*fiaba* (fable)
conoscere (know)	*latino* (Latin)	*storia* (story)
inventare (invent)	*crudele* (cruel)	*latino* (Latin)
evocare (evoke)	*medievale* (medieval)	*greco* (Greek)
apprendere (learn)	*saggio* (wise)	*dialetto* (dialect)
credere (believe)	*triste* (sad)	*mito* (myth)
sognare (dream)	*medioevale* (medieval)	*antico* (ancient)
insegnare (teach)	*romantico* (romantic)	*mestiere* (profession)
recitare (recite)	*napoletano* (Neapolitan)	*eroe* (hero)
sapere (know)	*italico* (Italic)	*poesia* (poetry)
tradurre (translate)	*eroico* (heroic)	*lingua* (tongue)
parlare (talk)	*nobile* (noble)	*poeta* (poet)
amare (love)	*parlato* (spoken)	*danza* (dance)
ispirare (inspire)	*indiano* (Indian)	*arabo* (Arab)
dipingere (paint)	*popolare* (popular)	*comico* (comic)
adorare (adore)	*orientale* (eastern)	*accento* (accent)
diventare (become)	*moderno* (modern)	*spagnolo* (Spaniard)
vivere (live)	*cinese* (Chinese)	*dramma* (drama)

Verbs	Adjectives	Nouns
compilare (compile)	*cartaceo* (of paper)	*fotocopia* (photocopy)
allegare (attach)	*elettronico* (electronic)	*copia* (copy)
spedire (send)	*allegato* (attached)	*certificato* (certificate)
corredare (equip)	*inviato* (sent)	*documento* (document)
inviare (send)	*apposito* (specific)	*ricevuta* (receipt)
inoltrare (forward)	*modulistico* (of form)	*modulo* (form)
stampare (print)	*leggibile* (readable)	*questionario* (questionnaire)
copiare (copy)	*firmato* (signed)	*autocertificazione* (self-certification)
archiviare (file)	*informatico* (of computer)	*pdf* (pdf)
ricevere (receive)	*digitale* (digital)	*documentazione* (documentation)
modulare (modulate)	*valido* (valid)	*informazione* (information)
certificare (certify)	*scaricabile* (downloadable)	*E-mail* (e-mail)
trasmettere (transmit)	*On-line* (on-line)	*dato* (datum)
consegnare (deliver)	*telematico* (telematic)	*posta* (mail)
depositare (deposit)	*redatto* (written)	*verbale* (report)
reperire (find)	*disponibile* (available)	*originale* (original)
redigere (write)	*lino* (of linen)	*scheda* (card)
sottoscrivere (sign)	*postale* (postal)	*certificazione* (certificate)
pervenire (reach)	*reperibile* (available)	*autenticazione* (authentication)
munire (provide)	*identificativo* (identifying)	*formato* (format)

Fig. 2. Two dimensions (first 20 lemmas), selected respectively as `info` and `phys` for both verbs and adjectives in the *DimsN* experiment

	Random	Manual	DimsN	DimsP	Lexit
P @ 5	0.18	**1.00**	**1.00**	0.40	**1.00**
P @ 10	0.18	**0.90**	**0.90**	0.40	0.80
P @ 15	0.18	**0.80**	0.73	0.33	**0.80**
P @ 18	0.18	**0.78**	**0.78**	0.33	0.72
AP @ k	0.18	**0.92**	0.89	0.46	0.89

