Proper Name Knowledge Acquisition for Text Understanding

Runli Guo

Submitted for the Degree of Doctor of Philosophy from the University of Surrey

Centre for Knowledge Management
Department of Computing
School of Electronics, Computing and Mathematics
University of Surrey
Guildford, Surrey GU2 7XH, U.K.

February 2002

© Runli Guo 2002
Abstract

Current work in proper name analysis is focused on identification and limited categorisation of names. Some research has been carried out in acquiring knowledge of proper names from the contextual information within texts. In this study, we investigate how to transform human-oriented compilations, which contain a rich knowledge of proper names, into formally-represented knowledge for computer consumption.

If a dictionary is considered the source of knowledge about common nouns, an encyclopaedia should be regarded as the source of knowledge about proper nouns. Considering the amount of work on knowledge acquisition from dictionaries, comparatively little work has been carried out on extracting knowledge about proper nouns from encyclopaediae. Here we discuss the knowledge with which proper names are related and present our methods for knowledge acquisition from dictionaries of biography.

Our analysis of biographical entries leads to the observation that there are indeed repetitive patterns used in biographies, in both English and Chinese. These patterns provide the possibility of transforming the knowledge of proper names from text descriptions to encoded knowledge. Taking a sublanguage approach, we report how the content of a dictionary of biography can be mapped onto a knowledge base with a minimum amount of human intervention.

ABKAS, Archetype of a Biographical Knowledge Acquisition System, has been implemented. It comprises the pre-processing unit, the sublanguage parser and the knowledge base constructor. The key component of ABKAS is its sublanguage parser which consists of a number of finite state machines. Based on the local grammar we have identified, the sublanguage parser achieves the syntax-semantic mapping. Biographical entries in both English and Chinese are parsed and their corresponding logical forms are generated and further represented in knowledge bases.

Key words: Proper Name Analysis, Knowledge Acquisition, Encyclopaedic Knowledge, Sublanguage.

Email: r.guo@eim.surrey.ac.uk
WWW: http://www.surrey.ac.uk/
Acknowledgements

I would like to express my sincere thanks to Professor Khurshid Ahmad for his continuous support and encouragement. I appreciate the combination of rich knowledge, great insight and witty humour with which he has motivated me and guided my research. My thanks also go to Dr Andrew Hippsley for sharing his linguistic knowledge and to Dr Lilian Tang for stimulating discussions.

Many see undertaking research as an isolating experience, even more so for a student coming from a country with a very different culture and background. I have been fortunate in forming great friendships during my time in the Department of Computing at the University of Surrey. Dr Tracey Bale, whom I thank for proofreading this thesis several times and for always being there when I needed help. Christine West, who has offered tremendous help in improving my English and my writing style. Duncan White, who helped me with Latex related problems and gave me advice on Perl programming.

I am grateful to the Department of Computing, School of Electronics, Computing and Mathematics for offering me an opportunity to work on a project in which I have a great interest, and also for their joint financial backing with CVCP who offered me the Overseas Research Scholarships.

My thanks are further extended to my current employer QAS and my colleagues, especially Jeremy Morris and Andrew Abraham, whose support has made the final stage of writing up easier.

I am lucky to have a loving and supportive family. My thanks go to my father and mother, my father and mother-in-law and my sister and brothers. My special thanks are due to my brother, Runxiang, who has placed great hope in me and guided me throughout my life.

Last but not least, words are not enough to express my thanks to my husband, Tim Southerwood, who has made my PhD study a life changing experience. It is his love and affection which has made me discover strength I never thought I had and kept me going all the way.
Contents

List of Figures .......................................................... vii
List of Tables ............................................................ viii

1 Introduction .............................................................. 1
  1.1 Unknown Lexicalisations for Text Understanding .................. 3
  1.2 Proper Name Analysis ............................................. 5
  1.3 Special Language in Dictionaries of Biography .................. 6
  1.4 ABKAS - Archetype of a Biographical Knowledge Acquisition System ... 9
  1.5 Outline of the Thesis ............................................ 12

2 Proper Name Analysis ................................................ 15
  2.1 Introduction ..................................................... 15
  2.2 Names and Naming Practice - Onomastics ...................... 16
  2.3 Linguistic Characteristics of Proper Names .................... 19
  2.4 Computational Work in Proper Name Analysis ................... 32
  2.5 Conclusion ..................................................... 43

3 A Sublanguage in Dictionaries of Biography ....................... 44
  3.1 Introduction ..................................................... 44
  3.2 Towards a Sublanguage of Definitions ......................... 46
  3.3 Structured Definitions in Biographical Dictionaries .......... 57
  3.4 Representing PN Definitions .................................. 65
  3.5 Conclusion ..................................................... 71

4 Archetype for Proper Name Knowledge Acquisition .................. 73
  4.1 Introduction ..................................................... 73
  4.2 System Specification .......................................... 74
  4.3 System Design .................................................. 79
  4.4 System Implementation ........................................ 92
  4.5 Conclusion ..................................................... 100
5 System Evaluation of ABKAS
  5.1 Introduction .......................................................... 102
  5.2 Assessing the complexity of the test data .......................... 104
  5.3 Evaluation of PN knowledge acquisition ............................ 107
  5.4 Evaluation of PN knowledge bases ................................. 113
  5.5 Conclusion ............................................................. 116

6 Conclusions and Future Work ........................................... 117
  6.1 Contributions of the Research ...................................... 118
  6.2 Future Work ........................................................... 120

A Organisations for Place Names .......................................... 123

B Penn Treebank Tag Set Used in LT.POS ................................ 124

C JASMINE Part-of-Speech Tag Set ........................................ 126

D Web Interface of LT.POS .................................................. 129

E Web Interface of JASMINE ................................................ 131

F Tagged Chinese Input ..................................................... 133

G Translated Chinese Output via BabelFish ................................ 136

H ABKAS for Chinese ........................................................ 140

I Example Entries ABKAS Failed to Parse ............................... 143
  I.1 Examples of Failed Entries in English ............................. 143
  I.2 Examples of Failed Entries in Chinese ............................. 145
# List of Figures

1.1 The frame representation for biographical entries ........................................... 8
1.2 Interactions between ABKAS and other servers for English ................................. 10
1.3 Interactions between ABKAS and other servers for Chinese ................................. 11
1.4 The distributed archetype in ABKAS ............................................................... 11

2.1 The taxonomy of the form of PNs by Quirk et al. (1985) ................................... 20
2.2 The semantic categorisation of PNs by Paik et al. (1996) .................................. 41

3.1 The workflow of term base analysis ................................................................. 52
3.2 A snapshot of ConceptBase ................................................................................. 71

4.1 Overview of ABKAS ......................................................................................... 74
4.2 Superordinate bootstrapping process .................................................................. 77
4.3 The system architecture of ABKAS ..................................................................... 79
4.4 The workflow of the sublanguage parser ......................................................... 83
4.5 The finite state machine for hinge ..................................................................... 84
4.6 The finite state machine for Type $A_1$ ............................................................. 84
4.7 The finite state machine for Type $A_2$ ............................................................. 86
4.8 The finite state machine for Type $A_3$ ............................................................. 87
4.9 The finite state machine for Type $A_4$ ............................................................. 87
4.10 The workflow of knowledge base construction ................................................. 89
4.11 Stage one of ABKAS - preprocessing ............................................................. 95
4.12 Stage two of ABKAS - parsing ......................................................................... 96
4.13 Stage three of ABKAS - knowledge base construction ..................................... 97
4.14 A demonstration of the generalisation ............................................................. 98
4.15 An example of the encapsulation ..................................................................... 99
4.16 An illustration of the deduction ....................................................................... 100

D.1 A screen shot of the LT.POS interactive demo .................................................. 129
List of Figures

E.1 A screen shot of the JASMINE interactive demo .................. 131

H.1 ABKAS for Chinese - preprocessing ............................. 140
H.2 ABKAS for Chinese - parsing .................................. 141
H.3 ABKAS for Chinese - knowledge base construction ............ 142
# List of Tables

2.1 The classification of naming practices by Allerton (1987) .......................... 17
2.2 The syntactic classification in Allerton (1987) ........................................... 21
2.3 The semantic classification by Allerton (1987) ........................................... 24
2.4 Classification of the Chinese language by Li & Thompson (1990) ................. 28
2.5 The representatives of the early IE systems ............................................... 33
2.6 The MUCs and their corpus domains ......................................................... 34
2.7 Comparison between SNOWY and ABKAS ................................................ 43
3.1 Definition types in Cobuild by Barnbrook and Sinclair (1994) ....................... 50
3.2 A comparison between entries from different English biographical dictionaries 55
3.3 A comparison between entries from different Chinese biographical dictionaries 57
3.4 The structure of Type $A^P_N$ ........................................................................ 63
3.5 The structure of Type $A^P_N$ ........................................................................ 63
3.6 The structure of Type $A^P_N$ ........................................................................ 64
3.7 The structure of Type $A^P_N$ ........................................................................ 65
4.1 Definition patterns of Type $A^P_N$ ............................................................... 75
4.2 Mapping between part-of-speech patterns to semantic interpretation ............. 76
4.3 Mapping table of the tag translation ............................................................. 82
4.4 Languages used in ABKAS ............................................................................. 93
4.5 Servers and their functions within the distributed archetype ......................... 94
5.1 Proportion across Type $A^P_N$'s in PODoB (1997) and ChinaOnInternet ...... 105
5.2 Content of the test bed one with known input .............................................. 108
5.3 Performance for acquiring superordinates ..................................................... 108
5.4 Performance for acquiring discriminators ..................................................... 109
5.5 Content of the test bed with novel input ....................................................... 111
5.6 Performance for acquiring superordinates from novel input ......................... 111
5.7 Performance for acquiring discriminators from novel input ......................... 112
5.8 Results of the evaluation of PNKBs ............................................................. 114
Chapter 1

Introduction

As Natural Language Processing (NLP) moved from a purely research-based field to one which has many industrially and commercially applied aspects, it has had to process naturally occurring unrestricted texts as opposed to artificially constrained examples. Names of persons, places and things dominate such texts.

Proper names, language expressions which are used for entities, usually including people, organisations, locations, commodities, dates and times, are sometimes referred to as "unknown lexicalisations" to indicate that they are "words that are not recognised as part of the language by the system" (Boguraev & Pustejovsky, 1996:14). The understanding of names is regarded as central to the analysis of free texts by most researchers in the NLP community (McDonald, 1996:21).

The development of systems in this area requires: (a) Understanding of the linguistic behaviour of this special class of nouns and (b) Access to machine-readable compilations of names of persons, especially from dictionaries of biography, gazetteers of place names amongst other compilations. My work focuses on the second aspect (b).

Some NLP programs, for example named-entity recognition systems, are expected to identify proper nouns in free texts. However, it is equally important to extract information from the texts once the names have been recognised; for example, the possible relationships the named persons may have with other persons and events, not necessarily included in the texts, but still significantly important in understanding the texts as a whole.

Proper names help to situate and contextualise a given text in which they occur. In order to elucidate the situation and contexts, it is important to understand the position of the name
within the network of the names and events. For us, if a dictionary of biography can be transformed onto data structures within a database for a named entity recognition program, this will help in the process of identification. Furthermore, if the context of the database can be mapped onto a knowledge base, then the knowledge can be used to represent, and then reason over, the professional, social and political network in which the referred person is situated. This for us will be for the next generation of named-entity IE systems.

As reference books, encyclopaedias share a number of characteristics with dictionaries. The similarities suggest that work on machine-readable dictionaries (see, for example, Boguraev & Briscoe, 1989; Barnbrook & Sinclair, 1994) can be adapted to deal with the analysis of encyclopaedia entries for the purpose of constructing lexical resources for proper names. As a special type of encyclopaedia, dictionaries of biography contain a wealth of information about the individuals recorded in them. This information can be used to make inferences about individuals referred to in free texts, such as newspapers, art reviews and magazines.

John Sinclair and Geoff Barnbrook (1994) have studied the structure of definitions used in a learner's dictionary, the COBUILD Student Dictionary (1990). These researchers regard 'definitions as a set of sentences in a sublanguage' and suggest that these sentences are 'declarative' and written in a restricted syntax. This implies that there may exist a 'special grammar', which governs the behaviour of the special language of definitions.

In this thesis, we adapt the sublanguage approach, which has been successfully deployed in dictionary analysis, into the analysis of the special language in dictionaries of biography. Our aim is to transform the knowledge embedded in biographical entries from its textual form to a knowledge representation formalism, for instance, Telos, an object-oriented knowledge representation schema.

Typically, a proper name analysis system has focused on one single language, which in itself is a challenging task. However, it is important to validate our hypothesis that there are, indeed, distinctive and repetitive patterns in dictionaries of biography even across different languages. Therefore, we reinforce our theoretical framework by studying Chinese proper names within the sublanguage used in Chinese biographical entries. Furthermore, the computational framework has been extended into Chinese.
1.1 Unknown Lexicalisations for Text Understanding

Proper names constitute a special part of the language positioning themselves as a boundary between the word knowledge and world knowledge. Amsler regarded proper names as the missing lexica which "are not quite lexical in nature" (1989:12). Language understanding requires, amongst other things, linguistic knowledge, as well as world knowledge. In order to understand a sentence that contains a proper name, the listeners or readers are required to apply their own knowledge about the referent of the proper name. This also applies to NLP applications which require deep understanding of texts.

The basic problem of NLP, as suggested by Carbonell & Heyes (1990), is its ambiguity. Five types of ambiguities have been identified: syntactic ambiguity; word sense ambiguity; case ambiguity, referential ambiguity, and literalness ambiguity. The difficulties in resolving such ambiguities are often due to a lack of world knowledge. Examples are given here to demonstrate the importance of world knowledge related to proper names in natural language understanding.

**Syntactic (or Structural) Ambiguity**

John saw the Grand Canyon flying to New York.

The ambiguity in the above sentence lies in resolving the subject for the action of ‘flying’. Is the subject ‘John’ or ‘the Grand Canyon’? Knowing that as a static geographical entity, it is impossible for ‘the Grand Canyon’ to fly may aid the disambiguation.

**Word sense ambiguity**

(1) Mary Smith is reorganising her cabinet.
(2) Tony Blair is reorganising his cabinet.

Now consider the above two sentences, the word ‘cabinet’ refers either to Sense 1 below, which refers to a case or cupboard with doors or shelves, or Sense 2, which is a government body with ministers, following the definitions of ‘cabinet’ in the Merriam Webster Collegiate Dictionary (2000):

**Sense 1:** a case or cupboard usually having doors and shelves.

**Sense 2:** a body of advisors of a head of state.
Chapter 1. Introduction

An NLP system, does not naturally have the world knowledge that ‘Tony Blair’ is the British Prime Minister. In order to simulate the way humans understand natural language we also need to provide computers with this kind of knowledge. Imagine if a computer system could process the biographical entries\(^1\) below, extract the encyclopaedic knowledge and represent it as part of its knowledge base, then it should be able to distinguish between the above two sentences as a human reader would.

**Blair, Tony** (Anthony Charles Lynton Blair; 1953- ) British politician, prime minister since 1997 and chief architect of Britain’s New Labour Party.

Furthermore, if we replace Smith in the first sentence with Robinson, the meaning of ‘cabinet’ in this sentence could again be changed into Sense 2 according to the biographical entry below:

**Robinson, Mary** (Mary Bourke; 1944- ) Irish lawyer and stateswoman who became the first woman president of Ireland (1990-1997).

Case

He ran the mile in four minutes.

the Olympics. (cf. Carbonell & Hayes, 1990)

The preposition ‘in’ can be used to indicate either temporal or locational relationships. The knowledge of ‘the Olympics’ plays the key role in the disambiguation process.

Referential

In October 1986, the political career of Jeffrey Archer, deputy chairman of the Conservative Party, came to a halt when allegations surfaced in the News of the World of his relationship with prostitute Monica Coghlan ... ... When the Daily Star repeated the story, he sued. The novelist won GBP 500,000 damages.

Taking the excerpt of a newspaper article above, it may be difficult for an NLP system to resolve the coreferential relationship between Jeffrey Archer and the novelist. However, with encyclopaedic knowledge about the proper name ‘Jeffrey Archer’, this problem could be resolved.

\(^1\)These biographical entries are taken from Who’s Who in the 20th of Century.
Literalness

Caspar Cazzo is no Pavarotti! (cf. Allan, 1995)

Proper names can be used connotatively. In the example above, ‘Pavarotti’, the famous opera singer’s name, is used in the description of the singing ability of ‘Caspar Cazzo’.

Proper names can also be used to represent other entities which are closely related. For example,

Wall Street is in a panic. (cf. Lakoff, 1987)

From the above discussion, one can see that knowledge about proper names, as part of world knowledge, plays an important role in language understanding as well as in communication. However, we are by no means claiming that the representation of world knowledge of proper names can resolve the problems in NLP. These examples are merely presented to illustrate how the bigger picture of natural language understanding could be improved by acquiring and representing the knowledge of proper names.

1.2 Proper Name Analysis

Computational work in proper name analysis started when researchers identified proper names as the important, but poorly represented, lexica in the computational lexical resources for larger scale NLP applications (Amsler, 1989; Coates-Stephens, 1992). The frequent occurrences of proper names in real world texts and their importance in text understanding makes it an essential task.

The identification of proper names marks the first step of proper name analysis. It involves recognising the boundaries between proper names and other types of language expressions and isolating them from the remaining texts (see, for example, Rau, 1991).

Based on identification, proper names are often categorised into several generic categories in order to capture their basic semantic properties. In MUC competitions, proper name analysis is often referred to as the named-entity recognition task. In this task, names including persons, places, organisations and time are required to be identified and categorised into these basic semantic categories (see, for example, MUC-6, 1995; MUC-7, 1998).
Chapter 1. Introduction

There has also been work undertaken to extract profiles of proper names from contextual information (Coates-Stephens, 1993; McDonalds, 1996; Radev & McKeown, 1997). Although contextual information has been used as practical means for obtaining knowledge of proper names in news analysis, it is not always found in typical texts. For instance, consider the excerpt of an art review article that follows:

A precursor of what was to be called the Dada movement, and ultimately its leading member, was Marcel Duchamp, who in 1913 created his first ready-made (now lost), the "Bicycle Wheel," consisting of a wheel mounted on the seat of a stool.

Moreover, the contextual information is useful for offering a snapshot of an named entity which can be either time-dependent or context-dependent. Knowledge acquisition from contexts is very useful for understanding of texts within the same context but the extracted knowledge may be difficult to integrate or transfer.

Two main types of textual resources have been used in the automatic knowledge acquisition for large scale lexica: MRDs and text corpora (Boguraev & Pustejovsky, 1996). However, text resources used in lexical knowledge acquisition for proper names are typically text corpora.

In Coates-Stephens' work (1992) on the acquisition of proper names, both the analysis and the experiments were based on a collection of news articles taken from UK newspapers and the Wall Street Journal. Message Understanding Conferences almost exclusively take news corpora as the test beds for the evaluation of Information Extraction (IE) systems (MUC-6, 1995; MUC-7, 1998). This phenomenon is largely due to the low coverage of proper names in machine-readable dictionaries (MRDs) as reported by Sampson (1989).

Work in lexical acquisition from MRDs leads us to consider whether reference books like encyclopaediae could be used in the same way for the automatic knowledge acquisition of proper names.

1.3 Special Language in Dictionaries of Biography

Dictionaries and encyclopaediae are closely related to each other. Just as dictionaries help us to understand a new concept, some encyclopaediae provide us with knowledge about proper nouns. According to Landau (1996:5), "a dictionary is a book which lists words in
alphabetical order and describes their meaning; an encyclopaedia is a collection of articles about every branch of knowledge.” It is further suggested that the focus of encyclopaedic entries, however, tends to provide knowledge about people, places and things.

A particular kind of encyclopaedia, a dictionary of biography, comprises encyclopaedic data; furthermore, each entry is very specific to a particular person. In other words, the lemma, that is, the entry word(s) in a biographical dictionary, is a personal name. Given that dictionaries of biography, or indeed any existing dictionaries, are compiled by humans, the assumption of a lack of structure or consistency may be well founded. However, there have been developments in lexicography with respect to the possible existence of structure in the definitions compiled by lexicographers for dictionaries used by the general public (Sinclair et al., 1994). Given the similarities between these compilations, there may be certain common structures within dictionaries of biography.


**Einstein, Albert** (1879-1955) German-born physicist, a thinker of astounding insight, author of the special and general theories of relativity, and winner of the 1921 Nobel Prize for Physics for his work on the photoelectric effect.

The entry appears to follow a certain template:

Surname, First Name (Date of birth - [Date of Death]) [Nationality at Birth] [Nationality] Profession [and secondary profession] Elaboration


巴金

中国政治协商会议全国委员会主席，中国作家协会主席，作家。

1904年11月生，四川成都人。 … …

The translation of this entry is also offered in the same dictionary of biography as:

**Ba Jin** (b. Nov 1904) Native of Chengdu, Sichuan

Vice Chairman of National Committee of CPPCC; Chairman of Chinese Writers Association; Writer … …
It may be observed that the Chinese biographical entry follows a similar templated structure but with a different ordering of the elements. There are also characteristics of this particular dictionary of biography. For instance, the positions of the referent in government were given special importance and listed before one's profession. This is due to the nature of this particular publication. Moreover, it is a Chinese national dictionary of biography. Nationality is replaced by the specific region the referent originally came from, for instance 'Native of Chengdu, Sichuan'.

Developing the theme of a template for biographical entries in dictionaries of biography and considering how proper names are defined, one can argue that the entries can be mapped onto a Minskian frame (Minsky, 1975) (see Figure 1.1).

![Figure 1.1: The frame representation for biographical entries](image)

'Frames' are used to represent prototypes. Compare this situation with a list or record where the data is essentially codified or encrypted. The data represents the descriptions of the objects explicitly and enables them to become amenable to computer-based inference of implicit knowledge. A frame representation will help in leveraging the data through reasoning mechanisms like inheritance and encapsulation. A frame-based proper name knowledge-based system will help in elaborating on a person (or place) and in interrelating persons and places. Such a system can be used to create a network of people and their relationships to their work whether in scientific circles or art communities. A Proper Name Knowledge Base System will facilitate the implementation of robust and accurate question answering and deep text understanding systems.
1.4 ABKAS - Archetype of a Biographical Knowledge Acquisition System

Before one embarks on a computer-based proper noun extraction/inferencing system, it is important to look at the evidence of such templates in other areas of lexicography and terminology. In particular, it is important to establish whether or not these slots and fillers in the frames have a linguistic basis.

The work of John Sinclair in the organisation of general purpose dictionaries, and of their constituent entries in particular has led to the publications of a series of CoBuild dictionaries. Barnbrook has worked with Sinclair on a project related to so-called a language of definitions. This work has been adapted to convert the contents of terminology dictionaries and databases into knowledge-bases (Ahmad & Al-Jabir, 1998). These researchers have exploited the fact that a terminology dictionary is comprised almost exclusively of nouns, and out of the nineteen templates in CoBuild, there are only two templates for nouns. In this thesis, we attempt to demonstrate how Sinclair's language of definition can be adapted for converting dictionaries of biography into a proper noun knowledge base.

More specifically, this thesis seeks to study the sublanguage in dictionaries of biography for the extraction, representation and the reasoning of lexical knowledge for proper names. The ultimate goal of this study is to build lexical resources on unknown lexicalisations for NLP applications. Thus, by understanding proper names, greater intelligence can be achieved in text understanding.

Our major hypothesis is when compiling a dictionary of biography, biographers use similar structures in writing biographies for a collection of named entities and these structures are governed by a sublanguage.

There are three main objectives in this research. The first is to broaden the scope of the claim made about the definitions of common nouns (Sinclair, et al. 1994) in that there are repetitive patterns within the dictionary definitions. In this thesis, we wish to investigate the possibility of extending such claim to the definition of proper names. In order to achieve this goal, we have studied the local grammar which governs the syntax in the definition part of biographical entries. The sublanguage and the local grammar allow us to develop a parser to draw mappings between the syntactic dependencies and the semantic relationships.

The second objective is to develop a computational framework to exploit the notion of sub-
Chapter 1. Introduction

Figure 1.2: Interactions between ABKAS and other servers for English

language existing within dictionaries of biography and integrate the language analysis with knowledge representation. It is our aim to build a computer system to take the advantage of the special grammar in the biographical entries for the purpose of knowledge acquisition. In short, we focus on developing a methodology for constructing knowledge bases for proper names from a special kind of textual resource - encyclopaediae.

Another objective is to test and validate both the theoretical framework and the computational framework using Chinese - a typologically unrelated language to English.

Definitions for commons nouns or proper nouns, consumed by human readers, are often written in free text form. However, these entries also follow a certain template in that the structure is repeatedly used by lexicographers, terminologists and as we found out, by biographers. Such templates have added a dimension of universality in that the entries in different languages use similar templates for the definitions of proper names.

With the increasing popularity of the World Wide Web, many linguistic tools are provided online for the public. We can find many stochastic taggers offering tagging services from a remote server; for instance, LT.POS for English and JASMINE for Chinese. For reason of conveniences and ease of maintenance, it is preferable to use these taggers. Indeed, these are the taggers we have chosen.

Figure 1.2 illustrates the interactions between ABKAS and the remote or local servers for the analysis of English biographical texts and the construction of Proper Name Knowledge Bases (PNKBs). In the diagram, the rectangle enclosed by the dotted line separates the internal components of ABKAS from the external servers. Input English texts are sent to LT.POS, which is physically located in Edinburgh University, for part-of-speech tagging (see
Appendix D). In order to build a hierarchical taxonomy of professions, definition analysis has been carried out in a recursive manner. This process is called Superordinate Bootstrapping in our work. Definitions of professions are retrieved from online dictionary Wordsmyth and analysed to generate a taxonomy of professions. The encoded knowledge is forwarded to the ConceptBase Server which is installed on a Solaris server - Atlas.

In Figure 1.3, note that we have used JASMINE to replace LT.POS for the part-of-speech tagging of Chinese texts. JASMINE only offers an online tagging service for other websites (see Appendix E). We need to submit the URL links of ChinaOnInternet to JASMINE in order to get the tagged biographical texts. BabelFish server is used for the automatic translation of the Chinese outputs in order to reuse the taxonomy generated for English biographical knowledge base. Finally, taking advantage of ConceptBase's client-server architecture, ABKAS acts as its client and feeds knowledge to the ConceptBase Server.

In short, ABKAS can be considered as a widely distributed system as shown in Figure 1.4.
Chapter 1. Introduction

It is implemented and tested under the Linux environment. The majority of the source code is implemented in Perl 5.6 and the knowledge base constructor developed in Java 1.2.

Classical IE systems are often developed for a given application in a specified domain. In recent years, active research has been carried out in *adaptive information extraction* to develop techniques and methods to move IE systems to new extraction domains automatically (Grishman, 2001). Adaptive IE is becoming increasingly popular due to the usage of Internet and Intranet technologies, especially for processing information gathered from heterogeneous sources (Kushmerick & Thomas, 2002). Extraction rules can be learned through tagged training corpora. Ciravegna (2001) introduces how rules can be induced in an incremental manner via an original algorithm called \( (LP)^2 \).

ABKAS does not learn on its own and it is restricted to the biographical domain. Its ability to adapt to multiple sources of texts relies on similar linguistics characteristics found in various dictionaries of biography, i.e. the existence of the local grammar.

ABKAS may be considered as a classical IE system which extracts information from texts. However, it is different from IE systems in many ways. Typical IE systems often analyse news report or journal papers. ABKAS focuses on a special type of reference books, dictionaries of biography. The output of a typical IE system is the template form of the information, whereas ABKAS first generates the templates as its intermediate representation form. The final output for ABKAS is a knowledge base of proper names. Furthermore, ABKAS is designed to process two typologically different language with few changes in its architecture.

In this thesis, we will focus on the architecture of ABKAS's theoretical framework, sublanguage characteristics of biographical entries and on the problem related to processing Chinese language and finally the problem and opportunities to create knowledge bases. Issues related to adaptive information extraction will be discussed towards the end of the dissertation.

1.5 Outline of the Thesis

This thesis examines the potential merit of constructing a computational lexicon for proper names from encyclopaediae. In this work, we have taken dictionaries of biography as textural resources for the construction of the knowledge of personal names.

The originality of this work can be considered in three respects. The first innovative point is due to the observation of the existence of a sublanguage in the definition part of dictionaries
of biography. This lays down the theoretical foundation which addresses the question of how the knowledge in encyclopaedic entries can be extracted and transformed into its logical form. Encyclopaediae, designed for human readers, usually exist in a textual format and are presented on an entry by entry basis following the alphabetical order of the headwords. Our analysis on biographical entries has suggested that similarly there are repetitive patterns in the language used to compile them.

These patterns have made it possible to extract and represent the knowledge of proper names which leads to the second innovative point - the development of the computational framework to utilise the local grammar for lexical knowledge acquisition and representation. This framework is implemented in ABKAS, a distributed archetype which transforms biographical entries into frame-based templates and further represented the templates in ConceptBase - an object-oriented knowledge base system. This framework may be used to generate lexical knowledge bases for proper names on a large scale, thus providing lexical resources for other NLP applications.

Our work in Chinese is used as a validation method for our hypothesis initially observed in English biographical entries. This constitutes the third innovative point by suggesting the distinctive and repetitive patterns we have used are not restricted solely to English. Furthermore, it indicates that it is possible to port our computational framework to other languages.

