Developing Literature Annotation Guidelines for Representing Normal Physiology in Biolink-Compatible Knowledge Graphs

Madeline Bittner, Willie Rogers, Dina Demner-Fushman,
Richard H. Scheuermann, Matthew Diller
Division of Intramural Research
National Library of Medicine, Bethesda, MD
firstname.lastname@nih.gov

Abstract

Much of our knowledge about anatomy and physiology is found in text format in research papers and medical textbooks. For an information system to have access to this knowledge, extracting and translating it into a computable format that can be stored in an ontology or knowledge graph is advantageous. Unfortunately, existing text mining corpora, which are needed to train and evaluate data mining models, are old and consist almost entirely of research papers, which rarely contain complete information needed to capture complex normal physiological processes and, subsequently, understand the pathophysiology of a disease. As a first step to filling in this gap, we have developed a guide for annotating medical textbooks for physiological events and entities involved in these events. In addition to providing our guidelines and describing the guideline development process, we analyze the coverage of normal physiology in existing ontologies.

1 Introduction

Anatomy and physiology are central components of biological systems, with knowledge of the relationship between the structures and functions of organisms and their parts driving most of what we know about health and disease. Most knowledge about healthy anatomy and physiological functions exists as text in medical textbooks and research papers, which is not easily accessible in a structured form for querying, analysis, and visualization. This lack of structured, computable knowledge representation has hindered the translation of large-scale multi-omics datasets into new knowledge about disease mechanisms and potential treatments. In addition, many ontologies and knowledge graphs (KGs) contain large gaps in coverage as they tend to lag behind novel research findings in terms of being up-to-date, and rely largely on manual curation (Li et al., 2023). To scale up curation efforts, ontology and KG developers will need to incorporate semi-automated tools and workflows for extracting knowledge about anatomical entities and physiological events from biomedical text.

Here we describe an annotation guide that can be used as a reference for annotating anatomical entities and physiological events in medical textbooks. Included in these guidelines are the scope of the annotation, a description of each entity and event of interest, and the reasons for why certain annotation decisions were made. To test the feasibility of translating textual physiology information into a structured format, we assess the coverage of extracted entities and events in existing biomedical ontologies.

2 Background

Currently, several tools exist to create structured representations of knowledge in ontologies and KGs that could be applied to anatomy and physiology, including Biolink (Unni et al., 2022), which aims to provide an upper-level model for interoperable KGs. However, there is an absence of up-to-date text mining corpora to support the extraction of anatomy and physiology information from text. The most recent annotated corpora—the Anatomical Entity Mention (AnEM) corpus (Ohta et al., 2012), the extended Anatomical Entity Mention (AnatEM) corpus (Pyysalo and Ananiadou, 2013), and the Multi-Level Event Extraction (MLEE) corpus (Pyysalo et al., 2012)—that broadly focus on mentions of anatomical entities and physiological events in the context of both healthy and diseased organisms, largely consist of annotated abstracts and full-text research publications. Although these corpora are annotated with ontology term identifiers, the length of time since these annotations were performed would warrant reannotation to ensure that active and the most precise terms are used. Additionally, since the annotated documents in these corpora consist mostly of research papers,

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 718–728

their structure and content will likely differ from the semantic and syntactic structures found in a medical textbook. Beyond the limitations of existing corpora, existing guidelines used for biomedical information extraction, including i2b2 (Uzuner et al., 2011) for clinical de-identification, GENIA (Kim et al., 2003) for biochemical entities in molecular biology literature, and CINEX (Reichenpfader et al., 2025) for high-level metadata extraction from clinical trial text, were not designed to capture normal physiology entities, as they are either too metadata-focused or operate at a biochemical level that does not translate to physiology annotation. To our knowledge, no annotation guidelines currently exist that are specific to normal human physiology, representing a gap that this work addresses.

3 Methods

In this exploration of approaches to annotation of physiological processes and events at several anatomy levels, we focus on annotating events that involve glucagon-like peptide 1 (GLP-1), which plays an essential role in processing nutrients after food ingestion. The effects of GLP-1 are also mimicked by agonistic drugs, such as exenatide (Byetta), liraglutide (Victoza), dulaglutide (Trulicity) and semaglutide (Ozempic/Wegovy), used for blood glucose control and weight management.

