

Semantic Relations in Compound Nouns: Perspectives from Inter-Annotator Agreement

Prabha Yadav^a, Elisabetta Jezek^b, Pierrette Bouillon^c, Tiffany J. Callahan^a,
Michael Bada^a, Lawrence E. Hunter^a, and K. Bretonnel Cohen^a

^a Computational Bioscience Program, University of Colorado School of Medicine, Aurora, Colorado 80045, USA

^b Department of Humanities, University of Pavia, Italy

^c Faculté de Traduction et d'Interprétation, Université de Genève, Switzerland

Abstract

Semantic relations have been studied for decades without yet reaching consensus on the set of these relations. However, biomedical language processing and ontologies rely on these relations, so it is important to be able to evaluate their suitability. In this paper we examine the role of inter-annotator agreement in choosing between competing proposals regarding the set of such relations. The experiments consisted of labeling the semantic relations between two elements of noun-noun compounds (e.g. cell migration). Two judges annotated a dataset of terms from the biomedical domain using two competing sets of relations and analyzed the inter-annotator agreement. With no training and little documentation, agreement on this task was fairly high and disagreements were consistent. The results support the utility of the relation-based approach to semantic representation.

Keywords:

Natural Language Processing; Evaluation Studies as Topic

Introduction

Linguists have tried to discover the basic building blocks of semantic relations between nouns for decades, but there is still little consensus about what the set of those building blocks might be. This is an important problem for biomedical natural language processing and biomedical ontologies, because those building blocks are at the heart of our information extraction tasks and the structure of our ontologies.

Compound nouns are crucial to practical tasks like knowledge representation and to theoretical problems like understanding compositionality in semantics [1; 2]. However, one of the most difficult problems in semantic representation and in language processing is the nature of the relations between the two parts of a compound noun [3-6] (see examples in Table 1). Compound nouns are formed by a sequence of two or more nouns [7]. In writing, they may appear as two tokens (*knockout mouse*), a hyphenated word (*nucleotide-excision*), or a single token (database) [7]. They are about twice as common in written English as they are in spoken English (248/million words in newswire text versus 123/million words in conversation)[7]. They are quite common in scientific writing. Linh (2010) [8] reviews a number of studies of the incidence of compound nouns in technical texts, reporting that one study found that 27% of words in scientific abstracts were in compound nouns; another study found that 11.86% of anaphors are compound nouns; and another found that

15.37% of a technical corpus was made up of compound nouns.

The study of compound nouns dates back to Pāṇini and Kātyāyana and Patañjali [9], but an enormous amount of work remains to be done, particularly on the semantics of the relations between the nouns in a compound, and they remain the topic of a considerable amount of research in both linguistics and natural language processing [3]. Various authors have attempted to describe the relation between the elements of compound nouns from a theoretical perspective [10]. Likewise, a number of studies in language processing have shown the difficulty of classifying the relations in these compounds automatically [11-15]. All of these studies are based on specific representations of the relations that can hold between the nouns in a compound. This raises a question: are those relations valid? One way to answer that question is by measuring whether humans can reliably label the relation that holds in any specific compound. If they cannot, then we must question the validity of the representation itself, and we must consider the possibility that any principled investigation of the relations in compounds, whether from a theoretical or a practical perspective, is impossible (see, for example, the logical positivist perspective and how inter-annotator agreement responds to the problems of “observing” semantics [16]. On the other hand, if they can, then it might be possible to train computers to do the same task, which could enable considerable advances in natural language processing.

We can address the reliability of labeling through examining inter-annotator agreement when two or more analysts label the relations in a sample of compound nouns. However, we are not aware of any studies that have looked at inter-annotator agreement in compound nouns. Identifying relations in compound nouns, whether done by humans or by computers, is a non-trivial task because there is an enormous amount of ambiguity in the correspondence between semantics and syntactic structure. For example, Table 1 gives a number of examples using the biomedical term *forceps*. We note that *forceps* can exist in at least five relations with another noun in a compound—that is to say, the same noun-noun syntactic structure can correspond to at least five relations between the first noun and *forceps*.

