10.12871/CLICIT2014142

Distributional analysis of copredication: Towards distinguishing systematic polysemy from coercion

Elisabetta Jezek
Università di Pavia
jezek@unipv.it

Laure Vieu
IRIT-CNRS - Université Toulouse III
vieu@irit.fr

Abstract

English In this paper we argue that the account of the notion of complex type based on copredication tests is problematic, because copredication is possible, albeit less frequent, also with expressions which exhibit polysemy due to coercion. We show through a distributional and lexico-syntactic pattern-based corpus analysis that the *variability* of copredication contexts is the key to distinguish complex types nouns from nouns subject to coercion.

Italiano *In questo contributo sosteniamo che il test di copredicazione utilizzato in letteratura per motivare l'esistenza di tipi complessi è problematico, in quanto la copredicazione è possibile, seppur con minor frequenza, anche con espressioni che esibiscono un comportamento polisemico a seguito di coercion. Attraverso una analisi distribuzionale che utilizza pattern lessico-sintattici mostriamo come la variabilità dei contesti di copredicazione è la chiave per distinguere nomi associati a tipi complessi da nomi soggetti a coercion.*

1 Introduction

Copredication can be defined as a "grammatical construction in which two predicates jointly apply to the same argument" (Asher 2011, 11). We focus here on copredications in which the two predicates select for incompatible types. An example is (1):

(1) *Lunch was delicious but took forever.*

where one predicate ('take forever') selects for the event sense of the argument *lunch* while the other ('delicious') selects for the food sense.

Polysemous expressions entering such copredication contexts are generally assumed to have a *complex type* (Pustejovsky 1995), that is, to lexically refer to entities "made up" of two (or more) components of a single type; it is thus assumed for example that *lunch* is of the complex type event • food.[1] Copredication as a defining criterion for linguistic expressions referring to complex types is, however, problematic, because copredication is possible, albeit less frequent, also with expressions which exhibit polysemy because of coercion, as in the case of the noun *sandwich* in such contexts as (2):

(2) *Sam grabbed and finished the sandwich in one minute.*

where the predicate *grab* selects for the simple type the noun *sandwich* is associated with (food), whereas *finish* coerces it to an event. The claim that the event sense exhibited by *sandwich* is coerced is supported by the low variability of event contexts in which *sandwich* appears (as opposed to *lunch*); see for example *"during lunch"* (780 hits for the Italian equivalent in our reference corpus, cf. section 3) vs. *"*during the sandwich"* (0 hits).

Our goal is therefore twofold: evaluate whether at the empirical level it is possible to distinguish, among nouns appearing in copredication contexts, between complex types and simple (or complex) types subject to coercion effects; and propose a method to extract complex type nouns from corpora, combining distributional and lexico-syntactic pattern-based analyses. Our working hypothesis is that lexicalized complex types appear in copredication patterns more systematically, and so that high variability of pair of predicates in copredication contexts is evidence of complex type nouns, while low variability points to simple (or complex) type nouns subject to coercion effects.

In the sections that follow, we will first raise the questions what counts as a copredication and what

[1] Dot/complex types have received different terminologies in the literature, particularly *nouns with facets* (Cruse 1995) and *dual aspect nouns* (Asher 2011).

copredication really tell us about the underlying semantics of the nouns that support it. Then, we will introduce the experiments we conducted so far to verify our hypothesis. Finally, we will draw some conclusions and point at the experiments we have planned as future work.

2 Copredication

2.1 What counts as a copredication?

In the literature, what exactly counts as a copredication is not clear. Typically, copredication has been restricted to classic coordinative constructions as in (3), where the adjective *voluminoso* 'bulky' selects for the physical sense of book, while *impegnativo* 'demanding' selects for the informational one.

(3) *È un libro voluminoso e impegnativo*.
'It is a bulky and demanding book'.

Research has shown, however, that copredication patterns based on coordination do not frequently mix different aspects but tend to predicate on a single aspect, as in (4), where both adjectives select for the same event aspect of *costruzione* 'construction' (Jezek and Melloni 2011):

(4) *La costruzione fu lenta e paziente*.
'The construction was slow and patient'.

Moreover, it has been claimed that constructions different from coordinative (or disjunctive) ones can be copredicative; for example, copredications with anaphoric pronouns (5)a, and structures where one of the predicates is located in a subordinative clause, as in (5)b and (5)c.

