Creating BLUE Formulations of the Refinement Phase Test Suite (RTS) Questions: Experience and Reflections

Peter Clark, Working Note 42
Last updated: Dec 7th, 2010

1 Introduction

Over the last few weeks we have created and debugged a set of more natural formulations of the Refinement Phase AP Test Questions, as part of the effort to improve the naturalness and variety of English which AURA is able to understand. The goal of this report is to summarize the experience of this activity, what the challenges were, what was achieved, and where the opportunities and bottlenecks lie for the future. In particular, the exercise has taken a considerable amount of time and effort, and this report offers a retrospective on where that effort was spent, what was achieved, where the difficulties remain, and ways to progress forward.

2 The Task

The original 128 RTS AP questions were originally posed to AURA in the somewhat stilted English of CPL (Computer-Processable Language), a controlled language. For example,

> **RTS-X-7C-BLUE-2:**
> The equatorial plate of the mitotic spindle is formed during _____.
> a. anaphase
> b. late prophase
> c. metaphase
> d. early prophase
> e. interphase
>
> **CPL:** There is a subevent of mitosis.
> The subevent has a collection of centromere.
> The site of the collection of centromere is a metaphase plate.
> What is the subevent?

The task was to create a more natural formulation of the question (which we will refer to as the "BLUE formulation") within the guidelines of the more expressive BLUE interpreter, and then check AURA could answer it, and if not then diagnose and fix whatever the problem was.

The task itself is ambiguous: Is the task to create a more fluent version of the CPL, or a BLUE-compatible encoding of the original AP question? These two approaches will produce different formulations, e.g.,:

1. A more fluent version of the CPL. In this case, working from the above CPL, we might get:

   BLUE: During which step of mitosis do centromeres collect at the metaphase plate?

2. A BLUE-compatible encoding of the original question. In this case, working from the original AP question, we might get:

   BLUE: When is the equatorial plate of the mitotic spindle formed?

These two extremes are quite different: For (a), as the CPL already is known to work, the (bulk of the) knowledge for this version of the question is demonstrably in the KB. Thus any BLUE failures are likely to be due to the language processor itself. (b), however, introduces new concepts and might require new knowledge not encoded in the KB, thus broadening the range of failures that might occur. In this exercise the BLUE formulations were generally closer to (b) than (a), in an attempt to simulate what a BLUE-

fluent user might enter without knowledge of the KB. Thus were were attempting the "high road", creating BLUE formulations that were more faithful, but required more knowledge engineering also.

Prior to this exercise, the 128 AP questions had been originally posed as 237 CPL questions (typically a multiple choice AP question may break down into more than one CPL question), and the KB had been engineered to make sure the CPL questions answered correctly (i.e., the CPL questions constituted unit tests of the KB). Of these 237, 30 remained not working for various reasons (e.g., out of scope). The remain 207 working CPL questions are of interest here. Each of the 207 working CPL questions was paired with a BLUE version, approximately its equivalent[1], and then tested, and then failures in the BLUE formulation diagnosed and fixed where possible. The initial and final[2] performance figures of the BLUE formulations were are as follows:

Original (before extensions): 81/207 = 39%
Final (after extensions): 160/207 = 77%

The original figure can be viewed, very approximately, as AURA's September performance for a naive end-user on a heavily engineered part of the KB. There are two salient questions here:

1. What problems were there with the original BLUE questions?
2. What problems remained with the final BLUE questions?

The answers to these are quite significant, as they will indicate where future effort is most needed. In addition, it is useful to examine whether we have fixed AURA only for these specific questions or whether the improvements are more general: To the extent that the repairs fix systematic rather than question-specific problems, one would expect the next original score on a new question set to be higher.

3 Analysis I: Repaired Problems and QI Improvements

We identified five classes of fixable problems in the original 79 BLUE questions that did not answer and that were repaired, as listed below:

Category	# questions
Lex: Lexical knowledge was missing (i.e., relating words to concepts)	41
KB: Biology knowledge was missing	35
QI: The question interpreter software required improvements	16
CLib: General knowledge was missing from the component library	3
BadHiFi: The original BLUE formulation was ungrammatical	2

(Note that they add up to more than 79 as a few questions had more than one problem).

We now provide some brief illustrations of these categories.