The thesis contains a further five chapters. In Chapter 2 the study of proper names is reviewed from a number of perspectives. Proper names are rooted in the local culture and often show the idiosyncrasies of the language of the culture. We have studied personal names and place names from an onomastic point of view. This is followed by the linguistic study of proper names including syntactic, semantic and pragmatic aspects. The linguistic characteristics of English proper names and Chinese proper names will be compared and contrasted. The significant part of the literature review concerns the computational work on proper name analysis. Current work on proper name analysis has concentrated on bridging lexical gaps. Knowledge gaps seem to remain as a challenging issue. Our solution to such problems is to use encyclopaediae as text resources in the knowledge acquisition for proper names. We propose to take dictionaries of biography, a type of encyclopaedia, as the text resource in order to acquire knowledge of personal names.

Chapter 3 starts with the study of the presentation and organisation of dictionaries of biography and the nature of the biographical entries. The language used by biographers when
compiling dictionaries of biography has been studied. Our analysis suggests that the existence of distinctive patterns shares a strong resemblance with the structure of definitions for common nouns. This observation forms the theoretical framework of this work.

Chapter 4 describes the specifications, design and implementation of ABKAS which aims to transform biographical entries from textual form to the logical form. The system architecture has been presented with the description of each component.

ABKAS will be evaluated from both a quantitative and a qualitative point of view. The results, which will be reported in Chapter 5, demonstrate that we have achieved automatic knowledge acquisition of proper names from encyclopaediae.

This thesis is concluded in Chapter 6 by summarising the three lines of inquiry we have made by carrying out this study. Future work is also proposed in this chapter, related to improving the performance and flexibility of the current sublanguage parser. Further proposals have been made in extending our sublanguage analysis in dictionaries of biography to other forms of encyclopaediae which may themselves have potential sublanguages.
Chapter 2

Proper Name Analysis

2.1 Introduction

The emergence and the growth of the Internet and the World Wide Web have rapidly increased the amount of information available to the public, particularly textual information. This information has been made electronically accessible in various ways, for instance, news wires, electronic journals, online dictionaries and encyclopaediae. It has led to dramatic changes in the area of natural language processing (NLP). There is a major demand for technologies to build efficient and intelligent search engines. Typical NLP applications are no longer limited to research-based prototypes which are designed to work with artificially constructed examples. New generations of NLP tools are geared towards tackling naturally occurring texts, including vast amounts of financial and political news articles, as well as academic and scientific material. These texts typically contain a significant proportion of proper names.

Proper names are language expressions describing entities, usually including people, organisations, locations, commodities, dates, times and so on. The study of proper names is an interdisciplinary enterprise. Since Plato, many philosophers, psychologists, linguists and even poets have contemplated the puzzle: "What is in a name?" Many philosophers believe that proper names are not merely rigid designations of individuals. Gottlob Frege (1949) demonstrated a rich variety of meanings that can be carried by proper names. Jerpersen (1965:67) supported Frege and argued that proper names are rich in connotation, even more so than common names.

From linguistic and socio-linguistic angles, Carroll (1985:164) argued that the "meaning" of proper names may be functionally, but not logically, grounded, whereas the meaning of
common terms and definite descriptions is grounded both logically and functionally. Proper names do contain knowledge but this knowledge is encyclopaedic rather than lexicographical in nature (Amsler, 1989). If the encyclopaedic knowledge about proper names could be automatically extracted and represented as part of a lexical knowledge base, greater machine intelligence could be achieved for text understanding.

This chapter offers a survey on the subject of proper names from various perspectives. The study of proper names, that is the study of unknown lexicalisations, requires some understanding of the language and the culture from which the names are taken. It is hoped that this study will bring insights to the analysis of proper names for text understanding. Names will be studied in relation to naming practices and names from different cultures and origins will be compared. A review of the linguistic characteristics of both English proper names and Chinese proper names will be provided. We will be discussing the syntactic, semantic and pragmatic aspects of proper names in comparison with common nouns aiming to seek answers to the following questions: 1) Do proper names have knowledge? 2) If so, what kinds of knowledge do they have? 3) How can we extract this knowledge? This is followed by state of art computational work on proper name analysis which suggests that despite the increased attention on proper name analysis, work needs to be done in automatic acquisition and representation of the knowledge of proper names. Finally, encyclopaedic texts will be proposed as the resources for the encoded knowledge for proper names.

2.2 Names and Naming Practice - Onomastics

The study of names and naming practices has formed its own discipline known as onomastics (Crystal, 1987). It is usually divided into anthroponomastics, the study of personal names, and toponomastics, the study of place names. In practice, however, the term onomastics is used for the former and toponymy for the latter. It is impossible to make a clear division between the two since personal names are often borrowed to name a place (e.g. Churchill, Washington) and vice versa. The Cambridge Encyclopaedia of Language offers an overview of onomastics (1987).

Many organisations have been founded to find out what is really in a name and to investigate the cultural insights, settlement history and linguistic characteristics revealed in names. Founded in 1951, the American Name Society is devoted to the study of names, especially place and personal names, but also proper names and terminologies of all kinds, including the
names in literature. Other significant societies include the Canadian Society for the Study of Names. A list of the organisations and their URLs can be found in Appendix A.

### 2.2.1 Personal names

Proper names are rooted in the local culture and sometimes show the idiosyncrasies of the language of the culture. Therefore, it is hardly surprising that naming systems vary greatly from language to language. European names often contain a 'given name' (or 'Christian name') and family name (or 'surname'). In most European languages, these names are also referred to as 'first' and 'last' names simply because the given name is followed by the family name. Hungarian, Japanese and Chinese names, however, are in the reverse order (e.g. Jiang Zemin). One or more middle names are widely used in some societies. In the US, people often prefer to only use the initial of their middle name (e.g. George W. Bush).

When there is a sequence of names, there may be variations in the levels of importance within different communities. In Britain, Timothy John Southerwood would usually be referred to as Timothy; in Germany, the more important name is put nearer to the surname, therefore, Carola Ruth Siegel would be referred to as Ruth.

According to Allerton (1987), the naming practice can be classified into five categories (see Table 2.1). The given names in Western society are usually picked out from a limited list, such as John and Mary. Such names tend to be semantically opaque. Although there are exceptions like Heather and Hazel which have relations to flowers or plants, the associations to their original semantic meaning are gradually blurred over the years. With globalisation, the name lists keep growing by borrowing from other languages. In some languages, given names can be descriptive, for example, the name Dances-with-Wolves. This naming practice

<table>
<thead>
<tr>
<th>Category</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Repertory-selected individual names</td>
<td>John, Mary</td>
</tr>
<tr>
<td>Descriptive individual names</td>
<td>Dances-with-Wolves</td>
</tr>
<tr>
<td>Patronymics/matronymics</td>
<td>Ivanovich</td>
</tr>
<tr>
<td>Provenance names</td>
<td>John of Gaunt</td>
</tr>
<tr>
<td>Family or clan names</td>
<td>Baker, McDonald</td>
</tr>
</tbody>
</table>

Table 2.1: The classification of naming practices by Allerton (1987)
is especially common in aboriginal languages.

In some societies, a name can be derived from the father's given name. In Russian, Ivan's son would be called Ivanovich. This naming practice is known as *patronymics*. The opposite of this phenomenon, *teknonymy* is common in the Arab culture where a parent is named after a child. For example, the name Abu Ahmad suggests that the person is the father of Ahmad. Patronymic names take different forms according to the specific culture. In Russian, the patronymic is placed between the child's given and family names. In Icelandic, the patronymic serves as the surname, which changes with each generation. In English, patronymic prefixes and suffixes appear in family names (e.g. Robinson). Similar cases happen in Scottish (*Mac/Mc-*), Irish (*O*), Welsh (*Ap*), Polish (*-ski*) and Greek (*-poulos*).

Another category of naming practice concerns provenance names which are often used as the names of English noblemen, such as *Richard of Gloucester*. In some oriental languages, family names are sometimes given after the original place where the family used to live, for example *Guo* means the boundary of the city in Chinese and *Yama* means mountain in Japanese.

### 2.2.2 Place names

Places names are another major topic in naming practices. A society's history, beliefs and values are reflected in the names which have been given to their surroundings. The early explorers of America introduced all kinds of names from different cultures including Spanish, French, Dutch, Indian and English. Similarly Celtic, Roman, Anglo-Saxon, Scandinavian and Norman names provide great insights into British history.

The study of place names covers both the natural geographical entities in the world, such as seas, mountains, rivers and lakes as well as the man-made places and institutions, such as streets, houses and bridges. Some place names can be interpreted either as geographical entities or as political or governmental institutions, such as countries, provinces and counties (Crystal, 1987).

The naming process is a rather creative process. A place name can be derived from a personal name which has religious or historical significance. For instance, *St Neots* and *Santa Cruz* are named after saints. Some place names are related to the royal family, e.g. *Victoria*, whereas others are named after their discoverer or conqueror, such as *Columbus, Washington*. There are also names following famous events; a town near Hastings is called *Battle* after the famous battle of Hastings.
Place names can vary greatly between languages both morphologically and semantically. A map of England produced in France would refer to England as Angleterre. Moreover, China in Chinese is Zhong Guo which means the country in the centre instead of a country being famous for producing ceramic china, as people outside China might understand it. When a name is introduced in another language, the original meaning can be lost. For example, Beijing means the northern capital. Nanjing means the southern capital. When used in English, the meaning in both of the names is no longer obvious.

The study of proper names is a complex subject because it is related to many other subjects including history, archaeology, geography and, more importantly linguistics. The next section looks into linguistic issues and studies the syntax, semantics and the pragmatic aspects of proper names.

2.3 Linguistic Characteristics of Proper Names

Studies in theoretical linguistics are usually the foundation of works in computational linguistics. Therefore, the linguistic study of proper names forms the baseline of a proper name analysis system. Early studies of proper names by linguists share the same topic areas with philosophers. They also deal with whether or not they have meaning, how they refer to entities and how they behave differently from common nouns. However, modern linguists tend to focus on the syntactic and semantic structure of proper names. In this section, we will first look at the syntactic attributes of proper names and then move on to the semantics. It is hoped that the detailed analysis of proper names from a linguistic point of view will provide insights for building a computational model.

2.3.1 Linguistic features of English proper names

Allerton (1987:62) suggested that proper names “fall partly inside and partly outside the lexicon and grammar of average speakers”. Compared with ordinary words and word-sequences, such as common nouns and common noun phrases, proper names possess special properties both grammatically and semantically. In this section, the grammatical properties of proper names will be examined regarding the internal structure of proper names and the grammatical constraints of the surrounding context. The semantic issues involve the meaning of names and the proposed semantic taxonomy for classification.
Chapter 2. Proper Name Analysis

2.3.1.1 The form of proper names - syntax

Formal linguistic characteristics of proper names include their internal grammatical structure as well as external grammatical structures.

Internal structure and taxonomy The grammatical structure of a proper name can be as simple as one word and it also can be very complex. In Quirk et al. (1985), English proper names with articles and names without articles were discussed separately. The category of names with no articles, personal names (e.g. Tony Blair), temporal names (e.g. Christmas) and geographical names (e.g. London) are discussed respectively. Names with articles have been divided into three types according to their internal structure; those without modification (e.g. The Times), those with pre-modifications (e.g. the English Channel) and those with post-modifications (e.g. the House of Commons) (see Figure 2.1).

Allerton (1987) studied proper names from both linguistic and socio-linguistic points of view. He categorised proper names into pure proper names (e.g. London), mixed proper names (e.g. London Bridge), common-based proper names (e.g. the Yellow River) and coded proper names which cover acronyms (e.g. IBM). The four types of proper names were further classified according to their syntactic structure. This taxonomy is illustrated in Table 2.2.

One of the special attributes of proper names is that each proper name tends to have more than one form. The form of proper names can vary through the "ellipsis of elements", a phenomenon mentioned by Quirk et al. (1985:295). Carroll (1985:143-160) described his in-depth investigation of the relationship between established names and their variant forms. He has observed that a place name usually consists of a category word and a namestem. The category word classifies the name into a particular category, e.g. square in Trafalgar Square.

Figure 2.1: The taxonomy of the form of PNs by Quirk et al. (1985)
<table>
<thead>
<tr>
<th>Type</th>
<th>Pattern</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pure PNs</td>
<td>Simple PN</td>
<td>London</td>
</tr>
<tr>
<td></td>
<td>Simple PN + article</td>
<td>the Hague</td>
</tr>
<tr>
<td></td>
<td>Simple PN + title</td>
<td>Mr. Tony Blair</td>
</tr>
<tr>
<td>PNs + CNs</td>
<td>PNs + CNs</td>
<td>Mexico City, Hyde Park</td>
</tr>
<tr>
<td></td>
<td>article + PN + CN</td>
<td>the Suez Canal</td>
</tr>
<tr>
<td></td>
<td>restrictive adjective + proper noun</td>
<td>Northern Island</td>
</tr>
<tr>
<td></td>
<td>PN + article + restrictive adjective or CN</td>
<td>Henry the Eighth</td>
</tr>
<tr>
<td></td>
<td>The + (adjective + )</td>
<td>the University of Surrey</td>
</tr>
<tr>
<td></td>
<td>CN + of + PN</td>
<td></td>
</tr>
<tr>
<td>Common-based proper names</td>
<td>the + adjective + head noun</td>
<td>the Red Sea</td>
</tr>
<tr>
<td></td>
<td>the + modifier noun + head noun</td>
<td>the Labour Party</td>
</tr>
<tr>
<td></td>
<td>adjective + noun</td>
<td>Green Park</td>
</tr>
<tr>
<td></td>
<td>modifier noun + head noun</td>
<td>Salt Lake City</td>
</tr>
<tr>
<td>Coded proper names</td>
<td>with the article</td>
<td>the FBI</td>
</tr>
<tr>
<td></td>
<td>without the article</td>
<td>IBM, NATO</td>
</tr>
</tbody>
</table>

Table 2.2: The syntactic classification in Allerton (1987)
The namestem is the non-classifying part which distinguishes the name from other names in the same category, e.g. *Trafalgar* in *Trafalgar Square*. Carroll observed that sometimes the category word could be omitted. For instance, *New York City* can be called *New York*. When such shortening process occurs, the namestem is called a *namehead*. He has given evidence to show that the namehead phenomenon was neither entirely lexical nor grammatical in nature.

The principle of the shortening process is that it only takes place when the reader or listener is provided with a clear referent with no ambiguity. This criterion contains two "pragmatically constrained" properties, familiarity and uniqueness. If a namestem does not provide a unique identifier within the context, the shortening process will not happen, and therefore the namestem cannot become a namehead.

Another aspect of the variation is the polymorphy of proper names which is caused via transliteration. For instance, *London* in French is *Londres* and in Italian is *Londra*. Compared to other parts of speech, proper names tend to have more or less constant forms across languages. This feature again positions proper names as partly inside and partly outside language. Allerton (1987:82) compared proper names from various semantic categories and suggested that the more language specific a name is, the more prone it is to translation. Temporal names tend to have language-specific forms. Organisation names are the second most prone to translation. Personal names rarely get translated, although alterations of the form do occur. Such alterations are often called transliteration.

**External structure and relation with common nouns** Both proper names and common noun phrases can act as references. Regarding the external syntactic structure, Quirk *et al.* (1985) suggested that proper names normally function as single units with respect to grammar and therefore behave differently from common nouns in many ways. Three aspects of the linguistic attributes are considered by Quirk (*ibid.*): number, determination and modification. Proper names usually lack article and number contrast. For instance, *London* cannot be *a London*, *the London* or *two Londons*. Once created, the internal structure of a proper name is stable and does not allow the insertion of modifiers, as in *King's famous College*. When a proper name is used in its normal circumstances, it can only be modified by non-restrictive modifiers; this may be either a non-restrictive relative clause or a non-restrictive apposition, for example:

- Dr Brown, *who lives next door*, comes from Australia.
Chapter 2. Proper Name Analysis

- Tony Blair, the British Prime Minister, arrived in Beijing Airport yesterday afternoon.

It has been reported that a proper name can behave as a common noun (Quirk et al., 1985:288-289). Under such circumstances, plural forms can be applied. Shakespeares means 'authors like Shakespeare' and Londons means 'cities called or resembling London'. When a surname following 'the' takes the plural form, it refers to the family or all the members of the family.

It is not clear when a proper name can be used as a common noun. The meaning of a proper name can be varied by taking articles and other determiners, such as possessive pronouns. For example, Shakespeare in his new Shakespeare means his copy of a book written by Shakespeare. The indefinite determiner 'a' followed by a proper name means an entity called this name.

When a proper name is used as a common noun, it can accept restrictive modifications:

- I know another John who went to Cambridge.

- I spoke to the younger Mr Hamilton, not Mr Hamilton the manager.

The determiner 'the' combined with restrictive modifications can alter the meaning of a proper name. Comparing 'in England' with 'in the England of Queen Elizabeth', the aspect of the referent has been changed.

2.3.1.2 The meaning of English proper names

In the past half century, linguists and psychologists joined the debate on the semantic issue of proper names which was initiated by ancient philosophers.

Many linguists have studied the semantic characteristics of proper names. Quirk et al. (1985) for instance defined proper nouns as names of specific people (Shakespeare), places (Milwaukee), months (September), days (Thursday), festivals (Christmas), magazines (Vogue) and so forth. Proper names tend to be associated with countable nouns, such as man, town, day, rather than uncountable nouns like milk, bread etc. However, many countable entities do not have a name, such as apple, door.

Allerton (1987) suggested proper names possess special properties in part grammatical and in part semantic. Allerton believed that proper names can be said to have meaning, though only of an individualistic kind. The meanings of proper names are stand-alone in that they are not integrated with other lexical meanings but fall into different categories according to the
referent. He has observed that being ignorant about names does not degrade one's language ability but only suggests that someone is lacking general knowledge. In turn this suggests that the meaning of proper names tends to be largely encyclopaedic. Because Allerton has based his analysis on a wide range of proper names, he has demonstrated that names can be semantically transparent in part if not all. *The United States of America* is one of such case. Furthermore, the level of descriptiveness can be unclear, for instance, *the Irish Sea* does not just belong to Ireland whilst *the Gulf* is used to replace *the Persian Gulf* to avoid controversies.

Allerton (1987) offered a semantic classification according to the entities they refer to. This semantic category includes a) human beings; b) vessels, vehicles and machines; c) geographical locations; d) social organisations; e) publications and works of art and f) languages and dialects. Table 2.3 illustrates the categories including their corresponding examples.

### Table 2.3: The semantic classification by Allerton (1987)

<table>
<thead>
<tr>
<th>Category</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human Beings</td>
<td>Aristotle, Socrates</td>
</tr>
<tr>
<td>Vessels, Vehicles and Machines</td>
<td>the Orient Express</td>
</tr>
<tr>
<td>Geographical locations</td>
<td>Switzerland</td>
</tr>
<tr>
<td>Social organisations</td>
<td>IBM</td>
</tr>
<tr>
<td>Publications and works of the art</td>
<td>The Times</td>
</tr>
<tr>
<td>Languages and dialects</td>
<td>Hebrew</td>
</tr>
</tbody>
</table>

Allerton (1987) offered a semantic classification according to the entities they refer to. This semantic category includes a) human beings; b) vessels, vehicles and machines; c) geographical locations; d) social organisations; e) publications and works of art and f) languages and dialects. Table 2.3 illustrates the categories including their corresponding examples.

#### 2.3.1.3 The use of English proper names

Having discussed both the syntactic and the semantic issues of proper names, we move on to the pragmatic characteristics of names. "Pragmatics is the study of communication as it is situated relative to a particular set of communication demands, speakers, hearers, times, places, joint surroundings, linguistic conventions, and culture practices" (Scha, Bruce and Polanyi, 1990:241).

Marmaridou (1989) has carried out an analysis of proper names from the communicative point of view. She claimed that there are two kinds of uses of proper names: the referential use and the connotative use. The following sentences (taken from the work of Jerpersen (1965)) have been used by Marmaridou to illustrate the two uses:
• Judas was Jesus Christ's disciple who betrayed him.

• Every great man nowadays has his disciples, but it is always Judas who writes the biography (Jerpersen, 1965:66).

The referential use is the function of identifying referents which we are more familiar with. With this function, proper names provide shortcuts for communication, which makes them more efficient compared to common nouns and definitive descriptions. In the first example above, 'Judas' was used to refer to the person who, according to the Bible, betrayed Jesus Christ which led to the crucifixion of Jesus. During the referential use of a proper name, the encyclopaedic assumptions about the referent are formed based on the retrieval of the associated information within one's memory. The goal is to achieve the optimal relevance for the precise and efficient identification of the referent. The hearer of this utterance of the name 'Judas' would activate the encyclopaedic information which is related to this person even though he or she has never met Judas.

In the connotative use, proper names function as metaphors. They offer shortcuts for ideas and thoughts which are conveyed implicitly. In the second example above, 'Judas' was not used to refer to anyone specifically, but to a kind of person who is likely to betray their masters. The connotative use of a proper name involves the retrieval of associated encyclopaedic information, which serves to yield a number of implications about some other entity which shares certain characteristics. The characteristics of being deceitful and dishonest has drawn the association. Connotations of proper names usually develop out of the cultural and historical significance that the named entity has acquired within a social community. A non-Christian person who speaks English but has no knowledge about the Bible would not be able to comprehend the second example. Marmaridou's study shows that whether used in their referential or their connotative form, the encyclopaedic information of proper names plays an important role in communication (1989).

The encyclopaedic data was called upon when extending a proper name like Kleenex to denote tissues and Hoover to denote vacuum cleaners.

Caspar Cazzo is no Pavarotti! (cf. Allan, 1995)

The above example illustrates that when names are used metaphorically, biographical and geographical data are activated to draw the comparison. Such data do not come out of a
lexicon. Instead, information about people is stored in the encyclopaedia and one means of access to the information is via a person's proper name. In this way, the whole encyclopaedia entry can be accessed through the part of the information in it enabling the name and further information about the referent to be retrieved. The encyclopaedic knowledge about Pavarotti must be accessed in order to deduce further information about Caspar Cazzo.

Pavarotti, Luciano (1935- ) Italian operatic tenor. He is widely known throughout the world on account of his many popular recordings.

Individuals in a community will have different mental encyclopaediae. When the sentence “I got a letter from Anna yesterday” is uttered by the speaker, the encyclopaedic knowledge about the person Anna or possibly several Annas will be activated in the listener's brain. Being able to identify the referent means being able to access information about the name-bearer which is stored in one's mental encyclopaedia.

Both the information from an actual encyclopaedia and that from one's mental encyclopaedia are encyclopaedic rather than lexicographical. We believe that an NLP system which requires deep understanding can only achieve the optimal performance when both the lexicographical and encyclopaedic information are given.

2.3.2 Linguistic features of Chinese proper names

The existence of unknown lexicalisations, mainly proper names, is a feature of many languages. Chinese, the language that has the largest number of speakers in the world, is one such case. Since modern linguistics is a relatively new topic for Chinese compared to English, applications in computational linguistics in Chinese are usually based on empirical studies. In this work, we intend to gain a greater understanding of Chinese proper names through a contrastive study of Chinese and English. First of all, an overview of the history and diversity of the language is offered. It is followed by the study of the naming practice in China. More importantly, linguistic features, such as the syntactic and semantic structure of Chinese proper names are examined.

2.3.2.1 Overview of the Chinese language

Chinese is a language spoken by over a billion people and is classified as a member of the Sino-Tibetan family. Linguists further classify Chinese into five groups based largely on the
geographical boundaries in which the different versions of Chinese are spoken. For many linguists, these different versions are given the status of a dialect (Li & Thompson, 1990). Table 2.4 shows the typical characteristics of Mandarin, Yue, Min, Wu and Hakka. Mandarin is the official language of the People's Republic of China and Taiwan and is one of the official languages of the United Nations. Much computational linguistic work for Chinese has been carried out on the processing of Mandarin. Typical applications include word processors, text retrieval systems, machine-readable dictionaries and machine translation systems.

2.3.2.2 Chinese names and naming

A Chinese personal name usually consists of a surname (姓) followed by a given name (名), which is opposite to the arrangement of names in many European countries. The surname is the family name which is inherited from one's father, or very occasionally from one's mother. Nowadays, Chinese women usually retain their family names even after marriage. There are more than 3,000 Chinese surnames, but the majority of Han Chinese have surnames which fall into a set of approximately 'a hundred family names', traditionally called 'bai jia xing' (百家姓) (Liu, 1999). Surnames, typically, have only one character, like Zhang (张), Wang (王) or Li (李); however, there are a number of surnames that have two characters such as Ouyang (欧阳) and Sima (司马).

Given names are coined by one's parents or other people when a child is born. A given name often contains one or two characters chosen from a large character set. Since there is no limitation as to which character is selected, Chinese given names are more prone to be descriptive. In other words, it is more likely for a person to have common nouns as his/her first name, for instance, Cheng Long (成龙). The first name Long (龙) means 'dragon'.

However, the naming practice can be more flexible. Parents reflect their good wishes or expectations in the given names. Changsheng (长生), meaning 'long life', is one such case. Children are sometimes named according to the day he (she) was born, such as someone called Yuanchun (元春) is likely to be born on the Chinese Spring Festival and a child born on the National Day could be called Guoqing (国庆).

The trend of naming can also reflect history. Immediately after the establishment of the People's Republic of China, many children were named Jianguo (建国) which means 'found the country'. During the cultural revolution (1966-1976), names like Weidong (卫东) meaning 'guard the East' and Yonghong (永红) meaning 'forever red' were popular because it was
<table>
<thead>
<tr>
<th>Dialect name</th>
<th>Status</th>
<th>Geographical boundaries</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mandarin (官话) is the English translation of ‘Guan Hua’ which means official language</td>
<td>Official language of the PRC and Taiwan. Used also by the United Nations. The native speakers of this group represent about 70% of the total Chinese population</td>
<td>The North China plain, the middle Yangzi plain, the Huai plain, the north-east plain, the Sichuan basin and most of Guangxi, Guizhou and Yunnan provinces</td>
</tr>
<tr>
<td>Yue (粤) is the alternative name of Guangdong province</td>
<td>Includes the well-known Cantonese, the language of Guangzhou (Canton)</td>
<td>Guangdong and Hong Kong, and the various Chinatowns in the United States, Europe and south-east Asia</td>
</tr>
<tr>
<td>Min (閩) is the alternative name of the Fujian province</td>
<td>85% of the people in Taiwan still speak a Min dialect as their native language</td>
<td>Taiwan and Fujian provinces and Hainan Island</td>
</tr>
<tr>
<td>Wu (吴) is the old name for the area that includes Shanghai and Suzhou</td>
<td>Provincial language</td>
<td>Jiangsu, Zhejiang and Anhui provinces including major cities like Shanghai, Suzhou and Wenzhou</td>
</tr>
<tr>
<td>Hakka (客家) means guest</td>
<td>The least well known Chinese language</td>
<td>Guangxi province</td>
</tr>
</tbody>
</table>

Table 2.4: Classification of the Chinese language by Li & Thompson (1990)
fashionable to show enthusiasm and loyalty towards Communism. Although there is no strict gender distinction in Chinese names, parents tend to choose characters which relate to feminine objects such as 花 (flowers), 玉 (jade) and 凤 (phoenix) for girls and more masculine characters for boys, such as 坚 (hard), 强 (strong) and 峰 (mountain).

2.3.2.3 The form of Chinese proper names

Internal structure of Chinese names Chinese proper names do not take articles as a part of the names whether they are definitive or nondefinitive. For instance, the Alps is translated into 阿尔卑斯山脉 and the article the is ignored during the translation. Regarding modifications, Chinese proper names can only take pre-modifications, such as 北海 (the North Sea) and 黄河 (the Yellow River). When a English name with post-modifications is translated into Chinese, the word order is changed. Isle of Wight is translated into 怀特岛, in which 怀特 is Wight and 岛 means island.

Following Allerton's categorisation, personal names fall into the first category of pure PNs, such as 江泽民 (Jiang Zemin) whilst most places names fall into the second category of PNs+CNs (refer to Table 2.2). They usually have a namestem followed by common noun descriptor, such as 洞庭湖 (the Dongting Lake), 波罗的海 (the Baltic Sea), 喜玛拉雅山 (the Himalayas), 虹桥机场 (Hongqiao Airport), 圣保罗教堂 (St Paul's Cathedral), etc. The determiners are absent in Chinese proper names.

Organisation names tend to be longer and more complex. A name of a commercial company is generally a namestem with possibly their service or product range, followed by their descriptor such as 公司 (Inc), 集团 (Ltd.). Provincial companies tend to have a place name at the beginning. For instance, 徐州丰宝矿业有限公司 (Xuzhou Fengbao Mining Co.,Ltd.). Governmental organisations also end with a descriptor, including 部 (ministry) or 局 (bureau): Ministry of Agriculture in Chinese is 农业部.

Other proper names relate to names of festivals or days of the week and tend to have relatively regular structures. Names of festivals normally have 节 (festival) as suffixes, such as 国庆节 (National Day). 星期 (week) appears as prefixes in 星期一 (Monday). When a date is mentioned in Chinese, the word order is the opposite to that of English. The 8th of September 1974 is 一九七四年九月八日 (Year 1974 Month 9 Day 8).