Fig. 3. Precision@5, 10, 15 and 18, Average Precision@k

Manual			Dimensions-nouns			Dimensions-preds			Lexit		
	C #	C %		C #	C %		C #	C %		C #	C %
lettera* (letter)	121	2.19	***rivista***	26	0.86	***libro****	147	1.35	***lettera****	140	2.54
documento* (document)	168	1.47	***testo***	78	0.86	***disco***	72	1.29	***documento****	288	2.53
rivista (magazine)	32	1.06	***giornale****	32	0.82	*tecnologia*	77	1.27	***pagina***	207	2.47
articolo* (article)	70	1.06	***documento****	92	0.81	*nome*	114	1.07	***carta***	166	2.44
pubblicazione (pubblication)	28	1.01	***disco***	40	0.72	*linguaggio* (language)	81	0.96	***testo***	209	2.31
libro* (book)	95	0.87	***libro****	77	0.71	*lato* (side)	72	0.90	***lista***	82	2.15
testo (text)	74	0.82	***pubblicazione***	19	0.69	***pagina***	72	0.86	***libro****	233	2.15
pagina (page)	66	0.79	***guida***	34	0.66	*faccia* (face)	34	0.86	***articolo****	135	2.04
memoria (memoir)	33	0.73	***lettera****	33	0.60	***rivista***	24	0.80	*decreto*	63	1.82
significato (meaning)	60	0.71	*prodotto* (product)	115	0.56	*stagione* (season)	47	0.79	*contratto*	88	1.53
guida (guide)	36	0.70	*tecnologia* (technology)	34	0.56	*stile* (style)	68	0.78	***giornale****	57	1.47
giornale* (newspaper)	27	0.70	*nome*	59	0.55	*epoca* (era)	18	0.78	***disco***	80	1.43
relazione (report)	93	0.59	***volume***	23	0.52	*attore* (actor)	30	0.77	***rivista***	43	1.43
carta (chart, paper)	37	0.54	***articolo****	32	0.48	***giornale****	30	0.77	*sentenza*	40	1.41
nome (name)	55	0.52	*titolo*	48	0.46	*essere* (being)	59	0.75	***programma***	245	1.36
programma (program)	92	0.51	***programma***	83	0.46	***lettera****	40	0.73	***volume***	60	1.35
lista (list)	19	0.50	***carta***	30	0.44	*posto* (place)	60	0.72	*calcio* (soccer, calcium)	26	1.33
titolo (title)	49	0.47	***pagina***	30	0.36	*prodotto*	144	0.71	*titolo*	137	1.32
decreto (decree)	13	0.38	***lista***	13	0.34	***testo***	61	0.67	***pubblicazione***	36	1.30
contratto (contract)	19	0.33	*relazione*	38	0.24	***pubblicazione***	17	0.61	***memoria***	53	1.17
rapporto (report)	65	0.28	***codice***	14	0.20	***documento****	65	0.57	***guida***	60	1.17
volume (volume)	12	0.27	***memoria***	8	0.18	***volume***	23	0.52	***codice***	77	1.12
disco (disk, record)	15	0.27	***legge***	14	0.13	***carta***	35	0.51	*relazione*	154	0.98
atto (act)	37	0.25	*atto*	18	0.12	***articolo****	29	0.44	*rapporto*	148	0.64
sentenza (judgment)	6	0.21	*contratto*	6	0.10	*contratto*	24	0.42	*atto*	80	0.54
legge (law)	19	0.17	*rapporto*	15	0.06	***memoria***	18	0.40	***legge***	39	0.35
codice (code)	8	0.12	*sentenza*	1	0.04	***codice***	25	0.36			
			decreto	0	0.00	***programma***	63	0.35			

Fig. 4. First 18 nouns ranked by relevant context ratio (C %), with number of such contexts (C #). Ranking of remaining 24 PSC nouns (first 10 for DimsP) below the double line. 18 gold nouns **in bold**, 5 seed nouns **starred***, 6 extra PSC nouns underscored. Translation only on the first occurrence of each noun.

sion we considered the sub-list of the remaining 18 nouns as gold for `phys`•`info`, but kept the 6 "extra" PSC nouns in our list of 100 nouns to be ranked. The gold list includes 5 nouns out of the 6 highly frequent seed nouns from [6].

V. Results

Results comparing the methods are assessed both quantitatively, in terms of Precision@k measured with respect to the 18 gold nouns from PSC (Fig. 3), and qualitatively, in terms of lists of best ranked nouns (Fig. 4). *Manual* gives the best results (0.92 Average Precision@k), closely followed by semi-automatic *DimsN* and fully automatic *Lexit* at a tie (0.89). Among these three best methods, the rankings of the whole list of 24 PSC nouns can be compared using Spearman's ρ: *Manual* and *DimsN* are the most similar (0.64, p=0.0007) followed by *Manual* and *Lexit* (0.58, p=0.0030).

Although it would seem that the fully automatic *Lexit* method is to be preferred, the semi-automatic *DimsN* one is the most flexible. The Lexit resource leverages on a large manually crafted lexicon, thus embedding manual work. But it makes use of 24 fixed semantic classes, not easily expandable without restructuring MultiWordNet's top-level, a highly difficult task [20]. Moreover, the 24 classes do not fully match the types involved in inherent polysemy, as can be seen for the type `info` and the class "communication" which includes speech act nouns in MultiWordNet.[5] We thus believe that semi-automatic *DimsN*, requiring little manual work to review 15 dimensions of 20 nouns, verbs and adjectives each in order to pick the relevant ones, is actually more promising. Our attempt with the *DimsP* experiment to further restrict the manual work in this semi-automatic method, by automatically picking the dimensions using manually selected seed predicates, doesn't work though, the precision of *DimsP* being much lower at 0.46 Average Precision@k. Further experiments shall compare *DimsN* with the semi-automatic clustering method proposed in [13].

VI. Conclusion

The evaluation procedure proposed, with high precision values (0.92 Average Precision@k for the best method), confirms that the whole method originally proposed in [6] based on the variability of pairs of predicates in copredication contexts is able to discriminate inherently polysemous nouns from those subject to coercion. The rankings also confirm our hypothesis that the PSC systematic polysemy relation considered, "PolysemySemioticArtefact-InformationObject", includes nouns subject to coercion, namely the 6 excluded ones (called here "extra") as well as *legge* (law) and *codice* (code) whose type probably is simply `info`. Context variability proves crucial, since a ranking based on the ratio of

[5]This is perhaps why the nouns *decreto* (decree) and *contratto* (contract), arguably not excellent examples of `info`•`phys` nouns, are ranked quite high with *Lexit* method as can be seen on Fig. 4.

relevant copredication hits instead of contexts systematically yields a lower Average Precision@k, with the best method's performance dropping from 0.92 to 0.60 (not shown on Fig. 3).