3.1 Annotation Schema Development

The annotation guide was developed using passages from the *Guyton and Hall Textbook of Medical Physiology* (14th Edition) (Hall et al., 2021), well-established reference in biomedical education and clinical medicine. This textbook was selected because it contains medically complex language representative of normal physiology text, is well-structured, and reflects the type of text that would be encountered when annotating other normal physiology domains beyond the GLP-1 use case. Passages were drawn specifically from sections covering GLP-1 normal physiology, with 17 paragraphs selected for annotation.

Three annotators with expertise in biomedical informatics annotated each text independently. Annotations were conducted iteratively, with label definitions and disambiguation rules continuously refined to improve clarity and ensure generalizability beyond GLP-1 physiology to other normal human physiology domains. Comprehensive coverage of relevant biological concepts while remaining broad enough to generalize to other normal human physiology processes. Text was considered out of scope if it describes disease pathways, pertains to processes at a purely biochemical level, or does not relate to normal adult human physiology. Entity, event, and attribute labels and their definitions are provided in appendix Table A1, and were informed by prior work on manual annotation of anatomical and physiological features in medical text (Ohta et al., 2012; Pyysalo and Ananiadou, 2013; Pyysalo et al., 2012). Counts for the number of occurrences of each entity, event, and attribute labels across the 63 annotated sentences from the 17 passages are provided Table 1. Target-Arg relations are used to formally link the events to their associated entities or other events, where the direction of the relation is defined as the source term to the target term. To reduce cross-sentence complexity, each sentence was annotated as independently as possible, with Target-Arg relations restricted within sentence boundaries when possible; the coreference attribute was used to capture topics that carries across sentences. Relationship serves as the catch-all event label, applied to any word or phrase that describes how entities are connected, interact, or influence one another. Three annotators independently annotated the corpus using BRAT (Stenetorp et al., 2012), as shown in Figure 1, with disagreements resolved through discussion and a predefined list of common multi-word expressions used to reduce complexity and increase annotation efficiency (appendix Table A2).

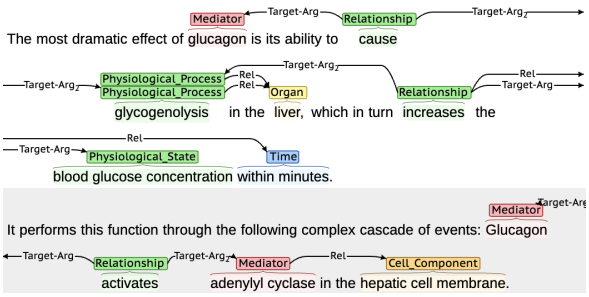

Figure 1: Example BRAT annotation.

3.2 Named Entity Recognition

To evaluate the feasibility of automated named entity recognition (NER) and to inform ontology selection for future work, we assessed the coverage of our annotation guide across a broad range of biomedical ontologies using the EMBL-EBI Ontology Lookup Service (OLS) (Côté et al., 2006) and

Category	Label	Count (N=420)
Entity	Mediator	33 (7.8)
	Measurement	24 (5.7)
	Chemical	22 (5.2)
	Organ	15 (3.5)
	Substrate	12 (2.86)
	Time	11 (2.6)
	Tissue	9 (2.1)
	Cell	7 (1.6)
	Organ System	2 (0.4)
	Cell Component	1 (0.2)
Event	Relationship	74 (17.6)
	Physiological State	50 (11.9)
	Physiological Process	47 (11.1)
Attribute	Rel	98 (23.3)
	Coreference	8 (1.9)
	is a	7 (1.6)

Table 1: Total counts (%) for each entity, event, and attribute label with percentages rounded to one decimal place.

the Unified Medical Language System (UMLS) (Bodenreider, 2004). A distinct list of Entities (N=29), Physiological Processes (N=23), Physiological States (N=13), and Relationships (N=55) was derived from our manual annotations. Multiword expressions and fragments were excluded prior to ontology lookup, as their structure is not directly represented in standard biomedical ontologies, though their constituent terms are covered individually. The terms were queried via the EMBL-EBI API (European Bioinformatics Institute (EMBL-EBI), 2026) and matched against available ontologies using exact string lookup. Coverage was calculated as the proportion of distinct manually annotated terms present in each ontology. The Entity labels *Time* and *Measurement* were excluded from this analysis, as these labels do not correspond to concepts represented within standard biomedical ontologies.

