Chapter 5: Experiments and Results

5.1 Outline

The first section of this chapter explains how the responses of the system were evaluated. Each of the next four sections describes an experiment: its aim, input data, system configuration, and results. Each section ends with a discussion of the results.

5.2 Evaluation of Answers

5.2.1 Evaluation Categories

In the first experiment, pattern instances were judged to be either Vital, Okay, or Wrong, following the categories used to assess answers in the definition question subtask of TREC-12. Generally, a Vital answer was an essential fact about the query term. An Okay answer had to be a general fact about the term, but one that would not suffice to explain the essence of the term. As mentioned, we used terms in the FishBase glossary in the second, third, and fourth experiments. This enabled evaluation against the definitions in the glossary. However, non-essential answers could sometimes be judged as Okay even when not constituting part of the FishBase definition.

In the second, third and fourth experiments, when we could not decide if an answer should be judged as being Okay or Wrong, we classified it in the new intermediate category Uncertain. For example, the system returned the following snippet in response to the query ‘abdominal cavity’: **`abdominal cavity`** `was opened and the kidney removed.` Obviously, this is not a proper definition, but it allows the reader to infer that the abdominal cavity contains a kidney. However, it is a fact that may be too specific to classify as Okay, yet it is not wrong and could be useful in some scenarios, so we judged it to be Uncertain. In rare cases, when we vacillated between Vital and Okay judgements, we classified the answer as being Okay. Typically, such an answer identified text which the authors of the paper intended to be a definition, but was insufficient when judged against a formal definition in FishBase. For example, when processing the term ‘diversity’, the system returned the answer number of taxa. The answer matched the pattern **TERM(DEF)** which is often used by writers for explicit definitions. However, the definition of ‘diversity’ in FishBase is ‘A parameter describing, in combination, the species richness and evenness of an assemblage of species. Diversity is often used as a synonym for species richness.’ Therefore, diversity is not simply the number of species but also relates to the composition of the species community, and the answer was judged as Okay. Table 5.1 shows an example of judgments in the first run.

Term	Judgement	Pattern	Text extract
Grilse	Vital	TERM (DEF	**fish maturing after 1.5 years in sea water**) were removed from the population, and 600 of the remaining fish were individually tagged with anchor T-bar tags beneath the dorsal fin (Floy (r) Tag, Seattle, WA)
	Okay	DEF (TERM)	Separating non-maturing and maturing fish before the **maturing fish**
	Wrong	TERM, DEF	and spawn opportunely
colour	Vital	TERM is DEF	**is a particularly important quality characteristic in salmonids, with uniform red colour being preferred by the consumer**
	Okay	TERM was DEF	**Also evaluated using a Roche Colour Card for Salmonids**
	Wrong	TERM (DEF	Lu et al., 1992

Table 5.1: Examples of answers to two queries judged as Vital, Okay or Wrong in the first experiment. The bold text is the relevant definition within the answer extract.

5.2.2 Recall and Precision

van Rijsbergen (1979) defines recall as ‘the proportion of relevant material actually retrieved in answer to a search request’ and precision as ‘the proportion of retrieved material that is actually relevant'. In this project, recall would have answered the question: What is the proportion of definitions in SOK-i that can be matched by our set of lexical patterns? Answering this question requires finding all the Vital and Okay definitions in SOK-i for each term. This is a difficult task because for frequent terms it would have involved evaluating thousands of sentences manually. For this reason we decided not to compute recall and concentrate on precision. Joho (1999) took a similar approach and focused on accuracy and distribution of patterns. Joho and Sanderson (2000) also suggested that users of a system retrieving descriptions of nouns are likely to prefer high precision to high recall, and that 'as long as a few are found relatively accurately, most will be satisfied'.

In the first and fourth experiments we calculated Average Precision of two types: Strict and Lenient. Average Strict Precision was the proportion of Vital pattern instances returned in each experiment; Average Lenient Precision was the proportion of Vital Okay and Uncertain pattern instances returned in each experiment (only Vital and Okay in the first experiment). In the first experiment we also computed average precision for the most common patterns.

In all the experiments we computed Average Strict Binary Responsiveness—the proportion of queries for which at least one Vital pattern instance was found—and Average Lenient Binary Responsiveness—the proportion of queries for which at least one Vital, Okay or Uncertain pattern instance was found.

Answers (pattern instances) were not penalised for exceeding any length allowance, because for many patterns the extraction of the definition text was coarse, merely identifying the region of a sentence containing the definition in relation to the query term.

5.2.3 Profile of the Intended User

When judging answers to definition queries, it is important to remember the target audience. The organisers of the TREC-12 question answering track suggested a scenario profiling the questioners, their goals and level of expertise (see Section 2.5.2). Similarly, in evaluating our system's answers we assumed that users were native speakers of

English, with an intermediate to high level of expertise and interested specifically in the fishery domain. For example, the users could be undergraduate biology students familiarising themselves with terms related to salmon fish. The users could also be professionals in the field who know the meaning of the terms they are searching for but want to survey their current definition in scientific literature perhaps to be consistent with their own writing. Another scenario could be of researchers for whom salmon biology is not the main area of expertise. In all these cases, the users are assumed to be sophisticated enough as to be capable of inference when reading definition snippets. We also assumed that the users would know they are interrogating a corpus on fish research.

5.3 First Experiment

5.3.1 Aim

To explore the effectiveness and distribution of an initial set of definition patterns and the environment of definitions.