3.1 Lex (Lexical knowledge was missing)

The majority of repairable failures were due to missing mappings between words and concepts. These are relatively easy to fix, although sometimes requires some thought as to which mappings are appropriate. In about half the cases the mappings were to general (CLib) concepts, in about half the cases they were to biology-specific concepts. For example:

RTS-X-25C: In mammalian cells, the first sign of prophase is the
b. separation of chromatids

[1] We originally made life hard for ourselves by first writing the BLUE, independent of the CPL, and then pairing it with the CPL, in an attempt to get "natural" BLUE formulations. A simpler approach would have been to write the BLUE direct from the CPL, although this makes it a less natural test of what a new user might enter given just the AP question alone.

[2] Including approximately 10 that are diagnosed and the repair verified, but awaiting a final KB update.

CPL [for option b]: Is it true that the prophase has detach of chromatids?

BLUE [for option b]: Do chromatids separate during prophase?

In this case, although AURA already knew "separate" might mean Move-Apart or Divide, it did not know this third sense of "separate", namely Detach. Thus a mapping from "separate" to this third sense was added[3].

A second example is the use of the word "substance" in

RTS-X-9E: Which of the following substances is found in RNA molecules but not in DNA molecules?
a. adenine
b. phosphate group
c. thymine
d. deoxyribose
e. ribose

BLUE: What substances are found in DNA molecules?

This BLUE formulation fails to find any substances because the parts of a DNA molecule (adenine etc.) are represented as molecular entities (Chemical-Entity), not substances (Substance). What is happening here is the word "substance" is being used in a looser way to refer to chemical "things" in general. The failure was repaired by adding "substance" as a word for Chemical-Entity, giving it a second meaning.

In other cases a word was simply not recognized, and had to be mapped to a CLib or Biology concept. Some examples (in italics) in the BLUE formulations, taken from the original AP questions, are:

- **RTS-X-11C:** Are the following enzymes *involved in* the replication of DNA...
- **RTS-X-15B:** spindle fibers *form* during prophase...
- **RTS-X-17C:** *Complementary* bases attach to each DNA strand...
- **RTS-X-56D:** ...the cell membrane is *composed of*...
- **RTS-X-95B:** ...the complex that *makes up* eukaryotic chromosomes

There were also occasional errors found and corrected in the existing word-to-concept mappings. In one case a problem was due to a bad WordNet synset: "cell wall" and "plasma membrane" are declared as synonyms in WordNet, although they are distinct concepts in the Biology KB. In a few other cases, some of the original synsets attached to CLib concepts caused problems. For example, CLib has WordNet synset command#v#2 connected to the concept of Command. However, this synset includes the word "require", allowing "A mitochondrion requires oxygen" to be undesirably interpreted as "A mitochondrion commands oxygen", a rather unusual reading (which AURA will select if no other reading is possible), and which should be dispreferred in the Biology domain.

3.2 KB (Biology knowledge was missing)

In many cases the KB was missing the required knowledge, primarily because the BLUE formulation touched slightly different knowledge to the CPL. In some cases this was because the CPL was significantly unfaithful to the original question, in other cases it was because of a more subtle shift in knowledge required. For example:

RTS-X-17C: In DNA replication, which of the following does NOT occur?
a. Helicase unwinds the double helix.

[3] The syntax for this is: (Detach has (wn20-synset ((:triple "separate" 0 "v"))))

CPL [for option a]: Is it true that helicase is the agent of detach of the dna-strands?

BLUE [for option a]: Does helicase unwind the double helix in DNA replication?

In this case, both the AP question and BLUE refer to the concept of a "double helix". However, this concept is absent in the KB, and needs to be added. Note that this is more than simply entering a symbol for Double-Helix, the KB also needs to represent its relation to DNA strands. In addition "unwind" is not adequately represented (it is only a synonym for Break-Contact). The KB needs to be extended to include the fact that unwinding a double helix results in separating its two strands. This type of missing knowledge was quite common in our work.

3.3 QI Improvements

There were some failures due to limitations in AURA's NLP engine, resulting in some extensions. These extensions are encouraging as they address systematic rather than question-specific problems, and thus are likely to be useful for questions in the future also. Three examples are:

3.3.1 Prepositional phrase (PP) attachment

In a small number of cases there were PP attachment mistakes made by the parser. For example:

RTS-X-65D: Cytokinesis is the portion of the cell cycle during which
d. a cell plate is formed in plant cells

BLUE [for answer d]: Is a cell plate formed during cytokinesis in plant cells?

The original and BLUE sentences are ambiguous, and can be read as:

(1) "formed during [cytokinesis in plant cells]"
(2) "formed [during cytokinesis] [in a plant cell]" (i.e., "[formed in a plant cell] during cytokinesis").

