GeneTUC

/j&’-ne-tük/

Anders Andenæs

April 28, 2000

Abstract

The GeneTUC system is an attempt to create an application capable of performing Information Extraction from bio-medical literature. GeneTUC is based on TUC, The Understanding Computer, and is based on lexical, syntactic and semantic analysis of the input material. GeneTUC is basically a TUC system with a data base containing the names of genes and proteins. This report describes the work and experiences made during the project.

Acknowledgments

I wish to thank my two supervisors. First, *Tore Amble*, for always being willing to help and stand by me, no matter how late or early. Also, *Tor-Kristian Jenssen*, for keeping me on a tight leash, and helping me retaining focus on the assignment, when I was drifting astray.

My fellow students, *Truong Van Le* and *Einar Fløystad Dørum*, also deserve mention. Thanks to you both for fruitful discussions in the computer lab.

Finally, *Kari Flaaten*, for providing a safe haven, away from computers and this project.

Preface

Working on this project has been a real challenge for me. GeneTUC has one foot planted in AI and NLP, while the other stands firmly in genetics and biomedicine. Covering all of these sciences has required a bit of reading up, both on my English and biology. Even still, I have perceived the work done on this project as fun and exciting, as well as being highly educational.

Although much work remains to be done, GeneTUC has grown to be a decent IE system in a surprisingly short time. I hope some of this work will be continued in the future, by me or others, as GeneTUC's potential has yet to be exploited to its full extent.

This report is written in GNU Emacs and typeset in Palatino 12pt, using LaTeX 2_ε.

Anders Andenæs, April 28, 2000

Contents

1 Introduction **1**
1.1 The assignment 1
1.2 This report 1

2 Background **3**
2.1 Text mining 3
2.1.1 Free text-search 3
2.1.2 Stored queries 4
2.1.3 Cross-matching 4
2.2 Knowledge-based approach 5
2.2.1 Knowledge Systems 5
2.2.2 General Methodology 5
2.2.3 Appeal 6
2.3 Natural language processing 7
2.4 TUC - The Understanding Computer 7
2.4.1 Inner workings 8
2.5 A crash course in genetics 10
2.5.1 DNA 10
2.5.2 Genes 12
2.5.3 Chromosomes 12
2.6 Extracting information about genetics 13
2.6.1 The domain 13
2.6.2 Existing work 13

3 Elementary linguistics **15**
3.1 The basics 15
3.1.1 Verbs 15

3.1.2 Nouns . . . 16
3.1.3 Adjectives . . . 16
3.1.4 Adverbs . . . 17
3.1.5 Pronouns . . . 17
3.1.6 Prepositions . . . 18
3.1.7 Conjunctions . . . 19
3.2 Sentence categories . . . 19
3.2.1 Declarative . . . 19
3.2.2 Imperative . . . 20
3.2.3 Interrogative . . . 20
3.3 Ellipsis . . . 20
3.4 Anaphora . . . 21
3.5 Naturally Readable Logic . . . 21

4 Learning “Goldilocks...” 23
4.1 Why “Goldilocks...”? . . . 23
4.2 From buses to bears . . . 23
4.3 New requirements . . . 24
4.3.1 Vocabulary . . . 24
4.3.2 Grammar . . . 26
4.3.3 Other changes . . . 27
4.3.4 Further imperfections . . . 29
4.4 Performance . . . 30
4.4.1 Input . . . 30
4.4.2 Questions . . . 31
4.4.3 Results . . . 31
4.4.4 Preliminary conclusions . . . 32

5 Genetical articles 33
5.1 Analysing the domain . . . 33
5.2 Ontology . . . 34
5.3 Meaningful relationships . . . 35
5.4 Further improvements to TUC . . . 37

6 Results and conclusions 39
6.1 Results . . . 39
6.1.1 Some examples . . . 40
6.2 Conclusions . . . 43

7 Future work **45**
7.1 GeneTUC 45
7.2 Information extraction 46

References **49**

A Questions and answers **51**

B Multiple categorisations **55**

C Original text **57**

D Adapted text **59**

1 Introduction

This chapter presents the assignment for the project, and this report.

1.1 The assignment

The assignment for this project, as given by Tore Amble at the Department of Computer and Information Science, was as follows:

> Current knowledge about genes and their interactions exist to a large extent only as free text. Searching and cross-linking such information rely to a large extent on existing indexes created either manually or by syntactic pattern-matching. As a first step we want a tool that is able to correctly recognize occurrences of a gene in free text, e.g. in a article abstract, and the context in which the gene is mentioned.

1.2 This report

This document reports on the current status on the endeavour to create a system conforming to the assignment given in the previous section, based on the TUC framework. The report opens with a chapter providing a general background, touching in on knowledge systems, natural language processing, genetics and the TUC system.

Chapter 3 provides an introduction into the principles of linguistics and grammar. This is crucial knowledge when dealing with natural language systems, and also required to understand some of the discussions later in the report.

To serve as an introduction to natural language processing and the

TUC system, a pre-project was carried out. The aim of this project was to enhance TUC's capabilities when treating prose. The "Goldilocks..."project is described in chapter 4.

The last three chapters concentrate on the GeneTUC project, describing what was done, the results and conclusions which can be drawn on the basis of the project so far. The last chapter outlines some of the work which lies ahead.

The appendices contain the dialog with the system and the original and adapted "Goldilocks..."text.

2 Background

This chapter will give a brief description of the background for the report. The aim of the first section is to try to familiarise the reader with some of the concepts of text mining. A few of the currently most popular techniques are described in brief.

The knowledge-based approach to text mining is discussed next. Knowledge systems in general are introduced, and the general methodology is explained in brief. Also, this section gives a short summary of why it is so appealing to use knowledge systems for text mining.

The chapter continues with a summary of what is meant by natural language processing, and a description of the TUC system. It concludes with some remarks regarding text mining biomedical articles.

2.1 Text mining

This section provides some background on some current methods for text mining: Free text-searching, stored queries and cross-matching.

Information Extraction (IE) is an application of natural language processing which takes a piece of free text and produces a structured representation of the points of interest in it. Input must be syntactically and semantically analysed in order to produce a sound output.

2.1.1 Free text-search

The straightforward way of searching for information in textual data, is by executing a free text-search in the data source. The large search engines on the Internet, e.g. *AllTheWeb* and *AltaVista*, are based on this technique, as well are search functions in application programs. Although this

method of searching is not considered to be IE, rather *information retrieval*, this method is so common, it deserves some mention in this chapter.

As the free-text searching is based on a purely syntactical analysis of the query-string and data base, the degree of relevance in the search result will vary. The search engine will rarely contain any domain knowledge, thus searching for a homonym[1] will possibly return undesirable results. A search for *temple* will find places of worship as well as information on the human anatomy.

In an effort to overcome this flaw, most search interfaces offer some kind of query language. The degree of user-friendliness inherent in these interfaces is different from case to case, but often the threshold for effective use is unnecessarily high. Constructing an efficient query-string requires the user to be competent in some proprietary query language, or regular expressions (RegExp's).

2.1.2 Stored queries

Another way of addressing the problem, is by trying to match the query at hand with some previously stored query. An example of this approach is *"Ask Jeeves"* [Ask]. This service catalogues the answer to all the searches it conducts, thus making itself more competent each time it is accessed.

The stored query-technique is essentially just an adaptation of and interfacing to free text-searching. Hence it suffers from the some of the same deficiencies as the latter. The strong point of the stored query-technique is the possibility to customise the stored queries in a way that provides the highest quality output.

2.1.3 Cross-matching

The two methods described above are examples of information retrieval techniques. What is more interesting are techniques based on IE. Some of these methods will be described in this section and in section 2.6.2.

Some efforts have been made in developing systems for cross-matching keywords in a cause-effect relationship. The *Arrowsmith* system [SS99], is basically an extension to the *MEDLINE* [MED] searching facility. It uses the results of MEDLINE searches to infer knowledge of causalities. As an example, consider the following scenario: The tabloids often claim that

[1] One of two or more words spelled and pronounced alike but different in meaning

excessive intake of caffeine causes headache. To investigate this, one could submit MEDLINE searches for the words “caffeine” and “headache”. These results would then be fed into the Arrowsmith system, which, in turn, would try to find some unknown factor X which is such that caffeine causes X and X causes headache.

The shortcoming of such a strategy is that the causality relationship one seeks, need not be a single-step one. If the connection between the cause and the effect only occurs through a multitude (and unknown number) of steps, this method fails.

2.2 Knowledge-based approach

Using knowledge systems is a powerful and flexible method for mining texts. This section briefly explains what is meant by the term knowledge systems, and how these are applied to the problem at hand. The section closes with some remarks on why using knowledge systems is so appealing.

2.2.1 Knowledge Systems

Knowledge Systems are a subfield of Artificial Intelligence (AI). According to [Nil98], the phrase *knowledge systems*, or, more accurately, *knowledge-based systems*, is used to describe programs that reason over extensive knowledge bases, containing facts and *rules*. This knowledge base is implemented in some formal knowledge representation language. In TUC’s case, this language is Sicstus Prolog, but Common Lisp is another widely used language.

2.2.2 General Methodology

Using the knowledge-based approach to perform text mining requires a natural language-capable knowledge system. Constructing such a system is beyond the scope of this report, hence focus will be on extending an existing system.

The first step is analysing the domain. The system’s vocabulary will have to be augmented to be sufficient to extract information and answer queries using the correct terminology. New concepts will have to be placed correctly in the existing hierarchy. (How the hierarchy is constructed is implementation dependent.)

The grammar must often be updated, reflecting how sound sentences describing the domain may be formed. This includes defining valid combinations of nouns and verbs (in “Goldilocks . . . “ [Sou71], do bears “speak”?).

Then the knowledge base will have to be fed with information. In the case of a natural language-competent system, this is most likely a straightforward task.

The system will readily accept texts written in plain language. Structured texts, i.e. texts using some kind of field-formatting, presumably not in natural language, must be reformatted upon entry. Rather, the information could be input directly into the knowledge base, if the internal format of the latter is known.

2.2.3 Appeal

The knowledge-based approach, using natural language processing, may seem cumbersome at first. The preparation of the system, building semantic nets and defining a sensible grammar, is both tedious and time-consuming, and this method would seem less intuitive than some of the methods described earlier in this chapter.

Still, basing retrieval on searching a knowledge base, holds potentially great advantages. The user friendliness of a well-constructed natural language interface, in lieu of a conventional[2] searching interface, needs not be stressed.

