Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 5456–5473
Marseille, 11–16 May 2020
© European Language Resources Association (ELRA), licensed under CC-BY-NC

WorldTree V2: A Corpus of Science-Domain Structured Explanations and Inference Patterns supporting Multi-Hop Inference

Zhengnan Xie[‡*], Sebastian Thiem[†*], Jaycie Martin[‡], Elizabeth Wainwright[‡], Steven Marmorstein[§], Peter Jansen[†‡]
School of Information[†], Department of Linguistics[‡], Department of Computer Science[§]
University of Arizona, Tucson, Arizona, USA
pajansen@email.arizona.edu

Abstract

Explainable question answering for complex questions often requires combining large numbers of facts to answer a question while providing a human-readable explanation for the answer, a process known as multi-hop inference. Standardized science questions require combining an average of 6 facts, and as many as 16 facts, in order to answer and explain, but most existing datasets for multi-hop reasoning focus on combining only two facts, significantly limiting the ability of multi-hop inference algorithms to learn to generate large inferences. In this work we present the second iteration of the WorldTree project, a corpus of 5,114 standardized science exam questions paired with large detailed multi-fact explanations that combine core scientific knowledge and world knowledge. Each explanation is represented as a lexically-connected "explanation graph" that combines an average of 6 facts drawn from a semi-structured knowledge base of 9,216 facts across 66 tables. We use this explanation corpus to author a set of 344 high-level science domain inference patterns similar to semantic frames supporting multi-hop inference. Together, these resources provide training data and instrumentation for developing many-fact multi-hop inference models for question answering.

Keywords: question answering, multi-hop inference, explanations

1. Introduction

Explainable question answering is the task of providing both answers to natural language questions, as well as detailed human-readable explanations justifying why those answers are correct. Question answering is typically approached using either retrieval or inference methods, where retrieval methods search for a single contiguous passage of text from a corpus or single fact in a knowledge base that provides an answer to a question. For complex questions, a single passage often provides only part of the knowledge required to arrive at a correct answer, and an inference model must combine multiple facts from a corpus or knowledge base to infer the correct answer. In addition to producing an inference that correctly answers a question, a frequent design goal of *multi-hop inference* algorithms is to use the set of combined facts as a human-readable explanation for why the model's reasoning is correct.

Successfully building multi-hop inference algorithms poses a variety of challenges. Combining facts to perform inference is an inherently noisy process that often drifts off-context to unrelated facts, a phenomenon referred to as semantic drift (Fried et al., 2015). As a result, most multi-hop inference models are generally unable to demonstrate combining more than 2 or 3 facts to perform an inference (Khashabi et al., 2016; Jansen et al., 2017; Das et al., 2017), and even then only infrequently. This is a significant limitation, as even the reasoning required to answer elementary science exams averages combining 6 separate facts, and as many as 16 facts, when building explanations that include detailed supporting world knowledge (Jansen et al., 2016; Jansen et al., 2018), with an example 8-fact explanation shown in Figure 1. Training data for multi-hop inference is expensive and difficult to generate, resulting in the few datasets available generally focusing on combining only two units of knowledge (Yang et al., 2018; Khot et al., 2019). While these are valuable resources, there are indications that many-fact inference may be much more challenging than two-fact inference (Jansen, 2018), and require different mechanisms to solve.

In this work we present a large corpus of extremely detailed multi-fact explanations to serve both as training data for multi-hop inference, as well as an instrument to evaluate and expand the information aggregation capacity of multi-hop inference models. This work is a continuation of the WorldTree project (Jansen et al., 2018), an effort to manually generate large science-domain explanations and a supporting semi-structured knowledge base, expanded with approximately 1000 hours of additional annotation to be nearly three times the size of the original corpus. We also include a new effort to generate large multi-fact inference patterns similar to schema or semantic frames to support many-fact multi-hop inference.

The contributions of this work are:

1. We provide a corpus of approximately 5,100 detailed explanations for standardized elementary and middle-school science exams represented as explanation graphs, whose facts are drawn from a manually-authored semi-structured knowledge base containing 9,216 facts across 66 tables. This represents more than half of all publicly available standardized science exam questions in the United States, and is approximately three times larger than the original corpus.

2. We provide a tool for authoring large multi-fact inference patterns that represent generic constraint-based semantic frames in the science domain, and use this

[*] Authors contributed equally to this work.

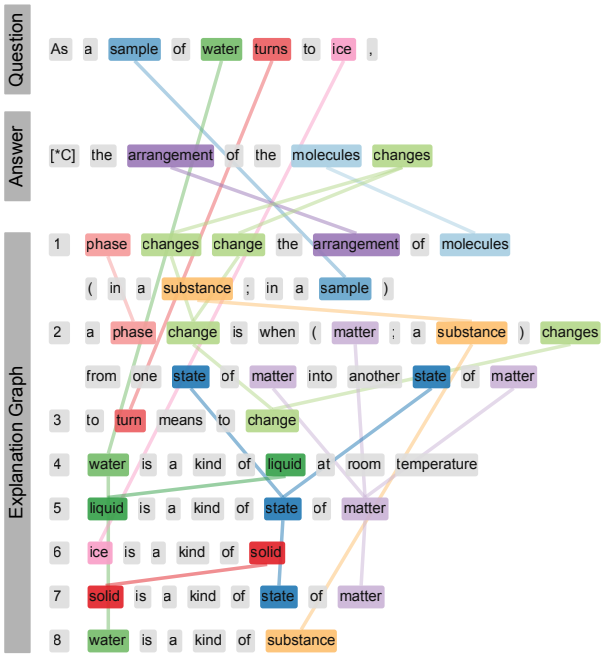

Figure 1: An example multiple choice science question (top), its correct answer (middle), and multi-fact explanation (bottom). The explanation forms an interconnected "explanation graph" with facts as nodes and edges based on shared words between facts and question or answer text.

tool to generate a corpus of 344 inference patterns containing a total of 3,650 facts drawn from the knowledge base.

In addition, we demonstrate fundamental properties of this explanation corpus as they relate to knowledge growth and reuse in Section 3.3., as well as provide baseline performance for using this corpus for performing the multi-hop explanation regeneration task (Jansen and Ustalov, 2019) in Section 3.4.

2. Related Work

A number of recent datasets have been made available to support multi-hop inference models. Because of the challenges, expense, and manual labour associated with generating these datasets, each includes either simplifications or non-ideal aspects, which may include: (a) Using artificial questions or tuple completion instead of natural language; (b) Having short reasoning chains; (c) Aggregating high-level structures, such as paragraphs, instead of sentences or axiomatic facts; (d) Annotating fact relevance (whether a fact is potentially relevant to answering a question) instead of explanation completeness (a complete set of facts that provide a detailed explanation); (e) Containing a small number of explanations, and/or a small supporting knowledge base or free text corpus. We examine the properties of 5 similar datasets below:

HotPotQA (Yang et al., 2018): A large dataset of 113k questions and supporting facts grounded in Wikipedia. *Benefits:* large size, open domain, supporting facts. *Drawbacks:* short aggregation (2 paragraphs, from which individual sentences are drawn), questions are generated to support facts.

Wikihop (Welbl et al., 2018): A large dataset of 51k tuple-completions grounded in Wikipedia. *Benefits:* large size, open domain. *Drawbacks:* Artificial simplified questions (represented as tuple completions), no multi-hop annotation to train or evaluate inference models, many questions likely do not require multi-hop inference to solve (Chen and Durrett, 2019).

ComplexWebQuestions (Talmor and Berant, 2018): A large dataset of 35k questions and associated SQL queries generated from an original set of 5k questions over Freebase. *Benefits:* Studies compositional questions that can be reduced to sets of simpler database queries. *Drawbacks:* Artificial questions generated from templates, limited semantics well-suited to Freebase.

OpenBookQA (Mihaylov et al., 2018): A dataset of 6k multiple choice elementary science-domain questions and a supporting knowledge base of 1,326 facts partially derived from Worldtree. *Benefits:* Tests world knowledge, includes supporting knowledge base. *Drawbacks:* Artificial questions generated to fit facts, only partial annotation for supporting facts (1 fact per question), unknown amount of aggregation required per question.

QASC (Khot et al., 2019): A dataset of 10k multiple choice science-domain questions and a supporting knowledge base. Questions generated by composing WorldTree and CK-12 corpus facts. *Benefits:* Tests world knowledge, includes a large supporting knowledge base (17M facts), crowdworkers generating questions had advanced degree qualifications, extended adversarial choices for language models. *Drawbacks:* Short aggregation (2 facts).

WorldTree V2 (This work): Containing 5k science exam questions, WorldTree V2 is the only multi-hop dataset using natural complex questions that are a benchmark of human reasoning rather than artificially generated questions. While most multi-hop datasets require aggregating 2 facts at most, Worldtree V2 explanations contain an average of 6 facts (and as many as 16 facts), making it the only dataset for investigating many-fact multi-hop inference. It is the only dataset constructed manually by knowledge engineers rather than crowdsourced, which also results in it having the fewest questions. WorldTree V2 explanations contain both scientific and world knowledge, which are both also tested by OpenBookQA and QASC. At 9k facts, WorldTree V2 contains the largest manually authored structured knowledge base of *n-ary* relations, but also the smallest overall knowledge base among comparable datasets. The WorldTree V2 knowledge base can be used both as structured text and as free text, allowing inference models requiring either type of knowledge to be directly compared. WorldTree V2 is the only corpus to contain a supplementary set of high-level many-fact inference patterns, though this is conceptually similar to the SQL query patterns of ComplexWebQuestions, only the patterns are larger, embedded in tables expressing complex *n-ary* relation semantics, and supporting inference over scientific and world knowledge.