5.3.2 Data

Forty-two terms suggested by salmon researchers were used (Table 3.2).

5.3.3 System Configuration

After observing the document collection, we expanded the initial list of definition patterns which were used in TREC-12. Table 5.2 lists the patterns used.

TERM is described as DEF

DEF—TERM--

TERM—DEF

TERM is the term for DEF

TERM is the term used to describe DEF

TERM is used to describe

TERM describes

TERM is/was/are/were defined as DEF

TERM is/was/are/were DEF

TERM, which is/was/were/are DEF

defines TERM as DEF

TERM defines DEF

Define TERM as DEF

TERM and other DEF

TERM or other DEF

TERM, and other

TERM, or other

TERM consists of DEF

TERM consist of DEF

TERM constitutes DEF

TERM constitute DEF

TERM comprise DEF

TERM comprises DEF

TERM include DEF

TERM includes DEF

DEF consists of TERM

DEF consist of TERM

DEF constitutes TERM

DEF constitute TERM

DEF comprise TERM

DEF comprises TERM

DEF include TERM

DEF includes TERM

TERM, a DEF

TERM, the DEF

TERM means DEF

TERM, or DEF

TERM (DEF

TERM refers to DEF

TERM WORD refers to DEF

TERM, i.e. DEF

TERM is WORD defined DEF

definition of TERM is DEF

definition of TERM as DEF

TERM: DEF

DEF, TERM

TERM, DEF

DEF called TERM

DEF called a TERM

DEF called an

DEF termed TERM

DEF such as TERM

DEF named as TERM

DEF named TERM

such DEF as TERM

DEF including TERM

DEF, including TERM

DEF especially TERM

DEF, especially TERM

DEF (TERM)

DEF known as TERM

Table 5.2: The set of definition (DEF) patterns used in the first experiment.

To discover which contextual characteristics correlate with the usefulness of sentences and documents we gathered the following statistics in the first experiment:

- the ordinal number of the instance of the term in the document (`doc_so_far_term_count`);
- the total number of instances of the term in the document (`doc_total_term_count`);
- the ratio of the above two numbers (`doc_so_far_term_proportion`);
- the length of the sentence containing a pattern (`doc_pattern_instance_sen_length`);
- the total number of sentences in the document (`doc_sentence_number`);
- the position of the sentence containing a pattern (`doc_so_far_sentence_number`);
- the relative sentence position (`doc_so_far_sentence_proportion`) calculated by dividing `doc_so_far_sentence_number` by `doc_sentence_number`.

5.3.4 Results

Three terms (‘gene duplicates’, ‘tetrasomic inheritance’, and ‘contemporary microevolution’) were not found in the document collection. An additional four terms had no answers. Table 5.3 presents the judgments and numbers of the answers to the remaining 35 term queries.

The number of pattern instances returned for each of the 35 queries ranged between 1 and 290 (mean=61, median=28). Vital instances were found for 13 terms. Okay instances were found for 23 terms. Eleven terms had only wrong instances. Appendix E lists a sample of Vital and Okay answers.

Term	Vital	Okay	Wrong	Term	Vital	Okay	Wrong
Isotypes	0	1	6	Atlantic	0	0	2
artificial photoperiod	1	0	0	quality	1	19	113
Chinook salmon	0	139	151	stunning	0	0	10
early maturation	0	2	2	colour	2	17	72
freshwater residence	0	0	1	filet	0	0	4
life history	0	2	11	astaxanthin	2	15	38
local adaptation	0	6	4	canthaxanthin	2	12	18
parr	1	17	164	phenotype	0	0	10
redd	2	3	13	otolith	1	3	24
alevin	0	4	0	watershed	0	0	37
smolt	2	6	58	hatchery	0	2	112
grilse	3	10	11	propagation	0	0	9
osmoregulation	1	4	14	stock	1	3	107
migration	0	11	106	aquaculture	0	6	116
Fry	0	21	185	DNA	2	1	208
wild	0	0	28	critical habitat	0	0	1
farmed	0	0	4	conservation	0	0	40
fat	0	9	119	**TOTAL**	21	313	1798

Table 5.3: Judgement of the answers to the query terms suggested by salmon researchers.

Table 5.4 shows the frequency and effectiveness of the different definition patterns.

Pattern	Total sentences	Percentage	Vital	Okay	Wrong	Number of queries with Vital answers
TERM is described as DEF	1	0.05%	0	1	0	0 (0%)
TERM—DEF	5	0.23%	0	0	5	0 (0%)
TERM is/was/are/were DEF	581	27.25%	8	78	495	7 (41.2%)
TERM, which is/was/were/are DEF	6	0.28%	0	1	5	0 (0%)
TERM and other DEF	3	0.14%	1	1	1	1 (5.9%)
TERM consists of DEF	1	0.05%	0	1	0	0 (0%)
TERM constitutes DEF	1	0.05%	0	1	0	0 (0%)
TERM constitute DEF	1	0.05%	0	0	1	0 (0%)
TERM include DEF	3	0.14%	1	0	2	1 (5.9%)
DEF include TERM	1	0.05%	0	0	1	0 (0%)
TERM, a DEF	11	0.52%	0	1	10	0 (0%)
TERM, the DEF	41	1.92%	0	3	38	0 (0%)
TERM, or DEF	5	0.23%	0	1	4	0 (0%)
TERM (DEF	603	28.28%	2	148	453	1 (5.9%)
TERM WORD refers to DEF	1	0.05%	0	0	1	0 (0%)
TERM, i.e. DEF	2	0.09%	0	1	1	0 (0%)
TERM is WORD defined DEF	1	0.05%	1	0	0	1 (5.9%)
definition of TERM as DEF	1	0.05%	0	1	0	0 (0%)
TERM: DEF	53	2.49%	1	1	51	1 (5.9%)
DEF, TERM,	65	3.05%	0	6	59	0 (0%)
TERM, DEF	704	33.02%	2	46	656	2 (11.7%)
DEF called TERM	1	0.05%	0	0	1	0 (0%)
DEF called a TERM	1	0.05%	0	1	0	0 (0%)
DEF termed TERM	1	0.05%	1	0	0	1 (5.9%)
DEF such as TERM	4	0.19%	0	3	1	0 (0%)
DEF including TERM	6	0.28%	0	4	2	0 (0%)
DEF especially TERM	2	0.09%	0	2	0	0 (0%)
DEF (TERM)	27	1.27%	4	12	11	2 (11.7%)
TOTAL	2132	100.00%	21	313	1798	17 (100%)