3.3 Relation Extraction

To evaluate the feasibility of automated relation extraction, we assessed whether Stanza (Qi et al., 2020) dependency and constituency parsers can structurally encode relationship types defined in our annotation guide. Both parsers were run on the 17 text paragraphs, with multi-word expressions adjoined with underscores to prevent term splitting. Manual relation annotations were mapped to parser outputs according to each parser type's structural properties: dependency parser spans were reduced to syntactic head words; while constituency parser, spans were aligned to leaf node sequences with Part-of-speech (POS) tags normalized to phrase-level equivalents. Each relation was parsed independently at the sentence level to avoid full-document parse context effects. Across the 162 annotated relation pairs, 29 distinct dependency patterns and 6 distinct constituency patterns were identified as the reference set. Coverage was assessed at two levels. Corpus-wide coverage determined whether each distinct relation pattern appeared anywhere in the full parser output across all 17 paragraphs, assessed for the dependency parser via direct UPOS $\rightarrow$ deprel $\rightarrow$ UPOS arc matching, and for the constituency parser via phrase label pair matching *(e.g., VP Target-Arg NP)* within the same sentence tree. Individual relation coverage determined whether each manually annotated pattern was present in its source sentence. For the dependency parser, using direct arc lookup between entity head words in both directions; for constituency parser, using the lowest common ancestor method to identify the smallest constituent dominating both entity head words and comparing its label to the annotated phrase type. Of the 162 annotated relation pairs, 2 were removed from the dependency pattern set and 18 were removed from the constituency pattern set prior to individual relation coverage, as these instances could not be reliably aligned to their corresponding sentence-level parse structures. Additionally, hierarchical agglomerative clustering (HAC) was performed to visually model and compare the relation pattern coverage of the dependency and constituency parsers, methods can be seen in appendix Figures A1 and A2.

4 Results

4.1 Inter Annotator Agreement

Mean pairwise F1 scores were computed to measure inter-annotator agreement, as shown in Table 2. Agreement for Entities, Relationships, which includes the labels Relationship, Physiological State, and Physiological Process, was measured at the character span level. For annotations with character span agreement, label agreement was then assessed. Agreement for Target-Arg relations was evaluated using the character spans for both the source and target terms, with label agreement then assessed for pairs that reached character span agreement. For Target-Arg label agreement, both the source and target labels were required to match. A token alignment leniency of two tokens was applied across all agreement levels. Attribute labels were excluded from inter-annotator agreement cal-

Section	Span	Label
Entities	0.626 (N=612)	0.814 (N=280)
Relationships	0.621 (N=809)	0.868 (N=365)
Target-Arg	0.354 (N=1202)	0.731 (N=260)

Table 2: Mean pairwise F1 agreement scores for span-level and label-level agreement across three annotators. N indicates matched spans/relations for label agreement.

culations, as their complexity and dependency on broader context make standard pairwise span-level evaluation unreliable. While span-level agreement is comparatively lower, this is expected given the nature of span boundary identification. Notably, label agreement among spans that all annotators agreed on is high, suggesting that annotators can reliably assign the correct label when a span is identified.

4.2 Ontology Coverage

Table A3 summarizes the coverage of each ontology from the OLS and UMLS. Among the OLS ontologies, OMIT ontology achieved the highest coverage for Entities (0.52), while NCIT achieved the highest coverage for Physiological Process, Physiological State, and Relationship (0.17, 0.38, and 0.24 respectively). Notably, PRIDE and INO shared the highest coverage of Physiological Process terms, and PRIDE also shared the highest coverage for Physiological State terms. Across all OLS ontologies, Physiological Process terms had the lowest overall coverage (0.17). In the UMLS, Relationship terms had the lowest coverage (0.11), while Entities had the highest (0.55). Overall, the UMLS achieved the highest coverage across all ontologies at 0.31, with NCIT second at 0.29.

4.3 Relation Coverage

Corpus-wide coverage was computed as the proportion of distinct manual annotation patterns found within the full parser output. The dependency parser achieved a corpus-wide coverage of 0.48, recovering 14 of 29 distinct manual relations, with a precision of 0.10, indicating that the 14 matched patterns account for only a small fraction of the 135 distinct patterns produced by the parser across the full corpus. The constituency parser achieved perfect corpus-wide coverage of 1.00, recovering all 6 distinct manual relations, with a precision of 0.19, meaning the 6 matched patterns represented a similarly small proportion of the 31 distinct patterns generated. At the individual relation level, the dependency parser covered 80 of the 160 eligible manual relation instances (0.50), and the constituency parser covered 89 of the 144 eligible manual relation instances (0.62). Results for the HAC dendrograms can be found in appendix Figures A1 and A2.

