Discovering and Modeling Knowledge Patterns from Tropes in Scientific Texts

Anna Sofia Lippolis[1,2,*,†]

[1]Dept. of Philosophy and Communication Studies, University of Bologna, Italy
[2]CNR Institute of Cognitive Sciences and Technologies, Rome, Italy

Abstract
In order to effectively convey concepts, scientific literature often derails or suspends the normal referentiality of language through figurative expressions. It is unsurprising, then, that science itself is rooted in metaphor and analogy for creating meaning. However, to understand the development of these phenomena and their consequences on society, most natural language processing solutions have tended to be merely based on prior quantifications of topics or lower level linguistic features. This work aims at bridging this gap by exploiting state-of-the-art knowledge extraction and representation techniques to discover and model knowledge patterns (KPs) in scientific texts. The hybridization of natural language processing and semantic technologies will foster the formalization and extraction of KPs from text used in a non-literal sense and abstractive form. Specifically, this work will: (i) detect tropes in a curated corpus; (ii) explore their relationship with other structural elements of the text; (iii) identify and formalize invariances into KPs and (iv) populate a knowledge graph based on this metamodel. The resulting insights and techniques will benefit knowledge representation and extraction techniques from texts in different research endeavors.

Keywords
Tropes, Knowledge Patterns, Science of Science, Scientific Articles

1. Problem statement

Combining science and rhetoric can sound oxymoronic. Nevertheless, it is a matter of fact that even the most objective-aspiring written artifact may leverage tropes.

These rhetorical devices, metaphor in particular, may convey complex notions, harmonize knowledge effectively and, ultimately, shape our scientific view of the world [1]. At the same time, tropes can constrain our understanding of natural processes. They may uphold outdated scientific paradigms, leading to misunderstandings among the general public.

For instance, as [2] notes, while the metaphor of genes as “blueprints” has long guided molecular biology research, critics argue that this view is deterministic, oversimplifies complex gene-gene and gene-environment interactions, and conflicts with recent advancements in developmental biology and epigenetics.

The language of science is indeed so trope-based that, according to T. Kuhn [3], even theory change must be expected to be accompanied by a change in metaphors and in the corresponding

Proceedings of the Doctoral Consortium at ISWC 2023 co-located with 22st International Semantic Web Conference (ISWC 2023)
annasofia.lippolis2@unibo.it (A. S. Lippolis)
0000-0002-0266-3452 (A. S. Lippolis)

parts of the network of language similarities through which terms attach to nature. When geneticists found themselves in need of a new way to frame their understanding of genetics, the language adopted by molecular biology, particularly the concept of the "genetic program", came to assume many of the roles that had previously been attributed to the "action" of individual genes.

As an increasing number of scientific literature is open access, more and more attention has gravitated towards linguistic and structural analysis. However, research in this field has often relied on the quantification of lower level linguistic features such as entities, word frequency or occurrences of parts of speech and other grammatical constructions. These aspects are not enough to gather a comprehensive insight on science communication.

This work explores the relationship between rhetoric and science by leveraging the adoption of the *Science of Science* [4] (SciSci) approach. To the best of our knowledge, this has been little explored so far. In such a framework, it is possible to refine existing state-of-the-art trope detection models for discovering and modeling related KPs in open access academic corpora.

KPs are structures that in different research areas are "used to organize our knowledge, as well as for interpreting, processing or anticipating information" [5, 6]. Using discovered and represented KPs as heuristics can simplify the coordination of universal invariances and localities, aligning with the human cognitive interpretation of the world. In this context, despite the concept of "frame" having been used in a range of different fields, a "frame" as defined by [7, 8], is a role or purpose that a cognitive process serves in acquiring and evaluating knowledge, and can be related to a KP. They can be both conceived as "primitives" [5], portions of background knowledge that connect language analysis with concepts and knowledge. Since tropes can be used to frame various cultural phenomena, examining patterns of frame variation in discourse is a crucial way to reflect how conceptualizations of societal issues change over time [9]. The unveiling of both universal and domain-specific patterns forms the foundation for the formalization of abstract schemata to organize scholarly knowledge into a knowledge graph (KG).