Knowledge Type	Prevalence (% of expl.)	Rows in Table	Avg. Row Freq.
Retrieval Types			
Taxonomic	73%	2,111	1.8
Synonymy	57%	1,224	2.4
Properties (Things)	13%	525	1.3
MadeOf	12%	216	2.1
PartOf	9%	230	2.1
Properties (Generic)	8%	96	4.0
Contains	7%	158	2.2
Measurements	5%	44	5.8
Resources (P)	4%	25	7.2
Examples	4%	102	1.7
InheritedTraits (P)	3%	43	3.1
StatesOfMatter (P)	3%	5	26.2
Conductivity (P)	2%	16	5.6
Measurement Units (P)	2%	43	2.0
What Animals Eat (P)	1%	35	1.9
Orbital Periods (P)	1%	4	9.3
Magnetism (P)	1%	59	0.6
Inference Supporting Types			
Actions	26%	598	2.2
UsedFor	13%	331	2.0
SourceOf	11%	165	3.5
Requires	11%	211	2.6
Affect	7%	124	2.8
FormedBy	6%	86	3.4
Opposites	6%	65	5.6
Affordances	4%	92	2.1
Complex Inference Types			
If/Then	19%	501	2.0
Cause	17%	377	2.3
Changes (discrete)	14%	130	5.4
CoupledRelationships	8%	242	1.6
Changes (vector)	7%	123	3.1
Transfer	6%	71	4.4
ProcessRoles	3%	18	4.3
Vehicles/Enablement	1%	22	2.0

Table 1: A subset of the most frequently used tables in the WorldTree V2 tablestore. *Prevalence* refers to the proportion of explanations containing at least one fact (row) from a given table. *Average row frequency* refers to the average number of unique explanations a given fact will be used in.

3. Explanation Corpus

3.1. Questions

We author detailed explanations for standardized science exam questions drawn from the Aristo Reasoning Challenge (ARC) corpus (Clark et al., 2018), which contains 7,787 multiple choice science exam questions collected from 12 US states. Questions range from early elementary school through middle school level (3^{rd} through 9^{th} grade), which are typically used to evaluate students aged 8 through 14 years old. Annotators sorted questions using the curriculum topic annotation of Xu et al. (2019) and worked through blocks of questions on similar topics to increase consistency and reduce topic switching. A breakdown of questions and explanations by curriculum topic is included in Table 7 (see *Appendix*).

3.2. Explanation Authoring

The original Worldtree corpus (Jansen et al., 2018) contains 1,680 structured explanations for elementary science questions. In this work, approximately 3,400 additional explanations and 4,200 additional supporting facts largely centered around middle school questions were authored following the annotation protocol of Jansen et al. (2018), which we describe briefly here.

For each question, annotators must author a detailed explanation that describes the reasoning required to arrive at the correct answer. Explanations take the form of "explanation graphs", or interconnected sets of facts, where each fact represents a piece of atomic knowledge required to explainably answer the question. Facts in an explanation are interconnected by having lexical overlap (i.e. shared words) with question words, answer words, and/or other facts in that explanation, which form the edges of the explanation graph (see Figure 1). In this way, inference algorithms making use of these explanation graphs have explicit training data for how knowledge interconnects in large explanations, and can use this to learn to better aggregate knowledge in multi-hop inference settings. Explanations are authored by one annotator, reviewed by a second annotator, and any revisions or suggestions to improve decomposition or consistency are implemented by the original annotator. This entire process takes approximately 15 minutes per explanation generated.

3.2.1. Authoring Supporting Explanatory Facts

Facts used in explanations are drawn from a manually authored knowledge base of semi-structured tables. Each table in our knowledge base represents a particular type of knowledge surrounding a particular kind of *n-ary* relation, such as *taxonomic relations, part-of relations, properties, changes, causes, coupled relationships, if/then knowledge, object composition, requirements for processes,* or *sources of things*. Our knowledge base contains a total of 66 tables whose relations and column structure were developed using a detailed data-driven analysis of the domain requirements (Jansen et al., 2016; Jansen et al., 2018), while several properties tables were drawn from the Aristo Tablestore (Khashabi et al., 2016). Example tables and the prevalence of knowledge across those tables is shown in Table 1.

Each table in the WorldTree tablestore contains between 2 and 16 content columns that form a detailed *n-ary* relation surrounding a given type of knowledge, and provides a fine-grained compartmentalization of the knowledge in each fact. Where tabular forms of expressing knowledge often include only key columns required for a given relation (e.g. including only 2 columns, *hyponym* and *hypernym*, for a table representing taxonomic knowledge), our tables also include marked "filler" columns that allow each row to be expressed as a natural language sentence. In this way the tablestore is designed to appeal to inference algorithms that require either structured or free text, as well as serve as a comparison instrument between structured and free-text multi-hop algorithms by allowing them to use the same knowledge base.

In order to facilitate automated analyses of knowledge reuse, and to enable automated discovery of multi-hop inference patterns, the annotation protocol requires that annota-

Question	Petrified palm trees are found in rock near glaciers. Their presence most likely provides evidence that:
Answers	[A] There was once more water in the area. [B] The area was once grassland. *[*C] The climate in the area was once tropical.* [D] There are active faults in the area.

Facts in the gold explanation:

Core Scientific Knowledge
1. Petrified plants are a kind of fossil.
2. If fossils of an animal or plant are found in a place then that animal or plant used to live in that place.
3. Climate is the usual kind of weather in a location over a period of time.

World Knowledge
4. Palm trees usually live in tropical areas.
5. Trees are a kind of plant.
6. Tropical is a kind of climate.
7. "A place" is synonymous with "a location" and "an area".

Table 2: Examples of both core scientific knowledge and world knowledge found in the gold explanation for a question about what can be inferred from the location of fossils.

tors attempt to re-use the same facts (i.e. specific table rows) across different explanations rather than author duplicate knowledge in the knowledge base. When constructing explanations, annotators make use of an authoring tool that allows them to use keyword-based queries to quickly search the knowledge base for existing facts to help construct an explanation. We show how this requirement for knowledge reuse allows analyzing the knowledge requirements for building large corpora of explanations in Section 3.3.If no existing facts in the knowledge base cover the required knowledge, annotators author new facts (i.e. one or more new table rows) in a Google Sheet[1] containing the tablestore, which the explanation authoring tool draws it's knowledge base from.

3.2.2. World Knowledge and Explanatory Depth

Machines generally lack the experience of being situated in an environment, the experience of growing up and progressively learning about language, inference, and the world developmentally like a child, and other experiences that provide humans with a wealth of world knowledge and capacities at common-sense reasoning. One of the guiding principles of the WorldTree project is that, in order for inference algorithms to perform detailed multi-hop reasoning like humans do, this common sense reasoning and world knowledge must be enumerated in the corpora we use to train and evaluate an algorithm's ability to perform multi-fact inference. As such, we endeavor to include not only information that would be meaningful to a *domain expert* (such as a science teacher) in our explanations, but also include information that would be relevant to a *domain novice* (such as a science student), and to further enumerate *world knowledge* that moves closer to providing first-principles axiomatic foundations for the explanations. In practice, this knowledge takes on a variety of forms, from core facts (e.g. *"animals are a kind of living thing", "objects are made of materials"*) to rules or principles (e.g. *"if something is outside during the day, it will likely receive sunlight", "absorbing sunlight causes objects to heat"*), and occasionally synonymy knowledge (e.g. *"to add something means to increase that something"*). An example of both core scientific principles as well as world knowledge in an explanation is included in Table 2.

[1] `http://sheets.google.com`

Question	As water starts to freeze, the molecules of water
Answer	[*D] decrease in speed.

Central role
1. Freezing means matter changes from a liquid into a solid by decreasing heat energy.
2. As the temperature of a substance decreases, the molecules in that substance will move slower.

Grounding role
3. Water is a kind of substance.
4. Temperature is a measure of heat energy.

Lexical glue role
5. "To slow/to slow down" means decrease speed.

Table 3: Example facts for each of the three explanatory role labels each fact in an explanation is annotated with.

In practice, including world knowledge, common-sense reasoning, and other axiomatic forms of knowledge in explanations proves to be one of the most challenging and time-consuming aspects of explanation construction – particularly as, in our experience, annotators require significant training to learn to identify and generate this knowledge. We use a pragmatic approach, where we target authoring explanations with the goal of making them "meaningful to a five year old child".[2] Functionally, the explanation authoring tool provides annotators with a live view of the explanation graph during the construction process, which allows annotators to identify important key terms that have not yet been covered by the explanation (i.e. that do not have any supporting facts, and thus lack lexical-overlap edges), and for the annotator to then include those additional supporting facts to solidify the explanation.

3.2.3. Explanatory Role Annotation

To increase the utility of explanations to serve as training data for inference algorithms, for each fact in an explanation we also annotate its "explanatory role" (Jansen et al., 2018). We simplify the original explanatory role protocol to include only three categories, *central, grounding, and lexical glue*, which are illustrated in Table 3 and described below:

Central Role: Central facts describe the core principles a given question is testing. For the example in Table 3, the question is testing *changes in states of matter*, so core knowledge for *phase changes*, such as *"freezing means*

[2] The goal of "making explanations detailed enough to be meaningful to a 5 year old" is often suggested as an informal goal of contemporary efforts towards explanation-centered inference. While we also follow this as a guiding principle, we make no claims that the explanations in our corpus would satisfy this criterion, and do not directly evaluate this.

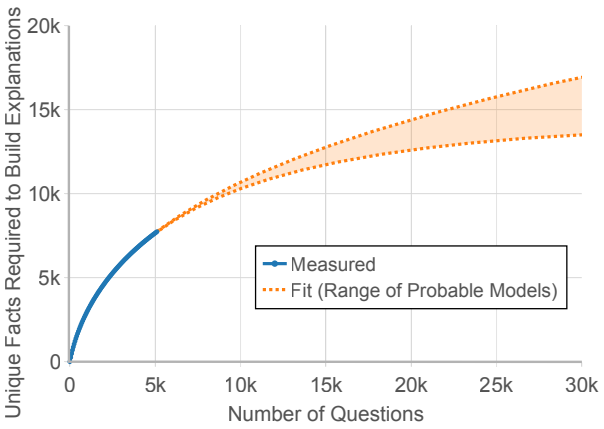

Figure 2: A monte-carlo analysis of knowledge growth in the explanation corpus. The X axis represents the number of questions subsampled from the corpus, and the Y axis represents the number of unique explanatory facts in those subsampled questions. The rate of adding new facts to explain additional questions in the corpus steadily decreases. Lines of best fit show upper and lower bounds for highly-probable extrapolations ($R^2 > 0.99$). Each point represents the average of 1000 subsampled corpora of a given size, subsampled in increments of 50 questions.