Table 1- Identical syntactic structures can reflect a wide variety of semantic relations

Relation	Example
Used_on	<i>Bone forceps</i>
Instrument_for	<i>Epilation forceps</i>
Shape_of	<i>Mosquito forceps</i>
Operated_by	<i>Thumb forceps</i>
Named for	<i>Kelly forceps</i>

Consider the term *chondrocyte development*. The relation between the two nouns is an activity/physical process—development—that is undergone by chondrocytes. *Motor activity* and *thrombin activity* have the same syntactic structure as *chondrocyte development* and *as each other*, but the relationships between the nouns in all three of them are different: an activity that is undergone in the first, the result of the second, and the action of the third.

This article investigates the ability of humans to reliably label semantic relations between the elements of noun compounds in the face of this semantic ambiguity in identical syntactic structures. The motivation for this is that inter-annotator agreement on labelling the semantic relations in compound nouns is a useful indicator of the validity of the proposed set of relations and can be used in choosing between two competing theories. The experiment was done using two sets of relations with two different contexts of theoretical status and computational applications—Generative Lexicon theory on the one hand, and a model of the domain on the other—holding constant the data set and the annotators. Good inter-annotator agreement for a given set of relations would lend some credibility to that set; bad inter-annotator agreement would detract from its credibility.

Materials

Generative Lexicon

The first set of relations is the Generative Lexicon relations described in Bouillon et al., 2012 [17]. We will refer to this set as GL relations in the following, although the set in Bouillon et al. includes extensions with respect to the original GL set [18], for example the tag *argumental*. GL theory is an attempt to explain how compositionality contributes to lexical semantics. Bouillon et al., 2012 [17] posit two basic elements of lexical semantic representations: Qualia and/or Argumental. Qualia relations involve predicates and their arguments, as well; we will re-visit this issue in the discussion. They identified four basic Qualia relations: Formal, Constitutive, Telic, and Agentive. We used the set of Generative Lexicon relations described in Bouillon et al., 2012 [17]. These relations are meant to be general and elementary, embodying a hypothesis about the fundamental building blocks of semantic representations.

Rosario and Hearst

Rosario and Hearst (2001) [4] identified 38 relations broadly inspired by linguistic theory, but the motivation for the relations is less language-theoretic and more application-oriented. Specifically, it is intended to represent the semantics and knowledge structures of biomedical literature. Where GL theory is meant to be cross-linguistically valid, the set of relations proposed by Rosario and Hearst is meant to be domain-specific—thus, it does not attempt to be valid even for all of the English language, but just for English-language scientific literature. In comparison to many other proposed sets of semantic relations, including that of Generative

Lexicon theory (Table 2), the Rosario and Hearst relations do not posit a set of semantic primitives. Rather, they embody a knowledge representation schema that is specifically tailored to biomedical science without making any claims about what relations might exist in other domains.

Table 2- The characteristics of the two sets of relations in terms of their size, goals, and generality/domain specificity

Relation set	Size	Goal	General	Domain-specific
Generative Lexicon	15	Theoretical	X	
Rosario and Hearst	38	Application-oriented		X

The sample of terms for annotation was drawn from the Gene Ontology (GO). The GO is a good test case for this study because its content is clearly relevant to the biomedical domain and because it contains a large number of noun compounds. The GO project (<http://geneontology.org/>) was founded in 1998. The GO is an open-access, community-based effort to create a unified representation of three biological concept domains (biological processes, cellular components, and molecular functions) that are used to describe the activity of genes and gene products in a species-independent manner. Each of the 41,775 GO concepts includes a term name [19; 20].

Methods

To select the sample, all of the terms in the GO were tagged with their part of speech using the CLEAR suite of language processing tools [21; 22]. Tagging errors are noted in the data set. All terms with exactly two words, such that both words are nouns—that is, compound nouns—were pulled from the full set. We then selected a random sample of 101 words from the compound nouns (intending 100, with an extra in case of tagging errors).