Table 5.4: Frequency and effectiveness of definition patterns in the first experiment.

Pattern	**Frequency** **(no. of specific pattern instances/ total no. of pattern instances)**	**Strict Precision**	**Lenient Precision**
TERM, DEF	33.02%	0.3%	6.8%
TERM (DEF	28.28%	0.3%	24.8%
TERM is/was/are/were DEF	27.25%	1.4%	14.8%

Table 5.5: Precision of the three most frequent patterns. Strict precision is based on the proportion of Vital answers, whereas Lenient Precision is based on Vital and Okay answers.

Average Strict Precision and Average Lenient Precision were 1.2% and 18.6% respectively. Lenient and Strict Precision were computed for the three patterns that together accounted for 88% of the answers (Table 5.5).

Average Strict Binary Responsiveness and Average Lenient Binary Responsiveness were 37.1% and 68.6% respectively.

To determine whether sentences that yielded answers in different judgement categories also differed significantly in the values of these features. We performed single factor analysis of variance (ANOVA) of the following:

- `doc_so_far_term_count,`
- `doc_total_term_count,`
- `doc_so_far_term_proportion,`
- `doc_sentence_number,`
- `doc_so_far_sentence_proportion,`
- `doc_pattern_instance_sen_length.`

Tables 5.6, 5.7, 5.8, 5.9, 5.10 and 5.11 present the results of each analysis.

SUMMARY						
Groups	**Count**	**Sum**	**Average**	**Variance**		
Okay	313	3310	10.6	234.3		
Vital	21	130	6.2	76.3		
Wrong	1785	28348	15.8	468.7		
ANOVA						
Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	8797.1	2	4398.5	10.2	3.85E-05	3
Within Groups	916911.6	2129	430.7			
Total	925708.7	2131				

Table 5.6: Analysis of variance of `doc_so_far_term_count` between answer judgement categories.

SUMMARY						
Groups	**Count**	**Sum**	**Average**	**Variance**		
Okay	313	7386	23.6	777		
Vital	21	475	22.6	982.9		
Wrong	1798	61683	34.3	1643.5		
ANOVA						
Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	32694.4	2	16347.2	10.8	2.1E-05	3
Within Groups	3215508	2129	1510.3			
Total	3248203	2131				

Table 5.7: Analysis of variance of `doc_total_term_count` between answer judgement categories.

SUMMARY						
Groups	**Count**	**Sum**	**Average**	**Variance**		
Okay	313	170.5	0.545	0.1		
Vital	21	10.9	0.518	0.1		
Wrong	1798	966.2	0.373	0.08		
ANOVA						
Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	0.023	2	0.01	0.1	0.9	3
Within Groups	184.44	2129	0.09			
Total	184.47	2131				

Table 5.8: Analysis of variance of doc_so_far_term_proportion between answer judgement categories.

SUMMARY						
Groups	**Count**	**Sum**	**Average**	**Variance**		
Okay	313	150929	482.2	99197.3		
Vital	21	10235	487.4	66721.5		
Wrong	1798	853472	474.7	84304		
ANOVA						
Source of Variation	SS	Df	MS	F	P-value	F crit
Between Groups	17878	2	8939.3	0.1	0.9	3
Within Groups	183778281	2129	86321.4			
Total	183796160	2131				

Table 5.9: Analysis of variance of `doc_sentence_number` between answer judgement categories.

SUMMARY						
Groups	**Count**	**Sum**	**Average**	**Variance**		
Okay	313	108	0.34	0.08		
Vital	21	4.6	0.24	0.02		
Wrong	1798	583	0.33	0.05		
ANOVA						
Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	0.2	2	0.11	2	0.14	3
Within Groups	120.4	2129	0.05			
Total	120.7	2131				

Table 5.10: Analysis of variance of `doc_so_far_sentence_proportion` between answer judgement categories.