1.0.1. Importance.

Understanding the evolution of scientific knowledge is essential, as it influences research directions, funding decisions, and policy-making. Tropes reflect this change in science: they reveal underlying frameworks and rhetorical strategies, aiding in the comprehension of scientific language, and can serve as powerful tools to uncover claims and arguments embedded in scientific discourse [10]. Identifying and organizing recurring metaphors or unconventional language usage can shed light on underlying conceptual frameworks and reveal connections between different scientific domains, contributing to interdisciplinary research. Moreover, rhetorical devices can be used as ways to uncover claims and argumentations embedded in scientific discourse (like in the analysis of COVID-19 discourse in [11]), and foster a more collaborative science for the development of new, shared tropes that align with contemporary scientific goals. For instance, as some metaphors can be more restrictive in conceptualizing complex scientific issues, it is crucial to consider how they may contribute to public misunderstanding and unintentionally reinforce social and political messages that undermine inclusive science. To identify, assess, and unpack tropes is a task beneficial to students as well, as they can deepen their understanding of scientific concepts and cultivate a sense of civic responsibility [2]. KP-modelled tropes can find practical applications in the "perspective web" [12]. In fact, they make the contextuality and fuzziness of statements explicit by including figures of thought that are based on conventions and personal experience. A natural use case for KP-modelled tropes are nanopublications, as they facilitate transparent and focused information integration in how scientists conceptualize and communicate findings. Hence, KP-modelled tropes foster trustworthy and interdisciplinary connections. Leveraging semantic technologies further enhances these capabilities by enabling machine-readable representations of scientific statements and their contexts. In addition to contributing to the SciSci field, the implementation of these functions may lead to other applications, such as improvement of recommender systems in suggesting related scientific articles (e.g. metaphor identification as a feature for poetry recommendation systems in [13]).

2. Related work

Despite the emergence of the SciSci discipline [4], there has been limited scientific research devoted to the presence and function of tropes in scientific literature corpora. This is fairly evident in terms of ontological representation. Early research in *metaphor processing* performed supervised classification with hand-engineered lexical, syntactic and psycholinguistic features [14]. Alternative approaches are corpus-based [15] or, more recently, work by training deep neural models [16, 17], and may leverage a KG-based approach [18]. Various methods of metaphor processing have also focused on the role the trope plays in communication, especially political discourse [19]. While machine-learning-based detection of metonymy has been explored to some extent [20, 21], limited attention has been given to tropes beyond metaphor. Among the *thesauri* dedicated exclusively to metaphors, the largest and most commonly used is the VU Amsterdam Metaphor Corpus [22] (VUA). However, existing corpora tend to hinder the analysis by not allowing the identification of different types of metaphors: the VUA has a high percentage of conventional metaphors, making it difficult to capture novel ones. For what concerns *tropes interpretation*, MetaNet [23] is the reference structured repository of conceptual metaphors, and provides alignments between conceptual metaphors and linguistic frames available in FrameNet [24]. In [25], the authors present the Amnestic Forgery Ontology, which relies on MetaNet, and is aligned with the Framester knowledge graph[1]. In this context, KPs are crucial in facilitating ontological reuse and clarifying formal models as sets of modular theories rather than mere formalization of axioms. To the best of our knowledge, most studies involving broad *semantic representations of rhetorical figures* have been developed independently of the rhetorical tradition and frame theory. Moreover, these models are not linked to frameworks of document representation or scholarly practices. Possible solutions in this direction are the SPAR ontologies [26], the Scholarly Ontology [27] (SO), the GRhOOT Ontology [28], and the Conference Ontology [29]. However, [12] define a pattern for the formal representation and extraction of perspectives, which can be the basis to outline a model that takes into account how researchers make sense of the tropes they use in scientific communication, and how readers consciously and unconsciously interpret them.