(1) is the intended reading, however the default attachment preference (namely, attach to the main verb) resulted in (2) being selected. Study of the question suite revealed that, in general, "in" prepositional phrases should attach to event nominalizations (e.g., "cytokinesis") in preference to the main verb, and this specific bias was added to BLUE.

3.3.2 Selectional Constraints

QI was sometimes selecting semantic relations where the relation's arguments violated its domain/range constraints. This was clearly undesirable and corrected.

3.3.3 Allowing Presuppositions in the Premise

Some questions include details not strictly necessary to answer the question, for example:

RTS-X-16A: BLUE: What happens during prophase I of meiosis?

AURA knows what happens during prophase I, but not that prophase I is (always) during meiosis. However, the question is not *asking* about this second fact, rather is presupposing it is true, and thus AURA needs to assume it, not prove it. QI was not accomodating such presuppositions, but was extended to do so.

3.4 CLib (General knowledge was missing from the component library)

(This category excludes missing word-to-CLib-concept mappings). An example of missing CLib knowledge is as follows:

RTS-X-33D: The nucleolus functions in the production of
a. Golgi apparatus
b. Microtubules
c. Mitochondria

d. Ribosomes
e. Endoplasmic reticulum

CPL: The nucleolus is site of synthesis. What is the result of the synthesis?

BLUE [for option d]: Does the nucleolus function in production of ribosomes?

In this case the KB (and CPL) talk about synthesis, but the AP question (and BLUE) talk about production. The BLUE originally failed because synthesis was not known to be a production event. Adding this fact to CLib[4] enabled the BLUE question to be answered.

In a few other cases, additional domain-general concepts needed to be added, e.g., Debris.

3.5 BadHiFi (The original BLUE formulation was ungrammatical)

There were 2 examples which originally failed due ungrammatical or nonsensical BLUE formulations:

RTS-X-82A-BLUE-2: Is smooth ER site of lipid synthesis?
(should be: Is the smooth ER the site of lipid synthesis?)

RTS-X-84C-BLUE: What is the site where mitotic cell division initiates?
(should be: What is the site where mitotic cell division is initiated?)

The formulations were subsequently corrected, but we note them as additional causes of original failure.

4 Analysis II: Remaining Problems

After making these extensions, 47 questions still did not answer. There were four primary categories of failure:

Category	# questions
Scope: The BLUE question is outside AURA's current scope	25
QI: Unresolved limitations of the current question interpretation module	10
BadQn: The original AP question was not coherent (according to the biologists)	5
QA: Unresolved limitations of the current question answering module	1
(Still to categorize):	10

Again we describe and illustrate these.

4.1 Scope (The BLUE question is outside AURA's current scope)

One reason for residual failures was scope, where the original question, hence the BLUE formulation, posed a question which AURA is known to be unable to answer. In these cases, the original CPL formulations succeeded largely because they were unfaithful to the original question. Although there were 25 failures, they were largely repetition of a smaller number of phenomena. Three salient examples are as follows:

(1) **RTS-X-9E:** Which of the following substances is found in RNA molecules but not in DNA molecules?
a. adenine…

BLUE [for part a]: Is adenine found in DNA molecule?

AURA is unable to deductively prove adenine is (always) part of a DNA molecule; the biologists tell us this is correct, that it's technically possible that there is no adenine there. Thus the original AP question should be viewed as an "Is it possible that...?" question rather than an "Is it true that...?" question. AURA is currently unable to answer "Is it possible that...?" questions.

[4] Syntax is: (Synthesis has (superclasses (Produce)))

This specific failure accounted for 15 of the 25 scope problems, caused by two AP questions each being broken up into multiple parts (one per molecule being queried about).

(2) **RTS-X-10C:** If a messenger RNA (mRNA) codon has the sequence UAC, which of the following would be the complementary anticodon triplet in the transfer RNA (tRNA) molecule?

a. ATG
b. AUC
c. AUG
d. ATT
e. ATC

AURA is currently unable to reason about base sequences.

(3) **RTS-X-11C:** All of the following enzymes are involved in the replication of DNA molecules except:

a. DNA helicase
b. DNA polymerase
c. RNA polymerase
d. RNA primase
e. DNA ligase

CPL: There is a DNA replication. What are the agents of the dna replication?

BLUE [for part b]: What is the role of DNA polymerase in DNA replication?

(This BLUE formulation happens to be rather poor as "role" is not mentioned in either the original or CPL questions). In this case the answer to the BLUE formulation would be a slot ("agent"), currently out of scope of AURA's QA.