It is important to note that there are identifiable differences in the Chinese language between proper names of Chinese origin and those of non-Chinese origin (transliterated names).
Chapter 2. Proper Name Analysis

Transliterated personal names in Chinese are generally long and the characters tend to be rarely used ones. When personal names in another language are translated into Chinese, the rule is to match the pronunciation. Each syllable can be translated to one character offering a similar sound. For instance, Tony Blair is translated into 突尼布莱尔. In Chinese, titles like Mr or Mrs are often placed after the name. Therefore, Mr Blair is 布莱尔先生 where 布莱尔 means (Blair) and 先生 means (Mr). Another important aspect of transliterated names is that the elliptical descriptors are often recovered through translation for clarity. For instance, the Sahara in Chinese is 撒哈拉沙漠 (the Sahara Desert) and the Atlantic is 大西洋 (the Atlantic Ocean).

Many Chinese proper names have more than one form. This may be caused by the shortening process as described in Carroll (1985), when the context offers enough information so that the omission of the descriptor does not cause ambiguities. The namestem can refer to an entity on its own and become the namehead. For instance, 江苏省 (Jiangsu Province) can be shortened into 江苏 (Jiangsu). When a name is too long, it can often be replaced with its acronyms. For example, 阿拉伯联合酋长国 (the United Arab Emirates) is often shortened to the first character of every word which is 阿联酋.

External structures compared to common nouns Having discussed the internal structures of Chinese proper names, we move on to their external structures. Chinese proper names show the following three characteristics which also apply to English proper names:

1. Chinese proper names, when used as references, lack number contrast relative to Chinese common nouns. In Chinese, singular or plural is expressed by adding the number phrase in front of a noun. Usually, one does not tend to find 一个 (one, a) or 许多 (many) appearing in front of a proper name.

2. In their referential use, neither do Chinese names take a determiner or possessive pronoun, such as 这个 (this) or 那个 (that) in front of them.

3. Proper names differ from common nouns in that modifiers usually cannot be inserted in the middle of a name whereas it is a common practice for common nouns. For example, 喜玛拉雅山 (Himalayas Mountain) cannot be written as 喜玛拉雅高山 (Himalayas high Mountain).

Unlike in English, non-restrictive modifiers do not exist in Chinese. Therefore, proper names can only be modified by restrictive modifications in Chinese. For example, the non-restrictive
relative clause in the following sentence:

Mr Brown, who lives next door, comes from Australia.

has to be converted into a pre-modification.

住在隔壁的布朗先生来自澳大利亚.

Live next door Brown Mr comes from Australia.

Alternatively, a modifier could be expressed as a separate sentence. Take the same sentence above, it could be translated as:

布朗先生来自澳大利亚. 他住在隔壁.

Mr Brown comes from Australia. He lives next door.

The same rules apply to non-restrictive appositions.

2.3.3 Similarities and differences between EPNs and CPNs

The differences between English proper names (EPNs) and Chinese proper names (CPNs) lie largely in their structures both internally and externally. In EPNs, determiners such as 'the' and propositions such as 'of' occur frequently. In CPNs, these close class words tend to be missing.

Another significant difference is the word ordering. In CPNs, titles or descriptors invariably follow the namestem, for instance, 张先生 (Mr. Zhang) and 怀特岛 (Isle of Wight). In EPNs, the order between the namestem and the descriptor can vary, for example, the Baltic Sea. This is also reflected in the external structures of CPNs in that post-modifying phrases or clauses for EPNs, when translated, have to be converted to pre-modifications or form a separate sentence (See the example of 'Mr Brown' in the previous section).

Despite differences between the forms of EPNs and CPNs, they often can be used to refer to the same set of objects and entities such as persons, places. Therefore, the semantics of proper names are largely language-independent. This suggests it is possible for the knowledge acquired for EPNs and CPNs to share the same knowledge base and the same conceptual model.

A further similarity lies in the use of proper names. Other than their basic reference use, CPNs can also be used connotatively. Let us examine the following two sentences:
Chapter 2. Proper Name Analysis

(1) Lei Feng will live in my heart forever.

Zhang Ming is a living Lei Feng.

In Sentence (1), the name ‘Lei Feng’ (Lei Feng) takes its referential use and refers to a person who was dedicated to serving other people and became a hero in peoples’ hearts. In Sentence (2), the same name was used to describe another person ‘Zhang Ming’ (Zhang Ming) who also works hard and helps other people.

Through the discussions on the linguistic features of both EPNs and CPNs, we have observed similarities and differences between the form, meaning and use of proper names in both languages. It helps to lay down the foundations for our comparative study between English and Chinese biographical entries and guides us to the design of our testing strategy. Current computational work on proper name analysis will be discussed in the following section.

2.4 Computational Work in Proper Name Analysis

Since the early 90s, proper name analysis has been applied in many fields in natural language processing including information retrieval (Paik et al., 1996), natural language generation (Radev & McKeown, 1997) and information extraction (MUC-6, 1995; MUC-7, 1998). It is often considered an indispensable stage in the processes of many NLP applications. Information extraction and lexical knowledge acquisition have been the two main areas which attracted and motivated many researchers working on the analysis of proper names.

2.4.1 Named-entity recognition in information extraction

Information extraction (IE) is a particular field in natural language understanding which takes unrestricted texts as input, finds useful information and encodes the information in a structured format (Cowie & Lehnert, 1996). The input of IE systems is often factual and event-oriented texts which describe the relationships between named entities, which makes the extraction and the understanding of proper names an important task. This argument is supported by a more specific definition of IE given by Grishman (1997:10) as “the identification of instances of a particular class of events or relationships in a natural language text, and the extraction of the relevant arguments of the event or relationship”.

32
Table 2.5: The representatives of the early IE systems

<table>
<thead>
<tr>
<th>Researchers</th>
<th>Project Starting Time</th>
<th>System Domain</th>
<th>System Attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zarri (1983)</td>
<td>early 1980s</td>
<td>social activities</td>
<td>Extract meeting of French historical figures.</td>
</tr>
</tbody>
</table>

IE was born in the early 1980s when researchers tried to apply NLP techniques into real world texts. Table 2.5 offers a list of early IE systems in a variety of domains. Naomi Sager (1981) presented a system which automatically converted patient records into database entries in an information system. FRUMP, implemented by Gerald deJong (1982), used a newswire as its data source and mapped news stories with scripts. DaSilva & Dwiggins's system was designed to analyse reports generated by satellite-flight monitors (1980). Zarri (1983) worked on a IE system which extracts information about the relationships and meeting between the historical figures in France. Cowie (1983) developed an IE system to extract information about plants and animals to populate a database with a fixed structure.

Early IE systems tended to be tailored for a specific domain and a particular task. It was difficult to evaluate or compare the performance of two different systems. This situation was changed during the early 90s by the government sponsored Message Understanding Conferences (MUCs) (MUC-3, 1991; MUC-4, 1992).

MUCs, instituted by DARPA in the late 80s, aimed to promote and evaluate research in IE. During these conferences, the performance of NLP systems was evaluated by extracting relevant information from free texts within a given domain. Driven by the military initiative,
The MUCs and their corpus domains

<table>
<thead>
<tr>
<th>Conference</th>
<th>Corpus Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>MUC-5 (1993)</td>
<td>News articles about joint ventures and microelectronics</td>
</tr>
<tr>
<td>MUC-6 (1995)</td>
<td>News articles about management changes</td>
</tr>
<tr>
<td>MUC-7 (1997)</td>
<td>News articles about space vehicle and missile launches</td>
</tr>
</tbody>
</table>

Table 2.6: The MUCs and their corpus domains

a set of naval messages were used as the corpus of the evaluation, for the first (1987) and second MUCs (1989) (see Table 2.6). For MUC-3 (1991) and MUC-4 (1992), the topic was changed into terrorist incidents in Latin American countries. The corpus has expanded into 1000 texts and at the same time, a template structure and associated extraction rules were to be followed. A semi-automatic scoring system was also provided for the evaluation.

MUCs have had a great impact in the field of IE as well as proper name analysis. The competition has encouraged the exploration of various NLP techniques and pushed the technology out of the lab and into the deployment of commercial products. The template-filling scheme has offered a well-defined evaluation system.

In order to improve proficiency, the design of the IE systems tends to be more modular. Since MUC-6 (1995), a subtask named named-entity recognition was specially designed for the extraction of proper names. This task tries to identify three types of expressions:

- ENAMEX: Proper names, including names of persons, locations and organisations.
- TIMEX: Temporal expressions such as dates and time.
- NUMEX: Numerical expressions such as money and percentages.

The MUC-6 and MUC-7 Named Entity (NE) Tasks and the Multilingual Entity (MET) Task have stimulated the creation of more than thirty name recognition systems for a variety of languages (English, Spanish, Chinese, Japanese and Thai), and doubtless many more systems exist as well. The approaches adopted in these systems range from the more or less purely statistical, for instance Identifinder™ by BBN (Weischedel, 1995), to the more or less purely rule based, for example LaSIE-II from Sheffield University (Humphreys et al.,
Chapter 2. Proper Name Analysis

1998) and Alembic from MITRE Corporation (Aberdeen et al., 1995). Furthermore, there are a number of hybrid systems which mix statistical with rule-based techniques. Examples are LTG from University of Edinburgh (Mikheev et al., 1998) and MENE (Borthwick et al., 1998) from New York University.

The developments in information extraction promoted by MUCs have led to improvements in proper name analysis in that the identification of proper names and precision of categorisation have been improved greatly. However, in order to ease evaluation, a set of rigid guidelines must be followed which brought certain restrictions to the field.

First of all, news articles have been the favourite source of the evaluation corpus in MUC competitions. Although the Named-Entity Recognition task is generally viewed as generic, assumptions have been made based on the characteristics of news articles. Poibeau and Kosseim (2001) compared IE systems based on journalistic texts and non-journalistic texts and reported a significant performance drop for non-journalistic texts, such as emails and telephone transcripts. While better scores are achieved by IE systems for news articles, the fact that languages used in different subject areas for different purposes have different characteristics has been overlooked.

Furthermore, only names which belong to certain categories were considered relevant, for instance, person, place, organisation, dates, time and money. Such classification strategy is necessary in order to ease evaluation. However, it oversimplifies the diverse nature of proper names, comparing to the categorisation suggested by Keith Allan (1995:298):

There are names for real and imaginary people; pets; newly discovered and cultivated biological specimens; places and topological features, buildings, institutions, businesses, radio stations, pop groups, orchestras, acting companies; events like wars and epidemics; computer files, books, newspapers, films, TV shows; manufactured products of all kinds; and other things.

The aim of an IE system is to filter out relevant information and store it in template form. In this kind of process, proper names are considered as building blocks for template filling rather than lexical entities within other NLP applications. As Wilks et al. (1996:110) has pointed out "this automatic process (IE) is intended to aid human analysts to sift through the huge numbers of documents that are now available to them electronically, but it does not attempt to automate a task that human analysts actually do", which is understanding.
An important challenge for a text understanding system is to make use of the encoded world knowledge which includes the knowledge of proper names (Amsler, 1989).

2.4.2 Lexical knowledge acquisition

Lexicons play an indispensable role in NLP. A person can only understand a specific language when one has gathered a certain amount of the vocabulary of that language. Likewise, applications in Natural Language Processing (NLP) usually rely on a lexicon in order to process texts from the real world.

One of the major obstacles for NLP systems is lexical knowledge inadequacy. Proper names in the lexicon and lack of the knowledge about proper names as part of the knowledge base are major contributors to this problem. These proper names, traditionally treated as *unknown lexicalisations*, have caused the lexical gap which has hindered performance of the NLP applications (Coates-Stephens, 1993). Research of proper name analysis should deal not only with lexical acquisition but also with knowledge acquisition. However, the review of the proper name analysis suggests that little work has tackled knowledge representation about proper names.

NLP applications range from fairly simple string manipulation like stemming to high-level applications like text understanding and generation (FOLDOC, 2001). Some applications only need the lexicon as a word list, like spell checking. Others need syntactic, semantic, even phonological information like speech understanding. Some applications can meet the requirement with a limited hand-crafted lexicon, whereas other require a large lexicon which would be too expensive to build by hand. The objective of high-level NLP tasks is the automation of the processes of language comprehension, production and acquisition in both written and spoken media. Such tasks tend to adopt a knowledge-based approach which attempts to incorporate appropriate linguistic and general knowledge (Boguraev & Pustejovsky, 1996).

2.4.2.1 Lexical knowledge acquisition for general lexica

In order to overcome the lexical knowledge bottleneck and provide NLP systems with the knowledge they need, the field of *lexical acquisition* was established. Lexical acquisition is concerned with the design, construction and implementation of serious lexical components for NLP systems (Boguraev & Pustejovsky, 1996). It can be carried out manually as well
as automatically from textual resources. There are two types of text resource for automatic acquisition: machine-readable dictionaries (MRDs) and text corpora (*ibid.*).

Early NLP systems were usually small scale and hand built word lists were sufficient to carry out the tasks. Although being a labour intensive method to populate a lexicon, manual acquisition was still frequently used for many reasons. First of all, the start up cost is low. No MRD or corpus needs to be created. No programs need to be written. All you need is a text editor. Secondly, it is suitable for applications which require a relatively small vocabulary ranging from a few dozen words to a thousand words and the vocabulary does not tend to change over time.

Manual acquisition was also applied to projects which aim to capture knowledge from which the text form is not obtainable. For example, the CYC project (Lenat & Guha, 1990) was dedicated to the acquisition of common sense knowledge in order to build a large scale knowledge base for natural language processing. Another example of the manual acquisition of lexical knowledge is the ONTOS project (Nirenburg, 1989), which provides an interface for acquisition of knowledge associated with words. It is assumed that the knowledge could not be found in any systematic form in MRDs or text corpora.

The growing complexity of language technology has increased the demand for fairly large scale computational lexicons. The availability of language references, such as dictionaries in electronic form has naturally led researchers to look into methods of utilising these resources. This field has developed into *computational lexicography* which is concerned with the design, construction and implementation of serious lexical components for NLP systems from MRDs (Boguraev & Briscoe, 1989). Published dictionaries are designed with the human reader in mind rather than machines and they lack formal representation. Therefore, much work has been carried out to study the nature of dictionary entries, representation and organisation by, for example, Alshawi (1989) and Sinclair *et al.* (1994).

Another text resource for lexical acquisition is text corpora where the use and patterns of word behaviour can be studied. Statistical analysis is usually applied to very large samples of text or speech to induce lexical properties. Corpus-based lexical acquisition can be efficient but a lot of effort is required for to design the learning mechanism and to create and balance the corpus (Boguraev & Pustejovsky, 1996).
2.4.2.2 Lexical knowledge acquisition for unknown lexicalisations

The development in proper name analysis can be seen as lexical acquisition for unknown lexicalisations. The conventional view within computational linguistics is that proper names are a kind of symbol, similar to numbers or dates. In this paradigm, proper names are considered as an insignificant part of language. There has been an assumption that one cannot gain deep insights by studying proper names. Most early NLP applications tended to be designed within a restricted domain; therefore, a very small hand-built vocabulary was able to meet the requirements.

When researchers realised that a manually compiled name list was not sufficient for NLP systems running against unrestricted texts, they turned to text corpora for lexical knowledge because the MRDs are based on the conventional dictionaries which do not offer much information about proper names.

The work in computational lexicography did not consider encyclopaedic entries - the entries for proper names either because MRDs do not offer encyclopaedic entries or the researchers believed that encyclopaedic knowledge is least related to computational lexicography (Boguraev and Pustejovsky, 1996). The work of Sampson (1989), and Seitz and Gupta (1990) showed that MRDs give poor coverage of PN. Proper names are often compiled as part of encyclopaediae. MRDs used in comparison to text corpora are usually based on a conventional dictionary which excludes proper names. It is hardly surprising that the inadequacy of static lexical resources for coverage of proper names was prominent.

As NLP systems moved out of the research labs and attempted to process large samples of naturally occurring language, lexical inadequacy became an increasingly prominent problem. This problem is partly due to the high percentage of proper names in unrestricted texts, especially in news articles, and partly due to the important role they play in these texts. Rau (1991:29) stated that "one of the major problems in the accurate analysis of natural language is the presence of unknown words, especially names. While names account for a large percentage of the unknown in a text, they can also be the most important piece of information in a text ...". However, their diversity in range and flexibility in structure make the extraction of proper names a difficult task.

The computational analysis of proper names include identification, classification and representation. Unlike many other lexical categories, a proper name could contain only one token or it could have substantial internal structure. The task of proper name identification focuses
on the detection of the boundary of proper names. In some systems, the standardisation of names is also considered in order to link variant names for the same named entity.

Classifying proper nouns involves deciding into which class the named individuals fall. Typical name classes include organisations, persons, locations, dates and monetary amounts. However, further classes can include book and movie titles, product names, restaurant and hotel names, ship names, etc. The task is made difficult by the unpredictable length of names (company names can be twelve or more words long), ambiguity between name classes (Ford can be a company, a person, or a location), embedding, where for example a location name occurs within an organisation name, variant forms and unreliability of capitalisation as a cue, e.g. in headlines in English, and all nouns in German.

The earliest approach of PN identification relied on the lexical lookup of names and name fragments in a name database. This approach has several drawbacks: firstly, a pre-existing name element database is required and the creation and update of such a database is labour intensive and time consuming; secondly, proper names form an open class which makes it impossible to produce a complete list.

Later, work on proper name analysis seems to have been carried out in an ad hoc manner, there is a lack of general theory describing proper names both from a syntactic and semantic angle. Allport (1988) described a text understander designed to process police traffic accident reports. In this work, title words, such as pc and sgt were used to classify a name into a personal name and locative keywords, such as lane and avenue were used to classify it into a place name.

The unreliability of keywords and the complicated embedding of names led many researchers to look for contextual information which can be utilised. Coates-Stephens (1993) suggested two kinds of syntactic patterns, named as within text descriptions (WTDs), in which proper names are usually introduced in news stories. They are:

- A proper name which is followed by appositives and which contains the description (e.g. Tony Blair, the Prime Minister of the UK).

- A descriptive word (or called keyword) is either proceeded or followed by a proper name. (e.g. the Chinese President Jiang Zemin, the Yangzi River)

A proper name extraction unit was built into a NLP system called FUNES. FUNES relies on both the internal structure of names and the descriptions (within text descriptions) in order to
detect proper names. A set of hand-coded rules was built to capture within text descriptions and to classify proper names into personal names, place names, corporation names, legislation names, information source names, events names, object names and origin names. Effort has been put into the analysis of corporate names especially due to the complexity of their structures. Conjunction, a big challenge, has been handled by applying heuristic rules.

McDonald (1996) has argued that "understanding names, and their patterns of initial and subsequent reference, is central to the analysis of the extended, unrestricted text". He has observed that the grammar of proper names is more lexical and less syntactic and also coined two terms, internal evidence and external evidence, to indicate, respectively, the structure of the word flow in a proper name and the context around which a proper name appears. PNF, the proper name facility, was developed as part of his NLP system; SPARSER. It was claimed to be the first time that context-sensitive rewrite rules were used to identify and classify proper names. Employing external evidence can be considered as progress compared to earlier systems that only rely on looking up gazetteers and name lists.

Mani and McMillan (1996) have broadened McDonald's external evidence approach by studying the discourse properties of names. There are three points of note in their study: firstly, a proper name as part of language can act as a syntactic head, a complement and an adjunct or attribute noun phrase. Sometimes they can also follow an indefinite determiner ('a' or 'an') behaving like an adjective. Secondly, semantically, a proper name can be made of primitive proper name elements, other proper names, non-proper names or sometimes it can be an open compound. Thirdly, a proper name tends to be descriptive, although in some of them the connotation has been lost by generation. The specialities of Mani and McMillan's approach are: Firstly, categorisation uncertainty has been treated by combining confidence of the hypotheses made according to the knowledge base. Secondly, appositive patterns have, in particular, been applied for generating new mentions (tokens of proper names in text). Thirdly, this system is able to normalise different expressions for one entity by means of a coreference procedure. Fourthly, 'shallow' ontology has been applied to associate entities with each other at a low level to tackle the coreference issue.

Paik et al. (1996) looked into the problem of proper names in the area of information retrieval. In this work, proper names were used as important resources for detecting relevant documents. A taxonomy is offered for proper names which contains base level classification as well as subclassification. Variant forms of proper names were standardised into their standard form so that different names for the same named-entity can be treated as one
name. Although a proper name categorisation and standardisation module was used to improve the performance of a information retrieval system, it can also be used in many other application areas. Although the research goal was to build a knowledge acquisition tool to deduce meaningful relations between information and store them in a knowledge base, no solutions were offered to integrate the information extracted from different sources.

Most proper name analysis systems to date considered the poor coverage of proper names in conventional dictionaries and steered towards the extraction of proper names relying on their surrounding contexts, for example in Coates-Stephens (1993) and McDonald (1996). Cues such as the title of a person can provide some information about the names to aid categorisation. When a category such as person is obtained for a proper name such as Bach, a significant amount of knowledge which will aid inference is lost due to the fact that the category is too generic. The knowledge related to proper names is critical in the understanding of the texts and thus in developing natural language understanding systems. Merely increasing size of a lexical database does not offer solutions for the problem due to the lack of represented knowledge.

An important source for information about proper nouns is the fact that a number of nations and states publish, or have published on their behalf, biographical data in dictionaries of biography, for example *Oxford Who's Who in the 20th Century* (1999). The reason that these dictionaries of biography have not attracted attention, perhaps, is that there is no guarantee that the information in a dictionary of biography will follow a typical structure.
Chapter 2. Proper Name Analysis

apart from the difference in type fonts used to indicate a headword - the name of a person - and the type fonts used to document other biographical information.

There has been extensive research in the area of computational lexicography which could be borrowed in the analysis of dictionaries of biography. The main objective of this thesis is to analyse and observe the language used in biographical texts, particularly the definition part of the entries. We propose to apply the knowledge and experience from computational lexicography into the lexical acquisition of proper names and extract the knowledge about the names from the dictionary of biography and further represent it in a formal knowledge representation scheme.

"There is no easy way to distinguish word knowledge from world knowledge, and no principled reason for dictionaries to remain distinct from encyclopaedias" (Wilks et al., 1996:254). Considering the complementary and interdependent relationship between dictionaries and encyclopaediae, extracting knowledge about proper names from encyclopaediae may be regarded as an extension of the work in computational lexicography which aims to extract and formalise encoded information from MRDs. In addition, it promotes the integration of word knowledge and world knowledge for NLP applications.

2.4.3 Comparing ABKAS with SNOWY

So far the only reported system which automatically acquires knowledge from biographical texts is SNOWY developed by Hull and Gomez (1999). Table 2.7 offers a comparison between the two systems. Firstly, they have different objectives. SNOWY was designed to be a question answering system whereas ABKAS aims to build computational lexicons for proper names.

The differences in their intentions dictate that the two systems focus on different parts of biographical entries. SNOWY takes the elaboration part of the biographical entries and focuses on verbal concepts. However, ABKAS studies the definition part of the entries and relies on the distinctive and repetitive patterns as the foundations for the analysis.

Furthermore, in SNOWY the parsing and the semantic interpretation are carried out at separate stages. The sublanguage parser in ABKAS, however, turns tagged texts directly into encoded knowledge. Despite their differences, the two systems can be considered as complementary in that ABKAS is able to construct the framework of the biographical knowledge base and SNOWY is capable of filling in the detailed information.
Table 2.7: Comparison between SNOWY and ABKAS

2.5 Conclusion

This chapter studied the onomastic behaviours and linguistic characteristics of proper names including syntax, semantics and pragmatic issues. From the review, we have learned that proper names are not just simple labels. Most proper names retain meaning. Such meaning is often associated with its encyclopaedic knowledge as well as its prototypical knowledge (Marmaridou, 1991). Until recently, work in lexical acquisition for proper names has been restricted to a corpus-based approach based on news analysis as reported in Boguraev and Pustejovsky (1996). We have also looked into the relationship between dictionaries and encyclopaediae and suggested that language references, such as encyclopaediae, provide rich knowledge for proper names and further propose to apply knowledge and experience gained in computational lexicography into the automatic knowledge acquisition for proper names.

In the following chapter, the language used in dictionaries of biography will be studied in order to extract knowledge about personal names. Corpora-based analysis will be carried out in order to observe the language used by biographers. The grammar which governs the encyclopaedic entries will be examined for the automatic acquisition of knowledge about proper names.
Chapter 3

A Sublanguage in Dictionaries of Biography

3.1 Introduction

If a dictionary is considered the source of knowledge of common nouns, an encyclopaedia may be regarded as the source of knowledge of proper names. Dictionaries of biography and some college dictionaries contain a wealth of information about individuals recorded in history. This information can be used as background information to make inferences about individuals who are referred to in free texts like newspapers, magazines and other documents.

A review of the state of the art in proper name analysis shows that little work has been carried out in representing the knowledge of proper names. Some existing work extracts knowledge of proper names from the contextual information within the texts (McDonald, 1996; Coates-Stephens, 1993). The challenge in knowledge acquisition for proper names lies in that knowledge of proper names is usually part of world knowledge and therefore can be difficult to obtain (Amsler, 1989).

The study of proper names from various disciplines, including linguistics, philosophy and psychology, suggests that an encyclopaedia is a repository with a rich knowledge of proper names. Encyclopaediae often contain all kinds of information in a complex order. Hence, systematic means are required to acquire such knowledge and represent it.

Indeed, dictionary definitions are ideally structured for taxonomic organisation since the defining words themselves appear in the headword list in the dictionary itself. These defining
words are structured in regular formats. John Sinclair and his fellow researchers noted that certain lexicographers define lexical items in "distinctive and repetitive shapes according to their meaning" (Barnbrook & Sinclair, 1994:7). The distinctive and repetitive shapes of dictionary definitions may be attributed to the classical structure of the definition in terms of a *genus* and a discriminating *differentia* (Landau, 1996:120). The genus is a phrase that is more general than the headword and corresponds to the core of the definition. The differentia corresponds to the definition modifiers and serves to differentiate the headword from other headwords with the same genus.

Ahmad and Al-Jabir (1998) have used Barnbrook and Sinclair's (1994) framework to convert terminology databases of seven specialised domains, ranging from paint technology to nuclear technology, into knowledge bases for each domain. They directly parsed the definitions of terms and then mapped the parsed output onto a λ-calculus notation. The λ-calculus propositions were asserted using frames. Their work can be considered as the natural extension of Barnbrook and Sinclair's work for specialist terminology which was largely made up of nouns.

Similar to a general language dictionary or a specialist dictionary, i.e. a term base, dictionaries of biography are compiled by human biographers. The difference lies in that a dictionary of biography comprises definitions of proper nouns rather than the definitions of common nouns. Therefore in our research, we investigated whether the sublanguage approach can be extended to biographical information with a view to creating a biographical knowledge base. Our study is focused on personal names and takes dictionaries of biography as the corpora of our experiments. Dictionaries of biography 'define' proper names: names of people here are the equivalent of headwords in conventional dictionaries and the definitions comprise details about the person whose name is specified as the headword. Each biographical entry is focused on a person's background knowledge regarding his or her experiences, which are often described as relationships with other entities, as well as events which have occurred in his or her life.

In this chapter, our intention is to study the language that biographers have used when compiling dictionaries of biography and to seek a possible correlation between the syntactic structures and semantic relations. We start this chapter by introducing the concept of sublanguage. This is followed by some work carried out in the analysis of conventional dictionaries and term bases which uses the sublanguage approach. Due to the similarities between dictionaries of biography and conventional dictionaries as well as term bases, the applica-
tions of the sublanguage approach in dictionary analysis and term base analysis provides a theoretical justification of taking such an approach in our research. This is followed by an overview of dictionaries of biography. We will be looking into the organisation and representation of the whole reference as well as the structure of a typical biographical entry. We argue that the nature of dictionaries of biography is as a reference book and even entries chosen from different dictionaries of biography share strong similarities in their language structure. Our preliminary examination suggests there could be a potential sublanguage in biographical texts.

In Section 3.3, the analysis guided us to identify the patterns within the sublanguage and construct the local grammar which governs the sublanguage. The structures in biographical dictionaries provide a way to acquire information automatically from a machine-readable dictionary of biography and hence convert free texts into their corresponding formal representation.

In Section 3.4, several knowledge representation schemes are reviewed in order to find an appropriate choice for a biographical knowledge base. Finally, this chapter will conclude that there are indeed distinctive and repetitive patterns which can map the syntactic structure to the semantic relationships.

3.2 Towards a Sublanguage of Definitions

A language typically defined has a vocabulary and a grammar together with systems to attribute meanings and communicate intent (semantics and pragmatics). A natural language is used for many purposes, for example, day to day language-based communication and the restricted language of specialist enterprises, for example, language of different professions, of science and technology; languages related to leisure pursuits, arts and crafts and so on.

These specialist varieties of language are usually distinguished from the original mother languages by the frequent use of restricted sets of grammatical structures found in the mother language. A number of scholars have remarked on special languages (Sager, Dungworth & McDonald, 1981), specialised varieties of language (Barnbrook & Sinclair, 1994), and sublanguages (Harris, 1991; Kittredge & Lehrberger, 1982).

Zellig Harris suggested that one should study the enclosed subsets of a language rather than trying to tackle the language as a whole: “Certain proper subsets of the sentences of a
language may be closed under some or all of the operations defined for the language, and thus constitute a sublanguage of it" (1991:272). Harris's view was shared by Barnbrook and Sinclair who suggested that a specialised variety of a language may not have all the features of a language as a whole, and it may have features which are not found in the language as a whole (1994).