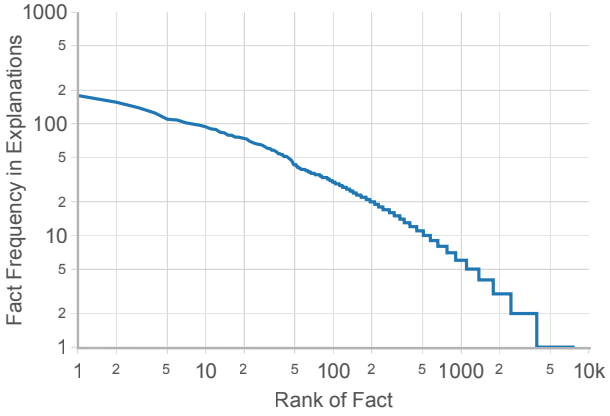

Figure 3: A frequency analysis of the number of unique explanations each fact in the knowledge base appears in. The distribution follows Zipf's law, where a small subset of core facts frequently recur across a large number of explanations, while a large set of highly-specific facts occur in only a single explanation.

matter changes from a liquid to a solid by decreasing heat energy" would be labeled as central.

Grounding Role: Grounding facts instantiate the core concepts the question is texting into specific examples, such as the specific liquid being frozen (in Table 3), or a specific animal (e.g. *a rabbit*) in a question testing environmental adaptations such as animals growing thicker fur to help stay warm in the winter season.

Lexical Glue Role: Lexical glue facts are an artifact of the requirement that when building explanation graphs, each core term in each fact must be explicitly connected through lexical overlap to either the question, answer, or other facts in the explanation. In this way, if one fact expresses knowledge using different terms, a lexical glue fact typically bridges those terms with a synonymy relation. In Table 3, the fact *"to slow down means to decrease speed"* serves a lexical glue role to bridge between the mention of *molecules moving slower* in the explanation and *decreasing in speed* in the answer.

Distribution of Roles: The average explanation contains 5.6 facts. Of these, 2.4 are labeled as *central*, 1.6 are *grounding*, and 1.3 are *lexical glue*.

3.3. Explanation Corpus Properties

The availability of a large corpus of semi-structured explanations allows analyzing the properties of explanations at scale. Here we describe the behavior of two properties, knowledge growth, and knowledge reuse.

Knowledge Growth and Domain Size: Manually authoring explanations and constructing a supporting knowledge base is expensive and labor intensive. By examining the number of new facts that must be authored to support constructing new explanations, one can arrive at an estimate of the domain size – the size of knowledge base that would be required to answer most of the questions that could be asked of a closed (but still large) domain, such as elementary and middle school science exams. To examine how knowledge grows with the number of questions in a corpus, following Jansen et al. (2018), we perform a monte-carlo analysis where we randomly subsample sets of questions from the full corpus and examine how many unique table rows are required to construct their explanations.

The results of this analysis are shown in Figure 2. In this corpus, knowledge growth tends to follow a power law, where as the number of questions increases, progressively fewer new facts are required in the knowledge base as the bulk of core scientific concepts becomes well-covered. A number of best-fit relationships have a high correlation to this distribution ($R^2 > 0.99$) that are equiprobable given the data available, and as such we include upper and lower bounds. For the science exam domain, given the natural variation that tends to exist within the questions, we estimate the domain size at between 14k and 25k unique facts, asymptoting between 40k and 300k questions. With our current knowledge base size of nearly ten thousand facts, 7,737 of them actively used in at least one explanation, this suggests that the WorldTree V2 tablestore likely covers between 30% to 55% of the domain.[3]

Knowledge Reuse and Zipf's Law: An open question in multi-hop inference is the general form that aggregating knowledge may take. On one hand, relatively small sets of knowledge arranged into customizable patterns or templates may be commonly reused to answer questions, implying

[3]This data-driven knowledge growth analysis assumes the same natural variation present within the standardized science exam questions. In principle a knowledge base could be many orders of magnitude larger than this by enumerating grounding knowledge (e.g. all possible objects, all possible animals, all possible substances) to generate slight variations of questions.

reasoning mechanisms similar to scripts or semantic frames (Minsky, 1974; Schank and Abelson, 1977; Wang et al., 2015; Ostermann et al., 2017), and that acquiring domain expertise is critical for multi-hop reasoning. At the other extreme, observing particular patterns of knowledge being reused may be rare, and multi-hop inference algorithms may need to acquire more domain-general mechanisms for connecting knowledge. Which point on this continuum a corpus of explanations exists at has strong implications for the design of multi-hop inference algorithms capable of generating similarly detailed explanations.

In this work, we observe a Zipfian distribution (Zipf, 1949) of fact reuse in explanations, with the results shown in Figure 3. Similar to how function words occur with high-frequency in free text while content words occur with decreasing frequency, we observe that both core scientific knowledge and world knowledge applicable to a large number of science curriculum topics (e.g. *"melting means changing from a solid to a liquid", "Earth is a kind of planet"*) occur with very high frequency. The most frequently reused fact, *"an animal is a kind of organism"*, occurs in 179 different explanations. Knowledge frequently reused in subdomains occupies the midsection of the distribution (e.g. *"the gravitational force of a planet does not change the mass of an object on that planet"*), while the long tail of facts used in only a single explanation includes a mix of core scientific facts (e.g. *"chromosomes contain thousands of genes"*) and highly-specific grounding knowledge (e.g. *"fiberglass is made of glass and plastic"*).

The results of this analysis suggest that rather than facts connecting with an average uniform probability, certain clusters of core scientific and world knowledge tend to be frequently connected and reused, while likely being augmented with the long tail of infrequently used facts that apply to highly-specific scenarios. We use this as motivation for constructing a corpus of structured inference patterns similar to schema or semantic frames that contain collections of commonly reused knowledge in Section 4.

3.4. Explanation Regeneration Task

The goal of multi-hop inference is to combine multiple facts in order to correctly answer questions, with a subgoal often being using those combined facts as an explanation for the reasoning process. Explanation Regeneration (Jansen and Ustalov, 2019) has been proposed as a stepping-stone task for many-fact multi-hop inference where given both a question and correct answer, a model must combine facts from a knowledge base to generate an explanation for that inference that is evaluated against a gold explanation manually generated by a human annotator. To enable a variety of methods to be used, the task is framed as a ranking task, where for a given question, a model must rank all the facts in a knowledge base in order of most-likely to least-likely to appear in the gold explanation. This can then be evaluated with standard ranking metrics, such as Mean Average Precision (MAP) and Precision@K. An example of the explanation regeneration task is shown in Table 4.

A variety of approaches have been applied to the explanation regeneration task using the original WorldTree corpus, including methods based on frames (D'Souza et al., 2019), information retrieval (Chia et al., 2019), and contextualized embeddings (Banerjee, 2019). Current state-of-the-art performance exceeds a strong BERT transformer baseline (Devlin et al., 2018) using an ensemble of BERT models trained on short chains of interconnected facts (Das et al., 2019). Here we provide baseline performance for both *tf.idf* and BERT models on the WorldTree V2 corpus in Table 5. The best performing BERT baseline is able to achieve a MAP of 0.52 on the test set, while 72% of the highest-ranked facts, and 41% (or 2) of the top 5 ranked facts are also found in the gold explanation. This baseline serves as a strong entry point for building more sophisticated models that use multi-hop inference to build detailed many-fact explanations, while the mid-level baseline performance highlights the difficulty of this task.

Question	The air in front of a meteor is compressed as it passes through the atmosphere of Earth. This causes the meteor to:
Answer	[*C] increase in temperature.

Gold Explanation

When a meteor enters Earth's atmosphere, the air in front of the meteor compresses.
When gas compresses, the temperature of the gas increases.
Atmosphere is synonymous with air.
Air is a kind of gas.
"Passing through" is similar to "entering".

Explanation Regeneration (top 10 ranked rows, tf.idf model)

1. * *When a meteor enters Earth's atmosphere, the air in front of the meteor compresses.*
2. A meteoroid is a kind of meteor.
3. Temperature rise means temperature increase.
4. Air temperature is a kind of temperature.
5. Shooting star is synonymous with meteor.
6. Warm up means increase temperature.
7. * *Atmosphere means air.*
8. If heat is transferred to the air , then the temperature of the air will increase.
9. Earth has air.
10. As temperature during the day increases, the temperature in an environment will increase.

Explanation Regeneration Scoring

Ranks of gold rows: 1, 7, 137, 215, 8650
Average Precision: 0.27
Precision@1: 1.0
Precision@5: 0.2

Table 4: An example of the explanation regeneration task, which ranks all facts in the knowledge base based on the likelihood that they appear in a gold explanation. This ranked list is then scored using MAP and Precision@K.

Model	fold	MAP	P@1	P@5
tf.idf baseline	dev	0.32	0.54	0.24
tf.idf baseline	test	0.29	0.48	0.24
BERT baseline	dev	0.53	0.71	0.41
BERT baseline	test	0.52	0.72	0.41

Table 5: Baseline performance on the explanation regeneration task for both *tf.idf* and BERT models.