(5) a. *He paid the bill and threw it away*. (Asher 2011, 63).
b. *La construction, qui a commencé hier, sera très jolie* (Jacquey 2001, 155).
'The building, which started yesterday, will be very nice'.
c. *Una volta completata, la traduzione si può caricare in una sezione apposita del sito* (Jezek and Melloni 2011, 27).
'Once completed, the translation may be uploaded in a special section of the site'.

These copredication patterns may be disputable from both a structural and semantic point of view because they involve pronouns and coreference, and one could argue that pronominalization leaves room for phenomena such as bridging and associative anaphora.

In our work we focus on what we argue is a less disputable copredication pattern, namely [V [Det N Adj]]. This pattern is instantiated in contexts such as the following, where for example the predicate *bruciavano* selects for the physical aspect of *book*, whereas *controversi* selects for the informational one:

(6) *... bruciavano i libri controversi*.
'... they burned the controversial books'.

2.2 What does copredication really tell us?

As referenced above, it has also been noted that copredication may actually involve coercion (Asher and Pustejovsky 2006; corpus evidence in Pustejovsky and Jezek 2008). Consider:

(7) *Aprire il vino rosso con 30 minuti di anticipo*.
'Open the red wine 30 minutes in advance'.

In (7), *vino* 'wine' appears to denote both drink and container in the same context, due to the two predicates *rosso* 'red' and *aprire* 'open'. Despite the apparent polysemy, the noun *vino* is generally assumed to be lexically associated with a simple type (drink), and to license a sense extension to container in specific contexts only, as a coercion effect induced by the semantic requirements of the selecting predicate.

We claim that a single occurrence of a relevant copredication context is not enough to identify a complex type, and we conjecture that a *variety* of copredication contexts appearing with enough regularity might constitute evidence. Indeed, one can observe that *vino* 'wine' displays a limited variability, since it cannot be coerced into a container type by any predicate that would felicitously apply to *bottiglia* 'bottle', as shown by (8):

(8) **Ho rotto il vino rosso*.
'I broke the red wine'.

3 The experiment

We conducted a corpus-based study to assess the possibility to empirically distinguish between complex types and simple (or complex) types subject to coercion effects through the analysis of copredication contexts. The concrete goal of the experiment was, for a given complex type, to extract a list of candidate nouns that do appear in some copredication context, and compute the variability of copredication contexts to order these nouns. The hypothesis is that nouns shall be ordered from most likely being of the complex type at stake to most likely being of some other type but subject to coercion. We exploited the SketchEngine (Kilgarriff et al. 2014) tagged Italian corpus It-

TenTen10 (2,5 Gigawords) and its tools. The complex type chosen for this first experiment was information_object • physical_object of which 'book' is taken to be the prototype in the literature, and as detailed above, the copredication patterns used are of the form [V [Det N Adj]].

3.1 Predicate extraction

The copredication contexts of interest are those based on a transitive verb and an adjective that each select for a different type. The first step was therefore to pick four lists of predicates: transitive verbs selecting for information_object (Info) or physical_object (Phys) as object complements and adjectives that modify nouns of either type.

The starting point was a list of 10 seed nouns[2] considered as good examples of the complex type. We extracted from the corpus predicates applying to these seed nouns, that are frequent and shared enough: on the most frequent 200 verbs (V) and adjectives (A) in the collocational profiles (*WordSketches*) of each of these seed nouns, we performed 2-by-2 intersections and then union, which yielded 427 V and 388 A. We manually doubly classified them into Phys and Info, avoiding predicates (too) polysemic, generic, or subject to metaphorical uses. We thus gathered 65 VPhys, 53 VInfo, 18 APhys and 127 AInfo.

3.2 Candidate extraction

Using a manually selected subset of 6-14 frequent predicates of each category, a series of concordance built on the copredication pattern with all context pairs ⟨VPhys, AInfo⟩ and ⟨VInfo, APhys⟩ produced nouns occurring in these contexts. We then manually annotated 600+ randomly taken hits, checking for actual copredication with both aspects, thus extracting 97 different nouns. The 5 seed nouns not present among these 97 were added, obtaining 102 nouns, as candidates for the complex type Info • Phys. For the rest of the experiment, since the relevant copredications are rather sparse, we focussed on the 54 nouns with frequency above 200,000, and selected 28 (52%) ones, aiming at covering most of the various types appearing among these and including 7 seed nouns (marked * in the table).