Although the set of sentences in a sublanguage is a subset of the set of sentences in the whole language, Harris argued that the grammar of the sublanguage is not necessarily included in the grammar of the whole language. Instead, the structures of these sublanguages have important new properties of their own which form their own special grammar (1991).

Some sublanguages are based on grammatical conditions, in that the subsets of the sentences satisfy certain grammatical conditions which are not satisfied by other sentences. These kind of sublanguages are called grammar-based sublanguages. There are also sublanguages which consist of sentences restricted to a closed subject matter. They are known as subject-matter sublanguages (ibid.).

The concept of sublanguage has been extended from a subset of sentences which satisfy closure properties to sets of sentences whose lexical and grammatical restrictions reflect their semantic restrictions. Through recent decades, work has concentrated on the subject-matter sublanguages. The sublanguage approach has been applied in various fields including information retrieval, machine translation and text understanding. Domains covered range from pharmacology, mechanical engineering, law, chemistry, weather reports and even recipes. A collection of work can be found in Kittredge and Lehrberger (1982).

A sublanguage is not simply an arbitrary subset of the sentences of a language. According to Lehrberger (1982:102), the distinctive features of a sublanguage include:

- limited subject matter;
- lexical, syntactic and semantic restriction;
- 'deviant' rules of grammar;
- a high frequency of certain constructions;
- unusual features of text structure; and
- the use of special symbols.
The sublanguage approach has also been successfully applied to some fields which do not have obvious subject-matter restrictions, such as dictionary definition analysis (Barnbrook & Sinclair, 1994) and term base analysis (Ahmad & Al-Jabir, 1998). As a type of language reference, dictionaries of biography have a similar nature to conventional dictionaries and term bases. A review of the sublanguage analysis in definitions and term base analysis might shed light on the analysis of biographical entries.

3.2.1 Language of definitions: general language dictionaries

Given that dictionaries of biography, or indeed any existing dictionaries, are compiled by human lexicographers, the assumption of a lack of structure or consistency across the dictionaries may be well founded. However, there has been a key development in lexicography recently, particularly with respect to the possible existence of structure in the definitions compiled by lexicographers for dictionaries used by the general public. John Sinclair, the pioneer of corpus-based dictionary analysis and his colleague, Geoff Barnbrook, have studied the structure of definitions used in a learner's dictionary, the Cobuild Student Dictionary (Barnbrook & Sinclair, 1994). The design of Cobuild is unique in that the targeted audience are initial learners of English. Barnbrook and Sinclair's work attempted to seek a sublanguage which was substantially simpler compared to implementing a general parsing strategy. Taking the sublanguage approach has improved the efficiency and eased the parsing task.

Initially, it was not obvious whether a limited subject matter applied to the definitions. On the one hand, definition entries form a subset of the language as a whole and are associated with a specialised area of human communication, that is defining new concepts. On the other hand, the headword list covers a wide range of subjects and definitions, and many explanations of the meanings of the headwords could contain various areas of knowledge. Despite the wide range of the subjects, definitions tend to be brief explanations with just sufficient information to enable the headword to be understood. In addition, the vocabulary chosen is often restricted to the simplest terms.

The first sentence of a definition can be parsed through a 'sentence grammar'. According to Barnbrook and Sinclair (1994), the sentence can be initially parsed into the left part and the right part. The two parts are related to each other through the relation of equivalence, in that the left and right parts essentially mean the same thing. Both parts are further connected through the relationship of explicitness in that the right part is more explicit than the left.
part.

The left and the right parts are connected to each other through a *hinge*, which comprises words that generally belong to the grammatical category of the words ‘to be’. Typically, a dictionary entry in Cobuild looks like:

*An antique is an old object which is valuable because of its beauty or rarity.*

The reader will note the rather idiosyncratic way of defining a word in that the *headword* is repeated in the definition. The left part in Barnbrook and Sinclair’s framework is essentially the phrase which merely introduces the headword in the definition and used as a pedagogic device. It is the right part which is of significance for us.

The right part is more important for us because it comprises a superordinate and at least one discriminator: “a superordinate requires a discriminator to express its relation to the headword, and a discriminator discriminates among hyponyms” (Barnbrook & Sinclair, 1994:20). Occasionally, the definitions may include a synonym or a synonymic phrase in the right part. Occasionally, the right part starts with a hinge.

\[
\text{[hinge]} \ (\text{superordinate discriminator}) \ / \ (\text{synonymic phrase}) \ / \ \text{synonym}
\]

There are further refinements to the grammar described above, suffice it to say here that using the above grammar the authors have found eight different definition types. These definition types relate closely to the grammatical category of the headword or lemma. For example, for nouns the authors suggest two local grammars covering a majority of nouns in the Cobuild dictionary. One of the structures of the local grammar for nouns is:

\[
\text{initial article} \\
\text{cotext1} \\
\text{headword} \\
\text{cotext2} \\
\text{hinge} \\
\text{matching article} \\
\text{initial discriminator(s) and superordinates(s)} \\
\text{subsequent discriminator(s)}
\]
### Chapter 3. A Sublanguage in Dictionaries of Biography

<table>
<thead>
<tr>
<th>Lemma Category</th>
<th>Example</th>
<th>Type</th>
<th>Local Grammar</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Nouns</strong></td>
<td>An abstract of an article or speech is a short piece of writing that summarises the main points of it.</td>
<td>A</td>
<td>initial_article cotext1 headword cotext2 hinge matching_article initial_discriminator(s) superordinate(s) subsequent_discriminator(s)</td>
</tr>
<tr>
<td></td>
<td>If you have charge of something or someone, you have responsibility for them. When there is an emission of gas or radiation, it is released into the atmosphere.</td>
<td>G</td>
<td>hinge cotext1 left_hand_hinge article headword matching_cotext right_hand_hinge remaining_definition_text</td>
</tr>
<tr>
<td><strong>Verbs</strong></td>
<td>If you apply a rule, system or skill, you use it in a situation and activity. When you design something new, you plan what it should be like.</td>
<td>B</td>
<td>initial_hinge cotext1 headword cotext2 matching_pronoun(s) for cotext1 remainder</td>
</tr>
<tr>
<td></td>
<td>To perform a task, action, or service means to do it.</td>
<td>C</td>
<td>initial_infinitive_marker headword hinge remainder</td>
</tr>
<tr>
<td></td>
<td>If you say that something depends on something else, you mean that it will only happen if the circumstances are right.</td>
<td>D</td>
<td>hinge projection cotext1 headword cotext2 remainder</td>
</tr>
<tr>
<td><strong>Adjectives</strong></td>
<td>Something that is debatable is not definitively true or not certain.</td>
<td>E</td>
<td>cotext1 left_hand_hinge headword right_hand_hinge remainder</td>
</tr>
<tr>
<td></td>
<td>If you are dependent on someone or something, you need them to survive.</td>
<td>F</td>
<td>hinge cotext1 left_hand_hinge headword cotext2 right_hand_hinge matching_cotext1 remainder</td>
</tr>
<tr>
<td></td>
<td>A steady person is sensible and reliable.</td>
<td>H</td>
<td>hinge cotext1 left_hand_hinge article headword matching_cotext right_hand_hinge remainder</td>
</tr>
</tbody>
</table>

**Table 3.1: Definition types in Cobuild by Barnbrook and Sinclair (1994)**

50
and the other structure is:

```
hinge
cotext1
left hand hinge
article
headword
matching cotext
right hand hinge
remaining definition text
```

Definitions in Cobuild have formed eight categories and each category has its distinctive structure. In addition to the two structures for the nouns, that is Type A and Type G, there are three structures for verbs and another three for adjectives (see Table 3.1 for details).

Given that words belonging to the category noun comprise a significant majority of words in Cobuild and indeed in any other dictionary, it is encouraging to see that the lexicographers are quite parsimonious in the number of structures they use to define nouns. This is also encouraging for us because we are only interested in proper nouns and expect the biographers to exercise the same level of parsimony as lexicographers.

### 3.2.2 Language of definitions: special language dictionaries

Special languages are good examples of sublanguages in which one finds a profusion of nouns, a limited number of sentence types and nouns with restricted semantics. Terminology dictionaries and the electronic variants of them called terminology databases, comprise definitions of terms of a specialist domain. These terminological resources are frequently multilingual and subject to frequent changes reflecting the evolution of the specialist domain. A large number of entries in a terminology resource are names of concepts, objects, devices and materials and a terminology dictionary essentially comprises definitions for a collections of nouns.

In Barnbrook and Sinclair's work (1994) there is an assumption that lexicographers use distinctive and repetitive patterns. This assumption has been further demonstrated in Ahmad and Al-Jabir's work (1998) to be applicable to specialist lexicographers, the terminologists. Methods have been developed for terminology databases of seven specialised domains, ranging
from paint technology to nuclear technology, to construct a knowledge base for each of the fields directly by parsing the definitions of terms and then mapping the parsed output onto an interactive notation, \( \lambda \)-calculus. The workflow of the system named DEARSys is illustrated in Figure 3.1 (Al-Jabir, 1999).

The methodology on how to map a definition entry in the term base into a frame representation is demonstrated taking a partial definition of the term *quark* in the domain of particle physics as an example. In a typical particle physics term base, the term *quark* is defined as:

**quark**: A fundamental fermion that has strong interactions.

First of all, the pre-tagging component recognises those terms which are already in the term base, especially those containing multiple tokens, for instance, ‘strong interactions’. The output of the pre-tagging unit is:

**quark (q): A fundamental fermion that has strong_interactions.**

The second task of pre-processing is to syntactically tag the above description into:

**quark/NN (/q/NN)/SYM:/: A/DT fundamental/JJ fermion/NN that /IN has/VBZ strong_interactions/NNS ./**

where the NN tag represents singular or mass nouns, SYM is a symbol, DT is a determiner, JJ is an adjective, IN is a preposition or a subordinating conjunction, VBZ is a third person singular present verb, and NNS is a plural noun. A full set of the part of speech tags are listed in Appendix B.
The definition is now ready for the logical analysis stage. This involves a morphological reduction of key verbs, identification of irregular forms and conversion into a formal structure. After the logical analysis, the definition is converted into:

\[
\text{all}(X, \text{quark}(X) \Rightarrow \text{fundamental}(X) \land \exists(B, \text{fermion}(B) \land \text{isa}(X, B)) \\
\exists(A, (\text{set}(A) \land \text{length}(A, set)) \land \text{all}(C, \text{member}(C, A) \Rightarrow \text{strong_interaction}(C) \land \text{have}(X, C)))
\]

Finally, the definition is represented as a frame as below:

\[
\text{frame}(644, \text{quark}, \text{isa}, \text{[fermion]}, \text{[chrc, have]}, \text{particle_physics}). \\
\text{slot}(644, \text{quark}, \text{chrc}, [], \text{[fundamental]}, [[]]) \\
\text{slot}(644, \text{quark}, \text{have}, [], \text{[strong_interaction]}, [[\text{number}([\text{strong_interaction}, \text{set}])]])
\]

Each frame contains parameters, including a unique reference, a frame name, a supertype link, a list of supertypes and a list of slots which represent the characteristics of the term. The knowledge in the definitions initially encoded in text format can be mapped onto frames. Therefore, it is possible to reason over it using FOL (First Order Logic) to dynamically link the various frames. Recall that the key conclusion of Barnbrook and Sinclair was that nouns, of various kinds, are defined almost as if the lexicographer had had a template available to him or her. The terminology of specialised disciplines again is largely comprised of nouns; hence, it is straightforward to extend the Barnbrook-Sinclair method to specialist domains.

3.2.3 Language of definitions: biographical dictionaries

A biographical dictionary contains a list of biographical entries. Each entry starts with a headword, that is a personal name. The entries are usually arranged according to the alphabetical order of the headword. The objective of a dictionary of biography is to provide, in a compact and readable form, a brief outline of the lives of the men and women from across the world. The study of the behaviour of the language used in dictionaries of biography is therefore the key to the extraction and representation of knowledge about personal names, a major part of the proper names.

There are many different kinds of biographical dictionaries. A typical biographical dictionary includes people from various disciplines, such as politics, religion, science, literature, etc. For
example, *Oxford Who's Who in the 20th Century* (1999) and *Pocket Oxford Dictionary of Biography* (1997). Some dictionaries of biography are designed for audiences in a specific subject area, and therefore contain a restricted group of people. For example, the *Baker's Biographical Dictionary of Musicians* (1992) discusses musicians and composers from ancient Greece to the present date.

Not all biographical dictionaries are monolingual. For example, *Who's Who in China: Current Leaders* (1994) contains biographical entries for Chinese leaders since the foundation of the People's Republic of China up to the year in which the dictionary was published. Each entry has its corresponding English translation.

Nowadays, the line between dictionaries and encyclopaediae is becoming increasingly blurred. Lexicographers have gradually developed an interest in proper nouns and this, in part, can be attributed to the emergence of college dictionaries. A good example is the *American Heritage Dictionary of the English Language* (AHD, 1992), which gives brief entries for all the presidents of the USA. It also provides gazetteer-like information about nation states, major cities of the world and most towns in the USA, along with entries on whom the editorial board of the AHD considers important people, such as musicians, writers, sports stars and scientists. Such dictionaries containing encyclopaedic entries are, sometimes quite logically, referred to as *encyclopaedic dictionaries*. Compared to the AHD, the *Oxford Dictionary of Music* (1994) is more similar to an encyclopaedia than a dictionary. It covers encyclopaedic knowledge about terms in music as well as the biographies of musicians.

With the Internet, as well as electronic and web publishing technologies, more biographical dictionaries have taken electronic form. The biographical entries are either put on a CD-ROM, wrapped with a user interface for searching or are hosted on a web server coupled with a web interface. The *Oxford Interactive Encyclopaedia*, available on a CD-ROM, covers the entire *Pocket Oxford Dictionary of Biography* (1997) and several other reference books. The change of media has improved their accessibility dramatically.

Although these biographical dictionaries have taken electronic form, they only offer machine-aided user access to biographical entries. During the searching process, each entry will be treated as a 'bag of words'. The knowledge contained in the entries is not represented in a form which a computer can understand in order to make inferences. However, the vast amount of data can be captured as corpora for empirical linguistic studies on the behaviour of a language.
<table>
<thead>
<tr>
<th>Dictionary</th>
<th>Einstein</th>
<th>Cage</th>
<th>Chomsky</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scientists (1994) [AHD]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Larousse Dictionary of Scientists</td>
<td>Einstein, Albert (1879-1955) German-Swiss-American physicist who ranks with Galileo and Isaac Newton as one of the greatest conceptual reservoirs of our understanding of the universe. Born in Ulm, Bavaria, of Jewish parentage, [...] he was awarded the 1921 Nobel Prize.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(1994)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Concise Oxford Dictionary of</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linguistics (1997)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3.2: A comparison between entries from different English biographical dictionaries
Before we go on to a detailed analysis of the structures of definitions for proper nouns in biographical entries, it is interesting to compare four dictionaries of biography (see Table 3.2). The entries of the physicist, Albert Einstein, and the composer John Cage and the linguist Noam Chomsky are considered. Similar structural templates can be identified across the four dictionaries named in Table 3.2, one reference dictionary, *American Heritage Dictionary* (AHD, 1992), two specialist dictionaries, the *Oxford Dictionary of Music* (ODM, 1994) and the *Concise Oxford Dictionary of Linguistics* (CODL, 1997) and one specialist biography, the *Larousse Dictionary of Scientists* (LDS, 1994).

There are nevertheless clear differences between the four dictionaries: AHD includes pronunciations; LDS notes more than one nationality; ODM includes the place of birth before the date of birth. The specialist dictionaries contain references to the key works of the individuals (books, musical arrangements, instruments) and references to other authors. In each of the dictionaries, the same templates are used with some variations, and across various dictionaries there is some consensus apparent in the entry about the template for entering and elaborating on a name. It is observed that almost all the entries have the following template structure:

- **Hinge:** Year of birth [Year of death]
- **Initial-discriminator(s):** Nationality [Nationality at birth (if different)]
- **Superordinate(s):** Profession [Other profession(s)]
- **Subsequent-discriminator:** Further attribute(s)

So it appears on the basis of the initial analysis that proper names generally fall into the Type A local grammar, classification of Barnbrook & Sinclair (1994).

Our analysis of Chinese dictionaries of biography further encourages us to believe that the type A structures are used by many biographers. Consider, for instance:

郭沫若作家，诗人，剧作家，历史学家，考古学家，古文学家，社会活动家。

Guo Moruo (1892-1978) writer, poet, playwright, historian, archeologist, social activist.

Here the initial-discriminator is missing which is not surprising in that this is a dictionary of Chinese biography, and the information related to the nationality is subsumed in the title.
Chapter 3. A Sublanguage in Dictionaries of Biography

<table>
<thead>
<tr>
<th>Chinese DoB</th>
<th>Entry in Chinese</th>
<th>English Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>China on internet</td>
<td>Ai Qing (1910-) poet. Zhejiang Jinhua origin.</td>
<td></td>
</tr>
</tbody>
</table>

Table 3.3: A comparison between entries from different Chinese biographical dictionaries of the dictionary. Our hypothesis is that distinctive and repetitive patterns exist in the definitions of proper names in dictionaries of biography.

The observation of the similarities amongst various biographical dictionaries suggests that the biographical entries, even those from different dictionaries, share special structural similarities. The structural characteristics could be used in parsing the texts containing knowledge about personal names for knowledge acquisition. There seems to be a potential sublanguage in the texts.

3.3 Structured Definitions in Biographical Dictionaries

The sublanguage approach has been successfully implemented into systems for automatically analysing general lexica (Barnbrook & Sinclair, 1994) as well as term bases (Ahmad & Al-Jabir, 1998). A dictionary of biography comprises definitions of proper nouns and hence it is of interest to see whether the sublanguage approach could be extended to biographical information with a view to creating a biographical knowledge base.
3.3.1 Sublanguage of an English biographical dictionary

Special characteristics have been observed in the language used in dictionaries of biography including grammatical features, special text constructions as well as the use of special symbols.

Consider the entry for 'Feynman':

Feynman /'feinmen/, Richard Philip (1918-88), American theoretical physicist.

The definition part of the entry does not have a valid predicate. The surname Feynman is followed by its phonetic correspondence separated from the person's first name and middle name by a comma. Two round brackets enclose two numbers connected via a hyphen which suggests the year of his birth and the year of his death. Following a comma, the noun phrase offers the definition or the classifying phrase placing the person in a certain category according to one's contribution and lifetime experience.

He worked in quantum electrodynamics and introduced important new techniques for studying the electromagnetic interactions between subatomic particles. This approach is expressed in diagrams that describe the exchange of particles (Feynman diagrams). He shared the Nobel Prize for physics in 1965.

Elliptical constructions occur in the definition part, which involves the omission of the indefinite articles a or an and copula be. This makes the first "sentence" of each biographical entry fragmented. At the same time, the construction with two numbers connected via a hyphen and surrounded by a pair of round brackets is interpreted as the year of birth and the year of death. Such constructions are used in almost every biographical entry.

Butterfield, William (1814-1900), English architect.

The grammatically complete sentence with no reduction would be:

Butterfield, William, who was born in 1814 and died in 1900, was an English architect.

The coordinating conjunction is used to link different professions or profiles about the person.
Ferrari, Enzo (1898-1988), Italian car designer and manufacturer.

When a person is from an aristocrat background, the definition is usually presented differently.

**Henry III** (1207-72), son of John, reigned 1216-72.

A restrictive relative clause is used to further discriminate the named-entity.

Pauli, Wolfgang (1900-58), Austrian-born American physicist who worked chiefly in Switzerland.

Appositions are frequently used as post modifications.

Kirchhoff, Gustav Robert (1824-87), German physicist, a pioneer in spectroscopy.

Typographical markers have been used in biographical entries. The headwords are often marked as bold and the title of a book, a piece of music or painting is marked in italics.

Montaigna /mamˈtaːnje/ ... He is noted for his altarpiece *Sacra Conversazione* (1499).

The structure of the headword is usually one's surname with the exception that if someone is a member of the royal family, the headword is the first name followed by the royal title.

Margaret, Princess, Margaret Rose (born 1930), only sister of Elizabeth II.

Sometimes, when more than one entry has the same surname, it would be marked up with numbered superscripts according to the alphabetical order of their first name.

Jackson\(^1\), Andrew (1767-1845), American general and Democratic statesman, ...

Jackson\(^2\), Glenda (born 1936), English actress and politician.

Jackson\(^3\), Jesse (Louis) (born 1941), American politician and clergyman.

Brackets are often used to insert the year of birth and death, and the year of a certain event or when some work was completed.

Paton, Alan (Stewart) (1903-88), South African writer and politician. He is best known for his novel *Cry, the Beloved Country* (1948), ...
3.3.2 Sublanguage of a Chinese biographical dictionary

Proper name analysis like many other fields in natural language processing tends to be Anglo-centric. Assumptions are often made based on linguistic and naming conventions in English language. The study of proper nouns in a typologically different language, Chinese, will not only help in extracting knowledge of proper nouns in languages other than English but also improve our understanding of the concept of sublanguage in a language independent manner.

Our analysis of Chinese biographical entries relies on a web-available dictionary of biography (DoB) http://www.chinaoninternet.com/famous. This DoB contains more than 270 entries, covering the time interval from 700 BC to the current day and a diversity of professions including philosophers, politicians and artists.

Despite its peculiar characteristics, the Chinese dictionary of biography shares many similarities with the Oxford DoB. Our analysis suggests that the sublanguage approach could be applied language-independently.

Much like the ODoB, the CDoB has a separate entry for each individual. The name of the person is followed by his or her date of birth (and date of death if appropriate). This is followed by professions of the named person. Then an elaboration part is presented.

Consider, a rather complex entry for the polymath ‘Guo Moruo’:

郭沫若 (1892 -1978) 作家，诗人，剧作家，历史学家，考古学家，古文学学家，
社会活动家。
Guo Moruo (1892-1978 writer, poet, playwright, historian, archaeologist, social
activist.

原名郭开贞，又名郭鼎堂。四川乐山人。
Originally named as Guo Kaizhen, also Guo Dingtang. Sichuan Leshan origin.

郭沫若出生在一个中等地主兼商人的家庭。早年就学于嘉定高等小学、嘉定中学堂、
成都高等学堂的分设中学。童年时便开始广泛接触文学作品[...]
He was born in a middle-class landlord and merchant family [...]

Each biographical entry in CDoBs has two distinctive parts: in the first part, the referent of the headword is assigned to one or more groups according to their profession or their achievements. Information is offered regarding the year of birth, year of death, profession, other names and origin. This part tends to be brief and fragmented.
A detailed description follows, covering education and career as well as activities and events one has participated in and the great works one has produced. In this section, the sentences tend to be more complete compared to the definition part.

Elliptical constructions are the most significant feature of a biographical entry in CDoB, especially the definition part of the entry. Articles, including *a*, *an*, *the* and copula, such as *is* are often ommitted on such occasions.

Guo Moruo (1892-1978) writer, poet, playwright, historian, archaeologist and social activist.

The grammatically complete version of the above sentence should be:

Guo Moruo was born in 1892 and died in 1978. He was a writer, poet, playwright, historian, archaeologist and social activist.

The elaboration part is more grammatically complete. The subject of a sentence is often omitted since the context suggests it refers to the headword.

(Guo Moruo) started to have a wide exposure of literature at an early age. In 1914, (he) studied in Japan, (he) started in medicine and converted to literature.

Phrases are often connected via a comma when listed.

writer, poet, playwright, historian, archaeologist, social activist.

A province name is followed by a town or a city's name and followed by 人 to specify one's origin or birth place as in 山西大同人 (Shanxi Datong origin).
3.3.3 Patterns in dictionaries of biography

According to Barnbrook and Sinclair (1994:17), a sublanguage is "more positionally restricted than the general language, and this leads to major simplifications in the grammar". The correlation between structure and function suggests the sublanguage may follow a special grammar which is much simpler compared to the grammar in the general language. The investigation is carried out on a cross-lingual basis in order to highlight both the similarities and the differences between different languages used for the same purpose.

The headword, the surname of a person, is the key identifier of a personal name. Although useful in many other applications including speech recognition and generation, the phonetic information is not the main concern in our study.

The digits connected by a hyphen and surrounded by a pair of round brackets conveniently positioned themselves between the headword and the definition body. This part of an entry acts structurally as a hinge. Moreover, it offers information about the years of birth and death of a person.

The final part of (the fragmented) text is called the definition body since the function of the text is to provide the position or the status of the referent in relation to the rest of the world. In this case, a person is often classified according to profession and nationality. The final component of a biographical entry contains several statements to emphasise a person’s significant contributions to the world.

Given that proper names are nouns, many of the dictionaries of biography entries in English and in Chinese have the template Type A as in Barnbrook and Sinclair (1994).

Headword (hinge) [initial-discriminator(s)] superordinate [subsequent-discriminator(s)]

**Type $A_1^{PN}$**: The most common and simplistic structure in the definitions of proper names does not contain subsequent-discriminators. Following the convention in Barnbrook and Sinclair (1994), this type of structure is named as Type $A_1^{PN}$. Two examples will suffice here (see Table 3.4):

**Feynman, Richard Philip** (1918-88), American theoretical physicist.

Similarly, an example of the basic pattern of Type $A_1^{PN}$ in the Chinese DoB:
Chapter 3. A Sublanguage in Dictionaries of Biography

<table>
<thead>
<tr>
<th>Source</th>
<th>Headword</th>
<th>Hinge</th>
<th>Discriminator(s)</th>
<th>Superordinate(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>OxfordDoB</td>
<td>Feynman, Richard Philip</td>
<td>(1918-88)</td>
<td>American</td>
<td>theoretical physicist</td>
</tr>
<tr>
<td>China on internet</td>
<td>Ai Qing</td>
<td>(1910 -)</td>
<td>浙江金华人</td>
<td>诗人</td>
</tr>
</tbody>
</table>

Table 3.4: The structure of Type $A_1^{PN}$

<table>
<thead>
<tr>
<th>Source</th>
<th>Headword</th>
<th>Hinge</th>
<th>Discriminator(s)</th>
<th>Superordinate(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>OxfordDoB</td>
<td>Aalto, Alvar</td>
<td>(1898-1976)</td>
<td>Finnish</td>
<td>architect, designer</td>
</tr>
<tr>
<td>China on internet</td>
<td>Song Qingling</td>
<td>(1893-1981)</td>
<td>Guangdong Wenchang origin</td>
<td>politician, social activist</td>
</tr>
</tbody>
</table>

Table 3.5: The structure of Type $A_2^{PN}$

艾青 (1910 ~) 诗人。 浙江 金华 人。
Ai Qing (1910~) poet. Zhejiang Jinhua origin.

Type $A_2^{PN}$: We have found variations of Type $A_1^{PN}$ in both English and Chinese which have more than one superordinate. This variant type is called Type $A_2^{PN}$. For example in Table 3.5


Similar cases can be observed in Chinese biographical entries:

宋庆龄 (1893~1981) 政治家, 社会活动家. 广东 文昌 人.

Type $A_3^{PN}$: Occasionally, some entries have subsequent discriminators as well as initial discriminators. The subsequent discriminator could be an adjectival phrase as well as an adjectival clause (see Table 3.6).

Haber, Fritz (1868-1934) German chemist who invented ...
Marc, Franz (1880-1916) German expressionist painter associated with the Blaue Reiter group of artists.

Chinese sentences rarely have post-modifications for nouns. Subsequent discriminators often take the form as another sentence or text fragment.

**Type \( A_3^{PN} \):** There are some entries in the *Pocket Oxford Dictionary of Biography* (1997) where we find a weaker form of a discriminator and superordinate structure in that the superordinate tends to be missing and Consider, for instance:

Elizabeth, I (1533-1603), daughter of Henry VIII.

The definition body is an apposition which explains the relationship of the headword with other named entities equally or even more well-known. In such circumstances, the superordinate may be considered as the generic superordinate which is 'person' in biographical context and the apposition therefore is regarded as the discriminator. Note that such instances often occur when the referent of the headword is or was a royal celebrity. Type \( A_4^{PN} \) also occurs frequently for a person who plays a significant role within a specific organisation. The entry for 'Wu Disheng' in Table 3.7 is such a case.

Semantically, the entry falls in the following structure:

**Headword (hinge) affiliation affiliate**

Definitions in Type \( A_4^{PN} \) sometimes contain multiple affiliations within the discriminator(s).

See the following Chinese DoB entry for example:

<table>
<thead>
<tr>
<th>Source</th>
<th>Headword</th>
<th>Hinge</th>
<th>Initial Discriminator(s)</th>
<th>Superordinate(s)</th>
<th>Subsequent Discriminator(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>OxfordDoB</td>
<td>Haber, Fritz (1868-1934) German</td>
<td>chemist</td>
<td>who invented...</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Marc, Franz (1880-1916) German</td>
<td>expressionist painter</td>
<td>associated with the Blaue Reiter group of artists.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Chapter 3. A Sublanguage in Dictionaries of Biography

<table>
<thead>
<tr>
<th>Source</th>
<th>Headword</th>
<th>Hinge</th>
<th>Affiliation(s)</th>
<th>Affiliate(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>OxfordDoB</td>
<td>Elizabeth, I</td>
<td>(1533-1603)</td>
<td>daughter of</td>
<td>Henry VIII</td>
</tr>
<tr>
<td>Who's who in China</td>
<td>武迪生</td>
<td>1935-10月生</td>
<td>市长</td>
<td>沈阳市</td>
</tr>
<tr>
<td>China on internet</td>
<td>武迪生</td>
<td>1935-</td>
<td>mayor of</td>
<td>Shenyang</td>
</tr>
<tr>
<td>China on internet</td>
<td>武迪生</td>
<td>(1899)</td>
<td>夫人及外交助手</td>
<td>蒋介石</td>
</tr>
<tr>
<td>Song Meiling</td>
<td>(1899- )</td>
<td>wife and diplomat</td>
<td>assistant of</td>
<td>Jiang Kaishe.</td>
</tr>
</tbody>
</table>

Table 3.7: The structure of Type $A_4^{PN}$

宋美龄 (1899- ) 蒋介石 夫人及外交助手.