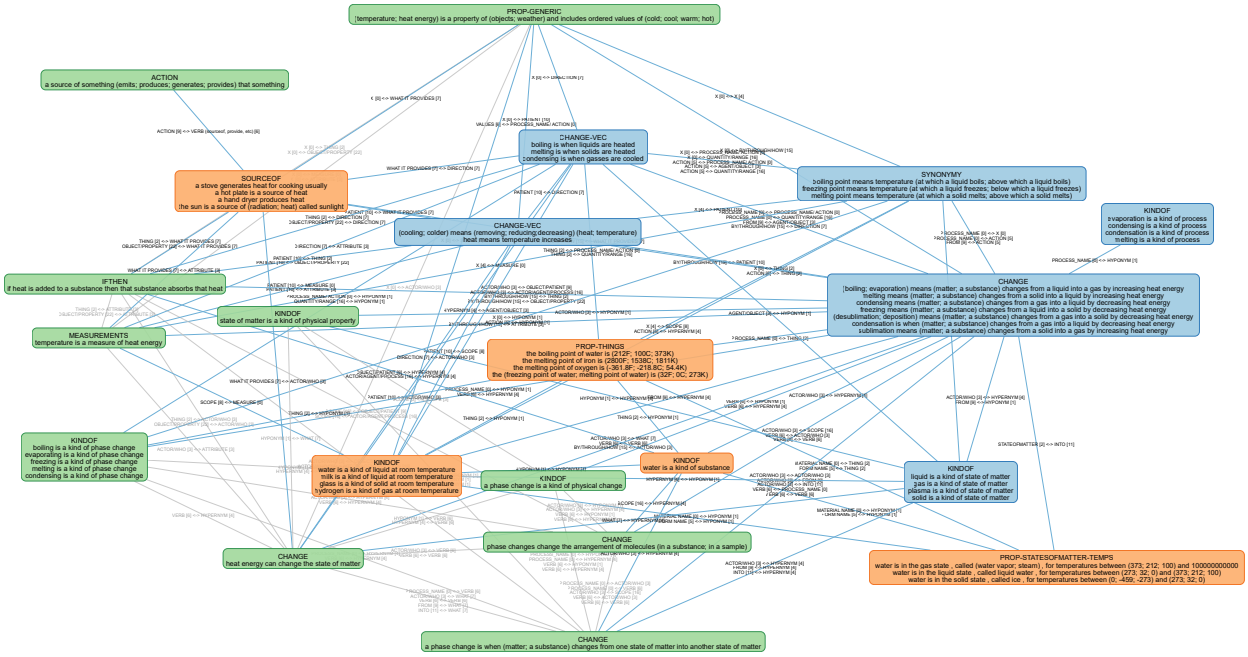

Figure 4: An example of one of the 344 inference patterns in this corpus, covering inferences involving *Changes of State of Matter*. Nodes represent "slots" filled by one fact, where edges represent constraints that must be satisfied between the knowledge contained in two nodes. Node color represents the explanatory role of a given node, while lists of rows provide examples of facts from explanations that could satisfy the edge constraints under certain circumstances. Edge labels represent lexical constraints between pairs of nodes, where for two nodes to contain specific rows, those rows must contain one or more of the same words in the table columns specified on the edge constraint. The full pattern contains 38 nodes and 76 edges, though for clarity, only nodes having *central (static), central (switchable)*, or *grounding* roles are shown. Nodes with more than one fact contain "hint rows" that provide clues for determining relevant edge connections (see *Appendix*).

4. Inference Pattern Corpus

Though elementary science questions generally require an average of 4 to 6 facts, and as many as 16 facts, to explainably answer (Jansen et al., 2016; Jansen, 2018), contemporary models for multi-hop inference generally struggle combining more than 2 pieces of knowledge, particularly from free-text. In a data-driven study of explanations drawn from the original WorldTree corpus, Jansen (2017) empirically demonstrated that it is possible in principle to regenerate most of gold explanations for unseen questions by combining abstracted versions of subgraphs from training explanations. Our own analysis in Section 3.3. shows that subsets of facts for specific subdomains are frequently reused across explanations. As such, a potential partial solution to the problem of many-fact multi-hop inference may be in using large multi-fact templates similar to schema or semantic frames (Schank and Abelson, 1977; Minsky, 1974), but derived from explanations, similar to approaches for explanation-based learning (DeJong and Mooney, 1986). Approaches to either use pre-existing frame resources such as FrameNet (Baker et al., 1998) for multi-hop inference (D'Souza et al., 2019), or to extract small subsets of knowledge graphs to serve as proxies for schemata (Lin et al., 2019), have shown initial progress for multi-hop reasoning. Here we explore generating large, detailed, domain-targeted and human-curated inference patterns at scale by analyzing the structure in this corpus of explanations.

One of the challenges associated with a frame-based approach to multi-hop inference is constructing the corpus of domain-relevant semantic frames, which is as labour-intensive a process as constructing structured gold explanations. Thiem et al. (2019) constructed a semi-automated tool that models inference pattern discovery as a graph construction process, first enabling a user to combine large numbers of explanation graphs from a specific sub-domain, then allowing the user to extract subgraphs of interconnected facts representing common inference patterns in that subdomain, such as *changes of state* or *electrical conductivity* for the *matter* subset of science questions. Unfortunately the manual annotation requirements of this method (approximately 2 hours per question) far exceed the ability to scale to large corpora. In this section we describe an alternate tool that models inference pattern generation as a bootstrapped binary judgement task, vastly reducing the time to generate inference patterns, and demonstrate this process at scale by extracting 344 high-level inference patterns from the WorldTree V2 training corpus.

4.1. Annotation Protocol and Tool

We describe the annotation protocol briefly here, with full details included in the *Appendix*. The inference pattern generation process starts with a set of seed facts surrounding a particular inference pattern theme, such as *"boiling means changing from a liquid to a gas"* and *"water in the gaseous state is called steam"* for an inference pattern about *changes of states of matter*. This set of seed facts are then used to bootstrap the generation of a larger list of candidate inference pattern facts by searching the training subset of the

Inference Pattern	Nodes	Edges
Astronomy		
Weather by Season	16	54
Heat from Sun by Distance	13	19
Gravity Factors - Mass and Distance	8	12
Tides by Gravity	13	32
Earth Science		
Rock Formation - Sedimentary	9	13
Habitat Destruction - Deforestation	15	15
Pollution - Burning Fossil Fuels	8	13
Earthquakes - Result in	10	19
Soil - Plants use up Nutrients	11	17
Erosion - Caused by Plant life	10	13
Energy		
Energy - Stays constant	7	50
Conv. - Chemical to Thermal Energy	7	21
Generators - How they work	9	5
Light - Color affects Reflection	13	42
Sounds - Speed in Different Media	12	33
Forces		
Friction - Effected By Texture	13	37
Gravity - Results In	14	6
Life Science		
Food for Repair and Growth	7	5
Animals - Hiding From Predators	14	19
Birds - Migration	15	24
Animals - Organ X has function Y	15	28
Photosynthesis - Through leaves	12	63
Plant - Reproduce through Pollination	17	53
Diseases - Curing is Positive	7	4
Decomposers - Role in Food Chain	12	32
Matter		
Change of State of Matter	38	76
Measuring - Temperature	10	9
Separation - Liquid-Solid Mixture	7	15
Objects have Mass	4	5

Table 6: A subset of the 344 inference patterns generated from the training corpus. *Nodes* represents the number of fact slots within a given pattern, while *edges* represents the number of constraints that exist between combinations of nodes. A full list of inference patterns is in the *Appendix*.

explanation corpus for explanations that contain one or more seed facts. If a matching explanation is found, *all* facts from that explanation are added to the list of candidate facts for the inference pattern. The annotator manually rates these candidate facts as to whether they are relevant to the inference pattern – that is, whether they are likely to be required when reasoning about a particular topic. The annotator also provides their assessment of the explanatory role of each fact, drawn from an expanded list of explanatory roles for inference patterns included in the *Appendix*.

Once the list of candidate facts has been rated, the ratings are then used to further bootstrap a longer candidate list of facts belonging to the inference pattern by again searching through the training subset of the explanation corpus for explanations that contain at least one fact rated as relevant in the candidate list. This iterative process continues for several iterations, until such time as the annotator decides the inference pattern has sufficient coverage of the topic. In this way, creating large generic multi-hop inference patterns is reduced to a large set of fast binary relevance judgements facilitated through a streamlined authoring tool, with the inference pattern authoring process generally requiring approximately 10-15 minutes per pattern. This is approximately an order of magnitude less than previous efforts, allowing pattern generation at scale.

4.2. Inference Pattern Representation

Following Thiem et al. (2019), inference patterns take the form of a graph that represents a series of constraints over knowledge base facts that must be satisfied in order for the inference pattern to generate a valid completion. In this way, nodes represent sets of facts drawn from a specific table in the knowledge base, while edges between nodes represent lexical constraints that must be satisfied between those nodes.

To illustrate this constraint satisfaction mechanism, in the *Changes of State of Matter* inference pattern shown in Figure 4, one node represents a core fact describing a particular change of state happening (e.g. *melting, boiling, freezing, or condensing*) drawn from the *CHANGE* table, while another node represents the initial state of the substance being changed (e.g. *solid, liquid, gas*) drawn from the *PROPERTIES-STATESOFMATTER* table. The constraint (edge) between these two nodes specifies that the initial state of matter in the change must match the state of matter of the substance. In this way, populating these nodes with the facts *"melting means changing from a solid to a liquid"* and *"water in the solid state is called ice"* would satisfy this constraint, and serve as a valid completion for these nodes. Similarly, *"freezing means changing from a liquid to a solid"* and *water in the gaseous state is called steam"* would not be a valid completion, and these nodes could never be simultaneously populated with these facts. By satisfying all the constraints of a given inference pattern, large meaningful multi-fact constructs can be constructed supporting the explanation-centered multi-hop inference tasks.

4.3. Inference Pattern Corpus Statistics

A set of 344 inference patterns were generated from the 2,582 training explanations in the WorldTree V2 corpus using the inference pattern authoring tool. On average, each pattern contains 11 nodes (sets of facts) and 26 edges (constraints between facts), representing a total of 3,650 annotated facts across all 344 inference patterns. Summary statistics for a subset of inference patterns is shown in Table 6, with the full list provided in the *Appendix*.

5. Conclusion

We present the WorldTree V2 corpus, a set of detailed multi-fact explanations for standardized science questions to support training and instrumenting many-fact multi-hop inference question answering systems. The corpus includes includes approximately 5,100 detailed explanations grounded in a semi-structured knowledge base of 9,216 facts across 66 tables representing fine-grained *n-ary* relations, as well as a set of 344 high-level common inference patterns containing a total of 3,650 facts. The corpus is is available for download at: `http://www.cognitiveai.org/explanationbank/`.