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

3.3 Computing the copredication context variability

For all 28 nouns we extracted all occurrences of the [V [Det N Adj]] pattern, N fixed. The hits of each lexico-syntactic pattern are grouped by pairs ⟨V, A⟩ that we here call "copredication contexts" for this noun. We then extract the *relevant* contexts ⟨VPhys, AInfo⟩ and ⟨VInfo, APhys⟩ combining selected predicates in our four lists. 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 lemma ability to jointly denote both Phys and Info referents.

The results, ordered from more variable to less variable, appear on Table 1, where **Hits** is the total number of hits of the lexico-syntactic pattern, **Cop. hits** are those hits with a relevant ⟨VPhys, AInfo⟩ or ⟨VInfo, APhys⟩ context, **Contexts** is the total number of ⟨V, A⟩ contexts, and **Cop. cont.** are the relevant ones. Ratios are in %.

Note that the hit ratio would yield a different order than the context ratio, since a single relevant context may have a large incidence. Indeed, with context ratio, the 7 seed nouns are ranked among the 10 first, while with the hit ratio, they would appear among the 14 first, and include at the very top *informazione* and *indicazione*, two nouns unlikely prototypes for the Info • Phys complex type.

4 Discussion

The copredication contexts extracted are sparse, and the ratio figures ordering the nouns are low (all below 3%). This might be due to the phenomenon of copredication across types being sparse, but obviously also because the 4 lists of predicates are by no means exhaustive. On the basis of a manual annotation of 200 (0,8%) hits on *libro*, the recall is estimated at 6%. A very high recall could not be reached without including polysemic or very generic predicates, thus lowering precision. Precision has been estimated for *libro*: 118 (86%) extracted copredication hits are indeed relevant cases. However, in the lower rows, precision drops: 9 (60%) for *volume* and even 0 for *fenomeno*, which means that if we had other means to screen the results, the ratio range would widen between top and bottom rows.

The method allows to distinguish four groups of lemmas (statistically significant partition, but finer-grained partitions could be drawn). At the

Lemma	Freq.	Hits	Cop. hits	Hit ratio	Contexts	Cop. cont.	Cont. ratio
lettera (letter)*	549552	13386	414	3.1	5513	130	2.4
giornale (newspaper)*	276139	6757	37	0.55	968	20	2.1
documento (document)*	547415	25615	313	1.2	11404	182	1.6
informazione (information)	1092596	68201	2635	3.9	18459	242	1.3
racconto (short novel)*	243777	7533	111	1.5	4418	56	1.3
capitolo (chapter)	218115	4982	60	1.2	2731	32	1.2
articolo (article)*	2458766	12885	104	0.81	6588	72	1.1
libro (book)*	968401	23958	137	0.57	10856	107	0.99
pagina (page)	716615	15850	111	0.70	8357	82	0.98
romanzo (novel)*	213778	7644	47	0.61	3844	35	0.91
testo (text)	528482	21080	108	0.51	9067	81	0.89
immagine (image)	641384	32097	256	0.80	19146	162	0.85
indicazione (indication)	279063	20831	651	3.1	6536	54	0.83
relazione (report)	744398	36274	467	1.3	15693	101	0.64
storia (story)	1505947	57074	235	0.41	21292	129	0.61
programma (program)	978951	39140	340	0.87	18029	103	0.57
parola (speech)	1087778	44619	139	0.31	16292	87	0.53
gioco (game)	637619	16815	60	0.36	8859	43	0.49
proposta (proposal)	716391	28007	149	0.53	12254	58	0.47
serie (series)	668564	12824	40	0.31	6872	31	0.45
dichiarazione (statement)	339720	13601	33	0.24	5817	25	0.43
fonte (source)	354620	20912	35	0.17	7692	33	0.43
riferimento (reference)	691282	18193	57	0.31	6705	27	0.40
ricerca (research)	1378351	25002	103	0.41	12228	46	0.38
carattere (character)	378986	45632	131	0.29	20504	70	0.34
volume (volume)	307808	6732	15	0.22	4445	15	0.34
pezzo (piece)	286093	13190	27	0.20	7201	23	0.32
prodotto (product)	837772	48285	72	0.15	20391	54	0.26
fenomeno (phenomenon)	342726	26876	20	0.074	11872	13	0.11