[1] http://etna.istc.cnr.it/framester_web/.

Based on this survey, it can be concluded that a SciSci-based quantitative model of semantic figures of speech needs to be complemented by a formal model of reference to account for the multi-layered nature of conveying meaning. Such a framework has not yet been developed for scholarly texts.

3. Hypotheses and research questions

This research relies on three hypotheses:

H1 Tropes can be found in science discourse and formalized in KPs;
H2 A KG can be constructed by using identified KPs as schema;
H3 KPs can be used as interpretive lenses over the scientific literature by querying the KG.

The proposed approach enables users to explore through a KG various ways of meaning-making, both synchronously (i.e. how and where a metaphor is used at a specific moment in time) and diachronically (i.e. the evolution of a specific metaphor's usage over time). The assessment of this objective consists in the evaluation of the following research questions:

RQ1 To what extent and through which knowledge extraction-based tools and algorithms can unstructured data of tropes in scientific texts be re-engineered into linked open data for a KG?
RQ2 How to formalize the relationship between tropes and other structural and cognitive elements of the text?
RQ3 To what extent a KG of tropes in scientific texts can be used to foster qualitative and quantitative studies on the topic?

The research will contribute as follows:

- A contribution to semantic rhetorical figures datasets, such as the VUA [22] and UniMet[30];
- Modelling of domain-specific KPs and subsequent outline of the Tropes Ontology (tropes schemata representation, addressing philosophical and cognitive theories);
- Creation of the Tropes in Scientific Texts KG and alignment to Framester;
- Definition of the method and evaluation system.

4. Research methodology and approach

In this section, methodology and approach are elaborated for each Research Question.

RQ1. To transform unstructured data on tropes in scientific texts into a KG, the primary objective is to establish a pipeline that combines literature review of tropes in science with existing models for trope detection. This pipeline aims at generating dataset of tropes and convert it into a KG according to an ontology. Therefore, the initial step involves comprehending existing approaches and improving or merging them. It is crucial to carefully select a corpus and conduct comparative studies to evaluate the performance of state-of-the-art models such as MelBERT [31], and GPT [32] on the chosen corpus. While focusing on papers in humanities-related fields may inadvertently neglect books, which are commonly used in scholarly production, it is essential to consider various criteria for corpus selection to ensure a consistent and well-balanced dataset. Additionally, it is important to acknowledge that even available open-access articles might not possess proper structure or contain relevant information.

The corpus selection criteria for this study aim at constructing a reliable and comprehensive dataset, focusing on factors such as authority, content, and design. Specific emphasis is given to the following parameters: i) text authority, which entails selecting articles from journals with a minimum of 5 years of impact factor, ii) currency, ensuring that the selected texts cover the chosen timespan for the analysis (2000-2023), iii) English language, iv) availability of full texts and open access. The dataset resulting from the tropes identification process will be transformed into the KG, according to the outlined ontology, using PyRML[2], with specific details described in RQ2.

RQ2. KPs are identified and formalised by following the approaches outlined in [33]. That is, formulating use cases and requirements elicitation, modelling key notions derived from the use cases and checking consistency of the ODP. This process implies leveraging: (i) statistical measures; (ii) existing ontologies on rhetorical figures; (iii) cognitive and philosophical theories of tropes; and (iv) document representation. The resulting KPs will be modularized and networked following the eXtreme Design methodology [34]. This framework will be the basis to develop the Tropes ontology and populate it with the obtained data to create a KG.

RQ3. The time-aware KG resulting from extracting information about the evolution of science communication can help researchers understand the correlation between argumentative discourse, rhetorical figures, and framing power of theories. The helpfulness of the outcomes explained in 1.0.1 will be explored during the project's course.