It is crucial to recognise that the knowledge based-approach relies on semantical, rather than syntactical, analysis of the data base. Unlike the conventional searching methods, do homonyms not pose a problem, given an adequate semantic net. Thus the query “Which *band* plays at Studentersamfundet next Friday”, will surely output the name of a musical group (or none) and not instructions on how to tie objects together.

Another strong point of the knowledge base is the ability to extract implicit knowledge. Returning to the caffeine and headache example, a correctly constructed knowledge base would find any connection between those two. Provided such a connection exists and is implied by the data base, of course.

[2] i.e. using RegExp’s or a proprietary query language

2.3 Natural language processing

The FAQ[3] for the comp.ai.nat-lang [NLP] newsgroup gives the following definition of Natural Language Processing (NLP), or *computational linguistics*:

> Computational linguistics (CL) is a discipline between linguistics and computer science which is concerned with the computational aspects of the human language faculty. It belongs to the cognitive sciences and overlaps with the field of artificial intelligence (AI), a branch of computer science that is aiming at computational models of human cognition. There are two components of CL: applied and theoretical. [...] The applied component of CL is more interested in the practical outcome of modelling human language use. The goal is to create software products that have some knowledge of human language. Such products are urgently needed for improving human-machine interaction since the main obstacle in the interaction beween [sic] human and computer is one of communication. [...] Natural language interfaces enable the user to communicate with the computer in German, English or another human language. Some applications of such interfaces are database queries, information retrieval from texts and so-called expert systems.

TUC, described in the next section, is an example of such a natural language processing system.

2.4 TUC - The Understanding Computer

The TUC project was initiated at NTH[4] in the early 1990's. It was based on a number of previous efforts in creating a natural language interface for querying data bases, among them CHAT-80, PRAT-89 and HSQL. The research goals for the project could be summarised as follows:

- Give computers an operational understanding of natural language
- Build intelligent systems with natural language capabilities
- Study common sense reasoning in natural language

[3] Frequently Asked Questions

[4] Norwegian Institute of Technology, now the Norwegian University of Science and Technology

The TUC project seeks to define a language denoted by NRL[5]. This language is as readable as plain English, but has well-defined syntax and semantics. In TUC, NRL serves as both a declarative knowledge definition language, and as a query language [HW94].

TUC relies on grammatical analysis for marking sentence elements. A sentence not being grammatically correct (according to TUC's internal grammar), will be rejected without further treatment. Enhancing TUC is thus both a question of adding to its vocabulary and semantics, *and* defining new grammatical constructs.

2.4.1 Inner workings

The language analysis in TUC is a five step process [Bra97], as shown in figure 2.1[6]:

- **Lexical analysis**
  The individual words of the input string is looked up in TUC's internal dictionary. If a word is not found in the dictionary, the lexical analyser tries to find it in a case specific data base containing words mentioned in earlier sentences. The lexical analyser also performs some spelling correction.
  If the input sentence is successfully analysed, the set of words are output as tokens in their inflective root forms, together with their possible word classes.

- **Syntactic and semantic analysis**
  The list of tokens is parsed using a differential attribute grammar. The parser builds a TFOL[7] formula representing the semantics of the sentence. It will output the first TFOL representation it finds that is syntactically and semantically satisfying.

- **Anaphora resolution**
  Anaphora[8] are replaced with the internal object they represent.

- **Optimising**
  The TFOL formula is skolemised and simplified into a TQL[9] formula.

[5] Naturally Readable Logic
[6] Figure taken from [Bra97]
[7] TUC First Order Logic
[8] Anaphora are words or phrases taking its reference from another word or phrase, e.g. she, it, then, see 3.4
[9] TUC Query Language

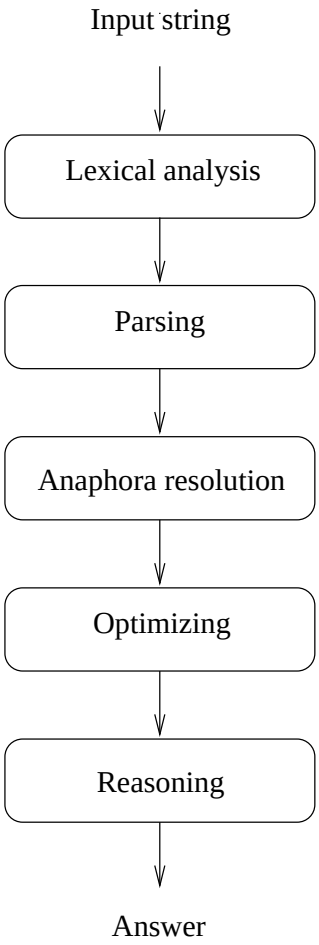

Figure 2.1: TUC's language analysis process

- **Reasoning**
  TUC uses the TQL formula to answer questions posed upon the system.

For clarity reasons, TUC's knowledge base is referred to in two ways. The *a priori* knowledge, i.e. the facts and rules hard-coded into the system, will be called its *permanent database*. The facts extracted from input text, is stored in the *semi-permanent database*.

2.5 A crash course in genetics

This chapter serves as a brief introduction into the field of human genetics. It is by no means a thorough discussion of the subject, but hopefully provides a minimal foundation for the discussions later in the report.

Much of what is written in this section is based on [HGP], the cell cycle illustration is adapted from [CCT].

2.5.1 DNA

The *genome* of an organism is the complete set of information describing the structure and activity of the organism. In humans, the genome is comprised of the DNA[10], and the associated proteins. These are organised into structures called *chromosomes* and are found in the *nucleus*, or core, of every cell. To understand how the DNA contains all information for building and maintaining life, information of its structure and organisation is needed.

The DNA consists of two strands wound tightly together, forming a "ladder". Each step of this ladder is made up of a base pair, either Adenine-Thymine (A-T), or Cytosine-Guanine (C-G). The *DNA sequence* is the order of the bases on the DNA. The DNA sequence specifies the genetic instructions needed to create the organism.

During the cell cycle, shown in figure 2.2, a cell divides into two daughter cells. During this process, the DNA is unwound, and the two strands are split apart from each other. Each strand synthesises a complement for itself, adhering to the base-pairing rules. *Mutation* occurs when there are errors in how the new strands are synthesised. Each of the new cells receive one of the "new" DNA molecules.

[10] DeoxyriboNucleic Acid

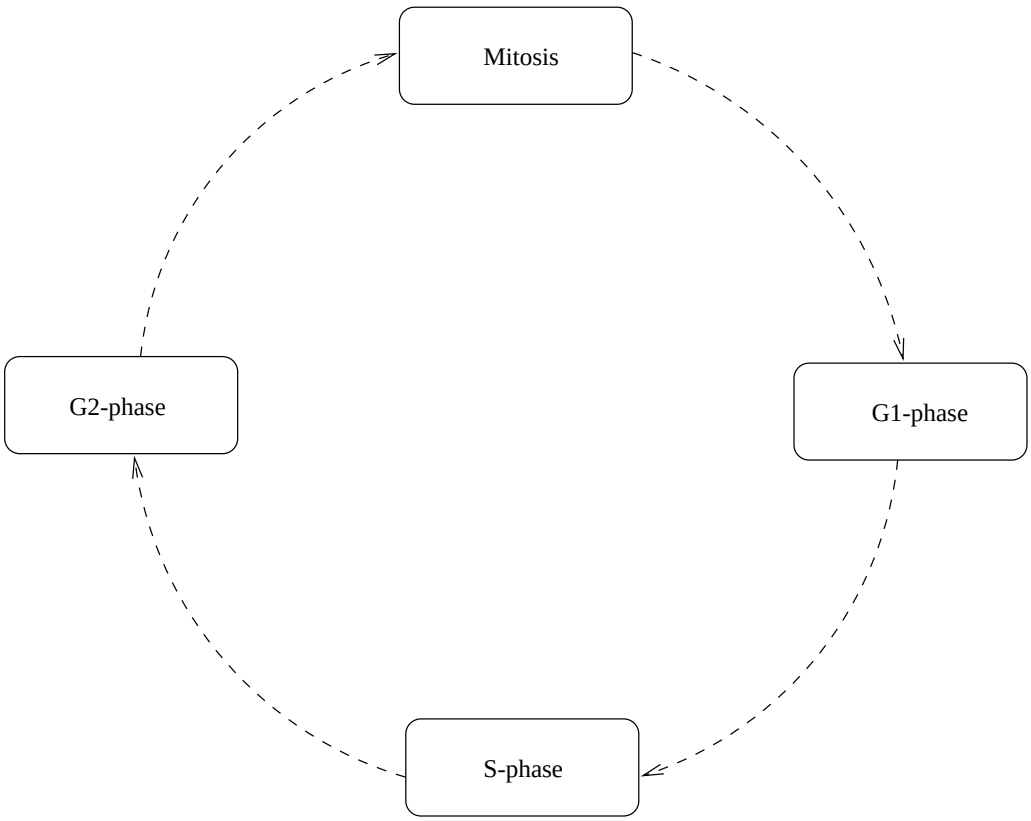

Figure 2.2: The cell cycle. During *mitosis*, cells are divided, giving each of the resulting cells identical complements of the number of chromosomes of the somatic cells of the species. The genes are expressed and the proteins synthesised in the G1 phase. In the DNA synthesis, or the S phase, the entire DNA content of the nucleus is replicated. In the G2 phase, the cell is tetraploid, i.e. having an extra chromosome.

2.5.2 Genes

The genes are the specific sequence of nucleotide bases in the DNA, and carry the heredity information. The gene contains the information required to construct *proteins*, the structural component of cells and tissues. The human genome is estimated to comprise over 100,000 genes.

The human organism can synthesise about 80,000 different kinds of proteins. The proteins are made up of large amounts of *amino acids*, usually about twenty different kinds. Three DNA bases, *codons*, direct the protein-synthesising process indirectly through the use of amino acids and mRNA[11]. For example will the base sequence ATG code for the amino acid *methionine*, which contains sulphur and is important in many body functions. The RNA resembles a single strand of DNA, and is indeed *transcribed* from the DNA in the cells nucleus. The mRNA moves from the nucleus to the *cytoplasm*, where the protein synthesis is performed. In laboratories, mRNA has been isolated, serving as a template for synthesising cDNA[12] strands for scientific purposes.

2.5.3 Chromosomes

The base pairs in the human genome are organised into 24 distinct units called chromosomes. Most human cells contain two sets of 23 chromosomes, 22 *autosomes* and an X or Y sex chromosome. The two sets are given from the parents.

When stained with dye, the chromosomes reveal a pattern of dark and light bands, visible when viewed through a light microscope. These bands are regional variations in the amount of the different base pairs. Some of the major chromosome anomalies can be detected using this process, e.g. trisomic Down's syndrome, in which the cells include a third copy of chromosome 21.