SUMMARY						
Groups	**Count**	**Sum**	**Average**	**Variance**		
Okay	313	11961	38.2	420.1		
Vital	21	655	31.2	344.7		
Wrong	1798	70297	39.1	441.5		
ANOVA						
Source of Variation	SS	Df	MS	F	P-value	F crit
Between Groups	1465.2	2	732.6	1.7	0.2	3
Within Groups	931365.9	2129	437.5			
Total	932831.1	2131				

Table 5.11: Analysis of variance of `doc_pattern_instance_sen_length` between answer judgement categories

The analyses of variance showed that

- There was a statistically significant difference (P<0.01) between the three answer judgement categories in `doc_so_far_term_count`. Wrong pattern instances had the highest average term count (15.8), and Vital ones had the lowest average (6.2);
- There was a statistically significant difference (P<0.01) between the three answer judgement categories in `doc_total_term_count`. Wrong pattern instances had the highest average count (34.3), and Vital ones had the lowest average (22.6);
- There was no statistically significant difference between the three answer judgement categories in `doc_so_far_term_proportion`;
- There was no statistically significant difference between the three answer judgement categories in `doc_sentence_number`;
- There was no statistically significant difference between the three answer judgement categories in `doc_so_far_sentence_proportion`;
- There was no statistically significant difference between the three answer judgement categories in `doc_pattern_instance_sen_length`.

5.3.5 Discussion

For ten query terms we returned over one hundred answers. Most of these terms occur more than 1,000 times in the SOK-i collection. However, the abundance of answers helped us in studying the weaknesses of the dominant patterns. Average Binary Responsiveness was much higher than Average Precision. This was expected because when we calculated Binary Responsiveness, a response to a query was judged Vital or Okay as long as there was one Vital or Okay pattern instance, regardless of how many incorrect pattern instances were returned for that query. Our Average Precision, both Strict and Lenient, is considerably lower than the 35% precision reported by Joho (1999) who evaluated descriptive phrases in unranked sentences which matched lexical patterns. Joho's higher precision may be due to a more precise and smaller set of patterns compared to our initial set. For example, his appositive pattern was **`TERM, DEF, is/was/are/were`** whereas we used the pattern **`TERM, DEF`** in this experiment. Also his version of the copular pattern is **`TERM is a DEF`** instead of just **`TERM is DEF`**.

The three most frequent patterns accounted for about 90% of the wrong pattern instances. Therefore, in the subsequent experiments we concentrated on eliminating wrong answers that matched these patterns. We observed that often the pattern **`TERM (DEF`** falsely matched citations and references to figures and tables. These could easily be eliminated by creating a list of stopwords which are typical to bracketed references (e.g., 'et al.', 'fig.', years). Section 5.4.3 describes the implementation of the stopword approach in the following experiments.

The frequency and precision of individual patterns as reported by Joho (1999) differ somewhat from our findings. Since he tested his patterns on 460 Mb of LA Times news articles, the difference might be due to the style of the text. Although his most frequent pattern (30%) is appositive, as it is in our case, he reported higher frequencies for the pattern **`DEF such as TERM`** (20%), **`TERM and other DEF`** (16%), and **`DEF, including TERM`** (11%). On the other hand, we found a higher frequency of the pattern **`TERM (DEF`**. This could be the result of the false matching to references, which is discussed above. Joho did not distinguish between strict and lenient precision. However, he did suggest his evaluation was liberal, so we may conclude that the precision of this pattern in our corpus, even with many false matches, is not lower than the 21% reported in his work.

The statistically significant difference between the three judgement categories in `doc_so_far_term_count` was predictable, because a term is more likely to be defined on one of its first occurrences in the document. The ordinal number of the term when it is defined should not necessarily be 1, because it can appear previously in the title of the document, the keyword list, and in the abstract. For example, in one of the documents a Vital definition for 'smolt' (`In Atlantic salmon culture, a` **`smolt`** `is usually defined as a juvenile salmon that is able to survive and grow normally in sea water)` appeared in the third paragraph of the Introduction after 13 occurrences of 'smolt' in the document: Three times in the abstract, once in the Author Keywords list, and nine times in the first two paragraphs of the Introduction. The highest `doc_so_far_term_count` of any Vital answer was 37.

Blair-Goldensohn, McKeown and Schlaikjer (2003) and Joho and Sanderson (2000) observed that definitions are likely to be found nearer the beginning of the document than its end. They relied on relative and absolute sentence position as a feature indicating the presence of definitions. However, our results suggest that at least in this collection of scientific papers, sentence position (relative or absolute) is not a good indicator of text containing definitions. This might be the result of the structured organisation of scientific papers, where each section is more self contained than paragraphs are in news reports. We expected to find most of the definitions in the Introduction but other sections yielded many definitions. For example, early in this project we considered discarding the References section during the document pre-processing stage. To our surprise we later found definitions in the References section of documents (e.g., `Canthaxanthin: a pigmenter for salmonids`). However, definitions from different sections of the paper may differ in nature and style. For instance, definitions extracted from the Methods are more technical (e.g., saltwater was defined in salt concentration value of parts-per-million). It is worth exploring whether certain types of terms are more likely to be defined in particular sections. A similar approach was suggested by Shah at al. (2003) for extracting keywords from full-text papers in genetics.

Our results also showed that definitions are more likely to be found in documents with lower `doc_total_term_count`. Perhaps when a term is abundant in the document or is its main topic, the authors assume that the readers are familiar with the term and avoid defining it.

The ANOVA of the total number of sentences in a document suggests that abstracts are not more likely to contain definitions than full-text articles. Abstracts have the highest density of keywords (Shah et al., 2003), but keywords are not necessarily more likely to be defined, especially not in concise text such as of abstracts.

Some of the terms we tested were ambiguous. For example, when searching for a definition for ‘quality’, the system retrieved Okay snippets which related to the quality of smolt, egg, gamete, flesh, and water. Therefore, even in a domain-specific corpus terms may be defined in different contexts. Other terms, such as ‘fat’ and ‘fry’, may be nouns or verbs. ‘stunning’ could be a verb, a gerund or an adjective. But in the SOK-i collection usually one part-of–speech assignation was preponderant (noun in those examples). All the terms we tested in the first experiment were nouns. Therefore, subsequent improvement of the system tended to focus on extraction of definitions to nouns.