5 Discussion

Through this work, physiology annotation guidelines were developed to support corpus development for automatic named entity and relation extraction from medical text using NLP methods, towards the goal of knowledge graph triplet generation. The evaluation results reveal both promise and limitations of using these guidelines as a foundation for that task. While the guidelines provide a structure for both entity and relation annotation that can be generalized to cover a broad range of normal physiology, the feasibility of automated triplet extraction depends on the performance of both NER and relation extraction, each of which presents with different challenges.

In terms of NER, the ontology coverage evaluation demonstrated that no single ontology is sufficient to capture the full range of terms present across all categories. This is consistent with prior work showing that ontologies often provide only partial representations of a domain (Abad-Navarro et al., 2025). The UMLS exhibited the best overall coverage, however relationship terms showed lower coverage, likely reflecting that UMLS is primarily organized around biomedical entities rather than relational predicates. Additionally, these ontologies are generally not designed to capture higher-level, abstract representations, but instead focus on domain specific-concepts. As a result, although our terms are similar to those found in existing ontologies, they are often not represented because they are broader in scope and intended to capture relationships across a wide range of physiological contexts. For example, in the UMLS, the generalized relationship term *Prevents* is not explicitly represented as a standalone concept; however, more specific phrases such as *Chair prevents rising*, *Prevents risky behaviors*, and *Prevents skin barrier interruption* are included. These findings support prior work demonstrating that UMLS does not fully cap-

ture all concepts and relationships within a given domain (Yu et al., 1999). Therefore, our future work will explore combining multiple ontologies, as integrating resources has been shown to support more complete domain representation (Osman et al., 2021), alongside mapping strategies to improve entity coverage across all categories.

For relation extraction, the dependency and constituency parsers show different results in relation coverage. The dependency parser's slightly higher individual manual relation coverage (0.50) relative to its corpus-wide distinct relation coverage (0.48) suggests it performs better on more frequently occurring relation types found in our corpus. The constituency parser, despite achieving perfect corpus-wide distinct relation coverage, showed a notable drop at the individual relation level (0.61), driven largely by a high false positive rate. Nevertheless, the constituency parser shows greater promise as a basis for automated relation extraction, and prior work has demonstrated that constituency-based approaches can be effective when combined with rule-based filtering to reduce false positives (Jiang and Diesner, 2019). In future work, we aim to expand the text corpus to additional open access medical textbooks and explore how post-processing of constituency parser outputs can be used to extract relevant relations.

6 Conclusion

We present an annotation guide for anatomical entities and physiological events in biomedical textbooks, designed to support the automated construction of knowledge graph triples from free text. To access the feasibility of translating free medical text into a structured format, we annotated a selected text corpus and evaluated ontology coverage of our annotations across OLS ontologies and the UMLS. We evaluated the coverage of relation extraction by assessing the coverage of annotated relation pairs within parser outputs. Our results demonstrate that while existing ontologies and parsers provide partial coverage, significant gaps remain. These findings motivate future work in multi-ontology integration and post-processing strategies for parser outputs. This work represents a step toward making rich knowledge present in medical textbooks accessible in a computable format that can support development of clinical knowledge discovery platforms.

7 Limitations

This study is a pilot study utilizing a smaller sample size with text specific to our use case of GLP-1 physiology. This use case was selected because it was believed that a schema developed from it could be generalized and applied to other physiological functions, though this remains to be validated.

8 Ethics

This work includes limited use of content from *Guyton and Hall Textbook of Medical Physiology* (14th Edition) (Hall et al., 2021) for purposes of non-commercial research and computational analysis. The authors have conducted a good-faith assessment and believe this use is consistent with fair use principles (Association of College and Research Libraries, 2025).

Acknowledgments

This research was supported by the Intramural Research Program of the National Institutes of Health (NIH) and utilized the computational resources of the NIH HPC Biowulf cluster (http://hpc.nih.gov). The contributions of the NIH authors are considered Works of the United States Government. The findings and conclusions presented in this paper are those of the authors and do not necessarily reflect the views of the NIH or the U.S. Department of Health and Human Services.