However, far from all changes in DNA can be detected using the *karyotype* method, as described above. Abnormalities due to mutations are too subtle for this method, but are still responsible for many illnesses, such as cystic fibrosis and predisposition to cancer.

[11] messenger RiboNucleic Acids
[12] complementary DNA

2.6 Extracting information about genetics

During the last years, extracting information from scientific articles about genetics has caught the attention of NLP specialists worldwide. Many undertakings have been set forth, some yielding quite good results.

2.6.1 The domain

The previous section has hopefully given a pointer as to how complex the domain of human genetics is. It is virtually impossible for any human to retain a complete overview of all the genes, proteins and chromosomes involved, or how they interact with each other.

At the same time, a vast amount of research is put into this field, all over the world. Conferences are held, articles publicised, books written; the quantity of information is excessive. Collecting all this information, and extracting the essence of it, is a task for computers. Correlating information from multiple sources, has already become a science in its own right.

The following section describes some of the ongoing initiatives.

2.6.2 Existing work

Finding a suitable ontology as a basis for the natural language system, is a key element in the process. [BCK+00] outlines a system which is based on UMLS[13]. UMLS is a gigantic system, comprising some 475,000 semantic concepts and 600,000 categorisations. The system described has a structure similar to that of TUC, displayed in figure 2.1. In addition, a new method of displaying the extracted knowledge using *keynets* is described.

The *EDGAR*[14] system, described in [RTWH00], uses a somewhat similar approach. Like the system in [BCK+00], EDGAR is also based on the UMLS ontology. Unlike the aforementioned system, EDGAR uses a stochastic part of speech tagger along with an under-specified syntactic parser to analyse the texts. The output of the parser provides input for a rule-based system that uses both syntactic and semantic information to extract factual assertions from the text.

Highlight, developed at SRI Cambridge and described in [TMO+00], is a multi-purpose NLP system customised into a system for performing information extraction from biological articles. (A strategy similar to

[13]Unified Medical Language System

[14]Extraction of Drugs, Genes And Relations

this report's way of customising TUC.) The Highlight system is highly adaptable, and, provided some simplifications of the source material, has yielded impressive results.

What sets all these systems apart from TUC, is TUC's heavy reliance on grammatical analysis. Whereas TUC rejects ungrammatical information, these systems perform template filling on partial sentences, thus making them more robust in terms of what input they accept. On the other hand, TUC may potentially extract more complex information from the material, given the deeper understanding of the semantics and grammar of the language.

3 Elementary linguistics

In order to comprehend a grammatical NLP system like TUC, a knowledge of the fundamental principles of grammar is required. The intention of this chapter is to (re-)introduce some of the key concepts of natural language, for which an understanding is crucial when dealing with TUC, or NLP systems in general. Much of what is written in this chapter is based on the excellent book, [HJL98].

The chapter concludes with a section on Naturally Readable Logic. This is a subset of natural language, with certain properties making it ideal as a basis for NLP systems. TUC, and consequently GeneTUC, is based on analysis and processing of NRL, rather than natural language.

3.1 The basics

The different word classes are the fundamental building blocks of the language. This section describes the most important word classes, their functions and use, according to [HJL98] and [GHb]. Although not exhaustive, in the sense that some classes are left out, this provides a foundation for the discussions later in the report.

3.1.1 Verbs

Verbs is the class of words used for denoting actions. These can be categorised further according to

- **Regularity**
  Verbs can be placed in the subclasses regular and irregular depending on how they are inflected in the past and past participle form.

The regular verbs are all inflected according to a general schema, whereas the irregular ones have individual patterns of inflection.

- **Transitivity**
  All verbs can be put into at least one of the three transitivity classes:
  - Intransitive - not taking object, e.g. "John *laughs*"
  - Copular - taking a subject predicative, e.g. "John *became angry*"
  - Transitive - taking one or more objects, e.g. "John *saw Mary*"

  Note that transitive verbs may or may not require an adverbial.

3.1.2 Nouns

Nouns give names to persons, places, things and concepts in general. *Common nouns* denote any member of a set of concepts, e.g. a car, thoughts, a girlfriend. *Proper nouns* give names to a specific member of the set, e.g. John, the Theory of Relativity, Oslo Airport Gardermoen.

Nouns can be derived from verbs and vice versa:

The noun *progression* is derived from *to progress*.
The verb *to house* is formed from *house*.

The *Gerund* is a type of word which can act as both a verb and a noun. It is formed as the present participle of the verb. Examples of usage:

As a noun: John likes *programming*
As a verb: John is *programming* his VCR

Nouns can also function as adjectives, modifying other nouns, as in

John likes *action* movies..
"...and a partridge in a *pear* tree."

3.1.3 Adjectives

Adjectives are words used to modify the noun, either as a part of a noun phrase or following a copular verb.

The adjective can be complemented, forming adjective phrases. These phrases are formed in four ways:

- **Adverb and adjective**
  John is *rarely late*
  This report is not *good enough*

- **Adjective and prepositional phrase**
  Mary is *fond of boxing*
  John is *sitting on the chair*

- **Compared adjectives**
  I am *taller than* you
  Mary thought *as hard as she could*

- **Adjective and subordinate clause, participle or infinitive clause**
  I'm *afraid John died*
  Mary is *good at doing nothing*
  This key is *supposed to fit*

3.1.4 Adverbs

The adverb is a word class that modify verbs, adjectives, other adverbs or complete sentences. Adverbs can be combined into adverbial phrases, with the same function as adverbs. The adverbs are grouped into three subclasses:

- **Simple**
  The first subclass is a simple modifier, e.g. "I am leaving *tomorrow*", "I'll eat my dessert *first*"

- **Interrogative**
  Interrogative adverbs are used for asking questions, e.g. "*Where* is my other sock?", "*When* was that?"

- **Conjunctive**
  The conjunctive adverbs connect independent clauses, e.g. "It was raining; *consequently*, John stayed at home", "I think; *therefore*, I am"

3.1.5 Pronouns

Pronouns are used in place of nouns. There are five principle groups of pronouns:

- **Personal**
  Personal pronouns point directly to a person or an object, e.g. "*He* is a good teacher", "Mary saw a film. *It* was scary"

- **Possessive**
  Possessive pronouns are pronouns showing ownership or possession, e.g. "Get off *my* lawn!", "The dog tried to bite *its* tail"

- **Demonstrative**
  Demonstrative pronouns focus the attention on the object pointed out, e.g. "*These* boots are made for walking", "Who's *that* girl?"

- **Reflexive**
  The reflexive pronouns point back at the noun or pronoun that has just been named, e.g. "Mary looked at *herself* in the mirror", "They've bought *themselves* a new car"

- **Relative**
  The relative pronoun joins a subordinate clause to a main clause, e.g. "John saw the girl with *whom* he was in love", "The parrot *that* I bought not half an hour ago, is dead"

3.1.6 Prepositions

Prepositions are words used to show a relationship between its object (noun or pronoun following the preposition) and another word in the sentence

In a galaxy, far, far away.
First *among* equals.

A prepositional phrase includes a preposition, the object of the preposition and a number of modifiers on the object. The prepositional phrase may have an adjectival or adverbial function

Adjectival function: The car *outside the house* is nice (the phrase gives more information on the subject)
Adverbial function: Mary looked *at the man*

3.1.7 Conjunctions

Conjunctions are employed to connect words, phrases or clauses, possibly indicating the relationship between the elements they connect in the sentence. There are three types of conjunctions:

- **Coordinating**
  Coordinating conjunctions connect elements having the same grammatical function, e.g. “Sticks *and* stones may break my bones”, “Many are called, *but* few are chosen”

- **Correlative**
  Correlative conjunctions act as coordinating conjunctions, but work in pairs to connect elements in a sentence, e.g. “*Neither* rain *nor* snow will stop him”, “I like *both* vanilla *and* chocolate”

- **Subordinating**
  Subordinating conjunctions connect two elements with different grammatical function, most commonly an independent and a dependent clause, e.g. “It looks *as though* it’s going to rain”, “*Since* you’ve been gone, I’ve been missing you”

3.2 Sentence categories

Sentences may be categorised according to function and structure. In terms of structure, the most important property is whether the verb is placed in front of or behind the subject. Also, sentences belonging to a certain category may have a function similar to a sentence from one of the other categories.

3.2.1 Declarative

The declarative sentence, in which the verb is placed behind the subject, is the most common of the major sentence groups. It is usually less marked in form and less restricted in function than the other groups. As implied by the name, declarative sentences state facts, as in

John likes to play guitar.
It was the best of times, it was the worst of times.
Mary has not eaten her peas yet.

Declarative sentences can be positive, i.e. affirm a fact, or negative, denying a fact. The first two examples above are thus positive declarative; the last one is negative.

3.2.2 Imperative

Imperative sentences are most often employed to issue a command. The imperative sentence often lacks an explicit subject and use the verb in its base form:

> Shut the door!
> Please keep your luggage with you at all times.

The subject of the sentence is, if not the addressee, often given from the context.

3.2.3 Interrogative

Interrogative sentences are used to query the addressee for information. In contrast to the declarative sentences, the verb often precedes the subject in interrogative sentences:

> Have you ever loved a woman?
> Where did all this mail come from?

Sometimes interrogative sentences have a non-interrogative function. Such *rhetorical questions* act as statements or commands, while avoiding having to use declarative sentences, which may seem blunt or obvious:

> Who cares?
> Do you mind closing the door when you leave?

3.3 Ellipsis

Ellipsis is a phenomenon often encountered in dialogue. Ellipsis is the omission of a phrase mentioned earlier in the discourse, a fact that can be inferred from the context.

Elliptic sentences are categorised as sentence fragments; words or phrases not included in a phrase structure but still carry a communicative function:

> Mary showed up late, but then again, she always used to [show up late].
> Do you know how to get there? Yes, I do [know how to get there].

3.4 Anaphora

Using personal pronouns for referring to persons or object mentioned earlier in the discourse, is called anaphora or anaphoric references. Anaphoric references require the speaker and addressee to share enough common knowledge to resolve the anaphor:

> John had no idea what *she* was talking about.
> Mary had given up on her fear of flying, and started to like *it*.

The pronoun *it* is extremely general, often making the resolution of the anaphor hard, or introducing ambiguities.

Cataphoric references is a phenomenon closely related to anaphora. A cataphoric reference is a reference dependent on something following the references, i.e. a forward pointer:

> *She* didn't know what to do with *it*, but Mary thanked politely for the gift, and placed it swiftly in the back of the closet.