We searched for the term in its exact form without any stemming. There were a few terms (e.g., 'stock') which could have benefited from adding the plural form to the search query submitted to dtSearch, but we observed that most terms are defined in the singular form. Otherwise, stemming is likely to add 'noise'. For example, when searching for 'stock', we did not want to match 'stocked' or 'stocking'. We also did not submit alternative spelling of terms, although we observed that British spelling, such as 'colour' and 'behaviour', was more common in the collection than American.

The case insensitivity of the index caused no observable problem. We noticed only one common noun term ('fry') which matched a proper noun ('Fry', a name of an author).

Even though the papers in the SOK-i collection seem to target a homogenous audience, it is possible to find definitions which are suitable for different levels of expertise. For example, the system retrieved a chemical formula in response to the query 'astaxanthin'. Such an answer, although incomplete, could satisfy an expert in biochemistry. Another answer was `Astaxanthin is an approved colour additive in the feed of salmonids.` The first definition was found in a paper on the digestability and accumulation of astaxanthin published in a journal titled *Comparative Biochemistry and Physiology - Part B: Biochemistry & Molecular Biology* (Bjerkeng and Berge, 2002). The second definition was taken from a paper published in *Fisheries Research* which discusses 'potential issues for human health and safety from net-pen salmon farming in the Pacific northwest region' (Fairgrieve and Rust, 2003). The readers of the second paper may be experts on fish biology but not necessarily on chemicals, food safety or even salmon farming, whereas the first paper is more limited to a single discipline. However, in both cases, we can assume that the target readers would be well educated and familiar with the style of scientific literature.

Patterns such as **`DEF including TERM`** tend to capture hypernyms which might be breaking the Principle of Usefulness suggested by Sarner and Carberry (1988). For example, the definition of parr as a 'life cycle stage' may be too general for most expert users. The users would be likely to know that parr is a stage in the life cycle of the salmon, but would be interested in what Swartz (1997) termed a 'precising definition', which for example could specify the onset of the parr stage (dispersal from nest) and its termination (seaward migration). However, if we imagine a future step of combining the answers into a coherent, grammatical response such general hypernyms can make the response more readable.

While evaluating hundreds of answers and sometimes checking their context in the full text documents, we encountered new phrases used for definitions, which we incorporated in the next experiment.

5.4 Second Experiment

5.4.1 Aim

To test the system with considerably larger set of queries and applying observations from the first experiment

5.4.2 Data

3,920 terms were extracted from the FishBase glossary.

5.4.3 System Configuration

We expanded the list of definition patterns to the set in Table 5.12.

In this experiment and the subsequent two, the answer output was restricted to sentences with a `doc_so_far_term_count` value of up to two. This means that the occurrence of a term in a sentence which was retrieved as an answer was at most the third one within the document.

For the pattern **`TERM(DEF)`**, we specified stopwords that the text enclosed in round brackets should not include. In addition, we also eliminated text in brackets which included years (19XX, or 20XX) and author names (e.g., `Beckman and Beckman`). Two capitalised words with ‘and’ between them, or a capitalised word followed by a full stop identified an instance of one re more author names. In the second experiment we began extracting the text which matched **DEF** in **TERM(DEF)** up to the closing bracket and not up to the end of the sentence, as we did in the first experiment. The list of stopwords grew with each subsequent experiment (Table 5.13).

TERM is described as DEF	TERM, a/an/the DEF
DEF--TERM--	TERM means DEF
TERM—DEF	TERM, or DEF
TERM is the term for DEF	TERM (DEF)[a]
TERM is the term used to describe DEF	TERM refers to DEF
TERM is used to describe DEF	TERM WORD refers to DEF
TERM describes DEF	TERM, i.e. DEF
TERM is/was/are/were defined as DEF	TERM is WORD defined
TERM is DEF	definition of (a/an/the) TERM is DEF
TERM are DEF	TERM: DEF
TERM was DEF	DEF, TERM,
TERM were DEF	DEF, a/an/the TERM,[b]
TERM, which is/was/are/were DEF	TERM, DEF[b]
defines (a/an/the) TERM as DEF	DEF called (a/an/the) TERM
TERM defines DEF	DEF is termed (a/an/the) TERM
define (a/an/the) TERM as DEF	DEF such as (a/an/the) TERM
TERM (,) and other DEF	DEF named as (a/an/the) TERM
TERM (,) or other DEF	DEF named (a/an/the) TERM
TERM consist(s) of DEF	such DEF as (a/an/the) TERM
TERM constitute(s) DEF	DEF(,) including (a/an/the) TERM
TERM comprise(s) DEF	DEF(,) especially (a/an/the) TERM
TERM include(s) DEF	DEF ((a/an/the) TERM)
DEF consist(s) of (a/an/the) TERM	DEF known as (a/an/the) TERM
DEF constitute(s) (a/an/the) TERM	DEF is/are/was/were considered TERM
DEF comprise(s) (a/an/the) TERM	DEF is/are/was/were considered a/an/the TERM
DEF include(s) (a/an/the) TERM	DEF is/are/was/were to be considered (a/an/the) TERM

[a] DEF cannot include specific stopwords, years, authors

[b] DEF is extracted up to the next comma. Otherwise, up to the end of the sentence

Table 5.12: The set of definition (DEF) patterns used in the second experiment.