5. Evaluation plan

For the automatic detection of tropes, we will employ widely used evaluation metrics such as precision, recall, and accuracy. These metrics will be applied to both established datasets like the VUA and an annotated sample of our dataset. Evaluating the KG entails assessing its cognitive soundness, which requires human involvement in the loop, in the form of gathering user feedback [35]. A goal-oriented evaluation will assess the effectivenes of KPs and the associated KG in facilitating the knowledge access process for humans based on specific goals. We will also employ standard KG evaluation metrics along with the three principles described in [36, 37]. In compliance with open science principles, all the code and results will be made publicly available as open source.

[2]PyRML library, available at https://github.com/anuzzolese/pyrml.

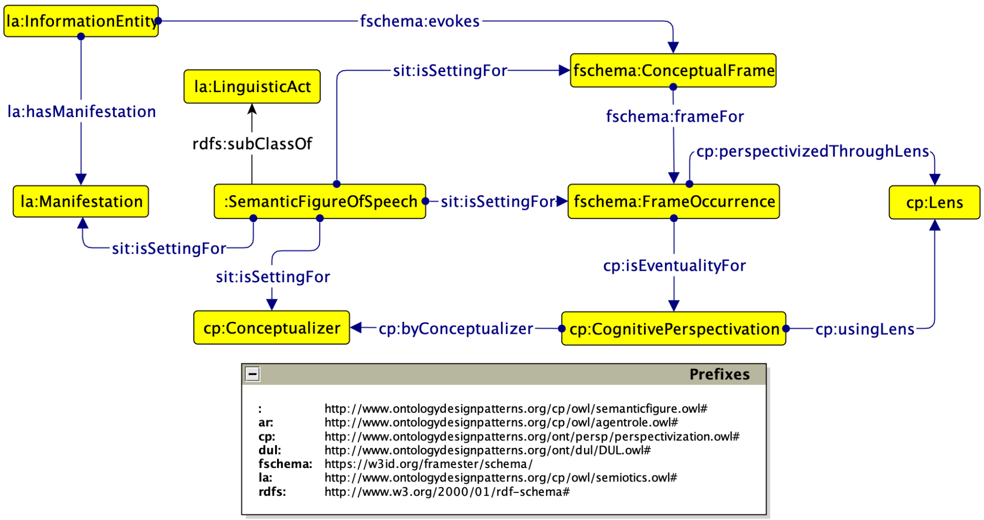

Figure 1: Pattern for Tropes.

6. Preliminary results

As this work is in its early stages, preliminary results are starting to address Research Questions 1 and 2, which are part of the project lifecycle's design phase. The primary goal is to establish a scalable and iterative pipeline that takes into account: (i) the corpus used, (ii) algorithms for identifying tropes, (iii) existing ontologies, and (iv) approaches to KP extraction. After conducting a literature review on algorithms for identifying tropes, we categorized them into different groups (as described in Section 2) and began testing them on a sample corpus derived from the PLOS One corpus[3]. It contains over 200,000 XML-formatted articles from 2006 to 2023, mostly in the biomedical field. To explicitly represent the relationship between rhetorical figures, argumentative structure of scholarly texts, and cognitive affordances, we reused the Semiotics ODP and Cognitive Perspectivation (CP) ODP [33, 12] to model information objects and their meaning. A first KP[4] has been drafted in Fig. 1 by using the Graffoo[5] notation. In this KP, a semantic figure of speech detected in an information entity is related to a frame, whose occurrence triggers a CP pattern by a conceptualizer, and uses a lens to produce a perspective.

7. Discussion and future work

This work aims at contributing to the field of SciSci by identifying patterns in the use of tropes in scientific texts and creating a KG of rhetoric in scientific texts. State-of-the-art algorithms were surveyed, and a sample dataset derived from the PLOS One corpus was built, using a pipeline that allows qualitative analysis of the results. The Semiotics ODP and the Cognitive Perspectivation ODP were extended to extract more domain-specific KPs. One of the challenges faced is the lack of heterogeneity in communication products across different disciplines, and the analysis is currently limited to English texts. To overcome these challenges, a wider dataset

[3] https://journals.plos.org/plosone/browse/text_mining.
[4] Available at https://raw.githubusercontent.com/dersuchendee/tist/main/knowledge-patterns/tropes-pattern.owl.
[5] https://essepuntato.it/graffoo/.

and continuous collaboration with scholars from different fields and experts in state-of-the-art algorithms will be considered.