3.5 Naturally Readable Logic

For a computerised system, natural language as such is far too complex to be fully understood. Rather than trying to do the (at this time) impossible, we focus on a subset of natural language, denoted by Naturally Readable Logic (NRL). As the name implies, NRL is founded on a well-defined syntax and semantic, while being as readable as natural language.

As defined in [HW94], the requirements for NRL are the following:

- NRL is definable
- NRL is acceptable as English (or other languages on ported systems)
- NRL is a logical language, all acceptable conclusions from a set of statements are verifiable
- NRL is sufficient, i.e. "everything" can be said in NRL

Note that the system needed for analysing complex semantic structures, like metaphors and analogies, is not a part of NRL.

4 Learning "Goldilocks..."

As a introduction to the field of natural language processing, it was decided to try make TUC comprehend the fairy-tale of Goldilocks and the Three Bears. This would serve as an appetiser for the later work, and also ease the process of acquiring a basis of knowledge regarding information extraction. Using a familiar domain instead of the field of genetics, was presumed to bring the focus toward learning natural language processing, and not be disturbed by the unfamiliar terminology and complex world of genetics.

4.1 Why "Goldilocks..."?

Choosing "Goldilocks and the Three Bears" as a basis for this project was quite random. But still, the story of Goldilocks is a classic fairy-tale, loved by children for generations. The language in the edition we chose [Sou71] is aimed at children just starting to read, thus the sentences and words should be quite simple to analyse and comprehend. (The former being the most important factor for our venture.)

4.2 From buses to bears

Making the leap from bus-tables in Trondheim (as is the BusTUC application's fort) to the make-belief forest, raised a number of issues which would have to be addressed. It was quite evident that the two domains were virtually disjoint. The vocabulary and the semantic net used in TUC [Bra97] would not be adequate when analysing the text in "Goldilocks...". Conversely, the semantic net required for "Goldilocks...", would not de-

pict the real world with the accuracy called for by TUC. In “Goldilocks...”, bears live in houses, sit in chairs and sleep in beds, which is surely not the case in the real world.

Even still, extending and enhancing the TUC system with new knowledge appeared more lucrative than replacing the existing knowledge. The information already in TUC provided a foundation for us to build on. Many words and terms needed were already defined and tested. This strategy worked fine, except in those cases where the added knowledge conflicted with the existing knowledge in the system. When this occurred, the path of generalisation was chosen, finding the least common denominator, rather than erasing existing knowledge. As an example: TUC only accepted *persons* to talk. In “Goldilocks...”, the three bears talk like humans. Hence, the new knowledge base accepts *agents* as such to talk.

TUC answers simple queries posed in single sentences. The language used resembles spoken language. Treating prose, as is the language in “Goldilocks...”, could complicate the matters. Would the “written” language require big changes to TUC?

4.3 New requirements

TUC was created as a multi-purpose natural language capable system. Still, system development has mainly been focused on making a competent program for answering inquiries into Trondheim Trafikkselskap’s bus tables. Extracting meaningful information from a written text, consisting of a number of interdependent sentences, put some new demands on the system.

4.3.1 Vocabulary

Extending the vocabulary meant reading through the text, marking out all verbs and nouns. Verbs not in the system had to be added, either in transitive or intransitive form, or both. These verbs were mainly “action verbs”, describing some concrete physical action, e.g. climbing, jumping, running. Also, the system had to be augmented with new verb complements, e.g.

> “Goldilocks *peeped through* the window.”
> “She *climbed onto* the bed.”

The nouns not currently in the system had to be integrated into the system’s semantic net. An excerpt of the new semantic net is shown in

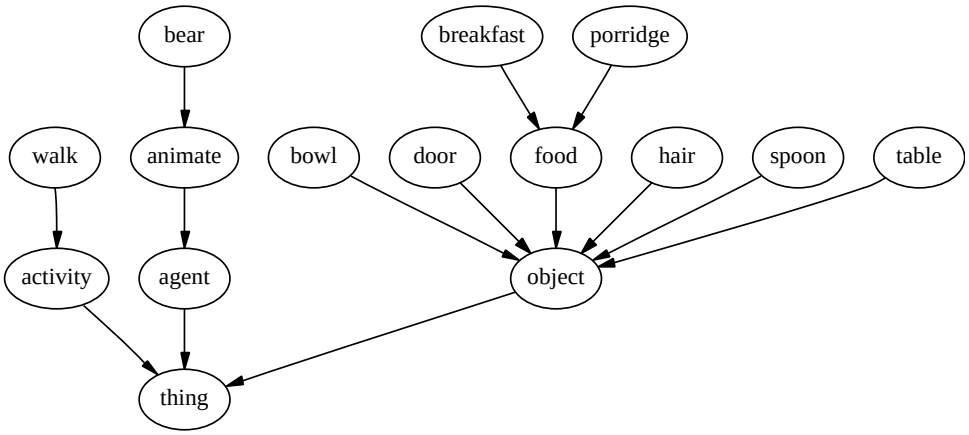

Figure 4.1: New words in the "a kind of"-hierarchy (1 of 2).

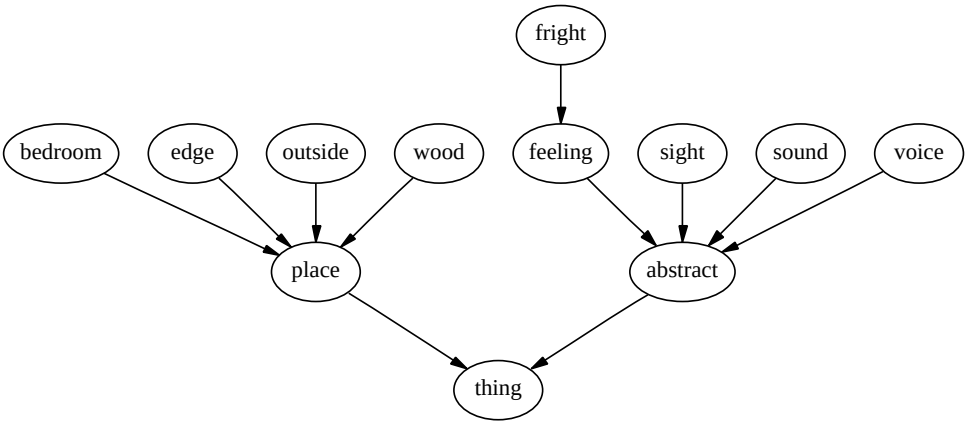

Figure 4.2: New words in the "a kind of"-hierarchy (2 of 2).

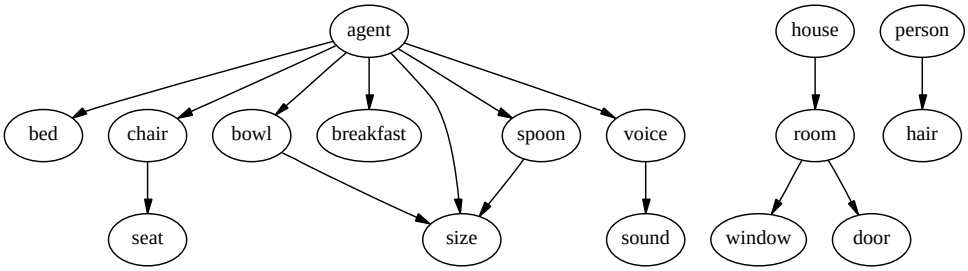

Figure 4.3: New “has a” relationships.

figures 4.1, 4.2 and 4.3. The arrows in the figures denote the *a kind of*-relationship between the two classes. Note that actual instances of an object class is kept outside the semantic net. Goldilocks *is a* girl, not *a kind of* girl, Father Bear *is a* bear and *a* father (thus making him *a person*, in lack of a species-neutral gender denomination). Effort was made to stay within TUC’s stratified hierarchy of nouns, and not go beyond the six level limit.

In addition to these, a number of adjectives and adverbs were added. These complementing the “new” words recently added, but also creating new combinations between existing words.

4.3.2 Grammar

As the project progressed, TUC’s grammar proved to be surprisingly advanced, and the changes that were required, were relatively few. On the other hand, some of the desired changes seemed to be quite difficult to implement in the current system.

The first decision that faced us, was to what extent we were to rewrite the original text. Should we require the grammar to “understand” the complete text as is, or should we rewrite it, in an effort to make it simpler for TUC to interpret? More on this in the following section.

Questions about bus tables, which is TUC’s speciality, rarely contains any reported speech. Quotations are seldom encountered, thus TUC’s mechanisms for handling these were poorly developed. On the other hand, “Goldilocks...”contains quite a lot of reported speech. Initially, it seemed tempting to try to develop TUC further in respect to this. However, when regarding the true objective of this project, information extraction from articles, understanding speech was considered less important. It was consequently decided to just disregard all occurrences of reported speech in “Goldilocks...”.

TUC handles mostly queries involving definite time references, e.g.

"When is the next bus from Gløshaugen to Torvet after 16.00?". Placing the events in "Goldilocks..."in some time frame is impossible. Still, we know that the events took place *one morning*, and we know the order in which the events occurred. These points on the time-line are represented internally as numbered Skolem constants. Potentially, one could compare the numbers of these constants to establish whether one event happened before or after another event. This feature is not implemented in the current version of TUC, but is a possible future enhancement.

4.3.3 Other changes

As stated earlier, TUC interprets single sentences separately. Any punctuation in the text is merely removed and deemed meaningless. Sentences are separated by newlines. When reading "Goldilocks..."and other large texts, having to separate each sentence by a newline character is inconvenient. Hence, a new optional flag, "textflag", was created. When "textflag" is true, TUC accepts multiple sentences, separated by periods, interpreting them one after another. This allows for the complete text to be entered in one run.

TUC has a built-in time-out limit, aborting the interpretation if the sentence is too difficult to understand, i.e. if the parsing was not successful within the given time. Analysing "Goldilocks..."'s sentences proved, in some cases, to be quite demanding on the system. All too often, the time limit expired before the system had provided any output. It was then decided to short-circuit the time-out mechanism altogether. This could leave the system exposed to eternal looping, but none were encountered. (Kudos to the programmers of TUC!)

But some of the sentences became just too complex for TUC to interpret correctly, and had to be rewritten. In plain language, the subject or verb is sometimes omitted, as it is (assumed to be) given implicitly by the context, see section 3.3. The sentence

> "She rushed to the window, jumped outside and ran quickly into the wood" ([Sou71], pg. 48)

contains a dependent clause which "inherits" the subject from the independent clause. Humans use common sense to resolve this, but, as [Nil98] points out, representing common sense in a computer system is very difficult. To help TUC out, the sentence could be rewritten as

> "She rushed to the window. *She* jumped outside and ran quickly into the wood."

splitting it into two independent sentences.