6. Bibliographical References

Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). The berkeley framenet project. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1*, pages 86–90. Association for Computational Linguistics.

Banerjee, P. (2019). ASU at TextGraphs 2019 Shared Task: Explanation ReGeneration using Language Models and Iterative Re-Ranking. In *Proceedings of the Thirteenth Workshop on Graph-Based Methods for Natural Language Processing*, TextGraphs-13, Hong Kong. Association for Computational Linguistics.

Chen, J. and Durrett, G. (2019). Understanding dataset design choices for multi-hop reasoning. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4026–4032.

Chia, Y. K., Witteveen, S., and Andrews, M. (2019). Red Dragon AI at TextGraphs 2019 Shared Task: Language Model Assisted Explanation Generation. In *Proceedings of the Thirteenth Workshop on Graph-Based Methods for Natural Language Processing*, TextGraphs-13, Hong Kong. Association for Computational Linguistics.

Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C., and Tafjord, O. (2018). Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Das, R., Dhuliawala, S., Zaheer, M., Vilnis, L., Durugkar, I., Krishnamurthy, A., Smola, A., and McCallum, A. (2017). Go for a walk and arrive at the answer: Reasoning over paths in knowledge bases using reinforcement learning. *AKBC*.

Das, R., Godbole, A., Zaheer, M., Dhuliawala, S., and McCallum, A. (2019). Chains-of-Reasoning at TextGraphs 2019 Shared Task: Reasoning over Chains of Facts for Explainable Multi-hop Inference. In *Proceedings of the Thirteenth Workshop on Graph-Based Methods for Natural Language Processing*, TextGraphs-13, Hong Kong. Association for Computational Linguistics.

DeJong, G. and Mooney, R. (1986). Explanation-based learning: An alternative view. *Machine learning*, 1(2):145–176.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

D'Souza, J., Mulang', I. O., and Auer, S. (2019). Team SVM$^{\text{rank}}$: Leveraging Feature-rich Support Vector Machines for Ranking Explanations to Elementary Science Questions. In *Proceedings of the Thirteenth Workshop on Graph-Based Methods for Natural Language Processing*, TextGraphs-13, Hong Kong. Association for Computational Linguistics.

Franz, M., Lopes, C. T., Huck, G., Dong, Y., Sumer, O., and Bader, G. D. (2015). Cytoscape. js: a graph theory library for visualisation and analysis. *Bioinformatics*, 32(2):309–311.

Fried, D., Jansen, P., Hahn-Powell, G., Surdeanu, M., and Clark, P. (2015). Higher-order lexical semantic models for non-factoid answer reranking. *Transactions of the Association for Computational Linguistics*, 3:197–210.

Jansen, P. and Ustalov, D. (2019). Textgraphs 2019 shared task on multi-hop inference for explanation regeneration. In *Proceedings of the Thirteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-13)*, pages 63–77.

Jansen, P., Balasubramanian, N., Surdeanu, M., and Clark, P. (2016). What's in an explanation? characterizing knowledge and inference requirements for elementary science exams. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2956–2965, Osaka, Japan, December.

Jansen, P., Sharp, R., Surdeanu, M., and Clark, P. (2017). Framing qa as building and ranking intersentence answer justifications. *Computational Linguistics*.

Jansen, P., Wainwright, E., Marmorstein, S., and Morrison, C. (2018). Worldtree: A corpus of explanation graphs for elementary science questions supporting multi-hop inference. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.

Jansen, P. A. (2017). A study of automatically acquiring explanatory inference patterns from corpora of explanations: Lessons from elementary science exams. In *6th Workshop on Automated Knowledge Base Construction (AKBC 2017)*.

Jansen, P. (2018). Multi-hop inference for sentence-level textgraphs: How challenging is meaningfully combining information for science question answering? *TextGraphs*.

Khashabi, D., Khot, T., Sabharwal, A., Clark, P., Etzioni, O., and Roth, D. (2016). Question answering via integer programming over semi-structured knowledge. In *Proceedings of the International Joint Conference on Artificial Intelligence*, IJCAI'16, pages 1145–1152.

Khot, T., Clark, P., Guerquin, M., Jansen, P., and Sabharwal, A. (2019). Qasc: A dataset for question answering via sentence composition. *arXiv preprint arXiv:1910.11473*.

Lin, B. Y., Chen, X., Chen, J., and Ren, X. (2019). Kagnet: Knowledge-aware graph networks for commonsense reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2822–2832.

Mihaylov, T., Clark, P., Khot, T., and Sabharwal, A. (2018). Can a suit of armor conduct electricity? a new dataset for open book question answering. *arXiv preprint arXiv:1809.02789*.

Minsky, M. (1974). A framework for representing knowledge. Technical report, Cambridge, MA, USA.

Ostermann, S., Roth, M., Thater, S., and Pinkal, M. (2017). Aligning script events with narrative texts. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, pages 128–134, Vancouver, Canada, August. Association for Computational Linguistics.

Schank, R. and Abelson, R. (1977). *Scripts, plans, goals*

and understanding: An inquiry into human knowledge structures. Lawrence Erlbaum Associates, Hillsdale, NJ.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, 13(11):2498–2504.

Talmor, A. and Berant, J. (2018). The web as a knowledge-base for answering complex questions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 641–651.

Thiem, S. and Jansen, P. (2019). Extracting common inference patterns from semi-structured explanations. In *Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing*, pages 53–65.

Wang, H., Bansal, M., Gimpel, K., and McAllester, D. (2015). Machine comprehension with syntax, frames, and semantics. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 700–706.

Welbl, J., Stenetorp, P., and Riedel, S. (2018). Constructing datasets for multi-hop reading comprehension across documents. *Transactions of the Association for Computational Linguistics*, 6:287–302.

Xu, D., Jansen, P., Martin, J., Xie, Z., Yadav, V., Madabushi, H. T., Tafjord, O., and Clark, P. (2019). Multi-class hierarchical question classification for multiple choice science exams.

Yang, Z., Qi, P., Zhang, S., Bengio, Y., Cohen, W. W., Salakhutdinov, R., and Manning, C. D. (2018). Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*.

Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort*. Addison-Wesley, Reading MA (USA).

Appendix

6.1. Explanations by Curriculum Topic

The distribution of explanations in the WorldTree V2 corpus by curriculum topic is provided in Figure 7. We use the curriculum topic annotation of Xu et al. (2019), which divides the ARC corpus questions into approximately 400 fine-grained categories that can be reduced to a set of 9 coarse categories, such as *Astronomy* or *Life Science*.

Curriculum Category	Questions with Explanations
Astronomy	606
Earth Sciences	1,880
Energy	938
Forces and Motion	142
Life Sciences	1,505
Matter and Chemistry	1,479
Scientific Inference	106
Safety	18
Other	89
Total Questions	5,114

Table 7: Summary statistics of questions broken down by curriculum topic. Note that questions may be labelled as belonging to more than one curriculum topic, and as a result the individual counts do not sum to the total number of questions.

6.2. Inference Pattern Annotation Protocol

Following Thiem et al. (2019), we model "inference patterns" as sets of interconnected facts similar to explanation graphs, where the edges define constraints on the knowledge (i.e. shared words) that must populate both facts on an edge in order for the inference pattern to be satisfied. We model inference pattern discovery as an iterative process that begins with a small collection of "seed facts" surrounding a particular inference pattern theme, such as *changes of state*, which are used to bootstrap the collection of a larger set of facts from the knowledge base. This bootstrapped collection of facts is then manually rated by an annotator using binary judgements to determine whether a given fact is generally relevant to answering questions on a specific theme. More specifically, our annotation protocol is as follows:

1. Seed Fact Collection: We take all explanations in the training corpus, and filter them to include only core scientific facts by including only those facts rated as having a *"central"* explanatory role (see Section 3.2.3.)This results in one set of seed facts per question. Duplicate sets of seed facts are removed first automatically, then manually for near-duplicates, whose sets of seed facts are merged. The annotator inspects each set of seed facts, and provides a thematic label (e.g. *measuring speed*, *electrical power generation*, or *effects of wind erosion*). This label and associated set of seed facts serves as the beginning of each inference pattern.

2. Bootstrap Inference Pattern: We search all explanations in the training corpus for cases where one or more seed facts are found in a given explanation. If one or more facts are found, we add *all* facts from that explanation to a list of candidate facts that may be relevant to the inference pattern.

3. Manual Rating: For each fact in the candidate list of facts, the annotator provides a binary judgement as to whether that fact is relevant to the inference pattern. Using a set of buttons, the annotator is able to provide a categorical judgement as to what explanatory role the fact takes in the inference pattern. As in the explanation annotation we include *central, grounding,* and *lexical glue* roles, as well as several variations described in Section 6.2.1.

4. Iteratively Bootstrap: The set of candidate facts that have been rated as relevant to an inference pattern now become the new set of seed rows for the bootstrapping process. A larger set of candidate facts are how added to the inference pattern, and new facts are manually rated. To prevent the annotator from being overloaded by irrelevant facts as the bootstraping process creates large lists of candidates, the annotator can sort candidates by either frequency or the number of times a given fact had a "central" role in explanations, to focus on locating core facts quickly. In practice, we found that only a few iterations of this bootstrapping process were required to obtain well-populated inference patterns.