Annotators

One annotator has a bachelor's degree in biology. The other annotator is a cardiovascular technologist with a doctorate in linguistics. They were instructed to base their annotations of the relation in a term on the definition of that term on the GO web site.

Evaluation

The inter-annotator agreement was measured using F-measure and Cohen's Kappa. In the calculation of kappa, P(e) (expected chance agreement) was calculated from the marginals of the confusion matrix. This is a conservatively high estimate of the P(e), and that should be kept in mind when interpreting the results.

Results

The Generative Lexicon relations from Bouillon et al.

The annotators used eight of the 15 possible relations to annotate the 101 GO terms (see Table 3). The most commonly used relation by annotator 1 was *argument* followed with *played_by*. The most commonly used relation by annotator 2 was *played_by* followed with *used_for*. Annotator 1 thought there were no proper relations available for four terms: *larval development*, *predatory behavior*,

lymphocyte anergy, and *lymphocyte homeostasis*. Table 4 shows the results in terms of true positives (1 for each match between the two annotators), false positives (1 for each mismatch between the two annotators), and false negatives (also 1 for each mismatch between the two annotators), and the corresponding measures of inter-annotator agreement. Cohen's Kappa value was 0.47 and the inter-annotator agreement, calculated as F-measure, was 0.58. The Cohen's Kappa value indicates a fair/good level of reliability according to the Green scale (1997). The annotators agreed that the Telic relation was the most frequent relation, followed by the Argumental relation. Annotator 1 thought 54.45% of the terms were Telic and annotator 2 thought 70.29% of the terms were Telic (Table 3). 36.63% of the terms were annotated as Argumental by annotator 1 and 19.80% of the terms were annotated as Argumental by annotator 2.

As defined by Cohen (1960) [23], unweighted kappa was calculated using the following equation (Eq. 1):

$$k = \frac{p_o - p_e}{1 - p_e} \quad (1)$$

Solutions were verified using the irr R package [24]

Table 3- Distribution of relations using GL relations

Relation	Annotator 1 (%)	Annotator 2 (%)	Examples
Formal [is_a]	0	0.99	<i>predatory behavior</i>
Constitutive [made_of]	0.99	0.99	<i>dynactin motor</i>
Constitutive [member_of]	0.99	0	<i>kinin cascade</i>
Telic [used_for]	0	20.79	<i>chondrocyte differentiation,</i>
Telic [aims_at]	21.78	18.81	<i>translation reinitiation,</i>
Telic	32.67	30.69	<i>GTPase activity,</i>
Agentive [played_by]	2.97	7.92	<i>chondrocyte hypertrophy, heart wedging</i>
Argumental [Argument]	36.63	19.80	<i>protease binding, p53 binding</i>
Un-annotated	3.96	0	<i>lymphocyte homeostasis</i>

Table 4 Overall inter-annotator agreement using GL relations

TP	FP	FN	P	R	F	κ
59	42	42	0.58	0.58	0.58	0.47

Table 5- Bouillon et al. GL categories, relations, descriptions, Relation Ontology equivalent(s), and examples

GL Category	Relations	Descriptions	RO equivalent	Examples
Formal	is_a	N2 is a kind of N1	is_a/ subclass of	<i>predatory behavior</i>
	shape_of	N1 has the shape of N2		
	holds	N1 holds N2	Contains	
Constitutive	made_of	N1 is made of N2	has_proper_part	<i>dynactin motor</i>
	part_of	N1 is a part of N2	proper_part_of	
	located_in	N1 is spatially/temporally located in N2	located_in	
	member_of	N1 is member of N2	member_of	<i>kinin cascade</i>
	has members	N1 has N2 as members	has_member	
Telic	predicate	N1 has the purpose of (Predicate) N2	has_function	
	used_for	N1 is used for the activity N2		<i>chondrocyte differentiation</i>
	aims_at	N1 has N2 as result/end goal		<i>keratinocyte development</i>
	played_by	N1 denotes the function which is N2.	has_function	<i>cholinesterase activity</i>
Agentive	caused_by	N1 is created/brought into existence/caused by N2		<i>heart wedging, chondrocyte hypertrophy</i>
	derived_from	N1 is derived/extracted from N2	derives_from,	
			transformation_of	
Argumental	argument	N2 is an argument of N1		<i>p53 binding</i>