4.2 QI (Unresolved limitations of the QI interpreter)

Some specific QI failures were due to more complex challenges with language processing. The two main categories of these were:

4.2.1 Special phrases

RTS-X-15B: Regarding mitosis and meiosis, one difference between the two forms of cellular reproduction is that in meiosis

a. there is one round of cell division, whereas in mitosis there are two rounds of cell division

BLUE [for part a]: Does the cell division happen 2 times in meiosis?

Here, QI is unable to translate "2 times" into a number count; this would require some special machinery for processing this phrase that has not yet been implemented. There is an easy rephrasing of this which is understood by AURA (“How many cell divisions occur in meiosis?”), but the original could not be processed.

4.2.2 Phrasal variants

There were two cases where an apparently redundant word occurred in the English that was not represented in the KB. Rather than represent the word, it was decided it would be better simply for QI to ignore it, i.e., performing a phrasal substitution. The two cases were:

- "fix errors" → "fix"
- "repair injuries" → "repair"

These occurred in:

RTS-X-67E-2: Which of the following functions can be attributed to DNA polymerase?

b. It fixes errors in the replication of DNA.

BLUE: Does DNA polymerase fix errors in DNA replication?

and

RTS-X-28C-2: Which of the following best describes the cells that result from the process of meiosis in mammals?

b. They can be used to repair injuries.

BLUE: Meiosis produces cells in mammals. Do the cells repair injuries?

QI does not currently support this kind of phrasal substitution.

4.3 BadQn (The original AP question was not coherent)

There were a few cases where the biologists disagreed with the AP question + answer itself:

(1) **RTS-X-15B-5:** Regarding mitosis and meiosis, one difference between the two forms of cellular reproduction is that in meiosis

c. chromosomes are replicated during interphase whereas in mitosis chromosomes are replicated during prophase

CPL [for part c]: Is it true that the interphase has a copy of chromosome?

BLUE [for part c]: During which phase are the chromosomes duplicated in mitosis?

Strictly, interphase is not a step of meiosis or mitosis, rather is a preceding step. Thus, the question is apparently based on a false presupposition. As the BLUE formulation explicitly asks only about phases of mitosis, it (correctly) fails to find interphase. We thus labeled this question as malformed; an alternative diagnosis would be to say "mitosis" could also be a synonym for "mitotic cell cycle", the macro-process which includes both interphase and mitosis.

(2) **RTS-X-58C:** To which part of the DNA strand does RNA polymerase attach in the synthesis of proteins?

a. the primer
b. the operator
c. the promoter
d. the regulator
e. the elongation site

According to the biologists, RNA polymerase is not involved in protein synthesis.

(3) **RTS-X-6C:** During which stage of mitosis is the nuclear membrane broken into fragments?

a. metaphase
b. anaphase
c. late prophase
d. early prophase
e. telophase

According to the biologists, the correct answer (Prometaphase) is not included as an option.

(4) **RTS-X-7C:** The equatorial plate of the mitotic spindle is formed during _____.

a. anaphase
b. late prophase
c. metaphase
d. early prophase
e. interphase

BLUE: An equatorial plate of the mitotic spindle is formed. During which phase of mitosis does the formation happen?

This question has somewhat loose English in it: The equatorial plate is an imaginary line, not a physical object, and so it is somewhat strange to talk about an imaginary line being "formed". Biologically what happens is that centromeres collect together in a line (the equatorial plate) across the cell. The KB represents this collecting, but the implication that an imaginary line is thus "formed" as a result is not encoded in the KB, and is a somewhat loose description of what is physically happening.

4.4 QA: Unresolved limitations of the current reasoner

There was one case where bounds placed on KM's reasoning resulted in QA failure. This was:

RTS-X-5A: In cells, which of the following can catalyze reactions involving hydrogen peroxide, provide cellular energy, and make proteins, in that order?

a. Peroxisomes, mitochondria, and ribosomes
b. Peroxisomes, mitochondria, and lysosomes
c. Peroxisomes, mitochondria,and Golgi apparatus
d. Lysosomes, chloroplasts, and ribosomes
e. Smooth endoplasmic reticulum, mitochondria, and ribosomes

BLUE: What catalyzes reactions involving hydrogen peroxide? [Answer: Peroxisome]

QI correctly interprets this as a "classes with properties" question ("For what class do all its members catalyze reactions involving hydrogen peroxide?"), a question type known to be computationally expensive to answer. For this type of question, QA only performs a lookup operation (scanning the KM prototypes), however in this case this is not adequate, as the peroxisome is not directly known to catalyze reactions; rather, a part of the peroxisome (a catalase) catalyzes reactions, and so reasoning would be required to infer that therefore the peroxisome as a whole can be considered to catalyze reactions. This was the only case we saw where reasoning bounds prevented an answer being found.