References

Francisco Abad-Navarro, Catalina Martínez-Costa, and Jesualdo Tomás Fernández-Breis. 2025. Ontology coverage analysis through language models. *Engineering Applications of Artificial Intelligence*, 162:112671.

Association of College and Research Libraries. 2025. Fair use. https://acrl.libguides.com/scholcomm/toolkit/fairuse. Accessed: April 17, 2026.

Daniel C Berrios, Sara C Jorgensen, Rachel K Shoop, Carlos De Los Santos, Danielle K Lopez, Alan E Wood, Gregory D Eley, Kathryn C Richards, Olga V Kieschnick, Sigrid S Reinsch, and 1 others. 2024. Increasing data discovery and re-use: The space life sciences ontology. In *Annual Meeting of the American Society for Gravitational and Space Research (ASGSR)*.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic Acids Research*, 32(suppl_1):D267–D270.

Richard G. Côté, Philip Jones, Rolf Apweiler, and Henning Hermjakob. 2006. The ontology lookup service, a lightweight cross-platform tool for controlled vocabulary queries. *BMC Bioinformatics*, 7:97.

Kirill Degtyarenko, Paula de Matos, Marcus Ennis, Janna Hastings, Martin Zbinden, Alan McNaught, Rafael Alcántara, Michael Darsow, Mickaël Guedj, and Michael Ashburner. 2007. Chebi: a database and ontology for chemical entities of biological interest. *Nucleic Acids Research*, 36(suppl_1):D344–D350.

European Bioinformatics Institute (EMBL-EBI). 2026. Ontology Lookup Service (OLS) API. Accessed: 2026-04-07.

John E. Hall, Michael E. Hall, and Arthur C. Guyton. 2021. *Guyton and Hall Textbook of Medical Physiology*, 14 edition. Elsevier.

Jingshan Huang, Christopher Townsend, Dejing Dou, Haishan Liu, and Ming Tan. 2011. Omit: a domain-specific knowledge base for microrna target prediction. *Pharmaceutical research*, 28(12):3101–3104.

Ming Jiang and Jana Diesner. 2019. A constituency parsing tree based method for relation extraction from abstracts of scholarly publications. In *Proceedings of the Thirteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-13)*, pages 186–191, Hong Kong. Association for Computational Linguistics.

J.-D. Kim, T. Ohta, Y. Tateisi, and J. Tsujii. 2003. Genia corpus—a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(suppl_1):i180–i182.

Harry Li, Gabriel Appleby, Camelia Daniela Brumar, Remco Chang, and Ashley Suh. 2023. Knowledge graphs in practice: characterizing their users, challenges, and visualization opportunities. *IEEE Transactions on Visualization and Computer Graphics*, 30(1):584–594.

James Malone, Ele Holloway, Tomasz Adamusiak, Misha Kapushesky, Jie Zheng, Nikolay Kolesnikov, Anna Zhukova, Alvis Brazma, and Helen Parkinson. 2010. Modeling sample variables with an experimental factor ontology. *Bioinformatics*, 26(8):1112–1118.

Todd Millecam, Austin J Jarrett, Naomi Young, Dana E Vanderwall, and Dennis Della Corte. 2021. Coming of age of allotrope: proceedings from the fall 2020 allotrope connect. *Drug Discovery Today*, 26(8):1922–1928.

Tomoko Ohta, Sampo Pyysalo, Jun'ichi Tsujii, and Sophia Ananiadou. 2012. Open-domain anatomical entity mention detection. In *Proceedings of the Workshop on Detecting Structure in Scholarly Discourse*, pages 27–36, Jeju Island, Korea. Association for Computational Linguistics.

Ibrahim Osman, Stefano F. Pileggi, Sadok Ben Yahia, and Gayo Diallo. 2021. An alignment-based implementation of a holistic ontology integration method. *MethodsX*, 8:101460.

Arzucan Özgür, Junguk Hur, and Yongqun He. 2016. The interaction network ontology-supported modeling and mining of complex interactions represented with multiple keywords in biomedical literature. *BioData mining*, 9(1):41.

Yasset Perez-Riverol, Jingwen Bai, Chakradhar Bandla, David García-Seisdedos, Suresh Hewapathirana, Selvakumar Kamatchinathan, Deepti J Kundu, Ananth Prakash, Anika Frericks-Zipper, Martin Eisenacher, and 1 others. 2022. The pride database resources in 2022: a hub for mass spectrometry-based proteomics evidences. *Nucleic acids research*, 50(D1):D543–D552.