Second Experiment	Third Experiment	Fourth Experiment
Fig/Fig/figure/Figure	Labs/Laboratories	pp/p
Et/Et	Table/table	University
Unpublished/Unpublished	Arrow	Society
	Personal Communication/personal communication/pers. Comm.	Institute
	Road	Center
	Street	Centre
	PO Box/P.O. Box	ed/eds/Editors
		edn.
		School
		Faculty
		Department
		Hospital
		Academy
		College
		Manuscript
		Unit
		Division
		Verlag
		Agency

Table 5.13: Stopwords that DEF should not include, when DEF is enclosed in brackets. Stopwords were added accumulatively in each experiment.

5.4.4 Results

Of the 3,920 FishBase terms submitted to the system in the second experiment, 3,156 matched no documents, and 310 matched no answers. Appendix F lists the remaining 454 terms which did yield answers. At the query level we judged fifteen of the responses as Vital (3%), 58 as Okay (12%), 30 as Uncertain (7%), and 351 as Wrong (78%) (see Table 5.18 which summarises the results for the second, third, and fourth experiments).

Thus, Average Strict Binary Responsiveness and Average Lenient Binary Responsiveness were 3% and 22% respectively. Appendix G shows a sample of Vital and Okay answers from this experiment.

5.4.5 Discussion

The lower responsiveness compared to the first experiment may be due to stricter judgement against the definitions of the FishBase glossary. In addition, this time we used a larger set of terms, many of which (e.g., 'data', 'demographic') were not specific to salmon biology or even the fish domain, whereas our original set was suggested by people who work in the field.

We did not check in advance if a term appeared in the collection. Therefore, it is not surprising that the system retrieved answers to only 11% of the FishBase terms. As mentioned in Chapter 3, FishBase is partly multilingual and contains non-English terms which were not removed before the run. Eight percent of the queries did not match any patterns even though their terms were still found in at least one document.

We observed that the term in the pattern **`DEF including TERM`** did not necessarily follow the word 'including'. For example, one of the definition answers to 'amphibians' contained the following text: `lower vertebrates including teleosts, amphibians.` The answer was retrieved because it matched the pattern **`DEF, TERM.`** However, to reduce false matching it would be preferable to use the first, more exact pattern but in a flexible form, allowing any number of words between 'including' and the term (the term would not be located further than the end of the sentence because we only extracted definitions between **`SEN`** tags).

We observed that many of the wrong answers matched the past tense version of the copular pattern (**`TERM was/were DEF`**). In scientific writing it is customary to use past tense when reporting original work and present tense when describing established knowledge (Day, 1998). Therefore, patterns in the present tense are more likely to match definitions. However, sometimes actions performed on or by the term can elucidate it. This is especially common in the Methods section of papers. For example, the term 'Seechi disc' is defined in FishBase as

'A 20 cm diameter disc marked in 2 black and 2 white opposing quadrants, lowered into the water. The average of the depth at which it disappears from sight and the depth at which it reappears when lowered and raised in the water column is the Secchi disc reading, a measure of transparency'.

We retrieved the Okay answer `Secchi disc was used to measure water visibility (m of visibility) at 1400h`. We failed to add the pattern **`TERM was used to DEF`** but in general, a more sophisticated approach than lexical patterns is needed to weed out uninformative verbal phrases.

5.5 Third Experiment

5.5.1 Aim

To test the effect of new patterns and rules on the query level precision.

5.5.2 Data

2,000 terms were extracted from the FishBase glossary. These terms were not the same as those used in the second experiment.

5.5.3 System Configuration

We removed the following patterns:

- TERM was/were DEF
- TERM, which was/were

and added the following ones:

- TERM, defined as DEF
- TERM, which have/has
- DEF including [Words], TERM
- DEF is/are/was/were defined as (a/an/the) TERM

- the term TERM DEF

We eliminated answers containing terms following a hyphen (-TERM) because in these cases the term was likely to be the second half of a hyphenated compound.

We also created a stopword list (Table 5.14) for the words immediately following the comma in the apposition pattern **`TERM, DEF`**.

At	is/was/are/were
when	then
where	respectively
there	however
whereas	while
thus	whereupon
therefore	but
we	it
and	although
Uppercase word	though

Table 5.14: Stopwords that should not appear after the comma in the pattern TERM, DEF.

5.5.4 Results

Of the 2,000 FishBase terms submitted to the system in the third experiment, 231 (listed in Appendix H) yielded answers. Ten of the responses were judged as Vital (4%), 37 were judged as Okay (16%), 41 as Uncertain (18%), and 143 as Wrong (62%) (see Table 5.18 which summarises the results for the second, third, and fourth experiments). Thus, Average Strict Precision and Lenient Precision were 4% and 38% respectively. Appendix I shows a sample of Okay and Vital answers from this experiment.

5.5.5 Discussion

The Average Strict Binary Responsiveness remained low, although the Average Lenient Binary Responsiveness improved considerably, from 22% to 38%.

The list of stopwords we introduced to eliminate false matching of the appositive pattern **`TERM, DEF`** was inadequate. Therefore, we decided that the pattern should be made stricter. Joho (1999) used **`TERM, DEF, is/was/are/were`**. We considered this pattern too restrictive and in the next experiment matched any verb that followed the second comma in the pattern.