5. Other Issues

5.1 Answer Fidelity and Precision

A few cases arose where AURA produced a deductively correct answer, but sometimes additional or alternative answers to the ones expected:

(1) **RTS-X-16A:** Crossing over occurs during which of the following phases in meiosis?

a. prophase I
b. metaphase I
c. anaphase I
d. prophase II
e. metaphase II

BLUE: During which phase of meiosis does the crossing over occur?

Here AURA (correctly) produces two answers, namely Meiosis-I and Prophase-I, although the user is really only interested in the Prophase-I answer.

(2) **RTS-X-84C:** Mitotic cell division is initiated in the _____.

a. centromere
b. centriole
c. nucleus
d. mitotic spindle
e. DNA

BLUE: Where does the first step of mitotic cell division occur?

AURA (correctly) produces three answers: Eukaryotic-cell, Nucleus, and Semiautonomous-Organelle, although the user really only is interested in the nucleus answer.

(3) **RTS-X-9E:** Which of the following substances is found in RNA molecules but not in DNA molecules?

a. adenine
b. phosphate group
c. thymine
d. deoxyribose
e. ribose

BLUE: What substances are found in DNA molecules?

As a design decision, AURA just lists the top-level parts (e.g., doesn't descend to atom, electron, etc.), in this case (correctly) answering: Carbon DNA-strand Functional-Group Hydrogen Nucleotide. These are correct, but not at sufficient partonomic depth to answer the original question.

(4) **RTS-X-33D:** The nucleolus functions in the production of

a. Golgi apparatus
b. Microtubules
c. Mitochondria
d. Ribosomes
e. Endoplasmic reticulum

BLUE (original): What produces ribosomes?

AURA's answer is Enzyme (as all Synthesis events has agent Enzyme). This is correct, but not specific enough to match the answers in the question options.

BLUE (revised): What does the nucleolus produce?

AURA's answer is Free-Energy (a result of all Synthesis operations), found by interpreting the nucleolus as the agent of the synthesis operations. AURA thus exits with this interpretation, although again it does not mention an answer in the question options (ribosome).

BLUE (re-revised): What organelle does the nucleolus produce?

Here the preferred interpretation (nucleolus as agent) no longer produces an answer (Free-Energy is not an Organelle), so QI looks for alternatives and finds a less preferred interpretation (nucleolus as site) which finally does produce an answer in the question options (ribosome).

Of course, the question can be rephrased to explicitly test for a given answer ("Does the nucleolus produce produce ribosomes?" → "Yes"). However, without this rephrasing, the answers may be misleading.

6 Discussion

6.1 Summary

Creating working, fluent BLUE formulations for the RTS questions has turned out to be more time-consuming than originally anticipated. Part of the reason for this has been developing a suitable process, and there were some inefficiencies at the start:

- We developed the BLUE independently of the CPL, making the task more difficult than it could have been (although also making the BLUE more natural). In particular, it meant there was a higher risk of hitting knowledge gaps, as there was no guarantee that the BLUE required the same knowledge used

to answer the CPL formulations. It also meant extra work in creating a post-hoc alignment of the BLUE and CPL formulations.
- It took a few iterations to develop suitable guidelines and a methodology for debugging and tracking progress efficiently.

However, although future iterations of this process will be faster, it is nevertheless a slow process, with each failed question taking approximately an hour on average to author, test, document, debug, correct, sometimes discuss, retest, and categorize. We now offer some reflections on this process.

First, we have made significant process. In particular, the improvements to the interpretation algorithm, and many of the extensions to CLib and the word-to-concept mappings, are general in nature and hence should help improve performance on new questions. In principle, we should expect a higher initial score if we were to repeat this exercise with a new area of the KB.

Despite this, there are still three major brittlenesses in the question interpretation and answering process that remain, and are likely to be a significant source of errors in future question-answering exercises:

(1) **word-to-concept mappings:** As we changed the question wording towards the original AP question, we invariably hit words or phrases that were not mapped to the KB. Although AURA's lexicon is reasonably complete, the mappings to concepts are not. The working vocabulary used in the text book is approximately 20,000 words, while the vocabulary understood by AURA is approximately 10,000 words. More specifically, there are about 3500 (different) words used in the text book that occur at least 5 times and are unknown to AURA. We have added maybe 400 extra word-to-concept mappings during this RTS exercise, and thus there are still a lot more that are needed. Non-understood words, or (more challengingly) words that are recognized but not in the sense that the user intended, were encountered in almost every BLUE question in this exercise.