Table 6¹ shows the confusion matrix for these relations. Note that disagreements between the two annotators are largely systematic. For example, Annotator 1 and Annotator 2 used the *aims_at* relation a similar amount of times (22 and 19), agreeing in the case of 10 noun compounds (Annotator 1 10/22 (45.45%); Annotator 2 10/19 (52.63%)). Of the disagreements, 8/22 times that Annotator 1 labelled instances of the *aims_at* relation, Annotator 2 labelled the *used_for*

relation; 7/19 times that Annotator 2 labelled the *aims_at* relation, Annotator 1 annotated the *argument* relation. Thus, refining the guidelines such that it is clearer when to use *aims_at* versus *argument* and *used_for* would have a large effect on the inter-annotator agreement for all three of these relation types.

The Relation Ontology relations corresponding to the Generative Lexicon relations are shown in Table 5. We observed that there were no corresponding ontology relations for most of the Generative Lexicon relations. This is consistent with the suggestion that the Relation Ontology is missing content that is fundamental to representing the biomedical domain.

¹ Table 6 is on the GitHub Repository: <https://github.com/KevinBretonnelCohen/SemanticRelationsCompoundNouns.git/>

Table 7- Rosario and Hearst relations and examples. There are no Relation Ontology equivalents for these relations.

Name	Examples	Name	Examples
Wrong parse	exhibit asthma, ten drugs	Time of (1-2)	morning headache, hour headache
Subtype	headaches migraine, fungus candida	Measure of	relief rate, asthma mortality, asthma morbidity
Activity/Physical process	bile delivery, virus reproduction, bile drainage, headache activity	Person/center who treats	headache specialist, headache center, diseases physicians, asthma nurse
Ending/reduction	migraine relief, headache resolution	Instrument (1-2)	aciclovir therapy, chloroquine treatment
Defect in Loc.	lung abscess, artery aneurysm	Instrument (2-1)	vaccine antigen, biopsy needle
Change	papilloma growth	Instrument (1)	heroin use, internet use, drug utilization
Produces (on a genetic level)	polyomavirus genome, actin mRNA, CMV DNA, protein gene	Object	bowel transplantation, kidney transplant, drug delivery
Cause (1-2)	asthma hospitalizations, AIDS death	Misuse	drug abuse, acetaminophen overdose
Cause (2-1)	flu virus, diarrhea virus	Subject	headache presentation, glucose metabolism
Characteristic	receptor hypersensitivity, cell immunity	Purpose	headache drugs, HIV medications
Physical property	blood pressure, artery diameter	Topic	time visualization, headache questionnaire
Defect	hormone deficiency, CSF fistulas	Location	brain artery, tract calculi, liver cell
Physical Make Up	blood plasma, bile vomit	Modal	emergency surgery, trauma method
Person afflicted	AIDS patient, BMT children	Material	formaldehyde vapor, aloe gel, gelatin powder
Demographic attributes	childhood migraine, infant colic, women migraineur	Frequency/time of (2-1)	headache interval, attack frequency, football season
Bind	receptor ligand, carbohydrate ligand	Activator (1-2)	acetylcholine receptor, pain signals
Research on	asthma researchers, headache study	Activator (2-1)	headache trigger, headache precipitant
Attribute of clinical study	headache parameter, attack study, headache interview	Inhibitor	adrenoreceptor blockers, influenza prevention
Procedure	tumor marker, genotype diagnosis		
Beginning of activity	headache induction, headache onset	Standard	headache criteria, society standard

Rosario and Hearst relations:

The annotators used 10/38 of the Rosario and Hearst relations to annotate the 101 GO terms (Table 7). The inter-annotator agreement, calculated via Cohen's Kappa, was 0.37 (Table 8). Table 9² shows the confusion matrix for these relations.