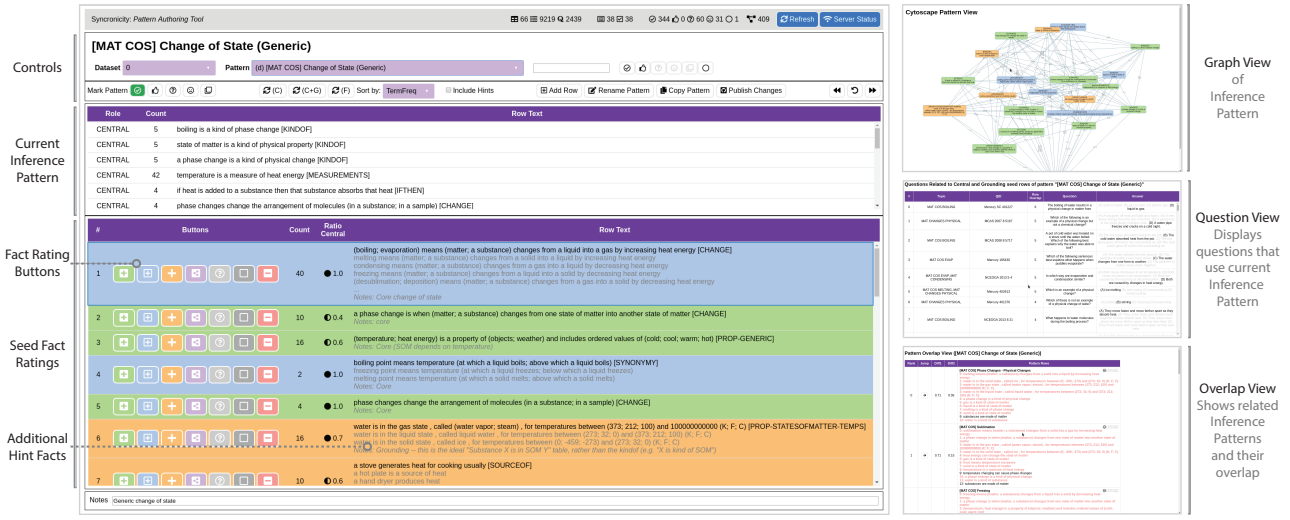

Figure 5: A screenshot of the inference pattern authoring tool, *Synchronicity*. *(Left)* The main dialog, including seed rows and row ratings for the bootstrapping procedure. Color represents the explanatory role each fact takes in the inference pattern. *(Right)* The graph view, question view, and overlap views, providing additional analyses to the annotator.

5. Populate "Connection Hint" Rows: To be maximally useful as templates for how knowledge interconnects, we allow the annotator to specify whether a given fact is a core scientific fact likely to always be present when using a given pattern, or whether that fact has a more grounding role and is likely to be swapped out for different related facts when applied to different questions. For cases where a fact is likely to be swapped, the annotator provides "hint rows", or examples of other facts that might take it's place. For example, if a given inference pattern contains the fact "a turtle is a kind of reptile", for which the annotator provided the hint rows "a gecko is a kind of reptile" and "an alligator is a kind of reptile", this additional annotation allows automated methods to infer that the most important edge upon which this fact connects to other facts in the inference pattern is on the shared word *reptile*. Without this annotation, automated methods for determining edges can be overconstrained, leading to poor generalization (Thiem and Jansen, 2019).

6.2.1. Inference Pattern Explanatory Roles

The explanatory roles for fact slots in inference patterns are a slightly expanded set of the explanatory roles used in the explanation annotation and described in Section 3.2.3.:

Central (static): Facts that are central to the inference pattern, and also unlikely to change across different instantiations of the pattern. For example, in the *Changes in State of Matter* pattern in Figure 4, the fact *"a phase change is when matter changes from one state to another state"* is unlikely to change whether a particular instantiation of the pattern is about *melting, boiling, freezing, or condensing*.

Central (switchable): Facts that are central to the inference pattern, but are likely to be swapped between a small set of related facts depending on the needs of an inference. For example, a given inference involving *Changes of State of Matter* may require knowing about the specific phase change happening *(e.g. melting, boiling, freezing, or condensing)*. The node representing this knowledge would be rated as *central (switchable)*, and given the value of a particular fact (e.g. *"melting means changing from a solid to a liquid by adding heat energy"*).

Grounding: As in the explanation annotation, grounding facts relate the central concepts of the inference pattern with specific examples that may be involved in a particular question, such as *"water in the solid state is called ice"*.

Lexical Glue: Also as in the explanation annotation, lexical glue facts express synonymy relations that bridge two facts in other nodes.

Peripheral: Peripheral facts are facts that may infrequently be relevant to a given inference, but were marked as potentially being relevant in some cases by the annotator.

Optional: The optional flag allows specifying that a given node is not necessarily required for all inferences, but may add extra information to an inference pattern when available. The optional flag is in addition to the explanatory role that a given node takes.

6.2.2. Annotation Tool

We developed an inference pattern authoring tool, *Synchronicity*, to facilitate the process of pattern generation, from bootstrapping facts, through manually rating rows, as well as evaluating patterns for actual usage utility in questions to ensure high relevance. A screenshot of this tool is shown in Figure 5.

The tool includes four main components. The main dialog shows the current inference pattern, and allows rating candidate facts obtained through the bootstrapping process, as well as performing the bootstrap procedure. Three separate pop-up dialogs show the annotator analyses helpful to the authoring process. The *graph view* shows the inference pattern as an explanation graph, including the explicit connections between facts in the pattern. The *question view*

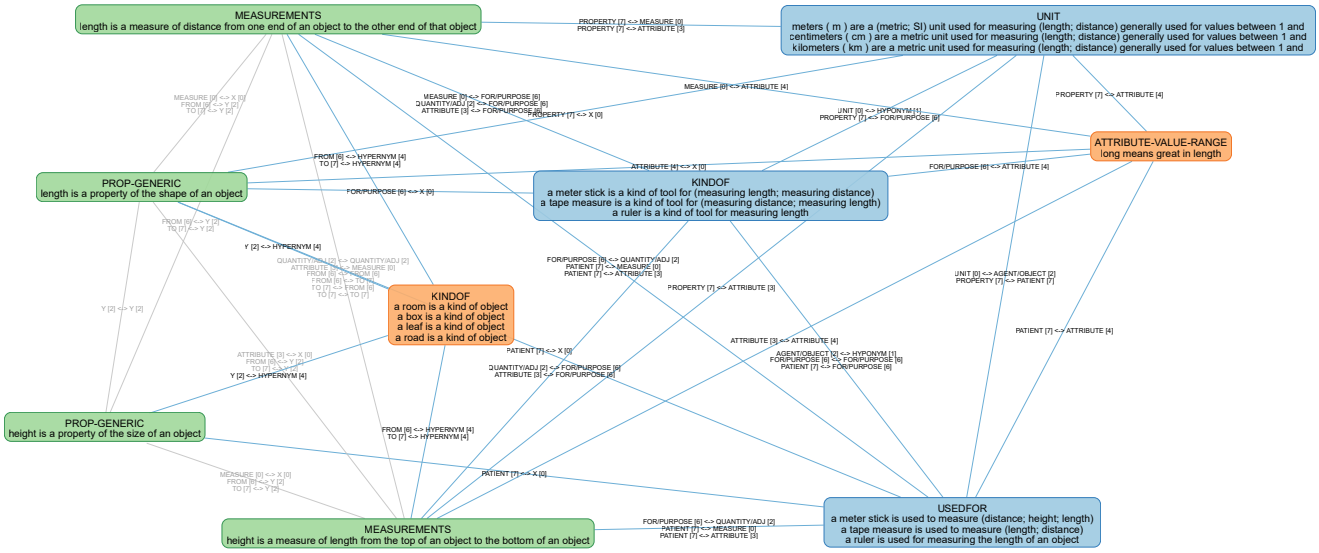

Figure 6: An additional example of one of the 344 inference patterns in this corpus, covering a subset of inferences involving *Measuring Length*. Nodes represent "slots" for one fact, where edges represent constraints that must be satisfied between the knowledge contained in two nodes.

shows a sorted list of questions and supporting summary statistics for where the inference pattern can likely be applied. Finally, the *overlap view* shows a sorted list of how the knowledge in the current inference pattern overlaps with knowledge in other patterns, to prevent the annotator from duplicating substantial portions of inference patterns. The total time required for generating a single inference pattern using this tool is approximately 10-15 minutes. The tool runs in the Chrome web browser, and was implemented in `node.js`[4], with the Cytoscape.js library (Shannon et al., 2003; Franz et al., 2015) used for graph visualization.

6.3. List of Inference Patterns

The full list of the 344 inference patterns included in this corpus is shown in Table 8 below.

[4] `https://nodejs.org/`

Inference Pattern Name	Nodes	Constraints
[CEL CYCLES] Daylight by Season	17	90
[CEL CYCLES] Solar Energy in Summer	15	51
[CEL CYCLES] Weather by Season	16	54
[CEL CYCLES] Years by Orbit	10	29
[CEL DISTANCES] Heat from Sun by Distance	13	19
[CEL DISTANCES] Lightyears as Units	9	13
[CEL FEATURES] Definition - Galaxies	11	33
[CEL FEATURES] Definition - Moons	6	1
[CEL FEATURES] Definition - Planet	13	76
[CEL FEATURES] Definition - Stellar	16	46
[CEL FEATURES] Definition - Sun	6	9
[CEL FEATURES] Energy Transfer - Comet Striking	12	77
[CEL FEATURES] Gravity Factors - Mass and Distance	8	12
[CEL FEATURES] Life Stage - Stellar	16	35
[CEL FEATURES] Light by The Moon	9	23
[CEL FEATURES] Properties - Sun	6	17
[CEL GRAVPULL] Weight by Mass	9	20
[CEL INSTRUMENTS] Instruments - Celestial	9	9
[CEL MOONPHASES] Moon Phases	8	42
[CEL ORBIT] Definition - Orbit	13	19
[CEL SPACEEX] Spacesuit Protection	12	9
[CEL TIDES] Tides by Gravity	13	32
[EARTH GEO] Classifying - Rocks	8	13
[EARTH GEO] Definition - Fossil	9	13
[EARTH GEO] Definition - Minerals	9	6
[EARTH GEO] Fossils in Rock Types	9	3
[EARTH GEO] Minerals - Hardness	13	17
[EARTH GEO] Rock Cycles	16	81
[EARTH GEO] Rock Formation - Igneous	10	4
[EARTH GEO] Rock Formation - Metamorphic	7	25
[EARTH GEO] Rock Formation - Sedimentary	9	13
[EARTH HUMIMP] Acid Rain	11	24
[EARTH HUMIMP] Deforestation	13	14
[EARTH HUMIMP] Fossil Fuels - Burns	7	10
[EARTH HUMIMP] Glaciers - Traces	11	11
[EARTH HUMIMP] Global Warming - Caused by Deforestation	15	38
[EARTH HUMIMP] Global Warming - Rising Sea Levels	11	65
[EARTH HUMIMP] Global Warming	8	17
[EARTH HUMIMP] Habitat Destruction - Deforestation	15	15
[EARTH HUMIMP] Habitat Destruction - Water Pollution	19	24
[EARTH HUMIMP] Nutrients by Farming	9	6
[EARTH HUMIMP] Nutrients from Soil	10	8
[EARTH HUMIMP] Organisms - Require Water	12	26
[EARTH HUMIMP] Origin - Coal	10	22
[EARTH HUMIMP] Origin - Fossil Fuels	6	2
[EARTH HUMIMP] Origin - Natural Gas	10	14
[EARTH HUMIMP] Origin - Nutrients in Soil	13	10
[EARTH HUMIMP] Plants - Protecting From Cold	11	22
[EARTH HUMIMP] Pollution - Burning Fossil Fuels	8	13
[EARTH HUMIMP] Pollution - Effect on Environment	12	16
[EARTH HUMIMP] Pollution - Energy Conservation	11	20
[EARTH HUMIMP] Pollution - Fertilizer	8	4
[EARTH HUMIMP] Pollution - Landfills	7	9
[EARTH HUMIMP] Pollution - Littering	9	6
[EARTH HUMIMP] Pollution - Solar Energy - Counteract	10	15
[EARTH HUMIMP] Pollution - Vehicles - Counteract	11	26
[EARTH HUMIMP] Pollution - Vehicles	10	19
[EARTH HUMIMP] Recycling - Landfills	14	21
[EARTH HUMIMP] Recycling - Resources	8	18
[EARTH HUMIMP] Renewable vs Non-Renewable Resource	11	22
[EARTH HUMIMP] Soil - Potential Nutrients	12	15
[EARTH HUMIMP] Soil - Quality by Nutrients	10	27
[EARTH HUMIMP] Solar Panels	13	63
[EARTH HUMIMP] Source of Freshwater	17	61