(2) **Missing KB knowledge**: A second major brittleness was knowledge missing in the KB, and again questions frequently failed because of this. Typically the omission was small, but nevertheless required the KEs to return and fill in the gaps. Our initial conjecture was that the working CPL validated that the KB had the required knowledge. However, except when the BLUE formulation was simply a more fluent, similarly worded version of the CPL, it often touched slightly different areas of knowledge. For example, in RTS-X-128A the CPL asked about the steps of a process (which were encoded in the KB), while the BLUE asks about the ordering of those steps (which was not), requiring backfilling the missing knowledge. Again this experience was common.

(3) **QI Brittleness:** The question interpreter defers commitment on some interpretation choices, in particular word senses and semantic roles, which turned out to be very advantageous. However, there are a number of other choices it makes eagerly, in particular:

- which parse to use (just the single best parse is used)
- whether to interpret a noun/verb as a concept or a relation (slot). While we can tune the interpreter to try to make good, eager choices, it remains a source of brittleness for processing new language. Ultimately we would like alternatives to these eager choices to be explored for more robust language interpretation.

Some of the existing mechanisms in AURA have turned out to be particularly useful, and some have not:

- The ~500 **"transfers thru" rules** in CLIb get used extensively when answering questions. (For example, one rule is: if a part of X does something, then X does that something). These turned out to be very helpful in bridging many granularity differences between the question formulation and the knowledge base.
- QI's **deferred commitment strategy** was very helpful, in particular for the (many) cases of subtly different word senses (e.g., whether "separate" means Divide, Detach, or Move-Apart). The KB is

able to nicely guide such disambiguations, a task that would be almost impossible using knowledge-poor sense disambiguation methods.

- QI's **paraphrase mechanism** was not particularly useful for the RTS questions (it was rarely invoked). While some paraphrases encode "slang" ways of expressing a fact, the majority encode world knowledge that can be (and was) more efficiently placed directly in the KB and/or word-to-concept mappings. For example, the paraphrase rule "IF X makes Y THEN X creates Y" can more directly captured simply by adding "make" as a word for the concept Create. Adding this knowledge to the KB both enriches AURA's knowledge and makes the paraphrase superfluous. This occurred several times. Thus, one might view paraphases as a fallback to more noisy knowledge when AURA's explicit knowledge does not find a non-null answer to a question.
- If the KB can answer the question, then it helpfully guides QI. However, if the KB cannot, then QI is currently left stranded and can produce a poor interpretation, making debugging harder. It'd be nice to also exploit the KB to better interpret questions which can't be answered.

6.2 Where the Debugging Time was Spent

Despite having now developed a reasonable methodology for authoring BLUE formulations, the process was still slow. The slowness was rarely due to language issues, but due to need to understand the biology representations being queried, ensure that they were complete, and understand how and where the language and knowledge should connect. In short, the task is not just "authoring questions", but developing the biology and language knowledge needed to support them. Compounding this is the fact that correcting/extending the word-to-concept mappings and the KB required a process in itself of writing up the problem for the biologists, waiting for them to make the change, downloading the updated KB, and retesting, adding latency to the task.

It might seem surprising that the process can be so slow, given that the CPL version of the question is working. In general adding word-to-concept mappings is relatively quick and easy (although currently requires a 1-day turnaround), but there are more severe challenges that can take substantially more time to resolve. There were six main reasons that BLUE questions took time to make operational:

1. Infidelity

If the CPL is too unfaithful, a completely new BLUE question may be needed, requiring extending the KB to support that new question, e.g.,:

> **RTS-X-77C:** Regarding mitosis and cytokinesis, one difference between higher plants and animals is that in plants:
> c. a cell plate begins to form at telophase, whereas in animals a cleavage furrow is initiated at that stage.
>
> **CPL:** What is the difference between the cytokinesis in animal cell and the cytokinesis in plant cell?
>
> **BLUE:** Does a cell plate begin to form in telophase in a plant cell?

Here a different part of the KB is required to answer the BLUE formulation, which (if not already present) needs to be developed and mapped to the words in the BLUE.