We observed that queries with common terms (e.g., 'family') still returned numerous answers. Many of the false answers matched the ambiguous, less precise patterns. Patterns that refer to the act of defining are usually unambiguous (e.g. **`TERM defines DEF`**, **`DEF named TERM`**, **`TERM is described as DEF`**). Therefore, it made sense to exploit the redundancy of common terms and limit the set of patterns associated with them to unambiguous ones (see Section 5.6.3).

In this and previous experiments, the term that was found in the document was sometimes the last or first part of a compound. For example, 'depression' in 'outbreeding depression', 'maturation' in 'early maturation' or 'rostrum' in 'rostrum degenerative disease'. We observed that elimination of answers based on the part-of-speech of the word preceding or following the query term could solve this problem in many cases (see Section 5.6.3). Deeper syntactic analysis could identify the head noun of a noun phrase, but we decided to improve our system incrementally and use the next simple method above lexical patterns.

5.6 Fourth Experiment

5.6.1 Aim

To test the effect of introducing shallow syntactical elimination patterns, term type-specific patterns, and a term-frequency based search strategy on precision.

5.6.2 Data

1,120 terms were extracted from the FishBase glossary. Most of the terms (900) were new terms.

5.6.3 System Configuration

We changed the search strategy so as to differentiate between three types of terms:

- Abbreviations (where the first two characters of the terms' first word are upper case letters; e.g., FAO, DNA),
- Common terms (appearing in more than 100 sentences; e.g., family, fry),
- Uncommon terms (e.g., kelt, caciques).

Each of the three types was associated with a different set of patterns. We used the patterns **`DEF (TERM)`** and **`TERM (DEF)`** to match definitions to abbreviations. Tables 5.15 and 5.16 list the patterns used to match definitions to common and uncommon terms, respectively. The list for uncommon terms included eleven new patterns. The patterns for common terms were patterns which we considered to be unambiguous.

We removed the patterns **`TERM, or DEF`**, and **`TERM: DEF`** which were not productive in previous runs.

The XeLDA tagger (XeLDA, 2003) was used to identify parts-of-speech when restricting some patterns and augmenting others (see below).

We added the words and phrases 'so', 'with', 'as', 'because', 'in addition', and 'both' to the list of stopwords which should not appear immediately after the comma in the pattern **`TERM, DEF`**. We restricted this pattern to **`TERM, DEF, VERB`** to reduce false matching.

As Tables 5.15. and 5.16 show, we also added optional adverbs to many of the patterns.

A pattern instance was eliminated if

- the word preceding the term in the pattern was a singular or mass noun, a proper noun, an adjective, or a preposition;
- the word after the term was a noun (singular or plural), a proper noun, or an adjective.

When searching for terms that are not abbreviations, we first submitted the query

(TERM is a OR TERM is an OR TERM is the OR TERM are) w/6 xfirstword, which specifies that the copular pattern **TERM be (article) DEF** should appear within six words of the first word in the document (sentence). The value six was chosen because the first three tokens in the document are the tag **SEN**, the sentence ordinal number, and the total number of sentences (as assigned during pre-processing of the document collection). If documents were found in response to this query, the system performed no further search, because we observed that copular patterns in the beginning of sentences are highly effective, as the term is likely to be the subject of the sentence. When no documents were retrieved in response to this query, the search continued as usual.

the term TERM DEF	definition of a/an/the TERM is
TERM is described (ADV) as DEF	TERM(,) such as DEF
TERM, (ADV) described as DEF*	TERM(,) including DEF
TERM is (ADV) the term DEF	DEF called (a/an/the) TERM
TERM, (ADV) the term*	DEF termed (a/an/the) TERM
TERM is (ADV) used to describe DEF	DEF named (as) (a/an/the) TERM
TERM is/was/are/were/ (ADV) defined as DEF	such DEF as (a/an/the) TERM
TERM, (ADV) defined as DEF	DEF known as (a/an/the) TERM
define(s) (a/an/the) TERM as DEF	DEF is/was/are/were/ (ADV) defined as (a/an/the) TERM
TERM defines DEF	TERM (ADV) refers to DEF
TERM (i.e. DEF	TERM (ADV) refers to DEF
TERM (e.g. DEF	TERM, i.e. DEF

Table 5.15: Patterns used in the fourth experiment to match definitions of common terms. Asterisked patterns are new. These patterns were selected for being the least ambiguous.

5.6.4 Results

Of the 1,220 FishBase terms submitted to the system in the fourth experiment, 35 terms yielded 55 answers in total (Table 5.17). The number of answers returned for each of the 35 terms ranged between 1 and 7 (mean=1.54, median=1).

Average Strict Binary Responsiveness and Average Lenient Binary Responsiveness were 5.7% and 60% respectively (Table 5.18). Average Strict Precision and Average Lenient Precision were 7.2% and 63.6% respectively. Appendix J shows a sample of

Okay and Vital answers from the fourth experiment.