[EARTH INNER] Earthquakes - Origin	8	12
[EARTH INNER] Earthquakes - Result in	10	19
[EARTH INNER] Volcanos - Block Sunlight Required by Plants	14	16
[EARTH INNER] Volcanos - Decreased Temperature	17	13
[EARTH OUTER] Atmosphere - Layers	9	12
[EARTH OUTER] Atmosphre - Greenhouse Gasses	9	43
[EARTH OUTER] Definition - Ocean	7	18
[EARTH OUTER] Origin - Water in Atmosphere	13	75
[EARTH OUTER] Ozone Layer - Protects Life	10	15
[EARTH OUTER] Pollution - SO2 in Ocean	11	19
[EARTH SOIL] Soil - Plants use up Nutrients	11	17
[EARTH WEATHER] Climate - Humidity	13	25
[EARTH WEATHER] Definition - Droughts	12	15
[EARTH WEATHER] Definition - Erosion	12	18
[EARTH WEATHER] Erosion - Caused by Plantlife	10	13
[EARTH WEATHER] Erosion - Inhibited by Plantlife	11	24
[EARTH WEATHER] Erosion - Negative Effects	6	2
[EARTH WEATHER] Erosion - Results in	11	8
[EARTH WEATHER] Floods - Caused by	13	2
[EARTH WEATHER] Landslide	10	8
[EARTH WEATHER] Measuring - Temperature	8	18
[EARTH WEATHER] Mountains Cause Low Humidity	19	43
[EARTH WEATHER] Ocean as a Source of Humidity	9	16
[EARTH WEATHER] Ocean Effect on Climate	9	7
[EARTH WEATHER] Ocean Effect on Hurricanes	11	13
[EARTH WEATHER] Origin - Clouds	14	43
[EARTH WEATHER] Precipitation	12	32
[EARTH WEATHER] Precipition - Examples of	6	10
[EARTH WEATHER] Properties - Weather	16	46
[EARTH WEATHER] Sediment - From Weathering	8	8
[EARTH WEATHER] Soil - From Weathering	7	8
[EARTH WEATHER] Storms - Result in	10	27
[EARTH WEATHER] Sun - Causes Evaporation in Water Cycle	13	65
[EARTH WEATHER] Sun - Gives energy to Earth	9	24
[EARTH WEATHER] Temperature - Effect on Forms of Precipitation	11	21
[EARTH WEATHER] Thunderstorms	10	6
[EARTH WEATHER] Weathering - Abrasion	6	3
[EARTH WEATHER] Weathering - Caused by Bio Acid	11	9
[EARTH WEATHER] Weathering - Caused by Ice Wedging	11	31
[EARTH WEATHER] Weathering - Result in	8	1
[EARTH WEATHER] Wildfires by Droughts	15	12
[EARTH WEATHER] Wind Erosion - Result in	8	5
[EARTH WEATHER]Direct Sunlight - Equator and Poles	10	17
[ENG CONSERVATION] Energy - Stays constant	7	50
[ENG CONV] Conversion - Chemical to Mechanical Energy	16	101
[ENG CONV] Conversion - Chemical to Thermal Energy	7	21
[ENG CONV] Conversion - Electrical To Mechanical Energy	4	22
[ENG CONV] Conversion - Mechanical to Electrical Energy - Hydropower	10	12
[ENG CONV] Conversion - Mechanical to Thermal Energy - Friction	17	62
[ENG CONV] Origin - Sound Energy - Musical Instruments	19	103
[ENG CONV] Origin - Sound Energy	13	50
[ENG DEVICES] Devices - Powered By Electical Circuits	18	148
[ENG DEVICES] Eelctrical to Sound	9	46
[ENG DEVICES] Electrical to Heat	9	42
[ENG DEVICES] Electrical to Light	10	78
[ENG DEVICES] Generators - How they work	9	5
[ENG ELEC] Generators	5	17
[ENG ELEC] Open vs Closed Circuits	13	99
[ENG FORMS] Chemical Energy - Is Stored in	11	32
[ENG FORMS] Electrical Energy - Static Electricity	15	8
[ENG FORMS] Energy - Examples of	7	51
[ENG FORMS] Examples - Mechanical Energy	4	4
[ENG INTERACTIONEM] Sun - Heats things	10	25
[ENG LIGHT] Light - Color affects Reflection	13	42
[ENG LIGHT] Light - Color by Reflection	10	43

[ENG LIGHT] Light - Properties	12	54
[ENG LIGHT] Light - Reflects more off shiny things	11	43
[ENG LIGHT] Refraction - Rainbows	12	71
[ENG LIGHT] Seeing - Light	11	42
[ENG POTENTIALKINETIC] Definition - Potential Energy	13	78
[ENG POTENTIALKINETIC] Kinetic Energy - Temperature	7	3
[ENG POTENTIALKINETIC] Potential to Kinetic Energy	10	36
[ENG SOUND] Echos - Caused By	10	25
[ENG SOUND] Sound Waves - Properties	9	37
[ENG SOUND] Sounds - Travel Speed in Different Media	12	33
[ENG THERM] Convection	12	38
[ENG THERM] Definition - Conduction	11	28
[ENG THERM] Radiation - Fire	13	34
[ENG THERM] Radiation - Sunlight	14	46
[ENG THERM] Thermal Conductors Conduct	16	55
[ENG WAVES] Refraction of Waves - Application	6	7
[ENG WAVES] Waves - Earthquake Waves	13	48
[ENG WAVES] Waves - Frequencies and Wavelength	9	53
[ENG WAVES] Waves - Properties	9	19
[FOR FRICTION] Definition - Friction	15	47
[FOR FRICTION] Friction - Effected By Texture	13	37
[FOR FRICTION] Friction - Effected by Water	9	25
[FOR FRICTION] Friction - Results in	13	34
[FOR GRAVITY] Gravity - Results In	14	6
[FOR MECH] Mechanical Forces - Result In	13	179
[FOR VELOCITY] Definition - Velocity	8	12
[LIFE CLASSIFICATION] Plants - Plant Cells	14	105
[LIFE CLASSIFICATION] Properties - Birds	8	7
[LIFE CLASSIFICATION] Properties - Fish	13	36
[LIFE CLASSIFICATION] Properties - Fungus	9	5
[LIFE CONTL] Food for Repair and Growth	7	5
[LIFE ENVADP] Animals - Adaptation for Hunting	13	24
[LIFE ENVADP] Animals - Bright Color Implies Toxic	11	8
[LIFE ENVADP] Animals - Camouflage	10	34
[LIFE ENVADP] Animals - Fat Protects From Cold and Starvation	14	47
[LIFE ENVADP] Animals - Fur Protects from Cold	9	23
[LIFE ENVADP] Animals - Hiding From Predators	14	19
[LIFE ENVADP] Animals - Hunt with X	11	25
[LIFE ENVADP] Animals - Sense X with Y	10	16
[LIFE ENVADP] Animals - Traits to attract a mate	16	23
[LIFE ENVADP] Bird Migration	15	24
[LIFE ENVADP] Plants - Plant does X to adapt to Y	12	22
[LIFE ENVADP] Plants - Using X to Defend from Y	8	7
[LIFE FUNCT] Animal - Breathing	7	6
[LIFE FUNCT] Animal Cells - Respiration	19	34
[LIFE FUNCT] Animals - Changing due to Habitat Changes	11	17
[LIFE FUNCT] Animals - Displaced by Habitat Destruction	11	17
[LIFE FUNCT] Animals - Eat X to get nutrients	12	25
[LIFE FUNCT] Animals - Organ X does Y	15	28
[LIFE FUNCT] Animals - Protected by part of body X	13	21
[LIFE FUNCT] Animals - React to Enviroment Change	9	41
[LIFE FUNCT] Animals - React to Habitat Changes	12	72
[LIFE FUNCT] Animals - Regulating water levels	14	49
[LIFE FUNCT] Animals - Seeking Shelter	8	6
[LIFE FUNCT] Animals - System X does Y	2	2
[LIFE FUNCT] Cell Theory - Things are made of cells	9	18
[LIFE FUNCT] Cells - Part of Biological Systems	6	5
[LIFE FUNCT] Cells - Sexual Reproduction	10	20
[LIFE FUNCT] Locomotion - Controlled by Nervous	11	36
[LIFE FUNCT] Oceans - Light by Depth	7	7
[LIFE FUNCT] Organs X part of System Y	2	0
[LIFE FUNCT] Parts of Organs	12	39
[LIFE FUNCT] Photosynthesis - Does X	14	57
[LIFE FUNCT] Photosynthesis - Process	14	52
[LIFE FUNCT] Photosynthesis - Through leaves	12	63