Sampo Pyysalo and Sophia Ananiadou. 2013. Anatomical entity mention recognition at literature scale. *Bioinformatics*, 30(6):868–875.

Sampo Pyysalo, Tomoko Ohta, Makoto Miwa, Han-Cheol Cho, Jun'ichi Tsujii, and Sophia Ananiadou. 2012. Event extraction across multiple levels of biological organization. *Bioinformatics*, 28(18):i575–i581.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Daniel Reichenpfader, Henning Müller, and Kerstin Denecke. 2025. Improving the reporting quality of studies on information extraction from clinical texts: Protocol for the development of a consensus-based reporting guideline. *JMIR Research Protocols*, 14:e76776.

Winifred Sewell. 1964. Medical subject headings in medlars. *Bulletin of the Medical Library Association*, 52(1):164.

Nicholas Sioutos, Sherri de Coronado, Margaret W Haber, Frank W Hartel, Wen-Ling Shaiu, and Lawrence W Wright. 2007. Nci thesaurus: a semantic model integrating cancer-related clinical and molecular information. *Journal of biomedical informatics*, 40(1):30–43.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. brat: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107, Avignon, France. Association for Computational Linguistics.

Deepak R Unni, Sierra AT Moxon, Michael Bada, Matthew Brush, Richard Bruskiewich, J Harry Caufield, Paul A Clemons, Vlado Dancik, Michel Dumontier, Karamarie Fecho, and 1 others. 2022. Biolink model: A universal schema for knowledge graphs in clinical, biomedical, and translational science. *Clinical and translational science*, 15(8):1848–1855.

U.S. National Library of Medicine. 2026. SNOMED CT International Edition, April 2026 Release. Release date: 2026-04-01; Accessed: 2026-04-07.

Özlem Uzuner, Yuan Luo, and Peter Szolovits. 2011. Evaluating the state-of-the-art in automatic de-identification. *Journal of the American Medical Informatics Association*, 18(5):552–563.

Danielle Welter, Simon Jupp, and David Osumi-Sutherland. 2018. Human cell atlas ontology. In *ICBO*.

Hong Yu, Carol Friedman, Andrey Rzhetsky, and Pauline Kra. 1999. Representing genomic knowledge in the umls semantic network. In *Proceedings of the AMIA Symposium*, pages 181–185. American Medical Informatics Association.

A Appendix

Category	Label	Definition	Examples
Entity	Cell	A material entity that is the basic structural unit of tissue.	Liver cells, alpha cells
	Cell Component	A subcellular structure or molecular complex that is associated with a cell and contributes to its function.	Nucleus, mitochondria
	Chemical	A molecule, substance, or compound involved in a physiological function where its specific role is unclear, unknown, or not relevant.	Blood glucose, glycogen in the liver
	Measurement	A data item that quantifies something.	140 mg, fivefold, one-third
	Mediator	A chemical that actively influences, regulates, or triggers a physiological reaction or process.	Glucagon
	Organ	A distinct anatomical structure made up of multiple tissue types that performs one or more biological functions.	Liver, heart
	Organ System	A group of distinct anatomical structures that work together to perform one or more biological functions.	Respiratory system
	Substrate	A chemical that is acted upon, consumed, or transformed in a biological process or reaction.	Blood glucose
	Time	A temporal descriptor for how long or what pattern a physiological process occurs.	Few minutes, within minutes, first hour
	Tissue	A group of similar cells organized together for a specific physiological function.	Blood, connective tissue
Event	Physiological Function	A biological activity or outcome carried out by cells, tissues, organs, etc. that contributes to normal body function.	Secretion, glucose uptake
	Physiological State	A condition, status, or internal environment of the body at a given time.	Blood glucose concentration, fasting, exhaustive exercise
	Relationship	A word or phrase that describes how two or more annotated entities are connected, interact, or influence each other within the physiological context.	Activates, decreases, rises, increases
Attribute	Coreference	When two entities refer to the same named entity.	Blood glucose (increase) *coreference* this (increase)
	is_a	One entity is a type or category of another.	GLP-1 *is_a* chemical
	Rel	When two entities are in relation to another.	Release *Rel* insulin

Table A1: Label definitions for the entities, events, and attributes included in the annotation schema. Contents within parenthesis are additional context not included in the named entity. Entities, Chemical, Substrate, and Mediator, were developed to distinguish molecules and compounds based on their functional role within the physiological context rather than chemical identity. Chemical serves as the catch-all label, substrate and mediator are more specific subcategories of Chemical. Three attribute labels, Coreference, is_a, and Rel, were developed to capture additional contextual information that cannot be fully expressed through entity or event labels alone.