TUC handles anaphora in a first-match manner, searching backwards in the text for the first object which is a likely candidate. As each sentence is parsed separately, inter-sentence anaphora pose a problem. The anaphor "it" in

> "One morning, Mother Bear cooked some porridge for breakfast. She put *it* into three bowls." ([Sou71], pg. 6)

cannot be successfully resolved by TUC. Firstly, the pronoun "it" is usually in reference to a lifeless thing, and TUC in its current version has no way of knowing what this "it" points to[1]. Also, enhancing TUC's anaphora resolution mechanism to resolve "far" anaphora was considered too difficult at this stage. The above sentence was therefore rewritten as

> "One morning, Mother Bear cooked some porridge for breakfast. She put *the porridge* into three bowls.",

removing the anaphor completely.

Phrases like

> "Goldilocks was rather *too* heavy *for* [the chair]." ([Sou71], pg. 22)

uses the adverb "too" to signal that some capacity has been exceeded, in this case the chair's carrying capacity. For simplicity reasons, TUC treats the words "rather" and "too" as noise, and disregards them. The phrase

> "Goldilocks was heavy for the chair."

is not semantically equivalent to the former (although the difference is subtle), since it states that Goldilocks was heavy, but the chair can stand it. Nevertheless, interpreting the two statements alike was deemed an acceptable compromise.

[1] In natural language, using "it" as an anaphor often introduces ambiguity. "John's has his cat on a leash. *It* looks nice." Does the leash or the cat look nice? Or is the view of John and his cat nice?

4.3.4 Further imperfections

TUC's treatment of sets of objects is less than perfect. There is no accessible connection between a set and its members, i.e. there is no way of specifying what objects comprise the set. Querying for properties of a set, e.g. its cardinality, fails. Also, when stating

> "Once upon a time, there were three bears who lived in a little house in a wood." ([Sou71], pg. 4),

TUC will still reply "No" to the question

> "Does Father Bear live in a house in a wood?",

since there is no way of telling TUC that Father Bear is one of the aforementioned three bears.

Human natural language contains a wealth of equivalent expressions. Apart from purely synonymical words, sentences can be re-phrased, while preserving the intended meaning. The following sentences mostly convey the same meaning:

Using adjectives:

> "Mother Bear was a *medium-sized* bear." ([Sou71], pg. 4)

Using noun complements:

> "Mother Bear was a bear *of medium size*."

Using nouns, relative phrasing:

> "Mother Bear was a bear *having medium size*."

Still using nouns:

> "Mother Bear was a bear and *her size was medium*."

TUC's way of treating such phrases is flawed. Actually only the first and last phrasing is interpreted alike. Hence a query should use the exact same phrasing as did the input statement. Ideally, all of these statements would be represented on the same way internally in the system, making retrieval of such information simpler and more unified.

As TUC removes any punctuation, sentences in which the punctuation is necessary for understanding, will not process successfully. Commas separating the elements in a series, as in

> "There was a big bowl for Father Bear, a medium-sized bowl for Mother Bear and a tiny, little bowl for Baby Bear." ([Sou71], pg. 6)

is not understood by TUC. One could rephrase this using multiple and's, as

> "There was a big bowl for Father Bear and a medium-sized bowl for Mother Bear and a tiny, little bowl for Baby Bear.".

Although this sentence is grammatically correct, it's readability is drastically impaired, and it cannot be considered "natural".

4.4 Performance

When the system was enhanced to a presumably satisfactory level, TUC was fed the input text and thereafter asked a number of questions. These questions would try to unravel how much TUC had really understood of "Goldilocks...", and give a pointer as to how good TUC's querying capabilities were.

4.4.1 Input

As it was unlikely that TUC could fully understand the original text, the text was re-written a number of times, each time easing TUC's task of processing the text. TUC was thus tested against the following versions of the original text:

- The original text, unaltered and unabridged. (In appendix C)
- All anaphora resolved manually beforehand.
- All possessive pronouns converted into the definite article "the".
- All dialog removed from the text, sentences containing speech truncated at our discretion. (In appendix D)

The versions were cumulative, i.e. the last version had all anaphora resolved, all possessive pronouns converted and all dialog removed. The last version of the text contained 13 % fewer sentences than the original text, due to removal of dialog.

Text	Sentences	Success
Original	70	39 (56%)
Anaphora resolved	70	41 (59%)
Pronouns removed	70	47 (67%)
Dialog removed	64	51 (80%)

Table 4.1: Success rates when processing "Goldilocks..."

4.4.2 Questions

To check on how well TUC had understood the text, a number of queries were posed upon the system. These were mostly quite simple questions, questions you would expect every five-year-old to be able to answer. At the same time, they tested two important things:

- Whether or not TUC grasps the intended meaning of the text.
- TUC's query capability.

This phase uncovered some deficiencies, among them some rather significant insufficiencies when treating sets. As mentioned earlier in this chapter, TUC's built-in way of treating sets is a weak-point, and the problems prevailed when trying to extract knowledge about the sets from the system. It became apparent that any and every query involving sets would ultimately fail. Upon closer inspection, we found that the facts had been successfully processed and were indeed in the system's knowledge base, but correlating the stored sets with a set or object mentioned in a query was impossible (in the current version of TUC; this is a highly attractive enhancement).

The questions were based on the parts of the story which TUC understood, as it would be of no purpose to make TUC answer questions it cannot possibly know the answer to.

4.4.3 Results

The story was, as stated earlier, re-written three times to make it easier for TUC to understand. The success rates of each run is displayed in table 4.1. The table shows that TUC has an overall success rate of 80% when given a little help, which must be considered satisfactory at this early stage.

As can be seen from the results in appendix A, TUC does quite a good job of understanding "Goldilocks...". It answers 25 of the 31 answers cor-

rectly, although this number can easily be manipulated by altering the collection of questions.

The questions are kept quite simple, as it would be of no interest showing the vast amount of queries TUC cannot answer correctly. Instead, a few questions illustrating some of the more important shortcomings of the current system, have been chosen. The level of difficulty ranges from the trivial “Is the porridge hot?” to the semantically more demanding “Did Goldilocks pick up the little spoon and taste the porridge in the big bowl?”.

4.4.4 Preliminary conclusions

All in all, TUC has proven to be a viable system for processing natural language. Still, in its current state, the information extraction part of the system seems to be more fully developed than the mining part.

The input system is easily extended to accept new words and word phrases. What is not so straightforward, is introducing new grammatical phenomena, or modifying existing grammatical rules. Changing TUC’s grammar requires extensive knowledge of the system, and this might impede the further development.

TUC understands more than it is able to express. Already, a large quantity of information regarding “Goldilocks...”lies within the system, but is impossible to put to use. One example often referred is the treatment of sets. TUC can extract information regarding sets, but it can not communicate any of this information to a user. Being able to answer questions about some properties of a set, like cardinality, and also give the possibility to treat sets as atoms, would be key functionality. The latter could be done without trying to connect the set to its members; the set is a separate entity.

Finally, in a future version, TUC should have a notion of synonymical phrases. Having to pose a question using the exact same verbs and adverbs as did the original text, is not satisfactory.

5 Genetical articles

The real intention of this project, according to the assignment in section 1.1, was to make a system able to perform IE from biomedical literature.

This chapter describes the work made to accomplish this. According to what is stated earlier in the report, much time was spent creating an ontology for this domain. The founding principles of GeneTUC's ontology is therefore presented in the first sections. Then, further improvements to the GeneTUC's system are described.

5.1 Analysing the domain

Extracting all meaningful relationships between the entities in biomedical literature is an almost insurmountable task, due to the complex nature of the field of biomedicine. In the early stages of such a project, focusing on a few key relationships, and perfecting (or at least trying to perfect) how the system interprets these, seems to be a attractive path to follow. When TUC "understands" these relationships fully, the vocabulary can be extended further, yielding a system which will incrementally produce deeper knowledge of the domain.

Some of the systems mentioned in section 2.6.2 use the UMLS as the ontology on which the systems are founded. GeneTUC's ontology is based on the HUGO[1] Gene Nomenclature data base and the SwissProt Annotated protein sequence database[2]. GeneTUC, in its present form, contains over 10,000 genes (with 10,000 aliases) and more than 5,500 proteins (with 14,000 aliases). This may seem a large number, but when the human genome is estimated to comprise about 100,000 genes, it is still many too

[1]The Human Genome Organisation, http://www.hugo-international.org/hugo/
[2]http://www.expasy.ch/sprot/sprot-top.html

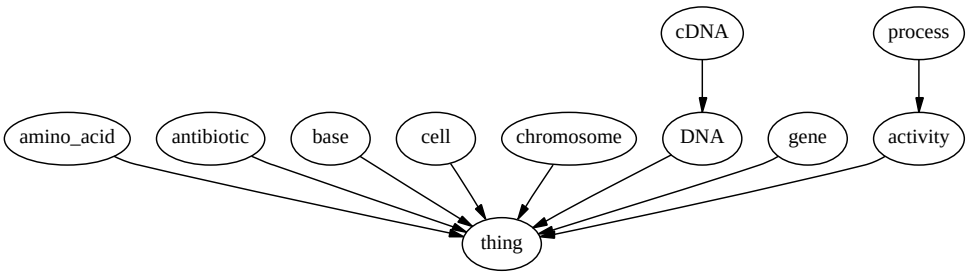

Figure 5.1: Key terms in GeneTUC (1 of 2)

few.

5.2 Ontology

The system's ontology is based on a few key terms. These terms, and how they are placed within TUC's existing semantic hierarchy, is shown in figures 5.1 and 5.2. (The choice of these terms is in part inspired by [HGP].) No further subdivision of the genes and proteins is performed.

As always in TUC, the root of the concept tree is the "thing". Most of the basic terms are placed as direct descendants of the "thing", i.e. a gene is a "thing", a protein is a "thing" and so on. This flatness of the tree is not ideal and the structure is by no means final; making a taller tree by introducing intermediary concepts will make the ontology better and the semantics more precise. Some of the concepts could have been placed differently in the tree, depending on how they are to be interpreted. The configuration in figures 5.1 and 5.2 is just an example of what was seemingly practical.

Figure 5.3 displays the "has a" relationships between the new key terms. Any and all connections between the new and old words, i.e. words in TUC's vocabulary prior to GeneTUC, is intentionally left out for the sake of graph readability.