Table 8: Overall inter-annotator agreement using Rosario and Hearst relations

TP	FP	FN	P	R	F	κ
71	30	30	0.70	0.70	0.70	0.37

The inter-annotator agreement calculated using the F-measure was 70.29%. The maximum number of terms were annotated as *Activity/Physical Process* followed by *Characteristics* and *Material*: *Bind*. The annotators observed that there was no good representation for movement terms (for example, *cilium movement*). The inter-annotator agreement is good, given the fact that no training was provided. Again, the disagreements were quite systematic. Of the 30 disagreements, 20 (2/3) were from one cell of the table: Annotator 1 classified 20 compounds as having the *Material*: *Bind* relation, while Annotator 2 classified the same 20 compounds as having the *Characteristic* relation.

Discussion

The inter-annotator agreement was much higher for the Rosario and Hearst relations than for the Generative Lexicon relations. This is surprising, since the set of Rosario and Hearst relations is much larger than the set of Generative Lexicon relations.

It's premature to say why this is the case, but we can propose some avenues for future investigation: (a) This result might be related to the fact that the Rosario and Hearst relations are domain-specific, while the Generative Lexicon relations are not; (b) this result might be related to the fact that the Generative Lexicon relations are abstract and theoretically motivated, while the Rosario and Hearst relations are concrete and motivated by practical considerations; (c) it might be related to the observation that the annotators only used 10 of the Rosario and Hearst relations implying that the difference in size might not be as big as it seems and the difference in IAA may not be quite as surprising; (d) the difference IAA might go away with actual annotation guidelines and training; and (e) we should also point out that the affordances of the two are different--in particular, the Rosario and Hearst relations might be better for defining information extraction tasks while the Generative Lexicon's relations might be better for supporting inference.

Conclusion

The assumption behind the methodology that was applied here is that inter-judge agreement on annotation task is capable of finding problems in a set of semantic relations. The inter-annotator agreement in the cases of both proposed sets of semantic relations approached that of many completed and published corpus annotation projects, even with very minimal guidelines and no real training. The agreement on this task was fairly high in both cases and disagreements were quite consistent, supporting the basic soundness of the relation-based approach to semantic representation and suggesting that it is not overly subjective. From a methodological perspective, the results suggest that higher levels of agreement and reliability can be reached with some training and refinement of the guidelines. IAA was different between

² Table 9 is on the GitHub Repository:

"<https://github.com/KevinBrettonelCohen/SemanticRelationsCompoundNouns.git>"

the two sets of relations, suggesting that IAA can differentiate between semantic representations, although a number of possible explanations for those differences should be pursued in future work.

The relatively high IAA suggests that the descriptions and examples of the relations in the Bouillon and Rosario and Hearst papers were easy to follow and that the annotators were able to clearly delineate the relations and the tags in most cases. This is consistent with the claim that they are precise and not overly subjective in their interpretability and applicability. Disagreements between the analysts were quite consistent. This suggests that a higher level of agreement and reliability can be achieved with a little training and refinement of the guidelines. This study should be replicated on a larger scale with proper guidelines and training to achieve a higher level of reliability.

An additional benefit of approaching the evaluation of a set of relations through an annotation task was that we uncovered some shortcomings of the relations. We noted that (a) there is no good representation of movement in the Rosario and Hearst relations, and (b) some of the GO terms were not representable at all with those relations. In the case of the Generative Lexicon relations, we observed frequent confusion between Qualia and Argument (especially *used for* and *aims at*). This suggests that there is a need to clarify the demarcation between the two. A fruitful direction for future work would be to evaluate the nature of any correspondences that might exist between the two sets of relations. The work reported here contributes to the basis for such an effort.

Acknowledgements

We thank NIH grants LM008111 and LM009254 to Lawrence E. Hunter and NSF grant IIS-1207592 to Lawrence E. Hunter and Barbara Grimpe for funding this work.