[LIFE FUNCT] Plant - Part X does Y	14	28
[LIFE FUNCT] Plant - Reproduce through Pollination	17	53
[LIFE FUNCT] Plant - Seed Dispersal	10	47
[LIFE FUNCT] Plants - Changing due to Habitat Changes	10	33
[LIFE FUNCT] Plants - Consume X for nutrients	9	13
[LIFE FUNCT] Plants - React to Habitat Changes	8	9
[LIFE FUNCT] Plants - Roots as a Part of the Plant	7	25
[LIFE FUNCT] Plants - Stems	7	40
[LIFE FUNCT] Plants - Transpiration	10	38
[LIFE FUNCT] Sense X is for Sensing Y	14	11
[LIFE FUNCT] Trait X Improves Pollination	16	54
[LIFE FUNCT] Trees - Block Sunlight	8	22
[LIFE FUNCT] Trees role in water cycle	7	21
[LIFE HEALTH] - Diseases - Curing is good	7	4
[LIFE HEALTH] Diseases - Food Poisoning	10	34
[LIFE INTERDEP] Animals - Mobility from Energy	4	2
[LIFE INTERDEP] Carnivores	7	6
[LIFE INTERDEP] Climate - Effect on Population	11	25
[LIFE INTERDEP] Consumer vs Producer	11	29
[LIFE INTERDEP] Consumers	14	37
[LIFE INTERDEP] Dead organisms spread disease	12	19
[LIFE INTERDEP] Decomposers - Consume X	8	20
[LIFE INTERDEP] Decomposers - Role in Food Chain	12	32
[LIFE INTERDEP] Food Chain - Dependancies	13	39
[LIFE INTERDEP] Herbivores	7	7
[LIFE INTERDEP] Living Things - Examples	6	7
[LIFE INTERDEP] Overconsumption of Resources Hurting Other Consumers	10	14
[LIFE INTERDEP] Predators and Prey	9	13
[LIFE INTERDEP] Producers - Definition	9	23
[LIFE INTERDEP] Producers - Examples	6	13
[LIFE INTERDEP] Sun - Gives energy to Living Things	15	44
[LIFE INTERDEP] X is a part of Y - X is abiotic	6	6
[LIFE LIVINGNONLIVING] Nonliving Things - Examples	9	31
[LIFE REPROD] Alleles	12	17
[LIFE REPROD] Amino Acids	3	1
[LIFE REPROD] Asexual Offspring	6	15
[LIFE REPROD] Characteristic X is a result of Y	7	7
[LIFE REPROD] DNA	12	25
[LIFE REPROD] Genes - Precentages	11	259
[LIFE REPROD] Genes Relating to Species	9	15
[LIFE REPROD] Incomplete Dominance	12	100
[LIFE REPROD] Inherited Characteristics	8	21
[LIFE REPROD] Instinctive Behaviours	6	5
[LIFE REPROD] Protein Formation - From DNA	8	27
[LIFE REPROD] Recessive Genes - Behaviour	12	72
[LIFE REPROD] Selective Breeding - Effect on diversity	5	7
[LIFE REPROD] Selective Breeding - Example	7	11
[LIFE REPROD] Sexual Reproduction - Outcomes of	8	35
[LIFE REPROD] Sexual Reproduction Process - Plants	11	57
[MAT CHANGES] Chemical Reactions - Fire	11	37
[MAT CHANGES] Chemical Seperation - From Energy	8	11
[MAT CHANGES] Chemicals - Properties	12	24
[MAT CHANGES] Compound - Definition	6	8
[MAT CHANGES] Compounds - Properties	10	35
[MAT CHANGES] Definition - Chemical Changes	9	11
[MAT CHANGES] Gas Containers - Behaviours	14	47
[MAT CHANGES] Iron - Properties	14	13
[MAT CHANGES] Physical Changes - Shape Change	7	33
[MAT CHANGES] Physical Changes vs Chemical Changes	11	57
[MAT CHEM] Atoms smallest compound	8	10
[MAT CHEM] Chemcial Reactions - Rusting	8	13
[MAT CHEM] Chemical Formulas	8	14
[MAT CHEM] Chemical Reaction - Example	9	27
[MAT CHEM] Compounds - H2O Example	8	14
[MAT CHEM] Compounds - Mass Conservation	4	1

[MAT CHEM] Conservation of Elements in Compounds	12	31
[MAT CHEM] Elements - Examples	3	1
[MAT CHEM] Elements - Ions	8	13
[MAT CHEM] Elements - Properties	7	7
[MAT CHEM] Endothermic Reactions	10	44
[MAT CHEM] Exothermic Reactions	10	21
[MAT CHEM] Organic Molecules - Properties	12	27
[MAT CHEM] Periodic Tables - Properties	16	79
[MAT CHEM] Protons Dictate Element	11	23
[MAT CHEM] Valence Electrons	8	21
[MAT CHEM] World Knowledge - Metals	7	8
[MAT COS] Change of State (Generic)	38	76
[MAT COS] Condensing - Example	17	46
[MAT COS] Condensing	15	35
[MAT COS] Evaporation - From Sunlight	15	43
[MAT COS] Evaporation - Kinetic Energy	11	18
[MAT COS] Evaporation - Over a Stove	14	33
[MAT COS] Evaporation	14	45
[MAT COS] Freezing - Change in Shape	15	61
[MAT COS] Freezing - Example	13	23
[MAT COS] Freezing and Melting - Molecular State	16	22
[MAT COS] Freezing	13	55
[MAT COS] Melting - Application	12	14
[MAT COS] Melting - Sunlight	11	19
[MAT COS] Melting Point Comparison	11	9
[MAT COS] Melting	14	35
[MAT COS] MeltingBoilingFreezing Point Independant of Mass and Volume	9	3
[MAT COS] Phase Changes - Physical Changes	12	29
[MAT COS] Sublimation	15	43
[MAT COS] Temperature Changes in a Medium	9	79
[MAT COS] Temperature, COS, and Molecular Speed	20	45
[MAT COS] Temperatures in Different States	8	23
[MAT ENVEFF] Mass Constant - Thermal Changes	7	2
[MAT ENVEFF] Physcial Changes - Temperature Changes	11	36
[MAT ENVEFF] Temperature - Effect on Particles	13	62
[MAT FUND] Atomic Theory	9	6
[MAT FUND] Matter - Properties	8	2
[MAT MEAS] Measuring - Length	29	31
[MAT MEAS] Measuring - Mass	10	8
[MAT MEAS] Measuring - Microscopic Things	9	11
[MAT MEAS] Measuring - Speed	10	22
[MAT MEAS] Measuring - Temperature	10	9
[MAT MEAS] Measuring - Weight	12	3
[MAT MEAS] Observing - Small Things	9	13
[MAT MEAS] Shape and Volume	11	22
[MAT MEAS] Volume - From Length	13	40
[MAT MIXTURES] Air - Madeof	11	15
[MAT MIXTURES] Compounds - Combining Elements	12	18
[MAT MIXTURES] Definition - Mixtures	8	13
[MAT MIXTURES] Separation - Liquid-Liquid Mixture	9	15
[MAT MIXTURES] Separation - Magnetic-Nonmagnetic Materials	9	23
[MAT MIXTURES] Seperation - Liquid-Solid Mixture	7	15
[MAT MIXTURES] Seperation - Solid-Solid Mixture	9	5
[MAT MIXTURES] Solution	6	17
[MAT PROPMATERIAL] Conductivity - Metal	18	102
[MAT PROPMATERIAL] Definition - Hardness	12	8
[MAT PROPMATERIAL] Density - Comparison Result	11	19
[MAT PROPMATERIAL] Density - Intensive	17	27
[MAT PROPMATERIAL] Electrical Insulators	14	34
[MAT PROPMATERIAL] Magnetic Attraction	17	19
[MAT PROPMATERIAL] Thermal Insulators - Examples	16	34
[MAT PROPO] Changing Shape by Action	5	4
[MAT PROPO] Rough Things - Examples	11	19
[MAT PROPO] Shape - Can be felt	8	6
[MAT PROPO] States with Definite Volume	7	8

[MAT PROPO] States with Variable Shapes	8	19
[MAT PROPO] World Knowlege - Objects have Mass	4	5
[MAT STATES] Chemical Properties - Examples	7	10
[MAT STATES] States of Matter - Kinetic Energy of Particles	8	11
[MAT STATES] Water - Examples	4	5
[MAT STATES] World Knowledge - States water in Equator and Poles	12	79
[MAT STATES] World Knowledge - States water is found in	9	48
[OTHER ENGINEERING] Weights effect on Transportation	8	7
[OTHER HIST] Computers - Effect on Communication	7	10
[OTHER HIST] Droughts effect on Farming	10	23
[OTHER HIST] Germs effect on Food Storage	8	10
[OTHER HIST] Invention of Telescope	12	20
[OTHER HIST] The Theory of Gravity	6	0
[OTHER] Light Bulb - UsedFor	5	7
[OTHER] Motives of Transportation	9	19
[OTHER] Plant based Products - Examples	3	1
[SAF EQUIP] Lab Safe Equipment	10	7
[SAF PROC] Multiple Plugs in one outlet is dangerous	7	9
[SCI INFERENCE] Greenhouses	10	20
[SCI INFERENCE] Identifying Soils	10	7

Table 8: The full list of 344 inference patterns generated in this work, sorted by curriculum topic category (in square brackets). Nodes represents the number of nodes a given inference pattern, while constraints represents the number of edges providing lexical constraints between nodes.