Chapter 3: Domain of Application

3.1 Outline

This chapter first explains the choice of fish-related terminology and salmon fish in particular as the research domain. The second section describes the creation of SOK-i, the document collection that served as the source of definitions in this project. The method used for obtaining the terms of terms is reported.

3.2 Why Salmon Fish Biology?

The question answering track in TREC is termed open-domain, because the test questions are of the trivia type and are not confined to a specific area. However, the journalistic content and style of the AQUAINT corpus which is used in TREC often dictates the depth and focus of the questions: American culture, history, sports, and geography. However, to be genuinely open-domain, questions should target every possible genre of text. Therefore, we tried to marry the interest in definition questions with a new domain and genre. These could be later subsumed in broad open domain question answering.

Recently, the natural language processing community has shown much interest in biological literature, or *bibliome* (Hirschman et al., 2002, Dickman, 2003) for the purpose of data and text mining (for example, recognising protein interactions, determining cellular location, and identifying trends and themes in the literature). The focus of the work in this area is proteomics and genomics (as the addition of the genomics track in TREC 2003 indicates). However, we decided to approach a sample of biological literature from a different angle. Instead of focusing on text in a particular discipline of biology (such as genetics, anatomy, physiology), we imagined a sophisticated user who might be interested in multi-disciplinary information on a particular species or taxon (taxonomic category). For our work on definitions we preferred a species that has commercial importance, because such a species is often associated with a voluminous body of literature and more importantly, a rich terminology.

One taxon which meets these criteria is the salmon. Total global supply of farmed and wild salmon has increased from 550,000 tons in 1980 to more than 2 million tons in 2002. The increase is mainly due to a rise in aquaculture production (GLOBEFISH, 2003). Since 1997 supply from aquaculture has exceeded supply from wild salmon fisheries and has encouraged extensive research on salmon, for example, the Norwegian Salmon Genome Project (Salmon Genome Project, 2003).

Most of the unique terminology related to salmonids is associated with their life cycle and stages, although our research is not restricted to such terms. The precise technical definition of salmon terminology is of interest to scientists and fishery managers. Allan and Ritter (1977) proposed in their paper a list of terms and definitions designating the different life stages of the Atlantic salmon and the migratory form of the trout (considered a species of salmon). Their purpose was to allow salmon fishery scientists to interpret correctly the terms used in countries other than their own and to suggest terminology for use by people involved in salmon fisheries. The list of definitions was restricted to terms whose basis was objective rather than subjective, because countries use different criteria for their own classifications. The difference is due to variations in growth rates and migration ages.

3.3 The SOK-i Test Document Collection

To create the test document collection, one thousand documents (70.5 Mb) in HTML format were downloaded from Science Direct (ScienceDirect, 2003) on April 2, 2003. These were the top references retrieved in response to the query term ‘salmon’. Science Direct is the online journal service of Reed-Elsevier and provides access to over 1,800 journal titles published by the company and its affiliated publishers in all fields of science. The database contains around five million full-text articles and 59 million abstracts. The term ‘salmon’ was searched for within the abstracts, titles and keywords fields, and in all sciences. Naturally, however, most of the documents retrieved were in agricultural and biological sciences, biochemistry, genetics and molecular biology and are related to fish research. We named the collection SOK-i (Salmon Of Knowledge and the letter I), pronounced the same as ‘Sockeye’ which is a species of salmon.

Seventy percent of SOK-i is full-text articles with the remaining 30% comprising abstracts. The journals represented include titles such as Aquaculture, Domestic Animal Endocrinology, Journal of Fish Biology, Molecular and Cellular Endocrinology, Biosystems, Fisheries Research, Comparative Biochemistry and Marine Pollution Bulletin. Appendices A1 and A2 show examples of a full-text article and an abstract respectively. Figure 3.1 shows an excerpt from both an abstract and a full-text paper which include definitions for the term ‘smoltification’.

Smoltification is a complex developmental phase encompassing hormonal, metabolic and osmoregulatory changes, enabling the Atlantic salmon to migrate from fresh water to sea water. The biological interface between fish and their aqueous environment is a mucus coat composed of biochemically diverse secretions, which have been implicated in respiration, disease resistance and osmoregulation. In this study mucus and blood were sampled from hatchery-reared Atlantic salmon over the period of smoltification. Our aim was to investigate changes in mucus proteins and enzymes and to monitor plasma thyroxine during smoltification.

(http://www.sciencedirect.com/science/article/B6T4D-485PC6D-2/2/4e67f09d9a6d8c3a4e3f99b03e8ece6f)

2.2. Salmon stock identification based on traits associated with smoltification

Anadromous salmonids migrate from rivers to marine habitats as a developmental stage termed **smolt**. **Smoltification** is a key physiological and ecological transition for salmon, and imposes strong selection pressures (Nicieza et al., 1994). Variation in the timing of smoltification cues and associated ontogenetic effects have been useful for salmonid stock identification.

(http://www.sciencedirect.com/science/article/B6T6N-3XNJYSC-K/2/7bf44db4994a9dd81c63091803b632)

Figure 3.1: Excerpt from an abstract (top) and a full-text paper (bottom) in the SOK-i collection with definitions for the term ‘smoltification’. The bottom excerpt contains also the definition for ‘smolt’.