The cell is said to "have" a nucleus, cytoplasm and chromosomes. Technically, one could argue that the chromosomes reside in the nucleus, and thus should the nucleus "have" the chromosomes. To retain the transitive connection between the cell and the chromosome, one would have do designate the nucleus to be "a part of" the cell, and the chromosome "a part of" the nucleus. But this would introduce some unwanted subtleties, as the cell would be made up of two "places" (see figure 5.2).

Proteins and DNA are the main building blocks of the chromosomes. The protein has a number of amino acids, and the DNA consists of nu-

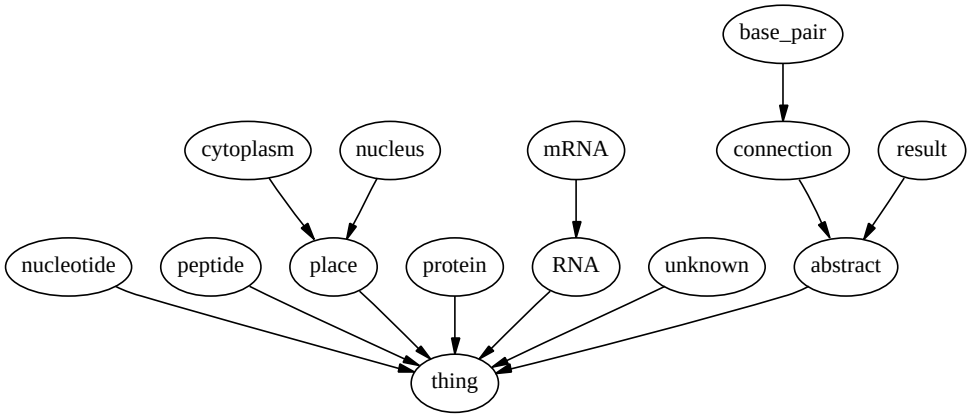

Figure 5.2: Key terms in GeneTUC (2 of 2)

cleotides (composed of sugar, phosphate and a base), genes and base pairs. The genes could be said to “have” nucleotides, this is left out in this ontology.

The reasons for choosing an ontology of such little complexity, are many. Keeping the ontology simple made it easier to maintain a complete overview on where the different concepts were placed in the hierarchy. When having to learn the basics on the human genome as the project progressed, a concise but sufficient ontology was convenient. Also, the field of human genetics is based on a limited number of concepts; although each class of concepts may include thousands of objects. At this point, going beyond the concept classes and into concrete proteins, genes or the like, would mean little more than crowding the concept tree.

5.3 Meaningful relationships

As mentioned in the previous section, focusing on a few key relationships between the key terms, would appear to be an appropriate way of dealing with the task at hand. Potentially, multiple verbs can be mapped to the same relationships. This would bypass some of the difficulties associated with equivalent statements, described earlier in section 4.3.4.

Choosing which relations to search for was somewhat arbitrary, and in some degree biased, taking the input material at hand into account. Since the number of abstracts used was limited, finding relationships with an many occurrences as possible was crucial to get an appraisal on how good the system worked.

Section 2.5 describes the process of cell division, wherein a single strand

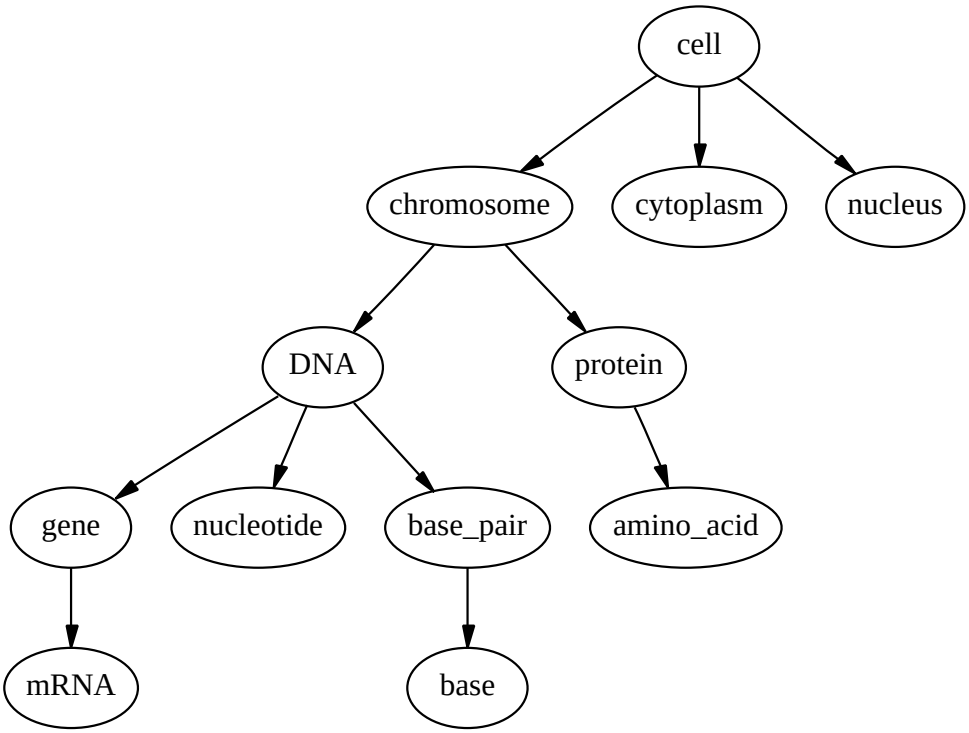

Figure 5.3: New “has a” relationships in GeneTUC

of DNA is created. This DNA strand is an exact replica of one of the DNA strands of the original cell (unless the DNA mutates). The original DNA thus *encodes* the new DNA strand. Also, when syntesizing new proteins, a strand of mRNA is created in the nucleus, through transcription. This mRNA is *encoded* by the DNA in the nucleus.

The proteins in an organism *regulate* a number of process. The progression of the cell cycle, displayed in figure 2.2, in which cells are fissioned into new cells, is *regulated* by CdK (Cyclin-dependent Kinases), MPF (Maturation Promoting Factor) and the proteins p27 and p53.

The "encode" and "regulate" relationships where thus chosen to test the initial version of GeneTUC. Testing was performed in small scale, and finding relations with numerous occurrences in the abstracts was imperative (to get a notion of the success rate).

5.4 Further improvements to TUC

When writing scientific papers, many people will have a tendency to revert to passive voice. Thus

> "Injection of RNA encoding Xenopus Bcl-2 *blocks* apoptosis induced by cycloheximide or alpha-amanitin."

becomes

> "Apoptosis induced by cycloheximide or alpha-amanitin *is blocked by* injection of RNA encoding Xenopus Bcl-2."

In the first sentence, GeneTUC would correctly identify *blocks* as the main verb, connecting the subject "injection" to the direct object "apoptosis".

The second sentence, however, would be interpreted differently, although it conveys the exact same meaning. The original version of TUC would see "is" as the main verb. Thus the adjective "blocked", which holds the essence of the sentence, is kept separate from the relation. This was revised, yielding a GeneTUC that recognises these two sentences as being equivalent.

Using adverbs to modify sentences or parts of sentences, e.g. to reduce the severity of the statement, is common in scientific literature. Bombastic statements is often considered to be bad form. Although these adverbs more often than not play a non-essential role in the statement (this is obviously not the case with adverbs like "never"), they should be kept on as modifiers for the relations for the sake of completeness. However, due

to the increased complexity of adverbs over adjectives, GeneTUC had to be revised from merely treating adverbs as "noise", i.e. unwanted and unnecessary words, to keeping them on as modifiers.

In section 4.3.3, the problem with omitted words was addressed. In the sentence

> "The catalytic activity associated with cyclin E2 complexes is cell cycle regulated."

the phrase "associated with cyclin E2 complexes" points back at the "catalytic activity". This sentence could also be written as

> "The catalytic activity *which is* associated with cyclin E2 complexes is cell cycle regulated."

emphasising this fact. Some small modifications had to be made to GeneTUC for it to be able to recognise this way of expression.

6 Results and conclusions

Although much work will still have to be done to make GeneTUC into a full-blown NLP system; equipped for performing IE on biomedical literature, some of the results seen so far are promising. GeneTUC has yet to exhibit any flaws or peculiarities making it unfit for the task intended.

This chapter comments on some of the results and observations, and draws some conclusions based on the project. See also the preliminary conclusions based on the experience with the “Goldilocks...”project, in section 4.4.4.

6.1 Results

As compared to the original TUC application, GeneTUC is bigger, in respect to the number of facts in the permanent database. Some uncertainty was connected as to how well the application would scale, in terms of memory usage and speed. What became evident, however, was that the application seemed to scale very well. The memory usage went up, as would be expected, but not alarmingly (23 MB versus 14 MB used).

The processing speed, was hardly impaired at all. Sicstus Prolog’s indexing mechanism proved to be highly efficient, even when the permanent database was extended with some 40,000 new facts. Although not measured specifically, the processing was perceived not to be more time-spending in GeneTUC than when processing “Goldilocks...”.

NRL is accurate and definable, natural language is not. The ongoing expansion of GeneTUC’s grammar, making it more and more like natural language, increases the risk of dissipating its accuracy, in terms of how well the intended meaning of the statements is caught on. GeneTUC was not tested extensively, given its current state as a project in progress. But

what has been seen so far, is quite reassuring; the accuracy of what is successfully processed, is still very high.

6.1.1 Some examples

To illustrate the current state of GeneTUC, some examples of sentences which do and do not analyse successfully are provided below.

```
E: The human cyclin E2 gene encodes a protein that is
related to cyclin E.

.....................................................
sk(5)isa protein
adj/related/sk(5)/sk(6)
event/real/sk(6)
srel/to/thing/cyclin_e/sk(6)
encode/ccne2/sk(5)/sk(7)
event/real/sk(7)
.....................................................
```

As can be seen from the system dialog above, GeneTUC successfully analyses this sentence (from [GPT+99]). The protein, which is unnamed, is represented by a Skolem constant (sk(5)).

The next sentence, which is somewhat adapted from the original text, also analyses successfully. More on the adaption further on.

```
E: The discovery of a novel second cyclin E protein
suggests that multiple unique complexes regulate
progression.

.....................................................
cyclin_e isa protein
sk(1)isa discovery
nrel/of/discovery/thing/sk(1)/cyclin_e
adj/new/cyclin_e/A
adj/second/cyclin_e/A
suggest/id/that/sk(1)/sk(2)/sk(3)
event/real/sk(3)
sk(4)isa complex
adj/multiple/sk(4)/A
```

```
adj/unique/sk(4)/A
sk(5)isa progression
regulate/sk(4)/sk(5)/sk(6)
event/sk(2)/sk(6)
.....................................................
```

A note on this sentence: As the discovery "suggests" something, GeneTUC does not feel completely confident that this indeed is the case. Therefore, GeneTUC will not tag the statement as a "real" event. (Compare lines 7 and 13 in GeneTUC's response.) Nevertheless, GeneTUC's analysis of the sentence is correct, the statement being true or not.