Song Meiling (1899- ) wife and diplomatic assistant of Jiang Kaishe.

广东 文昌 人.

Guangdong Wenchang origin.

Complex types Finally, there are a number of entries where more complex structures are found. Consider, for example,


Here we see a typical

Headword (Hinge) [Initial-discriminator(s)] Superordinate [Subsequent-discriminator(s)]

structure which fits in Type $A_4$ combined with Type $A_4$ whose discriminator describes an organisational relationship between the headword 'Lie' with 'United Nations'.

3.4 Representing PN Definitions

Type $A_4^{PN}$ entries can be represented as a text processing system through the markup language XML as:
Taking the entry of Richard Feynman as an example, the XML marked entry is:

<entry>
  <definition>
    <headword>Feynman</headword>
    <phonetic-info>/'feinmen/</phonetic-info>,
    <other-names>Richard Philips</other-names>
    <hinge>(1918-88)</hinge>,
    <def_body>American theoretical physicist.</def_body>
  </definition>
  <elaboration>
    He worked in quantum electrodynamics and introduced important new techniques for studying the electromagnetic interactions between subatomic particles. This approach is expressed in diagrams that describe the exchange of particles (Feynman diagrams). He shared the Nobel Prize for physics in 1965.
  </elaboration>
</entry>

Similarly, a Chinese biographical entry

<entry>
  <definition>
    <headword>艾青</headword>
    <hinge>（1910～）</hinge>
    <def_body>诗人。原名蒋海澄，笔名莪伽、克阿、林壁等。浙江金华人。</def_body>
  </definition>
  <elaboration>
<!elaboration>

1928年就读于杭州西湖艺术学院，次年留学法国 ...

</elaboration>

The above marked up texts will be able to generate print quality output formatted as found in a DoB. But this is the limit to which mark up language will work since one may be able to use the marked up entries within a lexical database for querying the database but not for generalisations, for aggregation or any other knowledge-based operations that many exploit the existence of the templates noted as Type $A^{PN}_{14}$.

Furthermore, for named entity recognition task it may be useful sometimes to establish chains of inference to relate to individuals through their professions or the fact they were contemporaries. We think that these may be instances when the system will have to make inferences. Within the object relational and knowledge representation literature, there is a discussion of data structures that can be used to show complex interrelated items of data from which one could derive new facts.

Object oriented frameworks provide two kinds of primitive units: *individuals* and *attributes*. Individuals are used to represent entities, which can be either named entities such as 'Cage', or conceptual entities such as 'person'. A named entity is usually defined as a *Token* because it usually represents a concrete object. A conceptual entity, however, is defined as a *Class*, *Metaclass* or *Metametaclass* depending on the level of abstractness. Attributes are designed to represent relationships between entities or other relationships. For example, the detailed information about 'John Cage', such as the year he was born and what he achieved in his life, can be represented as the attributes of object 'Cage'. Therefore, encyclopaedic knowledge about a personal name can be represented as a frame-like object; an object with its relevant attributes.

The basic operation in a knowledge base is to store facts. This involves checking the compliance of a fact to the syntax of the notation and its consistency with previously stored facts as well as constantly updating the knowledge base. The second operation is to query a knowledge base and retrieve facts from it. One could ask a *yes* or *no* question such as "Is Einstein a physicist?" or a retrieval question such as "Find me all the physicists in the 20th century". The third operation involves inference which derives extra knowledge based on existing facts and rules. A question like "Is Einstein a scientist?" can be be answered by searching the stored facts. It must be derived from "All physicists are scientists" and "Einstein is a physicist". It is preferred to deploy a knowledge representation scheme to perform
all three of the operations to host our knowledge base. For the construction, querying and updating of the knowledge bases (KBs), Telos offers a variety of operations. Whilst TELL, UNTELL and RETELL are used to keep a KB up to date, RETRIEVE and ASK can be used to access it (Mylopoulos et al., 1990).

The semantics of a knowledge base refers to the meaning of objects and statements, in other words, the relationship between objects in the representation and entities in the world. Regarding the semantics, an appropriate KR scheme must meet the following requirements: first, the consistency of the knowledge base must be ensured, i.e. no contradicting facts should be derived; second, the inference scheme derives only sentences which are true with respect to the model (soundness); third, every true statement expressible in the language can be derived from the knowledge base (completeness) (Kramer & Mylopoulos, 1990:883). Plexousakis (1993:28-30) offered the proof of the consistency, soundness and completeness of Telos KBs.

A good mechanism for organising knowledge should not only provide efficient retrieval and inference but also aid the understanding and maintenance of a large body of facts and rules.

One way to organise the knowledge is to construct semantic networks. Inspired by cognitive science, semantic networks were proposed as a representational mechanism to improve the speed of access (Quillian, 1968). A semantic network is a structure for representing knowledge as a pattern of interconnected nodes and arcs. Nodes represent the concepts of entities, attributes, events and states whereas arcs represent relationships between nodes.

A further kind of knowledge representation schema is inspired by Minsky's (1975) proposal of frame theory. A frame is a complex data structure representing a prototypical situation. For example, the following frame

\[
\text{frame PERSON} \\
\text{birth-year: YEAR} \\
\text{death-year: YEAR constraint death-year > birth-year if (death-year != NULL)} \\
\text{nationality: COUNTRY}
\]

represents a concept of a person. The frame defines the abstract data types for each slot.

The purpose of building a biographical knowledge base is to organise the knowledge about each individual. It is essential to distinguish between objects representing individuals such as 'Einstein' and objects representing classes such as 'physicist'. In a knowledge base, the
former are often called instances or tokens and the latter are called classes or types. There are several relationships considered to be important once the distinction between tokens and types are made.

The relation between an individual and its type is often called classification. This relationship classifies the objects into the concepts to which they belong. For example, 'Einstein' and 'Newton' are both classified as physicists but 'Tony Blair' and 'Bill Clinton' are classified as politicians. Instance-of is often used to label a classification relation. Classification can be used to assure knowledge base consistency. The frame 'physicist' defines its attributes as slots which take specific data structures. An instance of the frame 'physicist' has to comply with the structure and data types. Classification can also increase the efficiency of the retrieval by restricting the search space to a certain category of objects.

The relationship between a class and its subclass is called specialisation which often bears the label IS-A. For example, 'physicist' IS-A 'scientist', 'scientist' IS-A 'person'. Specialisation achieves great conceptual efficiency because the properties of the super-classes can be inherited by the subclasses. These properties do not need to be stored for the subclasses, thereby improving storage efficiency.

frame SCIENTIST
    IS-A: PERSON
    written-work: PUBLICATION

The specialisation relationship between 'scientist' and 'person' allows the properties such as birth-year, death-year and nationality to be inherited from the class 'person'.

Aggregation refers to the relationship between a frame and its slots or an object and its attributes. Each biographical entry offers information about a specific individual. A person can be viewed as a social object consisting of the aggregation of facts about him or her.

Classification, specialisation and aggregation are all considered important relationships in our working domain; therefore, only KR schemes which support all these relationships can be considered possible for the biographical knowledge base.

Typically, the literature on knowledge representation comprises different representations. The internal representation is propositional logic. Although individuals and attributes have different functionalities, internally they are all represented as propositions. Each proposition is a tri-tuple with a source, a label, and a destination. For example, [Cage, isa, composer]
is a proposition with the node 'Cage' as source, 'isa' as label and 'composer' as destination. Frame-based representation is used for input because it is descriptive and shares similarities with natural language. More importantly, such representation supports aggregation, specialisation and classification, which are the three important organisational relationships we have encountered in the domain under investigation.

**Inference schemes** In a declarative knowledge base, the basic kind of inference is based on the provability relation of first-order logic. For example, according to *modus ponens*, 'A implies B' and 'A' derives 'B'. There is another form of inference which involves testing the subsumption relationship in frame-based knowledge bases. One frame is considered a subsumption of another frame if all its instances are also instances of the other. For example, the frame 'physicist' subsumes the frame 'scientist' which subsumes the frame 'person'. We require that the KR scheme chosen is able to offer both types of inference mentioned above.

In one of the well-known knowledge representation schemes, Telos, the inference rules consist of the transitivity of the *IsA* relationship. *IsA* supports subsumption (Plexousakis, 1993:26). The relationship between 'physicist', 'scientist' and 'person' can be defined as

\[
\text{IsA(physicist, scientist)}
\]

\[
\text{IsA(scientist, person)}
\]

Telos meets the requirements above, but it also benefits from the ease of use which is a considered issue in the deployment of a knowledge base. Since its creation, Telos has been incorporated into a number of knowledge base management systems.

ConceptBase (Jarke *et al.*, 1998), a deductive object base management system, was designed to simplify the interaction processes and make the construction of a Telos knowledge base more manageable. It implements a client-server architecture and integrates techniques from deductive and object-oriented databases into the logical framework of the Telos language. ConceptBase provides graphical interfaces for both input and output. As shown in Figure 3.2, the input interface acts like an editor within which one or more frame-like Telos constructs can be manually entered. The knowledge contained in a frame-based construct can become part of the knowledge base when the TELL button is pressed and the constructs do not have any syntactic errors. Java API has also been provided in ConceptBase in order to access the knowledge base in batch mode.
Chapter 3. A Sublanguage in Dictionaries of Biography

3.5 Conclusion

Throughout this chapter we have described the theoretical framework on which our work is based. The framework is based on our observation of the sublanguage in dictionaries of biography. Our analysis of biographical entries leads to the observation that there are indeed repetitive patterns used in biographies, published both in English and in Chinese.

Typically entries in a dictionary of biography contain two parts: definition and elaboration parts. The definition part fits in the definitional sublanguage whereas the elaboration part has the characteristics of free texts. The analysis of the dictionary of biography provides a link between a proper noun and its superordinate(s). Following the chain of superordinates in a dictionary of definitions, the ontology can be automatically constructed. Therefore, world knowledge about individuals can be represented in a structured way.

This work is built upon the previous work on dictionary analysis and term base analysis, taking the sublanguage approach. We have found that a typical dictionary of biography has a specific structure, which can be mapped onto knowledge representation formalisms. We have used Telos, a much more structured knowledge representation language than say, simple frames. We have also expanded this framework into the analysis of the Chinese dictionary of biography.
Based on this theoretical framework, we have constructed an archetype which acquires biographical knowledge and constructs knowledge bases with little human intervention. This system is named ABKAS which stands for the Archetype for Biographical Knowledge Acquisition System. In the following chapter, we will discuss all the aspects of the archetype including the specifications, design, implementation as well as the evaluations.
Chapter 4

Archetype for Proper Name Knowledge Acquisition

4.1 Introduction

The characteristic of natural language is a lack of structure. The richness and flexibility brings ambiguities which mediate against computation at various linguistic levels including the syntactic, semantic and pragmatic levels. The previous chapters discussed the distinctive and repetitive patterns construed at a pragmatic level within dictionaries of biography across two typologically different languages. The tightly packed information is highly structured for publishing purposes. The challenge here is to achieve the transformation from the text description to represented knowledge.

This chapter presents the ABKAS which is being developed as an, Archetype for Biographical Knowledge Acquisition System. We begin this chapter with a discussion of the system specification focusing on the major challenges during the system development. Our solutions to the challenges will be presented (Section 4.2).

This is followed by the design of the ABKAS which consists of three structural as well as functional components: the pre-processing unit, the sublanguage parser and the knowledge base constructor. Details are offered for every unit.

The implementation in Perl and Java and reasons for choosing the language are explained. The distributed implementation is considered as an original feature of ABKAS. We have also walked through the procedures when running ABKAS, which we explained using with a series
of screenshots. The output of ABKAS, the knowledge bases, are discussed considering three important knowledge representation aspects: generalisation, encapsulation and deduction.

The evaluation strategy is discussed followed by the evaluation process and the analysis of the results. The system evaluation from both quantitative and qualitative aspects shows encouraging results.

In order to validate the soundness of the theoretical framework in languages other than English, we have adapted ABKAS to process Chinese biographical entries alongside the English biographical entries. Details related to Chinese are introduced in parallel in many sections. We end this chapter with a summary of what we have achieved through the development of ABKAS.

4.2 System Specification

Figure 4.1: Overview of ABKAS

Figure 4.1 offers an overview of ABKAS which demonstrates the workflow of the system and transformation of knowledge from text description to represented knowledge in knowledge
Chapter 4. Archetype for Proper Name Knowledge Acquisition

<table>
<thead>
<tr>
<th>Types</th>
<th>Characteristics</th>
<th>POS Patterns</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_1$</td>
<td>Initial-discriminator with one superordinate (Basic pattern)</td>
<td>NNP+ JJ* NN+</td>
<td>British industrialist.</td>
</tr>
<tr>
<td>$A_2$</td>
<td>Basic pattern with multiple superordinates</td>
<td>NNP+ JJ* NN (NN,)* CC NN</td>
<td>Finnish architect and designer.</td>
</tr>
<tr>
<td>$A_3$</td>
<td>Basic with post-modifying clause or phrase as subsequent discriminator</td>
<td>NNP+ JJ* NN+ WP$/PP$</td>
<td>German chemist who invented ...</td>
</tr>
<tr>
<td>$A_4$</td>
<td>Missing superordinate</td>
<td>NN IN NNP+</td>
<td>Daughter of Henry VIII.</td>
</tr>
</tbody>
</table>

Table 4.1: Definition patterns of Type $A_i^{P,N}$

bases. Two key stages during the transformation process, which have been highlighted, illustrate the two key challenges in this work. These two challenges guided us during the development of ABKAS. On the one hand, there were considerations to do with how the mappings between syntactic structure and semantic templates could be achieved. On the other hand, there were considerations to do with how to generalise the concepts and construct a taxonomy of people from the bottom up with minimum amount of human intervention.

4.2.1 Mapping process

The method of mapping biographical entries between their syntactic structures and their semantic templates is the key factor for the design of the sublanguage parser.

If we analyse the patterns of the part-of-speech tagging for every type listed in Table 4.1, we can find consistent mapping between certain syntactic sequences and their correspondent semantic interpretation (see Table 4.2). CD refers to numbers. It is straightforward to see when two numbers are connected via a hyphen, the first number is the 'year of birth' and the second, 'year of death'. Nationality is often tagged as NNP due to the tagger we use, which does not distinguish proper adjective from proper noun.

Professions, as superordinates, are either nouns, for example, 'professor', or compounds which consist of a noun as the root, such as 'theoretical physicist', which are often tagged as JJ NN
Table 4.2: Mapping between part-of-speech patterns to semantic interpretation

<table>
<thead>
<tr>
<th>Part-of-Speech Patterns</th>
<th>Semantic Mappings</th>
</tr>
</thead>
<tbody>
<tr>
<td>(CD-CD)</td>
<td>year of birth-year of death</td>
</tr>
<tr>
<td>(CD-)</td>
<td>year of birth</td>
</tr>
<tr>
<td>NNP</td>
<td>nationality</td>
</tr>
<tr>
<td>JJ* NN+</td>
<td>profession</td>
</tr>
<tr>
<td>JJ* NN+ (NN,)* CC NN</td>
<td>profession (profession,)* and (profession)</td>
</tr>
<tr>
<td>WP$ / PP$</td>
<td>additional information (major achievements)</td>
</tr>
<tr>
<td>NN+ IN NNP+</td>
<td>relationship of related entity</td>
</tr>
</tbody>
</table>

where JJ refers to adjectives and NN is the tag for nouns. Sometimes, the modifier of the compound is a noun, such as 'film director', tagged as NN NN which has the brief notation as NN+. CC stands for conjunctions such as ‘and’ used to concatenate multiple superordinates. Furthermore, WP$ and PP$ are either clauses or phrases used as subsequent discriminators to offer extra information about the referent of the headword. Such information often describes the major achievement in the person’s life. IN is used for tagging the preposition ‘of’. This tag is a cue for identifying relational patterns and generate the relationships between the headword and the related entity or entities.

The syntax-semantic mapping achieves the transformation of knowledge in text descriptions into templates with filled semantic slots. The method discussed here has laid down the foundation for the design of the sublanguage parser.

4.2.2 Superordinate bootstrapping process

For every biographical entry, we can extract related attributes of the named individual including year of birth, year of death and nationality as discriminating attributes. This information will be encapsulated within a frame. The person's profession will act as the classifying superordinate which clusters all people who share the same profession and, therefore, have similar knowledge and background. The frame-based templates successfully classify individuals into groups.

People who are 'physicist', 'chemist', or 'astronomer' have different specialities and interests. On the other hand, they share certain number of characteristics, which could be captured
at a more abstract level and which are often defined in their superordinate concept (their supertype), that is 'scientist' in this case. Thus, the challenge here is to find further abstract levels of supertypes for the superordinate of a person which is generated by analysing biographical texts. This kind of generalisation can be carried out in an recursive manner. The generalisation not only leads to economical storage for the knowledge representation but also facilitates knowledge sharing and deduction via inheritance.

Figure 4.2: Superordinate bootstrapping process

The diagram in Figure 4.2 illustrates the method we have used to carry out the definition analysis for the construction of a hierarchical taxonomy. In this work, we would like to call it the superordinate bootstrapping process. Recall that a superordinate is often a noun phrase. It could either be an individual noun or a endocentric compound. According to Spencer, in endocentric compounds, one element functions as the head and the other functions as the modifier (1991:310). We have observed frequent occurrence of endocentric compounds as superordinates; for example, ‘theoretical physicist’ or ‘aircraft engineer’. It is possible to take the advantage of the modifier-head morphological structure to extract the ‘head’ as its superordinate. For instance, ‘physicist’ is the superordinate of ‘theoretical physicist’ and ‘nuclear physicist’.

```plaintext
Extract Superordinate

Does Superordinate exist in KB?

Is Superordinate an endocentric compound?

Search online dictionary to find definition of the Superordinate

Parsing definition

Tagging definition

Superordinate

KB

n=n+1

Yes

No

Recall that a superordinate is often a noun phrase. It could either be an individual noun or a endocentric compound. According to Spencer, in endocentric compounds, one element functions as the head and the other functions as the modifier (1991:310). We have observed frequent occurrence of endocentric compounds as superordinates; for example, ‘theoretical physicist’ or ‘aircraft engineer’. It is possible to take the advantage of the modifier-head morphological structure to extract the ‘head’ as its superordinate. For instance, ‘physicist’ is the superordinate of ‘theoretical physicist’ and ‘nuclear physicist’.
```
More frequently, superordinates appear to be a single word, such as ‘chemist’, ‘painter’ and ‘poet’. The extraction of their superordinates is not a straightforward process. The process starts from a superordinate which is extracted from encyclopaedic texts. If the superordinate already exists as a represented concept, the frame which holds the knowledge about the personal name can be asserted. Let us note the current superordinate as \( \text{superordinate}^{(n)} \).

If the \( \text{superordinate}^{(n)} \) does not exist in the knowledge base (KB), we need to find out the superordinate of the current superordinate which is \( \text{superordinate}^{(n+1)} \). If \( \text{superordinate}^{(n)} \) is an endocentric compound, we take the head of the compound, which often is the last word of the compound, as \( \text{superordinate}^{(n+1)} \). If, however, \( \text{superordinate}^{(n)} \) is a word, it will be submitted to an online dictionary in order to retrieve its definition.

For instance, the definition of “physicist” is obtained from the dictionary:

- **Headword:** physicist
- **Superordinate:** scientist
- **discriminator:** whose speciality is physics

Further analysis of the discriminator can generate the following frame-based representation.

\[
\text{(defineType physicist}
\begin{array}{ll}
\text{superotype} & \text{scientist} \\
\text{speciality} & \text{physics}
\end{array}
\text{)}
\]

This process will be carried out in a recursive manner with the requirement that a frame cannot be asserted if its supertype does not exist. In this case, \( \text{superordinate}^{n} \) becomes ‘scientist’. From the following definition:

- **SYL:** sci-en-tist
- **DEF:** one who uses scientific procedures and is involved in science, esp. the physical or natural sciences.

we get the frame representation of ‘scientist’:
Chapter 4. Archetype for Proper Name Knowledge Acquisition

(defineType scientist
  (supertype person)
  (uses scientific_procedures)
  (involved-in science))

In the process of dictionary query, analysis is iterated until the superordinate reaches the concept 'person'. In the definition of 'person', we have specified the generic information (such as name, year of birth, year of death and nationality) which applies to people from all professional backgrounds.

(defineType person
  (supertype entity)
  (name String)
  (year Integer)
  (nationality String))

The bootstrapping manner of retrieving superordinates offers a solution for the automatic conceptual modelling, which is a key factor in the design of the knowledge base constructor. In the following section, the design of ABKAS will be presented.

4.3 System Design

This section describes the underlying architecture of ABKAS, the system that we have developed to analyse biographical entries and to automatically extract knowledge of personal names. ABKAS is designed to contain three major units: the pre-processing unit, the sub-language parser and the knowledge base constructor (see Figure 4.3).

![Figure 4.3: The system architecture of ABKAS](image)

The pre-processing unit performs a number of tasks including the creation of the input data file and submitting them for Part of Speech tagging. This component is discussed at greater length in Section 4.3.1.
Chapter 4. Archetype for Proper Name Knowledge Acquisition

As illustrated in Figure 4.3, the central unit of ABKAS is the sublanguage parser which contains a number of finite state machines (FSMs). Each FSM is designed to parse certain types of text constructs. The parsing task is carried out by delegating subtasks to the appropriate FSM. Taking a biographical entry with part of speech tagging, the sublanguage parser performs the logical analysis and extracts the semantic relationships within the entry and maps them onto a frame-based structure. A more detailed explanation is given in Section 4.3.2.

The third part of ABKAS is the knowledge base (KB) constructor. This unit converts every template generated by the sublanguage parser into Telos format. A key part of the conceptual modelling of the knowledge base involves the interaction with an online dictionary to request the definitions of the concepts which are required as a part of the biographical knowledge base. These definitions are parsed to generate the taxonomy required to model the domain. This is discussed further in Section 4.3.3. For every unit of ABKAS, processing in English will be discussed first, followed by a discussion of the corresponding unit in Chinese.

4.3.1 Pre-processing unit

Pre-processing for English biographical texts The pre-processing unit carries out the following tasks:

1. Collecting biographical entries as input data;
2. Segmenting each entry into two parts and keeping the definition part; and
3. Submitting input texts to an online tagger for Part-of-Speech (POS) tagging.

Recall that our analysis of biographical entries suggests that distinctive patterns largely occur in the definition part of the entries. Therefore, we needed to segment each entry into two halves and keep the first half which is the definition part. Since the definition part is often bounded within one sentence, the boundary between the definition part and the elaboration part is simply the first full stop in the paragraph.

The segmentation process is designed to combine with the next task in the preprocessing unit, that of part-of-speech (POS) tagging. POS tagging comprises the task of assigning tags to words to reflect their syntactic category (Brill, 1995). LT.POS, developed by the Language Technology Group of Edinburgh University, has been chosen as the tagger for our
Chapter 4. Archetype for Proper Name Knowledge Acquisition

experiments because of its high accuracy typically of 90 percent, and for its freely available web service.

Benefiting from the client-server architecture of LT.POS, ABKAS is designed to have a automated web submission module which sends requests to LT.POS without the necessity of having the system installed locally.

Syntactic taggings provides valuable linguistic information which can significantly enhance the exploration of input texts. An example of a tagged entry is presented below: NNP stands for proper nouns; CD refers to numbers; JJ means adjectives and NN represents nouns. In LT.POS, punctuation marks are simply repeated in their tags.

Feynman_NNP ,_, Richard_NNP Philip_NNP (_( 1918-88_CD )_),_, American_NNP theoretical_JJ physicist_NN ._.

LT.POS use the Penn Treebank tag set. Details of the tag set can be found in Appendix B.

Preprocessing for Chinese biographical texts  The preprocessing for Chinese biographical entries involves similar tasks to the processing for English but the order of the execution of these tasks is slightly different. The preprocessing unit for Chinese carries out the following tasks:

1. Submitting an index page which contains URLs to JASMINE for tagging and at the same time collecting the entries as input data;

2. Removing HTML taggings from input data;

3. Segmenting each entry and keeping the definition part; and

4. Converting JASMINE tagging into LT.POS tagging.

JASMINE is the Chinese equivalent to LT.POS which offers POS tagging services for Chinese texts, in particular, web pages. JASMINE is developed by the Department of Information Engineering of the Chinese University of Hong Kong. It is implemented based on a hybrid algorithm for automatic segmentation and tagging of Chinese texts (Qin & Wong, 1996). It offers a web service which takes the URL of a web page as input and produces the tagged texts. We have designed the pre-processor to scan the web page located at http://www.chinaoninternet.com/famous/ourtime/index.htm and find all the hyperlinks for every
biographical entry and submit the links to JASMINE. Since the input data are tagged web pages, it is necessary to remove the HTML tags. The HTML removal function is built in as part of the pre-processor.

The next stage is the separation of the definition part and the elaboration part of each biographical entry, and finally the retention of the first half of the entries. The word 人 (person) followed by a full stop, is taken as the boundary.

A tagged Chinese biographical entry appears as follows:

阿炳/PSN (1893-1950) 民间/NN 音乐家/NN。江苏/NPN 无锡/NPN 人/NN。

In the entry above, NPSN means the tagged word is a personal name, whilst NN refers to a noun. NPN suggests that the preceding word is a place name. A full listing of the tag set used by JASMINE can be found in Appendix C.

<table>
<thead>
<tr>
<th>Before</th>
<th>After</th>
<th>Positional change</th>
</tr>
</thead>
<tbody>
<tr>
<td>*_NPSN</td>
<td>*_NNP</td>
<td></td>
</tr>
<tr>
<td>*_NP *_NPN</td>
<td>**_NNP</td>
<td>Move to the beginning of the definition body</td>
</tr>
</tbody>
</table>

Table 4.3: Mapping table of the tag translation

The Chinese JASMINE POS tagger utilises a different tag set from the *Penn Treebank* tag set used by the English tagger LT-POS. In order to allow the input to be parsed by the same sublanguage parser, it is necessary to convert the JASMINE tags into their equivalent form in the *Penn Treebank* tag set.

Such conversions are carried out based on a mapping table illustrated in Table 4.3. The first step is to convert the delimiter between a word and its tag from ‘/’ into ‘_’. The personal name tag NPSN and place name tag NNP are both converted into NNP, the generic proper name tag. Punctuation marks are duplicated. The place names are moved to the beginning of the entry in order to form a similar pattern with English entries. The tagged entry above is hereby converted into the following.

阿炳_NNP (1893-1950) 江苏无锡_NNP 民间_NN 音乐家_NN 。
Abing (1893-1950) Jiangsu Wuxi folk musician.
4.3.2 Sublanguage Parser

In the system specification section, we discussed the mapping process between syntactic and semantic structures in biographical entries. The design of our sublanguage parser is based on that discussion. A Finite State Machine (FSM) has been designed to parse text constructs for a specific type of definition within Type $A_{4}^{PN}$ (see Figure 4.4).

![Diagram of the workflow of the sublanguage parser]

Taking the definition from the biographical entry, the segmentor divides each entry of the biographical entry into three parts: headword, hinge and definition body.

Following Maurice Gross's notation (1993) in his research in local grammar, we have designed the finite state automata using Directed Acyclic Graphs (DAGs) for each segment of the biographical entries.

The headword part of an biographical entry contains not only the surname of a person as the headword but also other names that occur. For instance, when the entry of 'Richard Feynman' is parsed, we get:

- headword: Feynman
- other_names: Richard Philip

Chinese surnames are usually written before first names and they are considered as being a single unit. Therefore, the full Chinese name is usually taken as the headword.

- headword: "T" "Philip"

In "T" "Philip", "T" is a surname and "Philip" is a first name.

By parsing the hinge, we get the year of birth and the year of death of a person.
The most important part of the processing is the parsing of the definition body. The significant characteristics of Type A_4 is that it contains an 'of' which is tagged with 'IN' and it expresses the relationship between the headword and other named entities. Type A_1-3 definition bodies do not contain 'of'; instead, they start with a proper name which is tagged with "NNP" and contain a noun tagged with an NN. Type A_1-3 takes the form of a classifying description. The patterns formed by the POS tags form the cues for the identification of types. Based on these cues, the type extractor scans the definition body and assigns them to the appropriate path.

Type A_1 The FSM for this type captures the basic structure of the Type A definition body and extracts the semantic relationships.

For example, the Type A_1 FSM takes the above tagged definition body as input and generates the following output:

```
super-ordinate    composer
nationality      'French'
```
Chapter 4. Archetype for Proper Name Knowledge Acquisition

The Type $A_1$ FSM parses the definition body below in Chinese in a similar way.