Common Multi-Word Expressions
Blood glucose (concentration/level)
(Increase/Decrease) in blood glucose
(High/Low) concentrations
Effects of (Chemical/Mediator/Substrate)
(Chemical/Mediator/Substrate) secretion
(Chemical/Mediator/Substrate) function
(Chemical/Mediator/Substrate) concentration
Blood glucose buffer system
Enzyme system
Glucose (stimulus/concentration)
Feedback mechanism
Feedback control system
Fasting level / Normal value
(Modifier) level
After a meal

Table A2: Common multi-word expressions identified in the GLP-1 physiology corpus. These multi-word expressions represent self-contained triples and were consolidated for simplicity, with the understanding that they can be decomposed into their constituent terms during post-processing. In cases where the surface representation of a phrase was disjoint, annotators were instructed to combine the constituent terms into a single fragment phrase capturing the full expression.

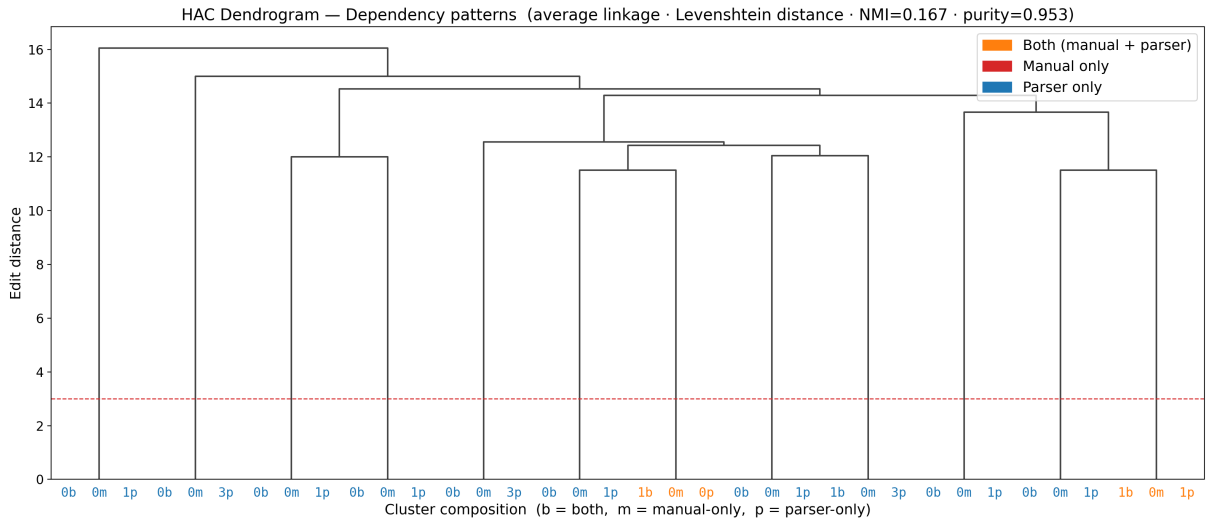

Figure A1: HAC dendrogram for dependency parser relation pattern clustering (average linkage, Levenshtein distance). The dashed red line indicates the cut height used to form final clusters; leaf labels show per cluster counts of patterns found in both sources (b), manual annotation only (m), and parser only (p). Clusters span edit distances up to 16, reflecting considerable structural diversity among dependency relation patterns.