References

- [1] J. Fan, K. Barker, and B.W. Porter, The knowledge required to interpret noun compounds, in: *IJCAI*, Citeseer, 2003, pp. 1483-1485.
- [2] A. Tribble and S.E. Fahlman, Resolving Noun Compounds with Multi-Use Domain Knowledge, in: *FLAIRS Conference*, 2006, pp. 122-127.
- [3] P. Nakov, On the interpretation of noun compounds: Syntax, semantics, and entailment, *Natural Language Engineering* **19** (2013), 291-330.
- [4] B. Rosario and M. Hearst, Classifying the semantic relations in noun compounds via a domain-specific lexical hierarchy, in: *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing (EMNLP-01)*, 2001, pp. 82-90.
- [5] M. Lauer and M. Dras, A probabilistic model of compound nouns, *arXiv preprint cmp-lg/9409003* (1994).
- [6] V. Nastase, P. Nakov, D.O. Seaghdha, and S. Szpakowicz, Semantic relations between nominals, *Synthesis Lectures on Human Language Technologies* **6** (2013), 1-119.
- [7] D. Biber, S. Johansson, G. Leech, S. Conrad, E. Finegan, and R. Quirk, *Longman grammar of spoken and written English*, MIT Press, 1999.
- [8] N.M. Linh, Noun-noun combinations in technical English, (2010).
- [9] S.D. Joshi and J. Roodbergen, *Patañjali's Vyākaraṇa-Mahābhāṣya Sthānavadbhāṣya: P. 1.1. 56-1.1. 57*, Bhandarkar Oriental Research Institute, 1990.
- [10] D.O. Séaghdha, Learning compound noun semantics, *University of Cambridge, Cambridge, UK* (2008).
- [11] B. Verhoeven, W. Daelemans, M. Van Zaanen, and G. Van Huyssteen, Automatic Compound Processing: Compound Splitting and Semantic Analysis for Afrikaans and Dutch, *ComAComA 2014* (2014), 20.
- [12] S. Tratz and E. Hovy, A taxonomy, dataset, and classifier for automatic noun compound interpretation, in: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, 2010, pp. 678-687.
- [13] L. Vanderwende, Algorithm for automatic interpretation of noun sequences, in: *Proceedings of the 15th conference on Computational linguistics-Volume 2*, Association for Computational Linguistics, 1994, pp. 782-788.
- [14] C. Dima and E. Hinrichs, Automatic Noun Compound Interpretation using Deep Neural Networks and Word Embeddings, *IWCS 2015* (2015), 173.
- [15] M. Lauer, Conceptual association for compound noun analysis, in: *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, Association for Computational Linguistics, 1994, pp. 337-339.
- [16] A.A. Leenaars and W.D. Balance, A logical empirical approach to the study of suicide notes, *Canadian Journal of Behavioural Science/Revue canadienne des sciences du comportement* **16** (1984), 249.
- [17] P. Bouillon, E. Jezek, C. Melloni, and A. Picton, Annotating qualia relations in Italian and French complex nominals, (2012).
- [18] J. Pustejovsky, The generative lexicon, (1995).
- [19] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, and J.T. Eppig, Gene Ontology: tool for the unification of biology, *Nature genetics* **25** (2000), 25-29.
- [20] G.O. Consortium, Gene ontology consortium: going forward, *Nucleic acids research* **43** (2015), D1049-D1056.
- [21] P.V. Ogren, P.G. Wetzler, and S. Bethard, ClearTK: A UIMA toolkit for statistical natural language processing, *Towards Enhanced Interoperability for Large HLT Systems: UIMA for NLP* **32** (2008).
- [22] S. Bethard, P.V. Ogren, and L. Becker, ClearTK 2.0: Design Patterns for Machine Learning in UIMA, in: *LREC*, 2014, pp. 3289-3293.
- [23] J. Cohen, A coefficient of agreement for nominal scales, *Educational and psychological measurement* **20** (1960), 37-46.
- [24] M. Gamer, J. Lemon, I. Fellows, and P. Singh, irr: Various Coefficients of Interrater Reliability and Agreement. R package version 0.84, Internet resource: [http://CRAN.R-project.org/package= irr](http://CRAN.R-project.org/package=irr). Verified April 10 (2012), 2013.

Address for correspondence

Kevin Bretonnel Cohen, e-mail id : kevin.cohen@gmail.com