丁玲_NNP (1904-1986) 湖南临澧_NNP 女作家_NN。

super-ordinate 女作家 (female writer)
native_of 湖南临澧 (Hunan Linli)

When there are two initial discriminators present, the first one is often a hyphenated word, such as “German-born” or “French-speaking”. The Type $A_1$ FSM decomposes this word and turns it into an attribute and value pair. For example, when the input

[ German-born_NNP US_NNP astronomer_NN ] ...

is processed, we get the following output:

superordinate: astronomer
nationality: ‘US’
born: German

Type $A_1$ FSM is able to handle complex superordinates, that is when a superordinate contains a pre-modifier. The pre-modifier can be either an adjective (tagged with a JJ) or an noun (tagged with a NN). This pre-modifier is combined with the key superordinate to form one superordinate. For example, with an input of:

American_NNP theoretical_JJ physicist_NN ...

we get the following output:

superordinate: theoretical physicist
nationality: American

Similarly,

Japanese_NNP food_NN technologist_NN ...

superordinate: food technologist
nationality: Japanese

With the following Chinese example:

阿炳_NNP (1893-1950) 江苏无锡_NNP 民间_NN 音乐家_NN ...

we can get a filled template as:

superordinate: 民间音乐家 (folk musician)
native_of: 江苏无锡 (Jiangsu Wuxi)
Chapter 4. Archetype for Proper Name Knowledge Acquisition

**Type A2** In Type A2 definition bodies, more superordinates are expressed via coordination. This FSM is designed to deal with coordination.

![Figure 4.7: The finite state machine for Type A2](image)

An example of Type A2 definition is illustrated below, taking the entry as follows:

Finnish\_NNP architect\_NN ,\_ designer\_NN ,\_ sculptor\_NN ,\_ and\_CC painter\_NN .\_

The NN tag is used for a singular or a mass noun, CC for a conjunction and NNP for a proper noun or a part of a proper noun. The tag CC signals the entry of a Type A2 and triggers the subtype extractor to forward this entry to the Type A2 finite state machine. The Type A2 FSM can handle multiple superordinates by generating a vector of superordinates.

- superordinate1: architect
- superordinate2: designer
- superordinate3: sculptor
- superordinate4: painter
- nationality: Finnish

成仿吾\_NNP (1897-1984) 湖南省新化县\_NNP 教育家\_NN, 
文学家\_NN and\_CC 翻译家\_NN .\_

- superordinate1: 教育家 (education expert)
- superordinate2: 文学家 (writer)
- superordinate3: 翻译家 (translator)
- native_of: 湖南省新化县 (Hunan, Xinhua)

**Type A3** If a definition body contains a post-modifying phrase, it will be forwarded to the Type A3 FSM. This FSM analyses the post-modifier and extracts an additional attribute for the referent of the headword. Type A3 is illustrated by the following example:
German-born NNP US NNP astronomer NN who WP proposed VBD the DT existence NN of IN two CD stellar JJ populations NNS ..

The following attribute is extracted:

proposed: the existence of two stellar populations

In Chinese, it is rare to see post-modifying clauses or phrases. Such kinds of constructs have not been found in the Chinese DoBs. Therefore, Chinese biographical entries do not need to be processed through the Type A3 FSM.

**Type A4** In Figure 4.9, the FSM captures an attribute which describes the relationship between the headword and the related entity. When Type A4 occurs on its own, no superordinate is generated. In such occasions, the default superordinate ‘person’ is supplied.

**Figure 4.9: The finite state machine for Type A4**

King NN of IN Saudi NNP Arabia NNP ..

superordinate: person

King of: Saudi Arabia

Although occurring rarely, some Chinese biographical entries follow the Type A4 format. The syntax of a Type A4 entry follows a regular expression of the form:
Chapter 4. Archetype for Proper Name Knowledge Acquisition

\[ \text{NPSN} (\,d^+ \sim [\,d^+] \,) \ \text{NPSN SDG NN, NPN NPN NN}. \]

which correlates to the semantic structure of: \text{headword} \ (\text{birth_year} \sim [\text{death_year}] \) \ \text{related_entity the relationship. province town 人}.

The trigger word of the Type A4 entry is the closed class word \text{的} which is the equivalent of the word 'of' in English. It connects the name of a person with the relationship word.

For instance:

宋美龄/NPSN (1899 ~) 蒋介石/NPSN 的/SDG 夫人/NN 及/CP 外交/NN 助手/NN。
广东/NPN 文昌/NPN 人/NN。
Song Meiling (1899-) wife and diplomatic assistant of Jiang Kaish.
Guangdong Wenchang origin.

The Type A4 FSM is designed to handle coordination within discriminators.

\begin{itemize}
  \item \text{native_of:} 广东文昌 (Guangdong Wenchang)
  \item \text{superordinate:} 人 (person)
  \item \text{夫人 (wife of):} 蒋介石 (Jiang Kaish.
  \item \text{外交助手(diplomatic assistant of ):} 蒋介石 (Jiang Kaish.
\end{itemize}

Finally, the semantic integrator combines the information extracted from the Headword FSM, Hinge FSM and the Definition FSMs and generates the complete list of semantic relationships.

4.3.3 Construction of Biographical Knowledge Bases

The sublanguage parser is able to parse the definition parts of biographical entries and extract information about personal names. Such information is presented in an intermediate template format. The templates can be furthered processed into various knowledge representation schemes via the template transformation process. Recall in Chapter 3, we have discussed the possible knowledge representation schemes and chosen Telos for the construction of biographical knowledge bases.

As Plexousakis (1993:3) noted, Telos offers "a formal notation, a deductive mechanism for drawing inferences from a body of statements represented by this formal notation (a knowledge base) and well-defined formal semantics that the deductive mechanism must respect in
drawing inferences.” ConceptBase (Jarke et al., 1998), the deductive object base management system which implements Telos, facilitates the construction of the biographical KBs as well as the interaction with them.

The knowledge base construction process is carried out in three stages (see Figure 4.10): first, the intermediate templates are transformed to Telos syntax; second, the conceptual taxonomies are modelled through the definition analysis; and third, a knowledge base is populated via the interaction with ConceptBase.

![Figure 4.10: The workflow of knowledge base construction](image)

### 4.3.3.1 Template transformation

In a Telos knowledge base, everything is represented as a proposition.

In a generic proposition like \( P(oid, x, l, y, tt) \), the oid is the key identifier, \( x \) is the source, \( l \) is the label, \( y \) is the destination and \( tt \) is the belief time of the proposition.

Telos propositions can be divided into *individuals* and *attributes*. The former represent entities and the latter represent relationships. Some classes are defined to impose the principles of relationships which include: Classification/Instantiation (*In/InstanceOf*), Specialisation/Generalisation (*IsA*) and Aggregation (Jarke et al., 1998).

In addition, *Token* is used to define *individuals* who cannot have instances. *SimpleClass* is used for *Classes* which only take Tokens as instances. A *SimpleClass* can be an instance of
a Metaclass. Class can take all the classes as instances. Other than the abstract data types, there are built-in primitive data types defined as Integer, String and Real.

<table>
<thead>
<tr>
<th>headword</th>
<th>Lippmann</th>
<th>Token</th>
</tr>
</thead>
<tbody>
<tr>
<td>superordinate1</td>
<td>physicist</td>
<td>Class</td>
</tr>
<tr>
<td>name</td>
<td>Gabriel Jonas</td>
<td>Attribute</td>
</tr>
<tr>
<td>year of birth</td>
<td>1845</td>
<td>Attribute</td>
</tr>
<tr>
<td>year of death</td>
<td>1921</td>
<td>Attribute</td>
</tr>
<tr>
<td>nationality</td>
<td>French</td>
<td>Attribute</td>
</tr>
</tbody>
</table>

Given an example of the output from the sublanguage parser in the intermediate template format, modifications must be made in order to reformat the template above into the Telos Syntax.

```
Individual Lippmann in Token, physicist with
attribute, other_name
  i_name : "Gabriel Jonas"
attribute, year
  born_in : 1845;
  died_in : 1921
attribute, nationality
  i_nat : "French"
end
```

From the template above for "Lippmann" in Telos, we can observe a set of undefined data types, such as "physicist" as Individual and "name", "year" and "nationality" as Attribute. In knowledge bases, concepts and attributes must be defined before being used to build up more knowledge. The acquisition of the conceptual knowledge leads to the conceptual modelling within the domain of biography. Next we look at how to extract the knowledge of the relevant concepts within the biographical domain.

### 4.3.3.2 Conceptual modelling

One of the key tasks for constructing a knowledge base is to analyse and model the domain in order to map real world knowledge into the logical formalisms. In ABKAS, this task is accomplished by taking a bootstrapping approach described in Section 4.2.2.
Every frame for an individual person is a 'Token' object. The definition of its superordinate is 'Class'. The superordinate concept captures information about a group of people which share the same superordinate. Defining the superordinate concepts in a knowledge base is a prerequisite of asserting the knowledge of proper names.

ABKAS is designed to query the KB in order to find out whether a concept is already defined or not. If yes, the Token object is TELLed. Otherwise, the superordinates will be passed on to an on-line dictionary - *Wordsmyth English Dictionary-Thesaurus*\(^1\).

The definition of "physicist" is obtained from the dictionary:

SYL: phys-i-cist
DEF: a scientist whose specialty is physics.

The definitions are also syntactically tagged and then semantically mapped onto the Telos format following a different set of sublanguage rules. Since the parsing of the definition entries is not the major concern of this work, detailed information about this respect will be limited in this thesis.

Individual physicist in Class is A scientist with
attribute
speciality: physics
end

The definition of scientist is also obtained in the same way.

SYL: sci-en-tist
DEF: one who uses scientific procedures and is involved in science, esp. the physical or natural sciences.

Words like "one", "someone" or "somebody" are often used as a replacement for 'A person' in definitions. It implies the superordinate of the concept is as generic as 'person'. Such definitions are standardised into 'A person who ..." form before it is analysed.

Individual scientist in Class is A person with
attribute
uses: scientific procedures
involved-in: science
end

\(^1\)http://machaut.uchicago.edu/cgi-bin/WEDT1.sh
In the process of dictionary query, analysis is iterated until the superordinate reaches the concept 'person'. In the definition of 'person', we have specified the generic information (such as name, year of birth, year of death and nationality) which applies to people from all professions.

```plaintext
Individual person in Class isA entity with
  attribute
  name : String;
  year : Integer;
  nationality : String
end
```

In the domain of biographical dictionaries, other than people, there exist entities such as places, organisations and artefacts. Consequently, we have defined 'person' as a subclass of entity. The top level concept 'entity' is defined as a 'Class' to be the root of the hierarchy by TELLing the following template into the knowledge base.

```plaintext
Individual entity in Class
end
```

Once the conceptual relationships exist in the knowledge base, knowledge about individual personal names can be TELLeed to the knowledge base. ConceptBase provides Java APIs for automated interaction. Taking advantage of this feature, the knowledge base constructor is able to call the Java API and send many records in a batch. So far, we have designed ABKAS to transform biographical entries into represented knowledge.

### 4.4 System Implementation

ABKAS was implemented in a combination of a number of programming languages, including Perl, Java and Perl Tk (see Table 4.4). For different modules we have chosen an appropriate language which, we believe, offers the suitable characteristics. The majority of the system is written in Perl and implemented under Perl v5.6.1. Perl offers a powerful regular expression engine which is particularly suitable for text manipulation. Therefore, we chose Perl for the implementation of the finite state machines for parsing and furthermore, there are many Perl modules for web interaction. Java was also used for interacting with the ConceptBase server via a Java API offered as part of the ConceptBase package for batch interaction.
Chapter 4. Archetype for Proper Name Knowledge Acquisition

<table>
<thead>
<tr>
<th>Module</th>
<th>Main Functionality</th>
<th>Programming Language</th>
<th>Important Characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preprocessing</td>
<td>web submission</td>
<td>Perl</td>
<td>web interaction modules</td>
</tr>
<tr>
<td>Sublanguage parsing</td>
<td>text manipulation</td>
<td>Perl</td>
<td>regular expression engine</td>
</tr>
<tr>
<td>Definition analysis</td>
<td>web submission &amp;</td>
<td>Perl</td>
<td>web interaction modules &amp; regular expression</td>
</tr>
<tr>
<td></td>
<td>text manipulation</td>
<td></td>
<td>engine</td>
</tr>
<tr>
<td>Knowledge base</td>
<td>interaction with</td>
<td>Java</td>
<td>integration with Java</td>
</tr>
<tr>
<td>construction</td>
<td>ConceptBase server</td>
<td></td>
<td>API offered by ConceptBase group</td>
</tr>
<tr>
<td>Graphical user interface</td>
<td>ease of use</td>
<td>Perl Tk</td>
<td>quick prototyping and</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>friendly interface with</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Perl</td>
</tr>
</tbody>
</table>

Table 4.4: Languages used in ABKAS

The graphical user interface of ABKAS was written in Perl Tk for its ease of use for quick prototyping. ABKAS was implemented under Linux operating system and currently has only been tested under the Linux environment.

4.4.1 A distributed archetype

One of the characteristics of ABKAS is its distributed architecture in that there are many interactions between ABKAS with remote servers, including web servers and non-web servers. Table 4.5 provides a list of servers which are accessed. Both LT.POS and JASMINE are web servers which offer POS (part-of-speech) tagging. LT.POS tags English texts and JASMINE tags Chinese texts. JASMINE can directly tag websites if the URL is given. It makes it possible for us to collect tagged Chinese biographical entries directly by submitting the URL of the website which contains biographical entries. China On Internet is a website from which we can download Chinese biographical texts. Furthermore, Wordsmyth is a web server which provides an online dictionary. We have taken advantage of this service for the recursive extraction of superordinates. Finally, BabelFish has been accessed to carry out automatic machine translation for the intermediate templates of Chinese output into English.
<table>
<thead>
<tr>
<th>Server Name</th>
<th>Resource Location</th>
<th>Purpose of Interaction</th>
<th>Necessary / Optional</th>
</tr>
</thead>
<tbody>
<tr>
<td>LT.POS</td>
<td><a href="http://www.ltg.ed.ac.uk/">http://www.ltg.ed.ac.uk/</a></td>
<td>English part-of-speech tagging</td>
<td>Necessary</td>
</tr>
<tr>
<td>JASMINE</td>
<td><a href="http://www.jansers.org">http://www.jansers.org</a></td>
<td>Chinese part-of-speech tagging</td>
<td>Necessary</td>
</tr>
<tr>
<td>China on internet</td>
<td><a href="http://www.chinaoninternet.com">http://www.chinaoninternet.com</a></td>
<td>Collecting Chinese biographical entries</td>
<td>Necessary</td>
</tr>
<tr>
<td>Wordsmyth</td>
<td><a href="http://machaut.uchicago.edu">http://machaut.uchicago.edu</a></td>
<td>Superordinate bootstrapping for conceptual modelling</td>
<td>Necessary</td>
</tr>
<tr>
<td>ConceptBase Server</td>
<td>atlas.surrey.ac.uk</td>
<td>Knowledge base construction and interaction</td>
<td>Necessary</td>
</tr>
<tr>
<td>BabelFish</td>
<td><a href="http://www.babelfish.com">http://www.babelfish.com</a></td>
<td>Machine translation of Chinese outputs to English</td>
<td>Optional</td>
</tr>
</tbody>
</table>

Table 4.5: Servers and their functions within the distributed archetype
4.4.2 Walking through ABKAS

Figure 4.11: Stage one of ABKAS - preprocessing

Running ABKAS is a straightforward process with the aid of the graphical user interface which takes a card layout. In the first card, users can select a data file which contains encyclopedic entries as input data and click the button ‘Preprocessing’ (see Figure 4.11). The pre-processing unit is activated and submits the input data to remote servers for POS tagging. Further necessary preparation work is also carried out before the tagged texts are presented to the users in the bottom pane of the window.

When this stage of processing is complete, users will be prompted to proceed to the parsing stage by clicking the ‘2-Parser’ to continue to the second card. Users need to press the ‘Start parsing’ button to activate the sublanguage parser. When parsing is complete, users are provided with three output sections. The left hand pane lists the superordinates extracted
Figure 4.12: Stage two of ABKAS - parsing

during the analysis. These superordinates are concepts of professions. The top right pane offers a template form of extracted information and the bottom right is the same information represented in Telos syntax.

The final layer of the card is the knowledge base construction stage. The definition analysis is carried out automatically in the background. The word list of superordinates are checked against the knowledge base and the concepts which exist in the knowledge base are separated from the non-existing ones. The top right section is for users to confirm the definition of superordinates in the knowledge bases against the definitions supplied by online dictionaries. The bottom right pane presents the represented knowledge of superordinates in Telos syntax.

Finally, when the user wishes to proceed to ‘populate KB’, a knowledge base will be constructed automatically in batch via repeated interactions with the ConceptBase (CB) server. ABKAS now acts as a client for the CB server.
4. Archetype for Proper Name Knowledge Acquisition

Please proceed through each tabbed section in order:

- Preprocessing
- Parser
- KB Construction

Unmatched Word List

- biochemist
- botanist
- computer_scientist
- industrial_chemist
- northern
- organic
- photographic_pioneer
- physician
- statistician
- zoologist

Matched Word List

- astronomer
- chemist
- logician
- mathematician
- nuclear_physicist
- philosopher
- physicist
- physiologist
- scientist
- theoretical_physicist
- writer

Conceptual Modelling

Word Definition

statistician:

a person who specializes in compiling and interpreting statistical data.

Concepts in Telos

- Individual scientist in Class isA person
- Individual physics in Class isA science end
- Individual science end
- Individual scientist physicist in Class isA scientist with attribute studies: physics end
- Individual physics end
- Individual nuclear_physics in Class isA

Populate KB

Figure 4.13: Stage three of ABKAS - knowledge base construction

4.4.3 Output of ABKAS - PN knowledge bases

The Proper Name knowledge bases constructed by ABKAS can be accessed by interacting with ConceptBase. As a knowledge base management system, ConceptBase, offers a user interface which allows humans to interact with knowledge bases by adding, deleting and updating records as well as retrieving knowledge. A graphical browser provides a visual presentation of how concepts, entities and their relationships are organised internally. There are three aspects of knowledge representation and reasoning worth noting including generalisation, encapsulation and deduction.

The attributes of objects and relationships within the biographical knowledge base can be presented as semantic networks generated by the graphical browser in ConceptBase. Figure 4.14 illustrates a part of the semantic network, where every person is clustered according
Figure 4.14: A demonstration of the generalisation to their profession. For instance, 'Clark Gable' and 'Michael Caine' are clustered as actors and 'Gabriel Jonas Lippmann' is an instance of 'physicist' and further clustered as 'scientist' with 'Walter Baade' as an astronomer. The object-oriented representation is efficient in that the knowledge about higher level classes can be shared amongst their subclasses and be instantiated by their instances.

Sowa (2000:143) suggested that "a knowledge representation language must have methods of grouping or nesting that can organise knowledge and package it in larger structures." With encapsulation, a set of features of a certain class are grouped into an abstract frame which is inherited and instantiated by its instances. Figure 4.15 demonstrates the encapsulation of the knowledge base where the abstract attributes such as 'name', 'nationality' and 'year' are defined as a group for 'person'. They are inherited by subclasses, for instance 'scientist' and sub-subclasses for example 'physicist' and 'astronomer'. Finally, these attributes are instantiated into the attributes of their instances, for example 'BARKHAUSEN' and 'JEANS'.

In the knowledge base, the prototypical knowledge is stored at the category level and the named entities can inherit such knowledge through inferences. Therefore, knowledge can be
shared by all the entities in the same category. In everyday life, people tend to view others by their ethnic group, gender, social class or profession. The knowledge about prototypes of social groups coupled with the encyclopaedic knowledge about well-known people can generate machine intelligence for natural language understanding systems.

Another important aspect of the knowledge base is its deductive power. For instance, in Figure 4.16, the answer to "Is BARKHAUSEN an instance of scientist?" is true. This is due to the deduction that

If \texttt{ISSUBCLASS(\(\text{physicist, scientist}\))}\nand \texttt{ISINSTANCE(\(\text{BARKHAUSEN, physicist}\))}\nthen \texttt{ISINSTANCE(\(\text{BARKHAUSEN, scientist}\))}\n
Similarly, another set of rules will be applied for the following inference:

If \texttt{ISSUBCLASS(\(\text{nuclear_physicist, physicist}\))}\nand \texttt{ISSUBCLASS(\(\text{physicist, scientist}\))}\nthen \texttt{ISSUBCLASS(\(\text{nuclear_physicist, scientist}\))}\n
This completes the description of the design and implementation of ABKAS, a system which is capable of transforming biographical entries from textual form into biographical knowledge.
bases. As the output of ABKAS, the knowledge bases are intended to be used as computational knowledge resources by other NLP applications for language understanding and other deep language analysis.

4.5 Conclusion

In this chapter, we have presented the computational archetype for constructing lexical knowledge resources for proper names based on the local grammar in dictionaries of biography. ABKAS has been implemented to establish this archetype.

Considering the workflow of the processing, ABKAS uses a modular architecture. It consists of a pre-processing unit which manipulates the input data and interacts with remote servers for part-of-speech tagging. The sublanguage parser is the key component which consists of a number of finite state machines, each of which is designed to process a specific type or subtype of linguistic construct in biographical entries. The knowledge base constructor takes the knowledge in template form and builds up the necessary conceptual relationships in the domain and finally feeds them into a knowledge base manager, the ConceptBase server.

It is important to note that the archetype implemented in ABKAS demonstrates strong
Chapter 4. Archetype for Proper Name Knowledge Acquisition

distributed characteristics. The pre-processing unit accesses remote servers, LT.POS and JASMINE, to obtain syntactic information. The Chinese input data is collected from remote web sites automatically and constructed automatically. Furthermore, ConceptBase has a client-server architecture which allows the knowledge base to be constructed across the network.

In the next chapter, the evaluation study of ABKAS will be carried out and results will be reported.
Chapter 5

System Evaluation of ABKAS

5.1 Introduction

An evaluation of an information processing system is an important process to validate the theoretical framework of a system and to verify its design and implementation. An evaluation of ABKAS can not only determine its performance but also examine how successfully the sublanguage approach has been for automatically converting free-text dictionaries of biography into a PN databases. At the same time, it indirectly validates the hypothesis that there are distinctive and repetitive patterns in the language used when dictionaries of biography are written.

Evaluating an NLP system is not a straightforward task: Sparck Jones and Galliers have suggested that "there is far too much variety in the situations and subjects of (NLP) evaluation to come up with a definite scenario [...] [therefore], [...] evaluations have to be designed for individual cases" (1996:193). However, general guidelines can be followed regarding how NLP evaluation should be approached.

According to Sparck Jones and Galliers (ibid.), in order to establish an evaluation strategy, a series of questions needs to be answered. For example, what is the motivation of the evaluation and what do you hope to discover through the evaluation? The motivation for evaluating ABKAS is to seek the scientific justification of the theoretical hypothesis we have made regarding the sublanguage in dictionaries of biography.

Through the evaluation, we hope to determine: (a) how successfully the system can analyse the biographical entries in both English and Chinese and extract knowledge from semi-
structured yet free natural language texts; (b) how well the extracted knowledge is mapped onto a Telos format and represented in a knowledge base.

Typical IE systems are designed to extract information from full texts, such as news articles and journal papers. News reports have templates for conveying information related to persons, locations and time of the events reported in the news. Similar formats could be found in journal papers. A typical IE system can rely on a scoring system provided by the Message Understanding Conferences (MUCs) (see MUC-7, 1998) to score hundreds and thousands of articles automatically. Once information is extracted from these free texts, one can set up evaluation strategy which can evaluate the performance of a IE system in terms of the popular metrics; precision and recall, for example.

Note that we intended to create a support system for extracting PNs by constructing knowledge resources from dictionaries of biography. It maps the content of individual entries onto a knowledge representation scheme - Telos. As discussed in previous chapters, a dictionary of biography is a highly structured document at a number of linguistic levels. First, the entry for each PN is distinct from the others as every headword is marked orthographically. Second, as a reference book, information is presented at an entry by entry basis. So, unlike free texts, news articles and journal papers we have discussed, it is unnecessary to distinguish between any two named entities. Furthermore, most entries are written following a distinctive and repetitive pattern and we have identified a local grammar to process the entries automatically.

The following sections offer some detail on how the evaluation was carried out. Evaluation results are reported on both the achievements in this study so far and limitations are discussed. In Section 5.2, our aim is to examine the complexity of the data sets. On the one hand, the exercise examines our theoretical framework from the point of view of the data. On the other hand, it aims to offer an guideline for constructing test samples during the evaluation of system performance.

In Section 5.3, two aspects of the performance of ABKAS are examined. One aspect is the acquisition of the superordinates, i.e. the generalisation of a headword, the proper name, into its professional category. The other is to examine how well ABKAS can extract the discriminators, in other words, the attributes of the headword. We have used in a proportionally smaller set of entries from the data set used in Section 5.2 to carry out this exercise. In this section, we also examine the records which ABKAS failed to parse and the slots in templates which were not filled correctly. The aim is to identify the problems and issues which may
have caused its failure in order to understand the limitation of our approach and provide possible directions for future exploration.

An important test of ABKAS would be to see if it can process entries from different dictionaries of biography. In the second half of Section 5.3, we describe the performance of ABKAS on random samples of entries taken from Larousse Dictionary of Scientists in English and Who's who in China: Current Leaders in Chinese.

Recall that ABKAS attempts to generalise some of the entries. The system uses a general purpose lexicon to find superordinates especially those related to professions. For example, if the entries comprises the fact that a person is a painter, ABKAS should be able to automatically infer he/she is an artist. Such information is useful for expanding potential queries for proper name knowledge bases. We have examined the performance using four human evaluators to assess the generalisation capability. Two English-native speakers and two Chinese-native speakers have interacted with the knowledge bases and made queries regarding the content of the knowledge bases built by ABKAS (Details of this interactions are reported in Section 5.4). Each section comprises: (a) Purpose of the evaluation; (b) Test data used; (c) Evaluation method; and (d) Results and discussions. Finally, this chapter is summarised and concluded in Section 5.5.

5.2 Assessing the complexity of the test data

Purpose: We have argued throughout this thesis that a local grammar is used by biographers to format and write the entries in a dictionary of biography; the entries are of persons and follow the template used in general language dictionaries for formatting/encoding (a specific class of) nouns - called Type A by Barnbrook and Sinclair (1994). In Chapter 3, the biographical entries were further categorised into Type $A_1^{PN}$, Type $A_2^{PN}$, Type $A_3^{PN}$ and Type $A_4^{PN}$. Different types have their own characteristics and the complexities during processing vary depending on which type of local grammar the entry falls in.

The performance of ABKAS relies on the repetitive use of the four types of grammar, in other words the sublanguage, within dictionaries of biography. In order to test its performance in a realistic manner, we need to select our test data carefully to reflect the complexity of the dictionaries of biography. Although the four types of entries are distributed throughout dictionaries of biography, the proportion of each type would offer indication as to how a
representative test set can be produced. Therefore, the first exercise in the evaluation is to examine a relatively big set of data in order to measure the percentage of each type in the local grammar so that we can select a smaller set of test samples following the same distribution. In this way, the evaluation can be carried out using a manageable data set without losing the complexity of the test beds.

**Method:** Every type of the local grammar has a distinctive syntactic structure which can be identified by examining the patterns of their part-of-speech tagging. We have developed a pattern analysis tool which can send texts to LT.POS and JASMINE for tagging and obtain the tagged texts. These tagged texts are furthered processed through a tag stripper which separates the original texts from their tags. The pattern analysis tool then examines the tags and assigns each entry into a specific local grammar type following the rules presented in Table 4.1.

**Data:** One thousand biographical entries in English taken from the *Pocket Oxford Dictionary of Biography* (PODoB) and have been examined through our pattern analysis tool. A similar experiment has been carried out with three hundred entries collected from the website: [http://www.chinaoninternet.com](http://www.chinaoninternet.com) which offers an indicator of the complexity of the Chinese biographical entries.

<table>
<thead>
<tr>
<th>Language</th>
<th>Type $A_1^{PN}$</th>
<th>Type $A_2^{PN}$</th>
<th>Type $A_3^{PN}$</th>
<th>Type $A_4^{PN}$</th>
<th>Others</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>36.8%</td>
<td>27.1%</td>
<td>14.7%</td>
<td>10.5%</td>
<td>10.9%</td>
</tr>
<tr>
<td>Chinese</td>
<td>39.4%</td>
<td>39.8%</td>
<td>9.1%</td>
<td>0.0%</td>
<td>11.7%</td>
</tr>
</tbody>
</table>

Table 5.1: Proportion across Type $A^{PN}$s in PODB (1997) and ChinaOnInternet

**Results:** The percentage of each Type $A^{PN}$ is presented in Table 5.1. From this table, we can see that nearly 37 percent of the English biographical entries fall in Type $A_1^{PN}$, which is the basic type with one superordinate. The second largest group is Type $A_2^{PN}$, which constitutes about 27 percent of the population. These entries have multiple superordinates concatenated either via a comma separator or through conjunctional words and as such, they demonstrate higher complexity. There are 10.5 percent entries belonging to Type $A_3^{PN}$, which have subsequent discriminators following the superordinates and nearly 15 percent fall in Type $A_4^{PN}$, which forms a structure where the superordinate is missing and the headword is described in relation to other named-entities.
Almost equal numbers of entries belong to either Type $A_{1}^{PN}$ or Type $A_{2}^{PN}$ categories. These two categories together make up of about 80 percent of the test population. Around 9 percent fall in Type $A_{4}^{PN}$ where a headword is defined in relation to another named-entity. Comparing the English test population to the Chinese test population, the distinctive difference is that Type $A_{4}^{PN}$ is missing in the Chinese set. This is because the post-modifying clauses and phrases used in English are often converted to pre-modifying constructions or simply expressed in separate sentences.