Ontology	**Entities** (N = 29)	**Physiological Process** (N = 23)	**Physiological State** (N = 13)	**Relationship** (N = 55)
OLS ontologies				
OMIT	**0.52**	(0.13)	0.31	–
MESH	0.48	0.13	0.31	–
NCIT	0.38	**0.17**	**0.38**	**0.24**
CHEBI	0.24	–	–	–
HCAO	0.21	(0.04)	–	–
SNOMED	(0.17)	0.13	(0.31)	0.16
EFO	(0.14)	(0.09)	0.31	(0.02)
PRIDE	(0.07)	**0.17**	**0.38**	(0.05)
INO	–	**0.17**	–	(0.05)
BIOLINK	(0.03)	(0.09)	–	0.09
SLSO	–	–	–	0.07
AFO	–	(0.04)	(0.15)	0.07
Unmatched (OLS)	7	11	7	29
UMLS				
Coverage	0.55	0.35	0.23	0.11
Unmatched	13	15	10	49

Table A3: Ontology coverage of terminology categories (Entities, Physiological Process, Physiological State, and Relationship) queried against OLS and the UMLS. Coverage values reflect the proportion of distinct terms matched within each ontology. Coverage counts that are inside parenthesis are not part of the top five ontologies for that category, but still contain some coverage. (—) indicates no coverage was found. Unmatched counts reflect terms not found in any OLS ontology or the UMLS. OMIT=Ontology for miRNA Targets (Huang et al., 2011); MESH=Medical Subject Headings (Sewell, 1964); NCIT=National Cancer Institute Thesaurus (Sioutos et al., 2007); CHEBI=Chemical Entities of Biological Interest (Degtyarenko et al., 2007); HCAO=Human Cell Atlas Ontology (Welter et al., 2018); SNOMED=Systematized Nomenclature of Medicine—Clinical Terms (U.S. National Library of Medicine, 2026); EFO=Experimental Factor Ontology (Malone et al., 2010); PRIDE=Proteomics Identification Database Ontology (Perez-Riverol et al., 2022); INO= Interaction Network Ontology (Özgür et al., 2016); BIOLINK=Biolink Model (Unni et al., 2022); SLSO=Space Life Sciences Ontology (Berrios et al., 2024); AFO=Allotrope Foundation Ontology (Millecam et al., 2021); UMLS=Unified Medical Language System (Bodenreider, 2004)

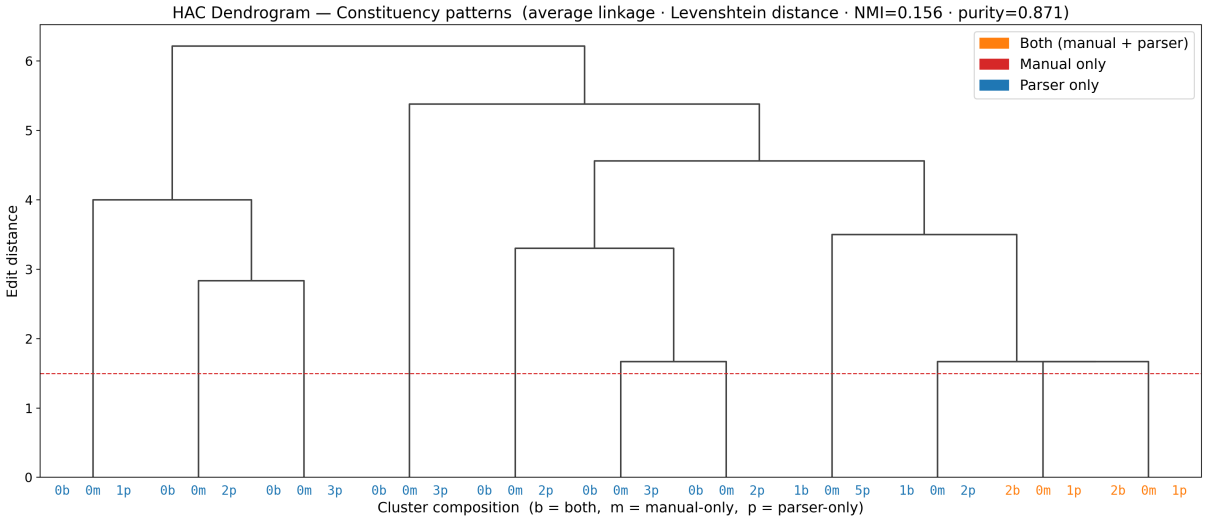

Figure A2: HAC dendrogram for constituency parser relation pattern clustering (average linkage, Levenshtein distance). he dashed red line indicates the cut height used to form final clusters; leaf labels show per cluster counts of patterns found in both sources (b), manual annotation only (m), and parser only (p). Cluster span edit distances up to 6, considerably lower than dependency relation patterns, likely reflecting the constrained phrase-label vocabulary from which the constituency relation patterns are constructed.