Acknowledgments

This work was supported by the PhD scholarship "Discovery, Formalisation and Re-use of Knowledge Patterns and Graphs for the Science of Science", funded by the Italian National Research Council, Institute for Cognitive Sciences and Technologies (ISTC-CNR) through the WHOW project (EU CEF programme - grant agreement no. INEA/CEF/ICT/A2019/2063229). The author is grateful to her supervisors Prof. Aldo Gangemi and Dr. Andrea Giovanni Nuzzolese for their helpful suggestions and comments.

References

[1] G. Lakoff, M. Johnson, Metaphors We Live By, University of Chicago Press, 1981.
[2] C. Taylor, B. M. Dewsbury, On the problem and promise of metaphor use in science and science communication, Journal of Microbiology & Biology Education 19 (2018).
[3] T. S. Kuhn, Metaphor in science, 2 ed., Cambridge University Press, 1993, p. 533–542.
[4] S. Fortunato, et al., Science of science, Science 359 (2018) eaao0185.
[5] A. G. Nuzzolese, A. Gangemi, V. Presutti, P. Ciancarini, Encyclopedic knowledge patterns from wikipedia links, in: Proc. of ISWC 2011, Springer Berlin Heidelberg, Berlin, Heidelberg, 2011, pp. 520–536.
[6] A. G. Nuzzolese, V. Presutti, A. Gangemi, S. Peroni, P. Ciancarini, Aemoo: Linked data exploration based on knowledge patterns, Semantic Web 8 (2017) 87–112.
[7] C. J. Fillmore, Frame semantics, in: Cognitive Linguistics: Basic Readings, De Gruyter Mouton, Berlin, New York, 2006, pp. 373–400.
[8] M. Minsky, A framework for representing knowledge, in: Frame Conceptions and Text Understanding, De Gruyter, Berlin, Boston, 1979, pp. 1–25.
[9] C. Burgers, Conceptualizing change in communication through metaphor, Journal of Communication 66 (2016) 250–265.
[10] E. A. Szymanski, Remaking yeast: Metaphors as scientific tools in saccharomyces cerevisiae 2.0, BioSocieties 14 (2019) 416–437.
[11] R. Alkhammash, Bibliometric, network, and thematic mapping analyses of metaphor and discourse in covid-19 publications from 2020 to 2022, Frontiers in Psychology 13 (2023).
[12] A. Gangemi, P. Valentina, Formal Representation and Extraction of Perspectives, Studies in Natural Language Processing, Cambridge University Press, 2022, p. 208–228.
[13] D. Zhang, et al., Through the eyes of a poet: Classical poetry recommendation with visual input on social media, in: Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM '19, Association for Computing Machinery, New York, NY, USA, 2020, p. 333–340.
[14] S. Rai, S. Chakraverty, A survey on computational metaphor processing, ACM Computing Surveys (CSUR) 53 (2020) 1–37.
[15] A. Buccheri, et al., Semantic analysis and frequency effects of conceptual metaphors of emotions in latin. from a corpus-based approach to a dictionary of latin metaphors, Journal of Latin Linguistics 20 (2021) 163–189.
[16] H. Gong, et al., IlliniMet: Illinois system for metaphor detection with contextual and linguistic information, in: Proc. of the 2nd FigLang, ACL, Online, 2020, pp. 146–153.