But the sentence is simplified from its original form, which appeared in [GPT+99]. The original version is a follows:

```
E: The discovery of a novel second cyclin E family
member suggest that multiple unique cyclin E-CDK
complexes regulate cell cycle progression.

.....................................................

--- Incomprehensible at  * ---

the discovery of novel second cyclin e family member
* suggest that multiple unique cyclin e cdk complexes
regulate cell cycle progression

.....................................................
```

There are a few noteworthy points about this sentence:

- **family member**
  Two issues here. First "family" is a noun acting as an adjective on the following noun, "member" (see section 3.1.2). Also, "member" is a meta concept, saying something about a concept rather than being a concept in its own right. As of now, GeneTUC has no elegant way of handling such meta concepts.
- **cyclin E-CDK complexes**
  Not dissimilar from the previous point. A "complex" is a meta concept, saying something about the structure of the preceding concept, namely cyclin E-CDK.

- **cell cycle progression**
  "Cell cycle" is a compound noun, with "cell" acting as an adjective on "cycle". This can be remedied by defining "cell cycle" as a text constant, although this solution is too verbose to be effective. (Imagine having to define all compound words explicitly!) Furthermore, "cell cycle" acts as an adjective on "progression".

Finally, a large example.

```
E: Our results suggest that the timing of the cell
cycle in the Xenopus embryo evolves from regulation by
 accumulation of mitotic cyclins to mechanisms
involving periodic G1 cyclin expression and inhibitory
tyrosine phosphorylation of Cdc2.

.....................................................

--- Incomprehensible at  * ---

our results suggest that the timing of the cell cycle
in the xenopus embryo evolves * from regulation by
accumulation of mitotic cyclins to mechanisms
involving periodic g1 cyclin expression and inhibitory
tyrosine phosphorylation of cdc2

.....................................................
```

This sentence is actually not as far from being accepted as one would think after taking a short glance. The difficulties are due to the following:

- **Xenopus embryos**
  Compound nouns have been addressed earlier in this section.
- **G1 cyclin expression**
  In the data base, the gene is denoted as "cyclin G1", not "G1 cyclin".
- **involving periodic G1 cyclin expression and inhibitory tyrosine phosphorylation of Cdc2**
  GeneTUC seems to have some problems dealing with the conjunction. Everything preceding "phosphorylation" is just marked adjectives, working on the latter. Furthermore, GeneTUC has some problems with "of Cdc2", for reasons yet to be established. (The Cdc2 gene is defined in the permanent base.)

6.2 Conclusions

The GeneTUC project is still a project very much in progress. Drawing any final conclusions would therefore be premature and possibly erroneous. However, some observations regarding the system can be made at this early point in time.

Updating the semantics is a fairly straight-forward task. New words and dependencies between them, in the concept hierarchy, can be readily added without having to change the grammar. Phrases and complements are added effortlessly, thus making GeneTUC highly versatile within the limits of its defined grammar.

Changing the grammar is more difficult. In contrast to the semantics, a deep understand on how the grammar works is costly, in terms of hours spent studying. However, the need for changes to the grammar is much smaller than the need for changes to the semantics. Furthermore, the grammar is beginning to become well-covering and robust.

Speed is also an issue. Adding a large number of facts to GeneTUC's internal data base meant no substantial decrease in speed. Analysing a sentence, if successful, will take about a second, although this has not been measured specifically. The system may be fed any number of sentences and left to work on its own, returning messages on the success or failure of each sentence.

In conclusion, GeneTUC would seem like a system fit for this task, extracting information from biomedical literature. The reliance on the input conforming to a "good" grammar, can imply that pursuing processing of article titles may seem futile. The titles are often ungrammatical, in the sense that a verb is commonly left out. Focusing on the abstracts, in which the information is highly condensed but grammatical, GeneTUC could prove to be a valuable supplement to the existing flora of NLP systems.

7 Future work

This chapter discusses some possible future enhancements of GeneTUC, given that the project is continued. The chapter is split in two; the first section discusses work related to TUC and GeneTUC, the second section treats issues pertaining to IE in general.

7.1 GeneTUC

Some improvements can, and will eventually have to, be made for GeneTUC to reach a state where it yields acceptable results. Some of these improvements are quite simple, others require some degree of redesign of the system.

GeneTUC's mechanism for dealing with sets, has been commented on earlier in this report. Though sets have not been encountered nearly as often in the biomedical literature as in "Goldilocks..."(which in essence is based on the adventures of the three bears), some way of relating a set to its members is desired. Somewhat related to this, is the notion of concepts which inherit the properties of other concepts. Consider the sentence

> "Multiple unique cyclin E-CDK *complexes* regulate cell cycle progression."

in which the concept "complexes" denotes compounds of cyclin E, a protein, and CDK, an enzyme. A "complex" here is no concept of its own, apart from stating that cyclin E-CDK is non-atomic, compounded in some way. The interpretation of "complex" is thus dependent on what the complex is made up of. As of now, TUC (or indeed, GeneTUC), cannot resolve this dependency.

It would also be convenient to have some method of marking the origin of the facts in the semipermanent database. This can possibly be accomplished be adapting the event calculus on which the system is based, or making a separate predicate for this use.

A more trivial enhancement, is making a better mechanism for treating compound nouns, where nouns act as adjectives for other nouns. As of today, this can only be done on a word to word (or phrase to phrase) basis. Whether to allow all possible compositions of nouns, or restricting them according to some rules, will have to be addressed, but the verbose manner in which this is done at present, is not satisfactory.

Synonymical phrases, rather than just words, are often encountered. “I went for a walk” means, in essence, the same as “I walked”. When filling the data base, the latter is the preferred phrasing, as it is more concise and “to the point”. Some way of mapping such equivalent phrases to the same fact in semipermanent memory, is much desired.

When creating an ontology, some concepts will inevitably fit in multiple places. In “Goldilocks...”, “Mother Bear” could be classified both as a “mother” and a bear. On encounter of “Mother Bear” in the input, TUC has to decide which of the two interpretations to use. Once this decision is made, it is final (in the context of the statement). Appendix B shows a sample dialog illustrating this. Making the system capable of resolving this more dynamically, would be a valuable enhancement of the system.

As for the mechanisms used for retrieving information from GeneTUC’s semi-permanent data base, two enhancements seem pertinent. Being able to retrieve the properties of a given concept, in other words being retrieving the adjectives and adverbs used to describe a noun, is quite useful, also seen in conjunction with the problem with compound nouns.

And finally, retrieving all relations between given concepts, i.e. listing all the verbs acting upon the specified noun, is necessary for the system to become complete. This is actually what is called *story summarisation*, and a field of NLP in its own right. Nevertheless, a system aspiring to become a useful addition to the existing biomedical NLP systems should include this capability.

7.2 Information extraction

Through the early stages of the GeneTUC project, an observation was made regarding the way of expression in article abstracts. It was noted that the language used was more condensed and less verbose, than in the article body. This is probably because the author’s wish to compress as

much information as possible into the abstract, without using much space. It would be interesting to analyse this further, is this an observation based on a true phenomenon, or is it just a coincidence?

Finding sources of information, other than scientific articles, could also be pursued. Retrieving information from books, conference proceedings etc., may also provide valuable information on the domain.

References

[Ask] AskJeeves.com. http://www.askjeeves.com.

[BCK+00] Kenneth Baclawski, Joseph Cigna, Mieczyslaw M. Kokar, Peter Mager, and Bipin Indurkhya. Knowledge representation and indexing using the Unified Medical Language System. In *Pacific Symposium on Biocomputing*, 2000.

[Bra97] Jon S. Bratseth. Bustuc - a natural language bus traffic information system. Master's thesis, Norwegian University of Science and Technology, 1997.

[CCT] The Cell Cycle and Mitosis Tutorial. http://www.biology.arizona.edu/cell_bio/tutorials/cell_cycle/main.html.

[GHb] Writer's Workshop - Grammar Handbook. http://www.english.uiuc.edu/cws/wworkshop-/grammarmenu.htm.

[GPT+99] J. M. Gudas, M. Payton, S. Thukral, E. Chen, M. Bass, M. O. Robinson, and S. Coats. Cyclin E2, a novel G1 cyclin that binds Cdk2 and is abarrantly expressed in human cancers. *Molecular and Cellular Biology*, January 1999.

[HGP] The Science Behind the Human Genome Project. http://www.ornl.gov/hgmis/resource/info.html.

[HJL98] Hilde Hasselgård, Stig Johansson, and Per Lysvåg. *English Grammar: Theory and Use*. Universitetsforlaget, 1998.

[HW94] Steffen Leo Hansen and Helle Wegener, editors. *Topics in Knowledge-based NLP systems*, chapter Tore Amble: Domain

Modelling and Automated Natural Language Interfaces. Samfundslitteratur, DK 1970 Frederiksberg C, 1994.

[MED] Medline. http://www.nlm.nih.gov/databases/freemedl.html.

[Nil98] Nils J. Nilsson. *Articicial Intelligence: A New Synthesis*. Morgan Kaufmann Publishers, Inc, 1998.

[NLP] Natural Language Processing FAQ. ftp://rtfm.mit.edu/pub/usenet-by-hierarchy/comp/ai/nat-lang.

[RTWH00] Thomas C. Rindflesch, Lorraine Tanabe, John N. Weinstein, and Lawrence Hunter. Edgar: Extraction of Drugs, Genes And Relations from the biomedical literature. In *Pacific Symposium on Biocomputing*, 2000.

[Sou71] Vera Southgate, editor. *Goldilocks and the Three Bears*. Ladybird Books Ltd, 1971.

[SS99] Don R. Swanson and Neil R. Smalheiser. Implicit text linkages between Medline records; using Arrowsmith as an aid to scientific discovery. http://kiwi.uchicago.edu/libtrends.html, March 1999.

[TMO+00] James Thomas, David Milward, Christos Ouzounis, Stephen Pulman, and Mark Carroll. Automatic extraction of protein interactions from scientific abstracts. In *Pacific Symposium on Biocomputing*, 2000.

A Questions and answers

- Where do the bears live?
  TUC's reply: (No reply)
  Incorrect. TUC cannot answer questions regarding the set of bears.

- Is Father Bear big?
  TUC's reply: `Yes`
  Correct.

- Is Mother Bear big?
  TUC's reply: `No`
  Correct. Mother Bear is medium-sized.