Table 1: Relevant copredication variability for 28 candidate Info • Phys nouns with high frequency

top, are those that arguably are prototypical examples of the complex type Info • Phys. Next comes a group of nouns with still classical examples of this dot-type, especially *libro*, as well as nouns of the simple type Info such as *informazione*. Since information objects generically depend on their physical realizations, coercion is readily available. What these data tell us is that the pure Info sense of *libro* (as in *il libro di Dante è stato tradotto in tante lingue* 'Dante's book has been translated in many languages') or *immagine* might prevail over their complex type sense. The next group gathers many nouns of a different complex type, Info • Event, such as speech act nouns, some of which, like *relazione*, do also have a lexicalized sense of document, while others, like *indicazione* and *dichiarazione*, are rather subject to coercion. The last group exhibits occasional coercion contexts, with the exception of *volume* which does have a standard Info • Phys sense but much less frequent than its spatial or sound quality sense.

We can therefore conclude that an experimental method to separate nouns of complex types from nouns subject to coercion appears possible. The proposed method constitutes the first attempt at semi-automatically extracting from corpus complex type nouns, something remaining elusive up to now. In addition, we learned that *letter* should be preferred over *book* as prototype of the complex type Info • Phys. In fact, this complex type is not the most straightforward since the dependence between the components of a dot object is not one-to-one. The case of Event • Food with *lunch* as prototype, in which there is such a tight symmetric dependence and no competition with separate simple senses, might prove easier to deal with. This will be tackled in a next experiment.

The predicate selection is a critical phase in the method proposed. It is difficult if not impossible to avoid polysemy and metaphorical uses, especially since the relevant copredications are sparse and we cannot rely only on highly specialized unfrequent predicates. In future work, we plan to experiment with fully automatic selection, exploiting distributional semantics methods. Dimension reduction through non-negative matrix factorization yields a possible interpretation of the dimensions in terms of "topics", which is confirmed by experiments (Van de Cruys et al. 2011). Building on this, we shall check whether "topics" for predicates correspond to selectional restrictions suitable to build our copredication patterns.

Acknowledgments

Thanks to Philippe Muller for help with programming issues and to the participants of the workshop on dot objects in May 2014 in Toulouse for feedback on early results of this work. We also acknowledge Tommaso Caselli and two anonymous reviewers for their useful comments.

Bibliography

N. Asher. 2011. *Lexical Meaning in Context: A Web of Words*. Cambridge: Cambridge University Press.

N. Asher and J. Pustejovsky. 2006. A type composition logic for generative lexicon. *Journal of Cognitive Science*, 6, 1–38.

D.A. Cruse. 1995. Polysemy and related phenomena from a cognitive linguistic viewpoint. In P. Saint-Dizier and E. Viegas *Computational Lexical Semantics*, CUP, 33–49.

E. Jacquey. 2001. *Ambiguités Lexicales et Traitement Automatique des Langues: Modélisation de la Polysémie Logique et Application aux déverbaux d'action ambigus en Français*. Ph.D. Dissertation, Université de Nancy 2.

E. Jezek, and C. Melloni. 2011. Nominals, Polysemy and Co-predication. *Journal of Cognitive Science*, 12, 1–31.

A. Kilgarriff, V. Baisa, J. Bušta, M. Jakubíček, V. Kovář, J. Michelfeit, P. Rychlý, and V. Suchomel. 2014. The sketch engine: ten years on. *Lexicography*, 1–30. `http://www.sketchengine.co.uk/`

J. Pustejovsky. 1995. *The Generative Lexicon*. Cambridge, MA: MIT Press.

J. Pustejovsky, J. and E. Jezek. 2008. Semantic coercion in language: Beyond distributional analysis. Distributional Models of the Lexicon in Linguistics and Cognitive Science. Special Issue on *Italian Journal of Linguistics*, 20(1), 175–208.

T. Van de Cruys, T. Poibeau, and A. Korhonen. 2011. Latent vector weighting for word meaning in context. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.