2. Lexical Debugging

> **RTS-X-88B:** During which stage of mitosis do chromatids separate to form two sets of daughter chromosomes?
>
> **CPL:** An event has detach of chromatid.
> The result of the detach is chromosome.
> The event is subevent of mitosis. What is the event?

BLUE: During which stage of mitosis do the chromatids separate into chromosomes?

In this example, we have to ensure a possible meaning of "separate" is the same thing that "detach" refers to, or one of its superclasses. To find what "detach" refers to, we have to first identify where the answer is in the KB (here, in Anaphase), and then the concept used in that answer (here, Detach), and then ensure "separate" also maps to that concept or one of its superclasses.

3. Generalizing a Specific Concept Name in the CPL

Sometimes the CPL names a very specific concept, implausible for a non-expert to enter, e.g.,:

RTS-X-83: Which of the following best corresponds to a site at which rRNA is formed?

CPL: There is synthesis-of-RRNA-in-Eukaryotes.
The result of the synthesis-of-RRNA-in-Eukaryotes is RRNA.
What is the site of the synthesis-of-RRNA-in-Eukaryotes? [Nucleolus]

BLUE: What is the site of rRNA formation?

To have the BLUE answer, the general reference "rRNA formation" (in the BLUE) needs to refer either the same concept as the CPL (synthesis-of-RRNA-in-Eukaryotes) or a similar one (e.g., synthesis-of-rRNA) that can answer the question. This might require checking if "rRNA formation" can be identified as synthesis-of-rRNA, and if not adding a trigger ("synthesis-of-rRNA ***isa*** Synthesis ***of*** rRNA"), or a synonym, and then making sure the wording works (e.g., "formation" maps to Synthesis or a superclass).

4. Missing General Knowledge

Sometimes a failure can be tracked down to missing CLib knowledge. For example, a working CPL question used the phrase "synthesizes proteins", while the non-working BLUE stated "produces proteins". After some debugging, and confirming "produce" mapped to Produce, the failure was tracked down to Synthesis not having Produce as a superclass.

In a more complex example (RTS-X-7), the working CPL asked "Do centromeres ***come together at*** the metaphase plate?", while the original AP and BLUE asked "Is the metaphase plate ***formed***?". Some analysis was needed to work out what exactly the original question was asking, and what general knowledge was needed to bridge from the KB/CPL to the original/BLUE: In this case, if things gather in a line (the "metaphase plate") then in a sense they create ("form") that line.

5. Recognizing out-of-scope questions

RTS-X-9: Which of the following substances is found in RNA molecules but not in DNA molecules? a. adenine...

BLUE: Is adenine found in a DNA molecule?

In this case (mentioned earlier also), the apparently simple BLUE failed (and the CPL was too unfaithful to test the KB). After some debugging we found the KB could not prove adenine was *always* in DNA, and the biologists confirmed this was correct (some very short DNA strands may have no adenine). Thus the question is really asking "Is it possible that adenine is in DNA?" rather than "Is it (always) true that adenine is in DNA?", a realization that took some time to reach.

6. Subtle mismatches between language and knowledge

On this same question, RTS-X-9 asks about "substances" but it turns out Adenine is represented in the KB as a Molecule, which in a strict sense is different from a Substance, leading to QA failure. This subtle discontinuity can be hard to find, and also takes some thought about how to fix, e.g., "substance" can also refer to molecules? "adenine" can also refer to some Substance made up of Adenine molecules?

In another subtle example, this BLUE question failed:

KB: Chromosomes are **copied** in interphase.

BLUE: Are chromosomes **duplicated** in interphase?

Copy and Duplicate are related taxonomically, and the word-to-concept mappings are there, so it would seem the BLUE should work. However, in the KB, Duplicate is a subclass of Copy: Thus, the fact chromosomes are ***copied*** (KB) doesn't logically imply they are ***duplicated*** (the question); e.g., they may have been copied in a different way other than duplication. However, clearly here these two words are being used synonymously in the question. The fix is to also allow "duplicate" as a synonym for Copy, but some debugging and analysis was required to find this.

A third example is:

RTS-X-95B: The **DNA-protein complex** that makes up eukaryotic chromosomes is b. Chromatin

BLUE: Is chromatin a **protein complex** which makes up chromosomes?

Here, the subtle shift from "DNA-protein complex" (original) to "protein complex" (BLUE) changes the meaning, as chromatin is considered to be a DNA-protein complex (a complex made up of DNA and proteins) but not a protein complex. Thus this subtle rewording has changed the polarity of the answer.