- How big is Baby Bear?
  TUC's reply: `little`
  Correct.

- Is Father Bear bigger than Mother Bear?
  TUC's reply:

  ```
  --- Incomprehensible at  * ---

  is father bear bigger than mother bear ? *
  ```

  Incorrect. TUC cannot compare adjectives, and the notion of "bigger than", "newer than", "fewer than" and alike, is unknown. (But still, TUC has a go at it; this query takes for ever to process.)

- What does Mother Bear cook for breakfast?
  TUC's reply: `porridge`
  Correct.

- Is the porridge hot?
  TUC's reply: `Yes`
  Correct.

- What is the girl called?
  TUC's reply: `goldilocks`
  Correct.

- Did Mother Bear cook porridge one evening?
  TUC's reply: `No`
  Correct. She cooked it one morning.

- Who cooked porridge?
  TUC's reply: `m_bear`
  Correct. (TUC's internal name of Mother Bear.)

- Was the door open?
  TUC's reply: `Yes`
  Correct.

- Did the porridge smell good?
  TUC's reply: `Yes`
  Correct.

- Did Goldilocks pick up the small spoon and taste the porridge in the big bowl?
  TUC's reply: `No`
  Correct.

- Did Goldilocks pick up the big spoon and taste the porridge in the big bowl?
  TUC's reply: `Yes`
  Correct.

- What size is Father Bear's bowl?
  TUC's reply:
  Incorrect. The adapted text does not explicitly state that a certain bowl belongs to Father Bear. (A bowl *for* Father Bear is not equivalent.)

- How many bowls are there?
  TUC's reply: `2`
  Incorrect. This is even more incorrect than it may seem. The answer

is not at all connected to the set of bowls, but rather TUC's enumeration of the objects which are a bowl in TUC's knowledge base.

- Was the porridge cool?
  TUC's reply: `No`
  Correct.

- Was the porridge hot?
  TUC's reply: `Yes`
  Correct.

- Was the porridge hot and lumpy?
  TUC's reply: `Yes`
  Correct. TUC cannot discriminate between the porridge in the three bowls. To TUC, it is all porridge.

- Who sat in the big chair?
  TUC's reply: `goldilocks`
  Correct.

- Was the chair high?
  TUC's reply: `Yes`
  Correct.

- Who was too heavy for the little chair?
  TUC's reply: `goldilocks`
  Correct.

- Did the big chair break?
  TUC's reply: `No`
  Correct.

- Did the small chair break?
  TUC's reply: `Yes`
  Correct.

- Did she go into the bedroom?
  TUC's reply: `Yes`
  Correct.

- Did she feel tired?
  TUC's reply: `Yes`
  Correct.

- Was she asleep?
  TUC's reply: `Yes`
  Correct.

- What did Father Bear look at?
  TUC's reply: `chair`
  Partially correct. He looked at both the chair and the bed.

- Did Father Bear look at the chair and at the bed?
  TUC's reply: `Yes`
  Correct.

- Did Father Bear see the chair?
  TUC's reply: `No`
  Incorrect. Although this is semantically (almost) equivalent with "Did Father Bear look at the chair?", TUC has no concept of equivalent modes of expression.

- What does Goldilocks do?
  TUC's reply:
  Incorrect. This is story summarisation. In its current version, TUC is not able to perform this operation. Thus, no answer.

B Multiple categorisations

Some of TUC's problems with dynamic resolution of an object's class, is illustrated by the following dialog with the system:

```
E: Mary is a woman.

.....................................................
mary isa woman
.....................................................



E: Mother Bear is a bear.

.....................................................
m_bear isa bear
.....................................................



E: Mother Bear thinks that she died.

.....................................................
think/id/that/m_bear/sk(3)/sk(4)
event/real/sk(4)
die/mary/sk(5)
```

```
event/sk(3)/sk(5)
.....................................................



E: Who died?

.....................................................
[which(A)::(A isa agent,die/A/B,event/real/B)]

.....................................................


mary


E: Is Mother Bear a woman?

.....................................................
[test::(m_bear isa mother,event/real/A)]

.....................................................


Yes
```

As can be seen above, the anaphora resolution in statement three incorrectly identifies Mary as “she”, and not Mother Bear, even though TUC knows that Mother Bear is also a woman, or, more precisely, a mother.

C Original text

Once upon a time there were three bears who lived in a little house in a wood. Father Bear was a very big bear. Mother Bear was a medium-sized bear. Baby Bear was just a tiny, little bear.

One morning, Mother Bear cooked some porridge for breakfast. She put it into three bowls. There was a big bowl for Father Bear, a medium-sized bowl for Mother Bear and a tiny, little bowl for Baby Bear. The porridge was rather hot, so the three bears went for a walk in the wood, while it cooled.

Now at the edge of the wood, in another little house, there lived a little girl. Her golden hair was so long that she could sit on it. She was called Goldilocks. On that very same morning, before breakfast, Goldilocks went for a walk in the wood. Soon Goldilocks came to the little house where the three bears lived. The door was open and she peeped inside. No-one was there so she walked in.

Goldilocks saw the three bowls of porridge and the three spoons on the table. The porridge smelt good, and Goldilocks was hungry because she had not had her breakfast. Goldilocks picked up the very big spoon and tasted the porridge in the very big bowl. It was too hot! Then she picked up the medium-sized spoon and tasted the porridge in the medium-sized bowl. It was lumpy! Then she picked up the tiny, little spoon and tasted the porridge in the tiny, little bowl. It was just right! Soon she had eaten it all up!

Then Goldilocks saw three chairs; a very big chair, a medium-sized chair and a tiny, little chair. She sat in the very big chair. It was too high! She sat in the medium-sized chair. It was too hard! Then she sat in the tiny, little chair. It was just right! But was the tiny little chair just right? No! Goldilocks was rather too heavy for it. The seat began to crack and then it broke. Oh dear! Goldilocks had broken the tiny, little chair and she

was so sorry.

Next Goldilocks went into the bedroom. There she saw three beds; a very big bed, a medium-sized bed and a tiny, little bed. She felt tired and thought she would like to sleep. So Goldilocks climbed up onto the very big bed. It was too hard! Then she climbed up onto the medium-sized bed. It was too soft! Then Goldilocks lay down on the tiny, little bed. It was just right! Soon she was fast asleep.

Soon the three bears came home for breakfast. Father Bear looked at his very big porridge bowl and said in a very loud voice, ”Who has been eating my porridge?” . Mother Bear looked at her medium-sized porridge bowl and said in a medium-sized voice, ”Who has been eating my porridge?”. Baby Bear looked at his tiny, little porridge bowl and said in a tiny, little voice, ”Who has been eating my porridge and eaten it all up?”. Next Father Bear looked at his very big chair. ”Who has been sitting in my chair?” he asked in a very loud voice. Then Mother Bear looked at her medium-sized chair. ”Who has been sitting in my chair?” she asked in a medium-sized voice. Then Baby Bear looked at his tiny, little chair. ”Who has been sitting in my chair and broken it?” he asked in a tiny, little voice.

Next the three bears went into the bedroom. Father Bear looked at his very big bed. ”Who has been lying on my bed?” he asked in a very loud voice. Mother Bear looked at her medium-sized bed. ”Who has been lying in my bed?” she asked in a medium-sized voice. Baby Bear looked at his tiny, little bed. ”Here she is!” he cried, making his tiny little voice as loud as he could. ”Here is the naughty girl who has eaten my porridge and broken my chair! Here she is!”. At the sounds of their voices, Goldilocks woke up. When she saw the three bears, she jumped off the bed in fright. She rushed to the window, jumped outside and ran quickly into the wood. By the time the three bears reached the window, Goldilocks was out of sight. The three bears never saw her again.

D Adapted text

Once upon a time there were three bears who lived in a little house in a wood. Father Bear was a very big bear. Mother Bear was a medium-sized bear. Baby Bear was just a tiny, little bear.

One morning, Mother Bear cooked some porridge for breakfast. She put the porridge into three bowls. There was a big bowl for Father Bear, a medium-sized bowl for Mother Bear and a tiny, little bowl for Baby Bear. The porridge was rather hot, so the three bears went for a walk in the wood, while the porridge cooled.

Now at the edge of the wood, in another little house, there lived a little girl. Her golden hair was so long that she could sit on the hair. She was called Goldilocks. On that very same morning, before breakfast, Goldilocks went for a walk in the wood. Soon Goldilocks came to the little house where the three bears lived. The door was open and she peeped inside. No-one was there so she walked in.

Goldilocks saw the three bowls of porridge and the three spoons on the table. The porridge smelt good, and Goldilocks was hungry because she had not had the breakfast. Goldilocks picked up the very big spoon and tasted the porridge in the very big bowl. The porridge was too hot! Then she picked up the medium-sized spoon and tasted the porridge in the medium-sized bowl. The porridge was lumpy! Then she picked up the tiny, little spoon and tasted the porridge in the tiny, little bowl. The porridge was just right! Soon she had eaten the porridge all up!

Then Goldilocks saw three chairs; a very big chair, a medium-sized chair and a tiny, little chair. She sat in the very big chair. The chair was too high! She sat in the medium-sized chair. The chair was too hard! Then she sat in the tiny, little chair. The chair was just right! But was the tiny little chair just right? No! Goldilocks was rather too heavy for the tiny little chair. The seat began to crack and then the tiny little chair broke. Oh dear!

Goldilocks had broken the tiny, little chair and she was so sorry.

Next Goldilocks went into the bedroom. There she saw three beds; a very big bed, a medium-sized bed and a tiny, little bed. She felt tired and thought she would like to sleep. So Goldilocks climbed up onto the very big bed. The bed was too hard! Then she climbed up onto the medium-sized bed. The bed was too soft! Then Goldilocks lay down on the tiny, little bed. The bed was just right! Soon she was fast asleep.

Soon the three bears came home for breakfast. Father Bear looked at the very big porridge bowl. Mother Bear looked at the medium-sized porridge bowl. Baby Bear looked at the tiny, little porridge bowl. Next Father Bear looked at the very big chair. Then Mother Bear looked at the medium-sized chair. Then Baby Bear looked at the tiny, little chair.

Next the three bears went into the bedroom. Father Bear looked at the very big bed. Mother Bear looked at the medium-sized bed. Baby Bear looked at the tiny, little bed. At the sounds of their voices, Goldilocks woke up. When she saw the three bears, she jumped off the bed in fright. She rushed to the window, jumped outside and ran quickly into the wood. By the time the three bears reached the window, Goldilocks was out of sight. The three bears never saw Goldilocks again.