It is observed that 10.9 percent of the entries in English and 11.7 percent in Chinese were not categorised according to the four types of grammar. This suggests that ABKAS will have difficulty parsing these entries and carrying out knowledge acquisition tasks. We have studied these entries and tried to examine the cause.

We found that for half of the unclassified entries, the failure was due to tagging errors. A list of words describing professions was tagged as ‘FW’ (Foreign Word) while it should have been ‘NN’ (Common Noun). For instance:

```
Calvin_NNP, Melvin_NNP (born_JJ 1911_CD), American_NNP
biochemist_FW.
```

Here we list a number of words which have been incorrectly tagged:

- churchman
- stateswoman
- biochemist
- animator
- geophysicist
- palaeontologist
- frontierswoman

These are also cases when a word has multiple senses and it could have different part-of-speech under different circumstances, for example:

```
Romanian-born_NNP American_NNP novelist_NN, and_CC academic_JJ.
```

where ‘academic’ is incorrectly tagged as an adjective.

We have also found entries which take a distinctively different form in the hinge part.
Chapter 5. System Evaluation of ABKAS

Jack the Ripper, unidentified (19th-century) English murderer.

This case occur more often for a person who is a mythological figure or who lived in ancient time. Appendix I contains a list of entries which could not be parsed.

In the pre-processing of Chinese entries, errors in text segmentation often result in incorrect part-of-speech tags get assigned, for instance:

黄公略/PSN (1898 - 1931) 平/PSN 江起义/PSN 领导人/NN。
湖南/PSN 湘乡/PSN 人/NN。
Huang GongLue (1898-1931) Leader of the Pingjiang Revolution.
Native of Hunan Xiangjiang.

where 平江起义 should not have been separated during the segmentation.

5.3 Evaluation of PN knowledge acquisition

Evaluating a set of data as big as 1000 entries is a time-consuming task. However, given the proportions of various types of local grammar, it is possible to draw a representative section of the larger set which reflects its complexity. In this section, ABKAS is first evaluated with the data which it is designed to process. That is followed by the evaluation of how well ABKAS can handle novel input.

5.3.1 Evaluation with a known source

Purpose: ABKAS processes biographical entries and extracts the superordinates and discriminators. In this section, the performance of acquiring such knowledge will be examined.

Data: Test bed 1 and test bed 2 are entries selected from the bigger sets we have used in Section 5.2 for the assessment of the local grammar. Set 1 comprises 100 entries from the Pocket Oxford Dictionary of Biography. Set 2 is a Chinese test bed consisting of 30 entries collected from the web site: http://www.chinaoninternet.com (Table 5.2).

The entries in both sets of data are selected according to the proportion reported in Table 5.1.
Table 5.2: Content of the test bed one with known input

<table>
<thead>
<tr>
<th>Set</th>
<th>Source</th>
<th>Type of reference</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td><em>Pocket Oxford Dictionary of Biography, 1997</em></td>
<td>General dictionary of biography</td>
<td>100</td>
</tr>
<tr>
<td>2</td>
<td><a href="http://www.chinaoninternet.com">http://www.chinaoninternet.com</a></td>
<td>General dictionary of biography on the Internet in Chinese</td>
<td>30</td>
</tr>
</tbody>
</table>

Method: Every parsed entry can be represented in Telos syntax such as:

Individual Aalto in Token, architect, designer, sculptor, painter with
attribute, name
    i_name: "Alvar Aalto"
attribute, year
    born_in: 1898;
died_in: 1976
attribute, nationality
    i_nat: "Finnish"
end

In this example, 'Aalto' has four different superordinates: architect, designer, sculptor and painter. Four attributes are encapsulated within the record which are his full name, his year of birth, year of death and his nationality.

Our evaluation aims to determine how much information is acquired for each individual by counting the number of correct and incorrect superordinates and discriminators (attributes) being generated.

Table 5.3: Performance for acquiring superordinates

<table>
<thead>
<tr>
<th>Set</th>
<th>$N^S_{key}$</th>
<th>$N^S_{correct}$</th>
<th>$N^S_{incorrect}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>145</td>
<td>134 (92.4%)</td>
<td>7 (4.8%)</td>
</tr>
<tr>
<td>2</td>
<td>44</td>
<td>38 (86.4%)</td>
<td>3 (6.8%)</td>
</tr>
</tbody>
</table>

Results: In table 5.3, $N^S_{key}$ refers to the number of superordinates which are expected in the given data set. $N^S_{correct}$ refers to the number of superordinates which have been correctly
identified by ABKAS. Similarly, $N_{\text{incorrect}}^S$ is the number of those which have been identified incorrectly.

More than 90 percent of superordinates have been correctly identified during the English test and the percentage of the correct identifications for the Chinese test is over 86 percent. Less than 5 percent of superordinates were extracted incorrectly for the English test and the incorrect rate for the Chinese test is less than 7 percent.

<table>
<thead>
<tr>
<th>Set</th>
<th>$N^D_{\text{key}}$</th>
<th>$N^D_{\text{correct}}$</th>
<th>$N^D_{\text{incorrect}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>403</td>
<td>376 (93.3%)</td>
<td>19 (4.7%)</td>
</tr>
<tr>
<td>2</td>
<td>110</td>
<td>89 (80.9%)</td>
<td>6 (5.5%)</td>
</tr>
</tbody>
</table>

Table 5.4: Performance for acquiring discriminators

The performance of ABKAS while extracting discriminators is reported in Table 5.4. Similar naming scheme is followed in this table. The ‘D’s in $N^D_{\text{key}}, N^D_{\text{correct}}, N^D_{\text{incorrect}}$ are used to refer to discriminators. For English, more than 93 percent of discriminators are correctly acquired and the incorrectly identified ones are less than 5 percent. For Chinese, more than 80 percent of the discriminators are correctly identified and the incorrect ones represent 5.5 percent. These results indicate that ABKAS is able to successfully parse biographical entries in both English and Chinese based on the local grammar we have defined.

Problems to do with linguistic coverage Some linguistic constructs require more substantial syntactic, semantic or even pragmatic analysis. We have looked into problems and issues regarding the records which are not correctly extracted and examined the cause of them. It is found that the majority of the errors are due to the fact that certain linguistic constructs were outside the scope of our local grammar. We have analysed the cause and made some suggestions on possible solutions.

**Elliptical conjunction**

Adul Rahman, Tunku (1903-90) First prime minister of Malaya (1957-63) and of Malaysia (1963-70).

In the entry above, 'prime minister', which may be expected to be repeated before 'of Malaysia', is omitted. This elliptical structure was not captured in our local grammar therefore ABKAS failed to parse it.
**Unique construct for headword with royal titles**  A headword for a typical entry would be Surname of a person, followed by the first name of a person with a comma as a separator. However, this rule does not apply if the headword is the name of a member of the Royal family. It is often written as firstname, followed by his/her royal title. For instance,

Anne, Princess (Anne Elizabeth Alice Louise, the Princess Royal; 1950- )  
Daughter of Elizabeth II of the United Kingdom.

In this case, Princess was not regarded as her title but misclassified as her first name.

One similar case is when a person is knighted and became Sir.

Coward, Sir Noel (1899-1973) British playwright, composer, producer, and actor.

In this case, the title should have been separated from his first name.

These kind of cases could be handed through a word list which offers a collection of titles.
Furthermore, a title attribute could be incorporated as a slot within the template.

**Alternative expression for nationality at birth**  We found that ABKAS is able to extract nationality_at_birth correctly when it follows the pattern as [Nationality]-born structure.

Baade, Walter (1893-1960), German-born US astronomer who proposed the existence of two stellar populations.

However, such information is expressed in an alternative way like:

Dahl, Roald (1916-90), British writer, of Norwegian descent.

In this case, the attribute of nationality_at_birth is not extracted. ABKAS could be improved under these circumstances by being given extra rules.

### 5.3.2 Extended evaluation with unknown source

**Purpose:** Test sets 1 and 2 allowed us to evaluate the performance of ABKAS against the same data upon which the system was built. Some may argue that the local grammar might be restricted to specific dictionaries of biography since a specific writing style for a reference book often ensures the existence of consistent patterns within the same dictionary. Extended evaluation is carried out to examine how ABKAS perform against novel input which is taken from totally different source.
Data: We carried out a similar exercise with a separate set of input taken from sources which ABKAS was not specifically designed to process. Set 3 consists of 100 entries taken from the *Larousse Dictionary of Scientists*. Another 30 entries of biographies of Chinese politicians were randomly selected from *Who’s Who in China: Current Leaders* to make up Set 4 (Table 5.5).

<table>
<thead>
<tr>
<th>Set</th>
<th>Source</th>
<th>Type of reference</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td><em>Larousse Dictionary of Scientists</em>, 1994</td>
<td>Specialist dictionary of biography</td>
<td>100</td>
</tr>
</tbody>
</table>

Table 5.5: Content of the test bed with novel input

The sources of Set 3 and Set 4 may be considered as specialist dictionaries of biography as *Larousse Dictionary of Scientists* collects biographical entries for scientists only and *Who’s who in China: Current Leaders* is focussed mostly on Chinese leaders, i.e. politicians. It is interesting to find out how these biographical entries are expressed differently. Since test sets 3 and 4 provide test material on which ABKAS’s design was not based, the extensibility of ABKAS can be tested. At the same time, the robustness of ABKAS and the ability to handle novel input can also be evaluated.

<table>
<thead>
<tr>
<th>Set</th>
<th>$N^S_{key}$</th>
<th>$N^S_{correct}$</th>
<th>$N^S_{incorrect}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>143</td>
<td>134 (93.7%)</td>
<td>8 (5.6%)</td>
</tr>
<tr>
<td>4</td>
<td>45</td>
<td>34 (75.6%)</td>
<td>5 (11.1%)</td>
</tr>
</tbody>
</table>

Table 5.6: Performance for acquiring superordinates from novel input

Results: We observed better performance for the acquisition of superordinates from Set 3 than that for Set 1. This could be explained in a number of ways. First, *Larousse Dictionary of Scientist* is a prestigious reference. Biographers must have followed stricter writing style during its compilation. Second, all the entries are about scientists. Therefore, less irregular forms occur such as entry like ‘Jack the Ripper’.

For Set 4, ABKAS acquired less superordinates correctly compared to the test for Set 2 since extra information regarding each person’s political role in the Chinese government was in most entries.
Additional information was found in *Who’s Who in China: Current Leaders*, which is the position of the person in government. Note that this dictionary of biography consists of people who are senior member of the government during early 1990s. Their political roles in the government is considered significant information. Therefore, they are often expressed before one’s real profession. For instance,

陈慧 中国人民政治协商会议河北省委员会副主席。教授。1933年1月生。浙江诸暨人。

The translation of the entry is

Chen Hui (b. Jan 1933) Native of Zhuji, Zhejiang. Vice-chairman of CPPCC Hebei Provincial Committee; professor.

where CPPCC stands for the Chinese People’s Political Consultative Committee. This information can be considered as a unique characteristic of the specific dictionary of biographical due to the fact that it is largely about politicians.

<table>
<thead>
<tr>
<th>Set</th>
<th>(N_{\text{key}}^D)</th>
<th>(N_{\text{correct}}^D)</th>
<th>(N_{\text{incorrect}}^D)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>452</td>
<td>376 (83.2%)</td>
<td>41 (9.1%)</td>
</tr>
<tr>
<td>4</td>
<td>101</td>
<td>87 (86.1%)</td>
<td>7 (6.9%)</td>
</tr>
</tbody>
</table>

Table 5.7: Performance for acquiring discriminators from novel input

During the acquisition of discriminators from unknown sources, we have observed a slight drop in performance in Set 3 in comparison to Set 1. This is largely due to extra information regarding birth place.

Set 3 has lower correct percentage for extracting discriminators compared to Set 1 is largely due to the fact that extra information is offered in the biographical entries in the *Larousse Dictionary of Scientists* (1994) compared to entries from the *Pocket Oxford Dictionary of Biography* (1997). For instance,

**Newton, Sir Isaac (1642–1727), British mathematician and physicist.**

**NEWTON, Sir Isaac (1642–1727) English scientist and mathematician,**

born in Woolsthorpe, Lincolnshire.
information regarding a person's birth place was in Set 3 but not in Set 1. In the *Pocket Oxford Dictionary of Biography*, information regarding a person's birth place is often in the elaboration part of the entry. However, in the *Larousse Dictionary of Scientists*, it is often part of the definition part of the entry.

The fact that the parser is able to cope with totally novel inputs without any specific tailoring of the program, provides strong supporting evidence for our hypothesis on the existence of the *distinctive and repetitive patterns*.

ABKAS is able to ignore structures which are outside its local grammar rather than failing completely, perhaps indicating robustness of the system. ABKAS's performance on Chinese biographical entries demonstrates the possible adaptability of the framework to other languages.

### 5.4 Evaluation of PN knowledge bases

**Purpose:** In the previous section, we have examined the performance of ABKAS on knowledge acquisition by checking how correctly ABKAS can parse the definition part of biographical entries and extract knowledge about individual named entities. These exercises have demonstrated that ABKAS is able to automatically transform free texts in a relatively restricted domain into its semi-formal template form. However, what makes ABKAS unique is that it is able to further process the information in template form into logical propositions taking a knowledge representation scheme - Telos and construct proper name knowledge bases automatically. It is important to evaluate the benefit we can gain through the construction of such knowledge bases. Furthermore, a knowledge base is able to infer implicit knowledge from the existing explicit knowledge through deduction. It is important to evaluate the deductive ability of the system.

**Data:** We have continued using the four sets of data as described in Section 5.3. These data have been through the pre-processing and parsing stages. The output takes the intermediate template form. These templates are further transformed into Telos format and populated in four different knowledge bases. One knowledge base is generated for each set of data. Note that data from different sources can be populated into one knowledge base. We chose to separate them for evaluation purposes.
Chapter 5. System Evaluation of ABKAS

Method: It is not a straightforward task to evaluate the performance of ABKAS on the construction of knowledge bases since such performance can be difficult to quantify. Furthermore, although the knowledge about the proper names were supplied by ABKAS, the management of knowledge base is carried out by ConceptBase, a knowledge base management system. It is impossible to evaluate ABKAS's performance without considering the role which ConceptBase plays.

The exercise was carried out in a simple interactive dialogue fashion. We have invited two English native speakers and two Chinese native speakers. Each user is asked to supply queries to ConceptBase and reported the answers which have been returned back to the author.

<table>
<thead>
<tr>
<th>Example queries</th>
<th>Queries in Telos</th>
<th>Answers returned</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Is Babbitt a writer?</td>
<td>IsInstance[Babbitt, writer]</td>
<td>FALSE</td>
</tr>
<tr>
<td>2. What is Yeltsin’s nationality?</td>
<td>FindiAttribute[Yeltsin, nationality]</td>
<td>Russian</td>
</tr>
<tr>
<td>3. Find me all the people who are physicists.</td>
<td>FindInstance[physicist]</td>
<td>Salam, Van_Allen</td>
</tr>
<tr>
<td>5. Is a nuclear_physicist a scientist?</td>
<td>IsSubclass[nuclear_physicist, scientist]</td>
<td>TRUE</td>
</tr>
<tr>
<td>6. Find me all the artists.</td>
<td>FindInstance[artists]</td>
<td>Aalto, Gabo, Kandinsky</td>
</tr>
<tr>
<td>7. How many writers are there in the knowledge base?</td>
<td>ECA rules required (more details to follow)</td>
<td>11</td>
</tr>
<tr>
<td>8. Do Einstein and Newton share the same profession?</td>
<td>Extra query required (more details to follow)</td>
<td>TRUE</td>
</tr>
</tbody>
</table>

Table 5.8: Results of the evaluation of PNKBs

Results: The queries we have collected are classified into three groups according to their complexity (see Table 5.8). Query 1, 2 and 3 belong to the basic level. The knowledge requested was explicitly represented. For instance the proposition \( P(\text{Babbitt, in, mathematician}) \)
exists in the knowledge base. However, it is not possible to answer query 4, 5 and 6 with just the existing knowledge. Deductive rules have to be applied. 'folk musician' is asserted as a folk musician in the KB. It is through two levels of deduction that 'a folk musician is a musician' and 'a musician is an artist'. Similar rules apply when nuclear_physicist is inferred as a physicist and then a scientist. When we search for all the artists, deduction has also taken in part to include subclasses of 'artist' such as 'painter', 'sculptor', etc.

Query 7 involves counting, which is not a straightforward task for a declarative knowledge representation language like Telos. New ECA rules have to be defined to count when a new object is inserted. For instance,

```plaintext
Class WriterCounter end
ECArule WriterCounterRule with
ecarule
  Counter : $ x/writer i,ii/Integer
  ON Tell(In(x,writer))
  IF i in writerCounter
  DO Untell(In(i,writerCounter)),
      CALL(increment(i,ii)),
      Tell(In(ii,writerCounter))
  ELSE Tell(In(1,writerCounter))
  $
end
```

The limitation of this method is that the counter has to triggered when a new entry is inserted in the knowledge base. It is important that the rule exists before the knowledge base is created. An alternative way of counting is to query the KB using its Java API and request all the instances of a class, 'writer' in this case, and count the number of the tokens which are returned.

In order to answer Query 8, we have added an extra ConceptBase query named as SameClass.

```plaintext
GenericQueryClass SameClass isa Proposition with
  parameter
    obj1 : Proposition;
    obj2 : Proposition
  constraint
    c : $ In(~obj1,this) and In(~obj2,this) and In(this,Class) $
end
```
With this query, the question of "Do Einstein and Newton share the same profession?" can be formally represented as $\text{SameClass}[\text{Einstein/obj1,Newton/obj2}]$ within ConceptBase and the answer we got was $\text{TRUE}$.

Through the query answering session, we have learned the advantages and limitations of the integration of ABKAS with ConceptBase. Both the strengths and the weaknesses have been demonstrated. It is encouraging to see that the knowledge bases which are automatically constructed by ABKAS can not only offer a certain amount of useful information, it can also reason over the explicit knowledge and deduce additional knowledge.

5.5 Conclusion

The evaluation of ABKAS has also been carried out and reported upon in this chapter. We started with a relatively large set of test data and examined the complexity of the data measured by the percentage of different types of local grammar. We then took 10 percent of the data and carried out the evaluation of ABKAS's knowledge acquisition performance in two respects: superordinates and discriminators. This exercise was extended with the supply of novel input from different sources compared to the dictionaries of biographies we used to design ABKAS. Furthermore, we have analysed the failed records and discussed the causes and possible solutions.

The results of the evaluation suggest that ABKAS can in many cases successfully parse the definition part of biographical entries and transform naturally occurring texts into a structured format. Furthermore, ABKAS is capable of integrating the formatted information into a knowledge base. The effectiveness of ABKAS for the automatic acquisition of encyclopaedic knowledge has suggested that there are indeed distinctive and repetitive patterns in dictionaries of biography. Further evaluation was carried out to examine the knowledge bases constructed by ABKAS. The results are encouraging.

In the next chapter, the contribution of this work and its future direction will be discussed.
Chapter 6

Conclusions and Future Work

In our work on proper name knowledge acquisition for text understanding, many aspects of proper names have been explored including their onomastic behaviour, linguistic characteristics and more importantly, the computational work which has been undertaken to date in proper name analysis. In Chapter 3, we have analysed the entries across dictionaries of biography from two languages and studied the language used by biographers to define personal names. Our study suggested there exists a sublanguage within the definition part of biographical entries which is shared across dictionaries of biography even in different languages. In order to examine the theoretical claim proposed in Chapter 3, a computational archetype was proposed and implemented into a system, ABKAS, to transform biographical entries from textual form to encoded knowledge (Chapter 4).

This archetype comprises three core functional units. The preprocessing unit prepares the input data and supplies it with syntactic information. Governed by a local grammar, the sublanguage parser derives semantic entities and relationships from the given syntactic information and stores them in intermediate templates. Finally, templates are analysed by a knowledge base constructor which translates the information in template form into a knowledge representation language and converts semi-structured information into its formal representation. This modular architecture was designed with the capacity to extend to other domains which possess a sublanguage and a local grammar. However, our focus here is concerned with the automatic construction of knowledge bases of proper names from texts, particularly encyclopaedic texts.

In this chapter, we conclude the thesis by suggesting that encyclopaedic texts can be used effectively as text resources in knowledge acquisition for proper names. Finally, we propose
future directions to be explored in order to take our research further.

6.1 Contributions of the Research

Our work has followed three lines of inquiry in the creation of lexical databases and complementary knowledge bases. NLP applications rely on the existence of appropriately formatted lexical databases and knowledge bases of objects and events. We have argued thus far that such databases and knowledge bases, if they exist at all, deal largely with common nouns. These nouns comprise a substantial part of any text corpus or indeed any digital library. The challenge for NLP applications is to deal with proper names of persons, places, objects and things. One can argue that the real challenge is to build lexical databases of persons and places and knowledge bases of objects and events in an arbitrary domain without relying too much on human effort. This is the challenge we have undertaken with some degree of success in this thesis.

One major source of information regarding proper names is the ubiquitous dictionaries of biography. Examples of dictionaries of biography include national dictionaries of biography, dictionaries of biography for scientists and biographical entries in a number of specialised dictionaries. Our contribution has been to show the manner in which entries in most of these human-oriented biographical resources are formatted following a templated pattern or a local grammar: these are distinctive and repeated linguistic patterns for encoding the knowledge of proper nouns. Barnbrook & Sinclair (1994) and Ahmad & Al-Jabir (1998) have formed Sinclair templates for entries in general language and specialised language dictionaries (of common nouns). Our work confirms their observations. Furthermore, we might argue that lexicographers, terminologists and biographers use a common frame-based structure to encode the knowledge of their respective domains.

The second line of inquiry relates to the automatic construction of lexical and knowledge-based resources for proper name analysis. We have argued that given that there are distinctive and repetitive patterns, it is possible to establish a computational framework to map the contents of these patterns onto a knowledge representation schema. A tentative framework has been created by Ahmad and Al-Jabir wherein they mapped a linguistic expression onto a Minskian-frame using arguments in λ-calculus. We have consolidated, expanded and refined this framework.
Chapter 6. Conclusions and Future Work

The third line of inquiry relates to the use of findings in lexicology and theoretical and applied linguistics in specifically building an NLP application or generally in building AI-applications. Typically, these applications are based on the findings of one language, usually English or a typologically related language such as French or German. We have used a typologically, and furthermore an orthographically different language, Chinese, to validate the results of our work.

We believe that we have explored the above three lines of inquiry with notable success. This is exemplified in the implementation of the computational framework to transform texts in dictionaries of biography into encoded knowledge. This successful deployment of the framework indicates that taking encyclopaedic texts as resources for automatic knowledge base construction is a promising avenue.

From this work, we have learnt that proper nouns and common nouns are interrelated entities. Proper nouns somehow interface lexical knowledge with world knowledge. It is important for NLP systems to have access to both types of knowledge as lexical resources. In NLP, linguistic knowledge and world knowledge have to be used in a combined way. The knowledge about proper nouns is the instantiation of the knowledge about common nouns. The representation of knowledge related to proper names should therefore follow an object-oriented approach which supports inheritance and encapsulation.

The implementation of ABKAS has been a challenging task. Without other systems and resources available on line, it would be difficult to achieve. We have learnt from the process of implementation that distributed system design is a way forward.

In conclusion, reference books, like dictionaries of biographical can be used successfully for automatic knowledge acquisition in order to bridge the knowledge gaps related to proper names.

Our approach has a number of limitations: firstly, considering the sublanguage parser, the local grammar is defined through the analysis of sample data and therefore may be dependent on the samples which are chosen. From our observations, the consistency of the patterns sometimes rely on the quality of the reference book or electronic resource; the second limitation concerns the coverage of texts which have been processed. Possible solutions for these limitations will be discussed later in the next section.
6.2 Future Work

The three lines of inquiry in areas as disparate as local grammars, automatic knowledge acquisition and contrastive linguistics have helped us to identify a number of promising areas of work: The first is concerned with utilising morphological cues for words and phrases and lexical resources to obtain better performance; The second focuses on extending our work towards an adaptive system in order to increase the flexibility of grammar and portability of ABKAS into other languages; The third direction involves applying the framework to a wider range of text input.

Utilising other linguistic knowledge We see the sublanguage within dictionaries of biography as a grammar-based sublanguage which is governed by a local grammar. There are other language markers that offer important information for extracting semantic knowledge, such as morphological knowledge and lexical knowledge. For instance, there are morphological cues for extracting professions, such as -ist in scientist and physicist, -or in director and actor, and -ian in musician, etc. In Chinese, such a phenomenon is more significant, for example: 科学家 (scientist) and 作家 (writer) and 学者 (scholar). These morphological cues can be used effectively for the extraction of certain semantic information. However, introducing morphological cues may result in extra overhead in performance due to the lexical comparison. It may also lead to maintenance cost since the lexical database has to be kept up to date.

Lexical knowledge can also utilised effectively for extracting semantic categories such as nationalities since the number of countries in the world is finite and fairly constant. Therefore, maintaining such list is not likely to be costly.

Adaptive information extraction Our sublanguage parser is implemented as rule-based finite state machines in which the local grammar is implemented as a set of rules utilising regular expressions. Much of the performance of ABKAS and the theoretical basis of this thesis relies on the exitence of local grammars. These are relatively simple grammars and a typical biographical entry may have used one of the four grammars preferentially.

We have also seen other dicitonaries of biography, for example the Chinese dicitonary of biography and the dictionary of sicentists. There are number of methods and techques by which system can be used to induce rules through learning from a set of data.
Such learning techniques in computational linguistics attempts to induce grammar from a set of utterances. Grishman (2001) emphasized the role of sublanguage analysis within the field of adaptive information extraction.

One extension of this thesis would be to adapt ABKAS for an unseen DoB by inducing the local grammar which was used to encode the entries in that dictionary of biography. Since the basic local grammar has been identified in this work, more rules could be learned incrementally in order to adapt to heterogeneous sources of biographical entries.

Expanding the coverage of the analysis  Our work has taken a significant step towards constructing knowledge bases from encyclopaedic texts taking personal names as the focal point of the study. The grammar-based sublanguage approach was dedicated to the analysis of the definition part of the biographical entries. Although the size of the definition part is considerably smaller than the elaboration part, it offers important information and builds up a profile of the named entities.

Once the framework of the knowledge base is constructed using ABKAS, more information can be added to increase the coverage and richness of the knowledge. Hull and Gomez (1999) have reported a complementary piece of research which focuses on the elaboration part of the biographical entries and extracts detailed information about a specific individual. Our work is able to construct the framework and oversee the structure and the organisation of a biographical knowledge base before the detailed information is filled in.

Although our experiments have been based on personal names and dictionaries of biography, the sublanguage framework could be extended to many other domains, such as gazetteers of place names. The framework may be further extended to process other semi-structured texts which are likely to follow certain templates, for instance address data. It could also be applied to knowledge management for analysing people's knowledge and skills presented in their résumés in textual form and constructing knowledge bases accordingly.