As a final example, another question (outside RTS) asks:

KB: ATP is used in the active transport of a transport protein.

BLUE: Which **source of energy** is used in the active transport of a transport protein?

Although the answer seems clearly ATP, the KB does not represent that ATP is an Energy-Source (in fact the concept of Energy-Source does not exist in the KB), and hence AURA cannot answer this question (it cannot prove that ATP is a source of energy). This example suggests an opportunity to provide a conditional answer to the user ("Yes, assuming ATP is an energy source").

6.3 Conclusions and Recommendations

The RTS BLUE exercise has substantially improved AURA's question interpretation facility, but has also been time-consuming and brittlenesses in QI still remain. Our experience suggests several ways that both the methodology and the language engine itself could be improved. Concerning the methodology:

(a) **Better quality "LoFi" CPL suites:**
Ensure that the "low fidelity" CPL truely tests that the knowledge required for the question (not some simplification of it) really exists in the KB. This may involve authoring several CPL questions, and further developing the KB to support them. While "slot language" English is acceptable, other aspects of the CPL caused problems. Some useful guidelines for future CPL formulations are:

- Ensure that the important words in the original question are used in the CPL (add a word-to-concept mapping, and/or maybe a new concept if not).
- Avoid including concept names explicitly in the CPL. Use a word or phrase instead.
- Avoid using artificially long concept names, or the names expressed as a phrase, in the CPL. Instead add synonyms or triggers to the KB if necessary to have a complex, specific concept recognized from the phrasal components in the CPL.
- If necessary, use several rather than just one CPL formulation to ensure all aspects of the original AP question is tested.
- Avoid reducing complex questions to generic questions (“What is an X?”, “What is the relationship between X and Y?”, etc.), instead try and be more faithful to the original question.

(b) More disciplined BLUE authoring:

The BLUE formulations should be based on the CPL, but as close to the original English as possible. In particular the BLUE formulations should not introduce yet more vocabulary not present in the CPL nor the original English.

(c) Systematic lexical and knowledge entry:
There is a strong case for systematic vocabulary entry rather than by a test-and-debug process: Through question debugging, it might take an hour to discover a single word-to-concept mapping is missing; in contrast, someone might step through all the concepts one by one and brainstorm words for each, doing maybe a hundred or so in an hour. Through this RTS exercise we added approximately 400 word-to-concept mappings over two months; in contrast, there are approximately 3500 words in the textbook that occur more than 5 terms and are unrecognized by AURA. Crowdsourcing/Amazon Mechanical Turk might be a resource that could be exploited for enumerating the vocabulary for known concepts. Similarly, systematically ensuring that the required concepts themselves are in the KB, and adequately representing them, is a major task and may be performed more efficiently with a systematic approach rather than stumbling on omissions one by one in debugging mode.

(d) Faster turnaround / Have biologists author BLUE:
The 1-day turnaround between the "language engineers" and biologists is time-consuming and a large source of inertia. It would perhaps be more efficient for biologists themselves to make a first pass at the BLUE formulations, and then pass just the difficult problems onto the "language engineers" to solve.

Concerning the language processing (QI) itself, there are also several obvious areas of improvements:

(a) More deferred commitment:
Although QI defers some commitments, it makes others eagerly, which in some cases can lead to irrecoverable failures. In particular, QI commits to a single parse, and eagerly commits to a decision whether a word denotes a concept or relation (slot). Deferring this commitment would remove some failures.

(b) Looser language interpretation:
QI is still brittle, and requires that the question be completely and fully provable from the KB in order to generate an answer. However, as we have described, sometimes knowledge is missing, or words are used in a loose or imprecise way, resulting in failures. Some looser/more error-tolerant way of handling language would be helpful. However, it should be noted that this is not a panacea: "Loose" language interpretation could itself lead to interpretation errors, and adding dialog with the user could be confusing/annoying instead of helpful.

(c) Better interpretation of failed questions:
QI currently does well on questions AURA can answer, but poorly on questions AURA cannot. This is because QI works by searching possible interpretations for one that answers: If it can find one, that usually is the intended interpretation and it scores well; however if it cannot, then it has nothing to guide it and so it makes a weak guess, knowing that whatever it guesses it won't be able to answer. In some ways, the weak guess does not matter (as all guesses will fail); but in other ways, the weak guess is problematic, as it makes debugging harder. Having QI do a better job on questions which AURA *can't* answer would help the debugging process.

These all suggest ways in which AURA's Question Interpretation could be more efficiently and systematically improved, and merit further discussion as we move forward.