DEF—TERM	definition of (a/an/the) TERM is
the term TERM DEF	DEF including [Word], TERM
TERM is (ADV) described as DEF	DEF such as [Word], (and) TERM*
TERM, (ADV) described as DEF*	DEF ([Words], TERM, [Words])*
TERM is (ADV) the term DEF	TERM(,) such as DEF
TERM, (ADV) the term*	TERM(,) including DEF
TERM is (ADV) used to describe DEF	TERM(,) especially DEF
TERM (ADV) describes DEF	TERM, that is DEF*
TERM is/was/are/were/ (ADV) defined as DEF	TERM, for example DEF*
TERM, (ADV) defined as DEF	TERM, for instance DEF*
TERM is/are a/an/the DEF	TERM, in general DEF*
TERM, which is/are a/an/the DEF	TERM, DEF, VERB*
define(s) (a/an/the) TERM as DEF	DEF called (a/an/the) TERM
TERM defines DEF	DEF termed (a/an/the) TERM
TERM—DEF	DEF such as (a/an/the) TERM
TERM(,) and other DEF	DEF named (as) (a/an/the) TERM
TERM(,) or other DEF	such DEF as (a/an/the) TERM
TERM (ADV) consist(s) of DEF	DEF including (a/an/the) TERM
TERM (ADV) constitute(s) DEF	DEF especially (a/an/the) TERM
TERM (ADV) comprise(s) DEF	DEF ((a/an/the)TERM)
TERM (ADV) include(s) DEF	DEF known as (a/an/the) TERM
DEF consist(s) of (a/an/the) TERM	DEF is/are/was/were (ADV) considered (to be) (a/an/the) TERM
DEF constitute(s) (a/an/the) TERM	DEF is/are/was/were (ADV) defined as (a/an/the) TERM
DEF comprise(s) (a/an/the) TERM	TERM is/are/was/were (ADV) considered (to be) (a/an/the) DEF
DEF include(s) (a/an/the) TERM	DEF, i.e TERM*
TERM, a/an/the DEF	DEF, e.g DEF*
TERM means DEF	TERM, i.e. DEF
TERM(,)(DEF)	TERM, e.g. DEF
TERM (ADV) refers to	TERM is (ADV) defined DEF

Table 5.16: Patterns used in the fourth run to match definitions of uncommon terms (terms for which fewer than 100 sentences were retrieved). Asterisked patterns are new.

Term	Vital	Okay	Uncertain	Wrong
abiotic	0	1	0	0
active metabolism	0	0	0	1
ad libitum	0	0	0	1
adipose	0	2	0	0
adsorption	0	0	1	0
aggregation	0	0	1	0
alevin	0	4	0	0
algae	0	1	0	1
Caciques	0	1	0	0
carrying capacity	0	0	0	1
catchability	0	0	0	1
catchment	0	0	0	1
erosion	0	0	0	1
caviar	0	1	0	0
Neartctic	0	0	0	1
nest	0	0	0	1
Amphibia	0	1	0	0
lineage	0	1	0	1
nominal species	0	0	0	1
nostril	0	0	0	1
notochord	0	1	0	0
laminarin	0	1	0	0
lapillus	1	0	0	0
lateral line	0	0	0	2
Length-weight relationship	0	0	1	0
lethargy	0	3	0	0
life cycle	0	0	0	3
Lm	0	0	0	1
locks	0	0	1	0
longevity	0	0	1	0
Ls	0	0	1	0
luciferase	0	2	0	0
lumen	0	0	0	2
grilse	3	4	0	0
cartilaginous fishes	0	2	0	0
TOTAL	**4**	**25**	**6**	**20**

Table 5.17: Judgement and number of answers in the fourth experiment.

	Experiment 2 (3,920 terms)	**Experiment 3 (2,000 terms)**	**Experiment 4 (1,220 terms)**
Vital	15 (3%)	10 (4%)	2 (6%)
Okay	58 (12%)	37 (16%)	13 (37%)
Uncertain	30 (7%)	41 (18%)	6 (17%)
Wrong	351 (78%)	143 (62%)	14 (40%)
TOTAL	454 (100%)	231 (100%)	35 (100%)
Average Strict Binary Responsiveness	3%	4%	6%
Average Lenient Binary Responsiveness	22%	38%	60%

Table 5.18: Summary of the final results of he second, third, and fourth experiments.

5.6.5 Discussion

Once again Average Strict Binary Responsiveness improved marginally, whereas Average Lenient Binary Responsiveness increased from 38% to 60%. If we assume that eliminating every Okay answer is not desirable, Vital answers are always likely to be outnumbered by Okay answers. We do not know, however, what the maximum ratio of Vital to Okay answers is in our collection (the Recall calculation problem, as discussed earlier).

In the fourth experiment we recorded a great reduction in the number of pattern instances returned for each query. In the first experiment the number of pattern instances ranged between 1 and 290, whereas the maximum number of instances returned in the fourth experiment was seven. This means that in the last experiment the output was equivalent to having all the answers ranked in the top ten. Expecting users to peruse ten answers to identify the correct one is a realistic scenario.

It is doubtful whether many more Wrong answers could be eliminated without resorting to deeper linguistic techniques (Fleischman, Hovy and Echihabi, 2003). For example, one of the definitions returned in response to the query 'nest' was `nest is a successful tactic that decreases the variation in reproductive success among males.` However, the complete sentence began with `Thus, sneaking into the nest is a successful…` The subject of 'is' was not 'nest' but the noun phrase 'sneaking into the nest'. In this case we require parsing of the sentence to eliminate the answer candidate, because the determiner 'the' was not included in the list of parts-of-speech which should not precede the query term.

5.7 Summary

This chapter reported and discussed the results of four experiments. We analysed the effectiveness of an initial set of patterns quantitatively and qualitatively using terms related to salmon and discovered that `doc_so_far_term_count` is a significant feature when predicting the occurrence of a definition in a sentence. In the subsequent three experiments we interrogated the SOK-i collection with a larger set of FishBase terms. We improved precision by adding and modifying patterns and their associated rules in each experiment. Changes to the system were implemented based on qualitative observations in preceding runs. In the next chapter we present our conclusions and suggest possible future steps.