[17] V. Dankers, et al., Modelling the interplay of metaphor and emotion through multitask learning, in: Proc. of the 2019 EMNLP and the 9th IJCNLP, ACL, Hong Kong, China, 2019, pp. 2218–2229.
[18] C. Peng, D. T. Vu, J. J. Jung, Knowledge graph-based metaphor representation for literature understanding, Digital Scholarship in the Humanities 36 (2021) 698–711.
[19] V. Prabhakaran, M. Rei, E. Shutova, How metaphors impact political discourse: A large-scale topic-agnostic study using neural metaphor detection, Proc. of the International AAAI Conference on Web and Social Media 15 (2021) 503–512.
[20] H. Li, et al., Target word masking for location metonymy resolution, in: Proc. of the 28th COLING, International Committee on Computational Linguistics, Barcelona, Spain (Online), 2020, pp. 3696–3707.
[21] K. A. Mathews, M. Strube, Impact of target word and context on end-to-end metonymy detection, 2021.
[22] G. Steen, et al., A method for linguistic metaphor identification. From MIP to MIPVU., number 14 in Converging Evidence in Language and Communication Research, John Benjamins, 2010.
[23] E. Dodge, J. Hong, E. Stickles, MetaNet: Deep semantic automatic metaphor analysis, in: Proc. of the 3rd Workshop on Metaphor in NLP, Association for Computational Linguistics, Denver, Colorado, 2015, pp. 40–49.
[24] C. F. Baker, C. J. Fillmore, J. B. Lowe, The Berkeley FrameNet project, in: Proc. of the 36th ACL and 17th COLING, volume 1, Association for Computational Linguistics, Montreal, Quebec, Canada, 1998, pp. 86–90.
[25] A. Gangemi, M. Alam, V. Presutti, Amnestic forgery: an ontology of conceptual metaphors, CoRR (2018).
[26] S. Peroni, D. Shotton, The spar ontologies, in: Proc. of the 17th ISWC 2018, volume 2, Springer-Verlag, Berlin, Heidelberg, 2018, p. 119–136.
[27] V. Pertsas, P. Constantopoulos, Scholarly ontology: Modelling scholarly practices, Int. J. Digit. Libr. 18 (2017) 173—-190.
[28] R. Kühn, J. Mitrović, M. Granitzer, Grhoot: Ontology of rhetorical figures in german, in: Proc. of the 13th LREC, European Language Resources Association, Marseille, France, 2022.
[29] A. G. Nuzzolese, A. L. Gentile, V. Presutti, A. Gangemi, Conference linked data: the scholarlydata project, in: Proc. of the 15th ISWC 2016, volume 2, Springer, 2016, pp. 150–158.
[30] T. Khishigsuren, et al., Metonymy as a universal cognitive phenomenon: Evidence from multilingual lexicons, 2022.
[31] M. Choi, et al., MelBERT: Metaphor detection via contextualized late interaction using metaphorical identification theories, in: Proc. of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2021, pp. 1763–1773.
[32] E. Liu, et al., Testing the ability of language models to interpret figurative language, 2022.
[33] P. Hitzler, et al., Ontology Engineering with Ontology Design Patterns: Foundations and Applications, Studies on the Semantic Web, IOS Press, Incorporated, 2016.
[34] E. Blomqvist, V. Presutti, E. Daga, A. Gangemi, Experimenting with extreme design, in: Knowledge Engineering and Management by the Masses, Springer Berlin Heidelberg, Berlin, Heidelberg, 2010, pp. 120–134.
[35] G. Tamasauskaite, P. Groth, Defining a knowledge graph development process through a systematic review, ACM Trans. Softw. Eng. Methodol. 32 (2023).
[36] A. Gangemi, et al., Modelling ontology evaluation and validation, in: Proc. of ISWC 2006, Springer Berlin Heidelberg, Berlin, Heidelberg, 2006, pp. 140–154.
[37] E. Kuric, et al., Knowledge graph exploration: A usability evaluation of query builders for laypeople, in: M. Acosta, et al. (Eds.), Semantic Systems. The Power of AI and Knowledge Graphs, Springer International Publishing, Cham, 2019, pp. 326–342.