With this semi-automatic method we are able to discriminate inherently polysemous nouns from nouns subject to coercion, assuming a particular complex type is given. We believe that this method can now be extended in order to compare systematic polysemy patterns (in terms of type pairs), not just nouns. This would allow telling empirically apart those pairs of types that form a complex type, and therefore belong to the class of inherent polysemy (like, as usually assumed, `physicalobject-informationobject` or `event-food`), from those that belong to classes of systematic polysemy based on coercion, such as the "pseudo-dots" `animal-food` and `container-containee`. Such an extension is conceivable only because the method proposed exploits a distributional semantics factorization technique and is scalable, as opposed to the original proposal made in [6].

Systematic polysemy studies that propose theories discriminating a variety of polysemy phenomena suffer from a lack of empirical data, as they use only a few manually crafted examples. On the other hand, existing corpus-based work collapses all cases of systematic polysemy into a single class. We have shown that more elaborate empirical methods can be developed to evaluate theoretical hypotheses regarding various systematic polysemy classes and foster further investigations in this field.

REFERENCES

[1] J. D. Apresjan, "Regular polysemy," *Linguistics*, vol. 12, no. 142, pp. 5–32, 1974.

[2] N. Asher, *Lexical Meaning in Context: A web of words*. Cambridge University Press, 2011.

[3] J. Pustejovsky, "A survey of dot objects," Technical report, 2005.

[4] D. A. Cruse, "Polysemy and related phenomena from a cognitive linguistic viewpoint," in *Computational Lexical Semantics*, Saint-Dizier and Viegas, Eds. Cambridge University Press, 1995, pp. 33–49.

[5] J. Pustejovsky, "From concepts to meaning. The role of lexical knowledge," in *Unity and Diversity of Languages*, Sterkenburg, Ed. John Benjamins, 2008.

[6] E. Jezek and L. Vieu, "Distributional analysis of copredication: towards distinguishing systematic polysemy from coercion," in *Proc. of CLiC-it 2014*, pp. 219–223.

[7] A. Kilgarriff, V. Baisa, J. Bušta, M. Jakubíček, V. Kovář, J. Michelfeit, P. Rychlý, and V. Suchomel, "The Sketch Engine: ten years on," *Lexicography*, pp. 1–30, 2014.

[8] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. of NIPS 2000*, pp. 556–562.

[9] T. Van de Cruys, "Using three way data for word sense discrimination," in *Proc. of Coling 2008*, pp. 929–936.

[10] P. Buitelaar, "Corelex: An ontology of systematic polysemous classes," in *Proc. of FOIS'98*, pp. 221–235.

[11] G. Boleda, S. Padó, and J. Utt, "Regular polysemy: A distributional model," in *Proc. of *SEM 2012*, pp. 151–160.

[12] L. Romeo, S. Mendes, and N. Bel, "A cascade approach for complex-type classification," in *Proc. of LREC 2014*, pp. 4451–4458.

[13] A. Rumshisky, V. A. Grinberg, and J. Pustejovsky, "Detecting selectional behavior of complex types in text," in *Proc. of Generative Approaches to the Lexicon GL2007*.

[14] M. Rooth, S. Riezler, D. Prescher, G. Carroll, and F. Beil, "Inducing a semantically annotated lexicon via em-based clustering," in *Proc. of ACL 1999*, pp. 104–111.

[15] T. Van de Cruys, "A non-negative tensor factorization model for selectional preference induction," in *Workshop on Geometrical Models of Natural Language Semantics*. 2009, pp. 83–90.

[16] M. Baroni, S. Bernardini, A. Ferraresi, and E. Zanchetta, "The wacky wide web: a collection of very large linguistically processed web-crawled corpora," *Language resources and evaluation*, vol. 43, no. 3, pp. 209–226, 2009.

[17] A. Lenci, "Carving verb classes from corpora," in *Word Classes: Nature, typology and representations*, Simone and Masini, Eds. John Benjamins, 2014, pp. 17–36.

[18] E. Pianta, L. Bentivogli, and C. Girardi, "Multiwordnet: developing an aligned multilingual database," in *Proc. of Global WordNet 2002*, pp. 55–63.

[19] N. Ruimy, M. Monachini, R. Distante, E. Guazzini, S. Molino, M. Ulivieri, N. Calzolari, and A. Zampolli, "CLIPS, a multilevel Italian computational lexicon: a glimpse to data," in *Proc. of LREC 2002*.

[20] A. Gangemi, N. Guarino, C. Masolo, and A. Oltramari, "Sweetening WordNet with DOLCE," *AI Magazine*, vol. 24, no. 3, pp. 13–24, 2003.