Final remarks  In this study, we have designed and implemented an archetype for the constructions of proper name knowledge bases. Our research has demonstrated the integration between various resources including system resources, such as on-line tagging services, and more importantly, knowledge resources, which consist of conceptual knowledge in dictionary definitions and encyclopedic knowledge in dictionaries of biography.
The integration between knowledge resources focused on the coalescing of biographical entries, the *definitions* of proper names, and dictionary definitions of common nouns, which describe professions in particular. The computational framework falls into a bootstrapping approach which has been suggested by Yorick Wilks et al. (1996:254), to explore the future of lexical knowledge acquisition. Wilks pointed out two promising avenues: first, "towards corpora-based studies of various sorts"; and second, "towards coalescing Machine Readable Dictionaries (MRDs) with other sorts of formatted texts, the encyclopaedia being the obvious choice". We believe that our work has made a step forward within the second avenue, which is to integrate MRDs with encyclopaedic texts. Such coalescing further marks the originality of this work. As Yorick Wilks et al. (ibid.:255) has stated "this is one avenue that has not been explored extensively, but it makes so much sense that it cannot be far away".
# Appendix A

## Organisations for Place Names

<table>
<thead>
<tr>
<th>Organisation</th>
<th>Website</th>
</tr>
</thead>
<tbody>
<tr>
<td>Geographical Names Data Base (Canadian Records)</td>
<td><a href="http://geonames.NRCan.gc.ca/">http://geonames.NRCan.gc.ca/</a></td>
</tr>
<tr>
<td>PLANSUS - Placename Survey of the United States</td>
<td><a href="http://www.wtsn.binghamton.edu/plansus/">http://www.wtsn.binghamton.edu/plansus/</a></td>
</tr>
<tr>
<td>Scottish Placename Society</td>
<td><a href="http://www.st-and.ac.uk/institutes/sassi/spns/spnstop.htm">http://www.st-and.ac.uk/institutes/sassi/spns/spnstop.htm</a></td>
</tr>
</tbody>
</table>
## Appendix B

### Penn Treebank Tag Set Used in LT.POS

<table>
<thead>
<tr>
<th>Tag</th>
<th>Description</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>CC</td>
<td>coordinating conjunction</td>
<td>and</td>
</tr>
<tr>
<td>CD</td>
<td>cardinal number</td>
<td>1, third</td>
</tr>
<tr>
<td>DT</td>
<td>determiner</td>
<td>the</td>
</tr>
<tr>
<td>EX</td>
<td>existential there</td>
<td>there is</td>
</tr>
<tr>
<td>FW</td>
<td>foreign word</td>
<td>d’hoevre</td>
</tr>
<tr>
<td>IN</td>
<td>preposition/subordinating conjunction</td>
<td>in, of, like</td>
</tr>
<tr>
<td>JJ</td>
<td>adjective</td>
<td>green</td>
</tr>
<tr>
<td>JJR</td>
<td>adjective, comparative</td>
<td>greener</td>
</tr>
<tr>
<td>JJS</td>
<td>adjective, superlative</td>
<td>greenest</td>
</tr>
<tr>
<td>LS</td>
<td>list marker</td>
<td>1)</td>
</tr>
<tr>
<td>MD</td>
<td>modal</td>
<td>could, will</td>
</tr>
<tr>
<td>NN</td>
<td>noun, singular or mass</td>
<td>table</td>
</tr>
<tr>
<td>NNS</td>
<td>noun plural</td>
<td>tables</td>
</tr>
<tr>
<td>NNP</td>
<td>proper noun, singular</td>
<td>John</td>
</tr>
<tr>
<td>NNPS</td>
<td>proper noun, plural</td>
<td>Vikings</td>
</tr>
<tr>
<td>Tag</td>
<td>Description</td>
<td>Example</td>
</tr>
<tr>
<td>-------</td>
<td>---------------------------------</td>
<td>--------------------------</td>
</tr>
<tr>
<td>PDT</td>
<td>predeterminer</td>
<td>both the boys</td>
</tr>
<tr>
<td>POS</td>
<td>possessive ending</td>
<td>friend’s</td>
</tr>
<tr>
<td>PRP</td>
<td>personal pronoun</td>
<td>I, he, it</td>
</tr>
<tr>
<td>PRP$</td>
<td>possessive pronoun</td>
<td>my, his</td>
</tr>
<tr>
<td>RB</td>
<td>adverb</td>
<td>however, usually, naturally, here, good</td>
</tr>
<tr>
<td>RBR</td>
<td>adverb, comparative</td>
<td>better</td>
</tr>
<tr>
<td>RBS</td>
<td>adverb, superlative</td>
<td>best</td>
</tr>
<tr>
<td>RP</td>
<td>particle</td>
<td>give up</td>
</tr>
<tr>
<td>TO</td>
<td>to</td>
<td>to go, to him</td>
</tr>
<tr>
<td>UH</td>
<td>interjection</td>
<td>uhhuhuhuhh</td>
</tr>
<tr>
<td>VB</td>
<td>verb, base form</td>
<td>take</td>
</tr>
<tr>
<td>VBD</td>
<td>verb, past tense</td>
<td>took</td>
</tr>
<tr>
<td>VBG</td>
<td>verb, gerund/present participle</td>
<td>taking</td>
</tr>
<tr>
<td>VBN</td>
<td>verb, past participle</td>
<td>taken</td>
</tr>
<tr>
<td>VBP</td>
<td>verb, sing. present, non-3d</td>
<td>take</td>
</tr>
<tr>
<td>VBZ</td>
<td>verb, 3rd person sing. present</td>
<td>takes</td>
</tr>
<tr>
<td>WDT</td>
<td>wh-determiner</td>
<td>which</td>
</tr>
<tr>
<td>WP</td>
<td>wh-pronoun</td>
<td>who, what</td>
</tr>
<tr>
<td>WP$</td>
<td>possessive wh-pronoun</td>
<td>whose</td>
</tr>
<tr>
<td>WRB</td>
<td>wh-abverb</td>
<td>where, when</td>
</tr>
</tbody>
</table>
# Appendix C

## JASMINE Part-of-Speech Tag Set

<table>
<thead>
<tr>
<th>No.</th>
<th>Tag Name</th>
<th>Chinese Name</th>
<th>English Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>NN</td>
<td>名词</td>
<td>Common noun</td>
</tr>
<tr>
<td>2</td>
<td>NLN</td>
<td>方位名词</td>
<td>Locative noun</td>
</tr>
<tr>
<td>3</td>
<td>NPN</td>
<td>专用名词</td>
<td>Proper noun</td>
</tr>
<tr>
<td>4</td>
<td>NTN</td>
<td>时间名词</td>
<td>Temporal noun</td>
</tr>
<tr>
<td>5</td>
<td>NRN</td>
<td>职务名词</td>
<td>Noun of official rank</td>
</tr>
<tr>
<td>6</td>
<td>NPSS</td>
<td>姓氏</td>
<td>Personal surname</td>
</tr>
<tr>
<td>7</td>
<td>NPSN</td>
<td>人名</td>
<td>Personal name</td>
</tr>
<tr>
<td>8</td>
<td>NTC</td>
<td>音译字</td>
<td>Character in translation</td>
</tr>
<tr>
<td>9</td>
<td>NAN</td>
<td>形名词</td>
<td>Noun as adjective</td>
</tr>
<tr>
<td>10</td>
<td>VT</td>
<td>及物动词</td>
<td>Transitive verb</td>
</tr>
<tr>
<td>11</td>
<td>VI</td>
<td>不及物动词</td>
<td>Intransitive verb</td>
</tr>
<tr>
<td>12</td>
<td>VA</td>
<td>情态动词</td>
<td>Modal auxiliary</td>
</tr>
<tr>
<td>13</td>
<td>VM</td>
<td>趋向动词</td>
<td>Motion verb</td>
</tr>
<tr>
<td>14</td>
<td>VL</td>
<td>系动词</td>
<td>Linking verb</td>
</tr>
<tr>
<td>15</td>
<td>VC</td>
<td>从句动词</td>
<td>Verb + a clause</td>
</tr>
<tr>
<td>No.</td>
<td>Tag Name</td>
<td>Chinese Name</td>
<td>English Name</td>
</tr>
<tr>
<td>-----</td>
<td>----------</td>
<td>--------------</td>
<td>-----------------------</td>
</tr>
<tr>
<td>16</td>
<td>A</td>
<td>形容词</td>
<td>Adjective</td>
</tr>
<tr>
<td>17</td>
<td>MM</td>
<td>数词</td>
<td>Numeral</td>
</tr>
<tr>
<td>18</td>
<td>MA</td>
<td>概数词</td>
<td>Numeral approximation</td>
</tr>
<tr>
<td>19</td>
<td>U</td>
<td>量词</td>
<td>Measure word</td>
</tr>
<tr>
<td>20</td>
<td>PP</td>
<td>人称代词</td>
<td>Personal pronoun</td>
</tr>
<tr>
<td>21</td>
<td>PD</td>
<td>指示代词</td>
<td>Determiner</td>
</tr>
<tr>
<td>22</td>
<td>PAD</td>
<td>副指代词</td>
<td>Adverbial pronoun</td>
</tr>
<tr>
<td>23</td>
<td>PA</td>
<td>特指代词</td>
<td>Definite article</td>
</tr>
<tr>
<td>24</td>
<td>FPN</td>
<td>名词前缀</td>
<td>Nominal prefix</td>
</tr>
<tr>
<td>25</td>
<td>FSN</td>
<td>名词后缀</td>
<td>Nominal suffix</td>
</tr>
<tr>
<td>26</td>
<td>FPM</td>
<td>数词前缀</td>
<td>Numeral prefix</td>
</tr>
<tr>
<td>27</td>
<td>FSA</td>
<td>形容词后缀</td>
<td>Adjectival suffix</td>
</tr>
<tr>
<td>28</td>
<td>DV</td>
<td>副词一</td>
<td>Verbal adverb</td>
</tr>
<tr>
<td>29</td>
<td>DS</td>
<td>副词二</td>
<td>Sentential adverb</td>
</tr>
<tr>
<td>30</td>
<td>DN</td>
<td>副词不</td>
<td>Nominal negative prefix</td>
</tr>
<tr>
<td>31</td>
<td>DA</td>
<td>动补副词</td>
<td>Complement predicate</td>
</tr>
<tr>
<td>32</td>
<td>I</td>
<td>介词</td>
<td>Preposition</td>
</tr>
<tr>
<td>33</td>
<td>SDG</td>
<td>结构助词的</td>
<td>De, genitive</td>
</tr>
<tr>
<td>34</td>
<td>SDA</td>
<td>结构助词地</td>
<td>De, adverbial</td>
</tr>
<tr>
<td>35</td>
<td>SDC</td>
<td>结构助词得</td>
<td>De, complement</td>
</tr>
<tr>
<td>36</td>
<td>SP</td>
<td>结构助词所</td>
<td>Suo3</td>
</tr>
<tr>
<td>37</td>
<td>SC</td>
<td>结构助词似的</td>
<td>Si4de</td>
</tr>
<tr>
<td>38</td>
<td>SO</td>
<td>结构助词等</td>
<td>Deng3</td>
</tr>
<tr>
<td>39</td>
<td>SY</td>
<td>结构助词以来</td>
<td>Yi3lai2</td>
</tr>
<tr>
<td>40</td>
<td>ST</td>
<td>时态助词</td>
<td>Temporal auxiliary</td>
</tr>
<tr>
<td>41</td>
<td>SM</td>
<td>语气助词</td>
<td>Auxiliary of mood</td>
</tr>
<tr>
<td>42</td>
<td>CC</td>
<td>从句连词</td>
<td>Clause connector</td>
</tr>
<tr>
<td>No.</td>
<td>Tag Name</td>
<td>Chinese Name</td>
<td>English Name</td>
</tr>
<tr>
<td>-----</td>
<td>----------</td>
<td>--------------</td>
<td>-----------------------</td>
</tr>
<tr>
<td>43</td>
<td>CP</td>
<td>词组连词</td>
<td>Phrase connector</td>
</tr>
<tr>
<td>44</td>
<td>T</td>
<td>叹词</td>
<td>Exclamatory word</td>
</tr>
<tr>
<td>45</td>
<td>E</td>
<td>象声词</td>
<td>Onomatopoeic word</td>
</tr>
<tr>
<td>46</td>
<td>O</td>
<td>习用语</td>
<td>Idiomatic word</td>
</tr>
<tr>
<td>47</td>
<td>Bx</td>
<td>标点(x=[A-Y])</td>
<td>Punctuation (25)</td>
</tr>
<tr>
<td>48</td>
<td>H</td>
<td>非汉字</td>
<td>Loan word</td>
</tr>
</tbody>
</table>
Appendix D

Web Interface of LT_POS

LT POS and LT CHUNK: the demo

LT POS is the part-of-speech tagger and LT CHUNK is a chunker developed at the Edinburgh Language Technology Group. The part of speech tagger assigns parts of speech to tokens in a text; the chunker chunks sentences into noun groups or verb groups.

Elsewhere you can find more elaborate information on LT POS and information on LT CHUNK -- technical information, as well as information about licensing options.

Here you can try out the software for yourself.

Figure D.1: A screen shot of the LT_POS interactive demo

As shown in Figure D.1, users can paste texts in the text area to be tagged by LT_POS. The web interface is convenient for tagging a small sample of texts. However, for a large corpus, it is too laborious to submit texts one by one. To overcome this problem, we have
implemented programs to interact with the LT-POS CGI script which uses the POST method. The programs were written in Perl to take the advantage of Perl's HTTP and LWP modules for web automation. Sample code is shown below:

```perl
sub submit {
    my $ua = LWP::UserAgent->new();
    my $req = POST 'http://www.ltg.ed.ac.uk/~mikheev/cgi-bin/tagger_demo.cgi',
                   [do => 'Start Tagging', sentence=>$head, tagornot=>'checked', variants => 'checked',
                    dict => 'checked', highlight=>'checked', tagdelim =>'_',
                    vardelim=>'=', ng=>'checked', vg=> 'checked', ngstart=>'[', ngend=>']',
                    vgstart=>'<', vgend=>'>', layout=>'break'];
    my $content = $ua->request($req)->as_string;
}
```
Appendix E

Web Interface of JASMINE

JASMINE offers POS tagging services for Chinese web pages (See figure E.1). It is designed to take the URL of a web page as input and produce the tagged output.

The input of the pre-processor is an index page with the URL http://www.chinaoninternet.com/famous/ourtime/index.htm. The pre-processor scans the index page and finds all the
hyperlinks for every biographical entry listed using the HTML::LinkExtor module.

    my $parser = HTML::LinkExtor->new(undef, $baseurl);
    $parser->parse_file($infile) || die "Cannot parse input file";
    my @links = $parser->links;

The function of the following submodule is to submit the contents of a web page to JASMINE for tagging.

    sub submit
    {
        my $ua = LWP::UserAgent->new();
        my $link = $_[0];
        my $req = POST 'http://www.jansers.com/cgi-bin/new/service.cgi',
          [service_func => 'Tagging (GB)', user_url => $link];
        my $content = $ua->request($req)->as_string;
        (my $plain_txt = $content) =~ s/[
]*///gs;
        return () unless $plain_txt;
    }

Appendix F

Tagged Chinese Input

阿英/NPSN（1900～1977）文艺/NN理论/NN批评家/NN、文学史家/NN、作家/NN。原名/NN钱德富/NPSN，又/DS名/NN德赋/NPSN、钱/NN杏村/NN，笔/NN名/NN阿/FPNonym/NPSN、钱谦吾/NPSN、张若英/NPSN等/SO。安徽/NPN芜湖人/NN。


丁玲/NPSN（1904～1986）女作家/NN。原名/NN蒋伟/NPSN，字/NN冰/NN之/SDG，别名/NN蒋玮/NNN 丁玲之/NPSN，笔名/NN彬芷/NN、晓芸/NN等/SO。湖南/NPN临澧/NPN人/NN。

郭沫若/NPSN（1892～1978）作家/NN，诗人/NN，剧作家/NN，历史学家/NN，考古学家/NN，古文字学家/NN，社会/NN活动家/NN。原名/NN郭开贞/NPSN，又/DS名/NN郭鼎堂/NPSN。四川/NPN乐山/NPN人/NN。


成仿吾/NPSN（1897～1984）教育家/NN、文学家/NN、翻译家/NN。湖南省/NPN新化县/NPN人/NN。

徐特立/NPSN（1877～1968）无产阶级革命家/NN，教育家/NN。1877年/NTN2月/NTN1日/NTN生/VT 于/I湖南/NPN长沙/NPN。

艾思奇/NPSN（1910～1966）哲学家/NN。云南/NPN腾冲/NPN人/NN。
Appendix F. Tagged Chinese Input

艾中信/NPSN（1915～）油画家/NN，美术/NN评论家/NN。上海/NN 川沙/NN人/NN。
徐悲鸿/NPSN（1895～1953）画家/NN，美术/NN教育家/NN。江苏/NN 宜兴/NN人/NN。
吴作人/NPSN（1908～）画家/NN，美术/NN教育家/NN。祖籍/NN 安徽/NN 泾县/NN，
生/NN 元/NN于/NN江苏/NN 苏州/NN人/NN。

爱波斯坦/NPSN（1915～）新闻/NN 记者/NN，报刊/NN主编/NN。1915年/NN 4月/NN 20
日/NN 生/VT。原籍/NN 波兰/NN。
巴金/NPSN（1904～）作家/NN。原名/NN 李尧棠/NPSN，字/NN 蓑甘/NN。笔/NN 名/NN 有/VL
佩竿/NN，笔/NN 名/NN 有/VL 范用/NN，笔/NN 名/NN 有/VL
欧南/NN，欧阳/NN 佩南/NN，欧南/NN 佩彝/NN，欧南/NN 佩南/NN，欧南/NN 佩彝/NN，欧南/NN 佩彝/NN。

白寿彝/NPSN（1900～）历史学家/NN。河南/NN 开封市/NN 人/NN。

白瑞/NN，NPSN（1939～）芭蕾/NN舞蹈家/NN。女/NN，辽宁/NN 新宾/NN 人/NN。

白先勇/NPSN（1937～）作家/NN。广西/NN 桂林/NN 人/NN。

白杨/NPSN（1920～）电影/NN、话剧/NN 演员/NN。原名/NN 杨成芳/NPSN，女/NN。原籍/NN
湖南/NN 湘阴/NN。

白云鹏/NPSN（1874～1952）京剧/NN 大鼓/NN 表演/NN 艺术家/NN。河北/NN 霸州/NN 人/NN。

曹禺/NPSN（1910～1996）剧作家/NN，戏剧/NN 教育家/NN。本/PN 名/NN 万家宝/NPSN，
字/NN 小石/NPSN。祖籍/NN 湖北/NN 潜江/NN。

老舍/NPSN（1899～1966）作家/NN。原名/NN 舒庆春/NPSN，字/NN 舒予/NPSN。满/A
族/FSN，北京/NN 人/NN。

陈景润/NPSN（1933～）数学家/NN，中国科学院/NN 院士/NN。1933年/NN 5月/NN 22
日/NN 生/VT 于/I福建/NN 福州/NN。
何香凝/NPSN（1878～1972）国民党/NN 革命派/NN 杰出/AA 代表/NN，画家/NN。原籍/NN 广
东省/NN 南海县/NN。

胡蝶/NPSN（1906～1989）电影/NN演员/NN。女/NN。原名/NN 胡瑞华/NPSN。生/NN 于/I
上海/NN。

华罗庚/NPSN（1910～1985）数学家/NN，中国科学院/NN 院士/NN。1910年/NN 11月/NN 12
日/NN 生/VT 于/I江苏/NN 金坛/NN。

梅兰芳/NPSN（1894～1961）(扮演/VT 京剧/NN《宇宙/NN 锋/NN》中/NN 红/NN 花/NN 梅/NN)
京剧/NN表演/NN 艺术家/NN。工旦/NN。原籍/NN 江苏/NN 泰州/NN。

田汉/NPSN（1898～1968）现代/A 戏剧/NN 著名人/NN，作家/NN，诗人/NN。原名/NN 常
昌/NPSN。湖南/NN 长沙县/NN人/NN。
叶圣陶/NPSN (1894~1988) 作家/NN，教育家/NN，出版家/NN，政治/NN 活动家/NN。原名/NN 叶绍钧/NPSN，笔/NN 名/NN 叶圣陶/NPSN。斯提/NPSN 等/SO。江苏/NPN 苏州/NPN 人/NN。

张爱玲/NPSN (1921~) 作家/NN。笔/NN 名/NN 梁京/NPSN。女/NN，河北/NPN 丰润/NPN 人/NN。
Appendix G

Translated Chinese Output via BabelFish

headword -> 阿英
born in -> 1900
died in -> 1977
native of -> 安徽芜湖
instance of -> theory of literature and art
instance of -> Criticism
instance of -> Literature historian
instance of -> writer

died in -> 1981
native of -> 浙江桐乡
instance of -> writer
instance of -> Political activist

headword -> 丁玲
born in -> 1904
died in -> 1986
native of -> 湖南临澧
instance of -> Female writer

headword -> 艾青
born in -> 1910
native of -> 浙江金华
instance of -> poet

headword -> 郭沫若
born in -> 1892
died in -> 1978
native of -> 四川乐山
instance of -> writer
instance of -> poet
instance of -> playwright
instance of -> historian
instance of -> archaeologist
instance of -> Ancient philology
instance of -> social activist

headword -> 阿炳
born in -> 1893
died in -> 1950
native of -> 江苏无锡
instance of -> Folk musician

headword -> 茅盾
born in -> 1896
<table>
<thead>
<tr>
<th>Headword</th>
<th>Born in</th>
<th>Died in</th>
<th>Native of</th>
<th>Instance of</th>
<th>Headword</th>
<th>Born in</th>
<th>Died in</th>
<th>Native of</th>
<th>Instance of</th>
</tr>
</thead>
<tbody>
<tr>
<td>吴作人</td>
<td>1908</td>
<td></td>
<td></td>
<td></td>
<td>付香</td>
<td>1909</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>现代</td>
<td>1909</td>
<td>1963</td>
<td></td>
<td></td>
<td>田</td>
<td>1939</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>画家</td>
<td>1909</td>
<td>1966</td>
<td></td>
<td></td>
<td>刘</td>
<td>1910</td>
<td>1945</td>
<td></td>
<td></td>
</tr>
<tr>
<td>艾冰</td>
<td>1909</td>
<td>1965</td>
<td></td>
<td></td>
<td>杨</td>
<td>1910</td>
<td>1937</td>
<td></td>
<td></td>
</tr>
<tr>
<td>徐悲鸿</td>
<td>1895</td>
<td>1953</td>
<td></td>
<td></td>
<td>黄</td>
<td>1909</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>艾中信</td>
<td>1915</td>
<td></td>
<td></td>
<td></td>
<td>艾</td>
<td>1916</td>
<td>1917</td>
<td></td>
<td></td>
</tr>
<tr>
<td>艾思奇</td>
<td>1910</td>
<td>1966</td>
<td></td>
<td></td>
<td>艾</td>
<td>1915</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>艾梅</td>
<td>1916</td>
<td>1966</td>
<td></td>
<td></td>
<td>艾</td>
<td>1939</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>艾冰</td>
<td>1909</td>
<td>1966</td>
<td></td>
<td></td>
<td>艾</td>
<td>1909</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>艾冰</td>
<td>1909</td>
<td>1966</td>
<td></td>
<td></td>
<td>艾</td>
<td>1909</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>艾冰</td>
<td>1909</td>
<td>1966</td>
<td></td>
<td></td>
<td>艾</td>
<td>1909</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>艾冰</td>
<td>1909</td>
<td>1966</td>
<td></td>
<td></td>
<td>艾</td>
<td>1909</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>艾冰</td>
<td>1909</td>
<td>1966</td>
<td></td>
<td></td>
<td>艾</td>
<td>1909</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>艾冰</td>
<td>1909</td>
<td>1966</td>
<td></td>
<td></td>
<td>艾</td>
<td>1909</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>艾冰</td>
<td>1909</td>
<td>1966</td>
<td></td>
<td></td>
<td>艾</td>
<td>1909</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>艾冰</td>
<td>1909</td>
<td>1966</td>
<td></td>
<td></td>
<td>艾</td>
<td>1909</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>艾冰</td>
<td>1909</td>
<td>1966</td>
<td></td>
<td></td>
<td>艾</td>
<td>1909</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>艾冰</td>
<td>1909</td>
<td>1966</td>
<td></td>
<td></td>
<td>艾</td>
<td>1909</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Headword</td>
<td>Born in</td>
<td>Died in</td>
<td>Native of</td>
<td>Instance of</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---------------------</td>
<td>---------</td>
<td>---------</td>
<td>-----------------------</td>
<td>--------------------------------------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>白云鹏</td>
<td>1874</td>
<td>1952</td>
<td>上海</td>
<td>film actor</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>生于 1874, died 1952</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>沈阳</td>
<td>1910</td>
<td>1996</td>
<td></td>
<td>scholar</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>生于 1910, died 1996</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>赵丹</td>
<td>1899</td>
<td>1966</td>
<td>北京</td>
<td>Play educationalist</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>生于 1899, died 1966</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>陈景润</td>
<td>1933</td>
<td>1968</td>
<td>湖南长沙县</td>
<td>scholar</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>生于 1933, died 1968</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>何香凝</td>
<td>1878</td>
<td>1972</td>
<td>广东省南海县</td>
<td>Revolutionary faction</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>生于 1878, died 1972</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>胡蝶</td>
<td>1906</td>
<td>1989</td>
<td></td>
<td>painter</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>生于 1906, died 1989</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
headword -> 张爱玲  
instance of -> writer

born in -> 1921

native of -> 河北丰润
Appendix H

ABKAS for Chinese

Figure H.1: ABKAS for Chinese - preprocessing
Figure H.2: ABKAS for Chinese - parsing
Figure H.3: ABKAS for Chinese - knowledge base construction
Appendix I

Example Entries ABKAS Failed to Parse

The evaluation results reported in Chapter 5 demonstrate that ABKAS can successfully parse the definition part of the biographical entries. However, there are certain cases where ABKAS was unable to parse the entries. In this appendix, some examples of the failed entries are listed and the cause of the failures are examined and reported for both English and Chinese tests.

I.1 Examples of Failed Entries in English

A list of example entries that ABKAS failed to parse an entry in Pocket Oxford Dictionary of Biography is given the table below. The items highlighted indicate where the system failed.

<table>
<thead>
<tr>
<th>Name</th>
<th>Hinge</th>
<th>Entry</th>
<th>Cause of failure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adams, John</td>
<td>(1735-1826)</td>
<td>Second American President (1797-1801).</td>
<td>‘Second’ was tagged as an ad-verb so it was not recognised.</td>
</tr>
<tr>
<td>Caedmon</td>
<td>7th century</td>
<td>English poet and monk.</td>
<td>Hinge form unexpected.</td>
</tr>
<tr>
<td>Calamity Jane (born Martha Jane Cannary)</td>
<td>(1852-1903)</td>
<td>American frontierswoman.</td>
<td>Fails to recognise ‘frontierswoman’ as a profession due to tagging error</td>
</tr>
<tr>
<td>Edison, Thomas Alva</td>
<td>(1847-1931)</td>
<td>The most famous and versatile US inventor.</td>
<td>Cannot parse the intensifier ‘most’ and the adjectives ‘famous and versatile’.</td>
</tr>
<tr>
<td>Name</td>
<td>Hinge</td>
<td>Entry</td>
<td>Cause of failure</td>
</tr>
<tr>
<td>----------------------------------</td>
<td>----------------</td>
<td>-------------------------------</td>
<td>----------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Fairfax, Thomas, 3rd Baron Fairfax of Cameron</td>
<td>(1612-71)</td>
<td>English Parliamentary general.</td>
<td>'Parliamentary general' was not recognised as a profession in ABKAS.</td>
</tr>
<tr>
<td>Gandhi, Indira</td>
<td>(1917-1984)</td>
<td>Indian Stateswoman.</td>
<td>'Stateswoman' is tagged as &quot;FW&quot; (foreign word).</td>
</tr>
<tr>
<td>Jack the Ripper</td>
<td>(19th century)</td>
<td>English murderer.</td>
<td>Unexpected hinge form.</td>
</tr>
<tr>
<td>Jenkins, Roy, Baron Jenkins of Hillhead</td>
<td>(born 1920)</td>
<td>English Labour and Social Democrat and scholar.</td>
<td>Failed to parse the complex noun phrase 'Labour and Social Democrat' as a superordinate.</td>
</tr>
<tr>
<td>Keble, John</td>
<td>(1792-1866)</td>
<td>English churchman.</td>
<td>'Churchman' is tagged as a foreign word.</td>
</tr>
<tr>
<td>MacDonald, Flora</td>
<td>(1722-90)</td>
<td>Scottish Jacobite.</td>
<td>'Jacobite' is not recognised as a profession.</td>
</tr>
<tr>
<td>Park, Nick</td>
<td>(born 1960)</td>
<td>English animator.</td>
<td>'Animator' is not recognised as a profession.</td>
</tr>
<tr>
<td>Roberts, Frederick Sleigh, 1st Earl Roberts of Kandahar</td>
<td>(1832-1914)</td>
<td>British Field Marshal.</td>
<td>Fails to recognise military titles 'Field Marshal'.</td>
</tr>
<tr>
<td>Warburg, Otto Heinrich</td>
<td>(1883-1970)</td>
<td>German biochemist.</td>
<td>'Biochemist' is tagged as a foreign word.</td>
</tr>
</tbody>
</table>
Appendix I. Example Entries ABKAS Failed to Parse

1.2 Examples of Failed Entries in Chinese

The following table offers a list of example entries that ABKAS failed to parse in ChinaOnInternet. Items which have been highlighted indicate where the system failed.

<table>
<thead>
<tr>
<th>Name</th>
<th>Hinge</th>
<th>Entry</th>
<th>Cause of failure</th>
</tr>
</thead>
<tbody>
<tr>
<td>毛泽东</td>
<td>(1893-1976)</td>
<td>中国人民的领袖，马克思主义者，伟大的无产阶级革命家、战略家，理论家，中国共产党、中国人民解放军和中华人民共和国的主要缔造者和领导人，诗人，书法家。</td>
<td>Discriminators are too complex to parse.</td>
</tr>
<tr>
<td>孔祥熙</td>
<td>(1880-1967)</td>
<td>中国官僚资产阶级代表之一，与蒋介石、宋子文、陈果夫兄弟合称“四大家族”。山西省太谷县人。</td>
<td>No significant superordinate found.</td>
</tr>
<tr>
<td>江青</td>
<td>(1914-1991)</td>
<td>文化大革命期间，林彪、江青反革命集团的首要分子。山东诸城人。</td>
<td>No significant superordinate found.</td>
</tr>
<tr>
<td>黄公略</td>
<td>(1898-1931)</td>
<td>平江起义领导人。湖南湘乡人。</td>
<td>Segmentation and tagging error occurred at 平/NPSN 江 起义/NPSN.</td>
</tr>
<tr>
<td>杨森</td>
<td>(1884-1977)</td>
<td>国民党军将领。四川广安人。</td>
<td>Segmentation and tagging error occurred at 国民党/NPN 军/NN.</td>
</tr>
</tbody>
</table>
Bibliography