Most of the full-text articles in the collection follow the structure known as IMRD (Introduction, Methods, Results, Discussion). This is the most common organisation of scientific papers that report original research (Day, 1998). Most of the papers in the SOK-i collection are of this type. Review papers and short communications may somewhat differ in their structure. The guidelines to authors submitting papers to the journal Aquaculture (Elsevier Author Guide, 2003) specify the following required sections: Abstract, Keywords, Introduction, Methods and Materials, Results, Discussion, Conclusion, Acknowledgments, and References. Other journals which are represented in the collection have similar requirements.

Apart from content and typical structure, other features characterise text of scientific papers such as the ones in our collection. These features include:

- Long sentences, often with parenthetical clauses (e.g., ‘The degenerate RT PCR analysis of the ventral aorta tissue (which includes thyroid follicles) using the degenerate primers (P1 and P2) gave rise to a single amplicon composed of a single cDNA species, which was identical to the ovarian cDNA’),
- Low-frequency words, mostly specific technical terms (e.g., ‘redd’, ‘kelt’) or jargon,
- Abbreviations,
- Numerical data,
- References to figures and tables,
- References to published work in Harvard author-year citation style. Most of the citations are silent (i.e., enclosed in brackets). When there are more than two authors, only the first is mentioned and followed by ‘et al.’,
- Passive voice. Guides to scientific and technical writing encourage authors to use active voice, which is clearer and more concise. However, passive voice is still prevalent in scientific writing (e.g. ‘samples, which on average weighed 50 kg, were taken back to the laboratory’, or ‘125 I-calcitonin (salmon) was obtained from Amersham Life Sciences’).

In their original HTML format, the documents included underlined hyperlinks (see Appendix A1). A section titled Article Outline consisted of a list of hyperlinks to different segments in the paper. Readers online require such a navigational aid, because when a paper is published on the Web, it loses its traditional page structure and normally consists of a long scrollable window (Day, 1998). References to published work and to figures and tables were also hyperlinked in the text.

Section 4.3 describes how the documents in SOK-i were pre-processed.

3.4 Obtaining Terms for Testing

At the start of this project we contacted salmon researchers in different disciplines (such as genetics, aquaculture, nutrition and ecology) in different countries (Canada, Ireland, New Zealand, Norway, United States) and asked them to suggest 14 terms related to their field of study. Half of the terms were meant to remain unseen until the testing phase of the project and therefore were sent to a separate email account. Seven researchers responded to the request and together suggested 49 terms for the first exploratory experiment. Table 3.1 lists the 42 terms that were left after removal of seven duplicates.

tetraploidy	local adaptation	stunning
gene duplicates	parr	colour
immunoglobulin M (IgM)	redd	filet
immunoglobulin D (IgD)	alevin	astaxanthin
isotypes	smolt	canthaxanthin
pseudogenes	grilse	fat
tetrasomic inheritance	osmoregulation	phenotype
artificial photoperiod	migration	otolith
Chinook salmon	fry	watershed
contemporary microevolution	wild	hatchery
early maturation	Farmed	propagation
freshwater residence	Atlantic	stock
life history	quality	aquaculture
DNA	critical habitat	conservation

Table 3.1: Terms suggested by salmon researchers. Each researcher was asked to provide 14 terms specific to their field. Only half of the terms suggested were used initially.

After the first experiment, we realised that we needed many more terms, especially considering the small size of the test corpus. We identified the glossary of FishBase (FishBase, 2003) as an appropriate and rich source of terms related to the fish domain. FishBase is a relational database with information on all fish known to science. It was developed by the World Fish Center in collaboration with the Food and Agriculture Organization of the United Nations and many other partners, and with support from the

European Commission to provide information on fish to professionals such as research scientists, fisheries managers and zoologists. The advantage of extracting terms from the FishBase glossary was that we could evaluate the system's responses more objectively against the definitions in the glossary.

Naturally, FishBase includes terms specific to salmon (e.g., kelt, smolt, parr, redd), but these constitute a fraction of the entire glossary. Many of the terms have other, often more common meanings outside the fish domain (e.g., release, run, satellite, anonymous, family) or are general scientific terms (e.g., recombinant DNA technology, Ordivician). However, because most of the papers in the collection are on salmon, any definitions to the terms are likely to be in this context. Table 3.2 lists examples of such terms and their FishBase definitions.

Term	**FishBase definition**
parr	A young salmonid (salmon or trout) with parr-marks before migration to the sea and after dispersal from the redd.
homogenous	Uniform; in ichthyology used to describe egg yolk in larval fishes as opposed to segmented.
vertebrae	The bones of the axial skeleton; divided into two sections, precaudal and caudal vertebrae.

Table 3.2: Examples of terms and definitions in the FishBase glossary (FishBase, 2003).

Some of the definitions for English terms in the glossary include the equivalent term in other languages (French Spanish Portuguese Russian), sometimes along with the definition in these languages. The non-English term is hyperlinked often to the definition in English. This means that the glossary consists of English and non-English terms.

3.5 Summary

In this chapter we identified the increasing interest in the processing of biological literature, the need to expand the domains used in question answering, and the richness of fish/salmon-related terminology as the main reasons for selecting salmon fish biology as the domain of application. We described the creation of the appropriate document collection, named SOK-i, and the sources of the terms used in testing the system. The next chapter will present the system and its implementation.