



A Starter's Guide to Linguistic Corpora Building

Pei-Ying Li, Chih-Ming Chiu,
Huo-Tsen Kuo, Shu-Chuan Tseng,
Chu-Fang Huang, Ching-Hsun Chan,
Su-Chuan Tsai, Chiu-Jung Lu,
Su-Ying Hsiao, Chia-Min Lai,
Hao-Yi Tai, Pei-Yu Hsieh,
Hsiu-Fang Su, Chinese WordNet Group

Taiwan e-Learning and Digital Archives Program
Taiwan Digital Archives Expansion Project

Publisher's Preface

After the “National Digital Archives Program” was initiated in 2002, members of numerous institutional projects and request-for-proposals projects joined our team to engage in digital work that covered countless categories and massive amounts of content. The first phase of the five year project was successfully completed in 2006. The following year, the “National Digital Archives Program” and “National Science and Technology Program for e-Learning” were integrated into the “Taiwan e-Learning and Digital Archives Program (TELDAP, <http://teldap.tw/>)”, striving to achieve the ultimate goal of “presenting Taiwan’s cultural and natural diversity” as it continued to expand digital resources in various fields, and systemically promoted digital achievements in education, research and industries. TELDAP is preparing to actively collaborate with the private sector to drive growth in related industries, not only preserving important cultural assets, but also accelerating the development of a new culture in the digital age of today.

Originally named the “Content Development Division” during the first phase, we were renamed “Taiwan Digital Archives Expansion Project” (<http://content.teldap.tw>) as a subproject of TELDAP, and took more active measures to expand the sources of digital content, extending our reach to the collections of private institutions and even individuals. We have widely requested proposals for digitization projects related to archives, archeology, philology, geography, ethnicity, art, daily life, animals and plants, and hope to better integrate digital content with different characteristics, to develop them into fun and inspiring digital materials, and to provide them free of charge to the public for education and research; this will also help firms and public or private holding institutions to find cooperation opportunities in value-added applications. Collaboration between the “Taiwan Digital Archives Expansion Project” and other projects under the “Taiwan e-Learning and Digital Archives Program” will help speed up development of educational, research and commercial value-added applications of digital content, which will benefit the presentation of Taiwan’s cultural and natural diversity, and allow people everywhere around to understand and appreciate the richness of our history and culture, as well as the beauty of our natural ecology.

While collecting and developing value-added applications of digital content, whether it may be during the “Content Development Division” or “Taiwan Digital Archives Expansion Project” period, members of this project have continuously followed up on digital workflow related technologies used by public and private institutions and open request-for-proposals projects, and compiled a series of “Digitization Procedures Guideline Books” that introduce various international standards and provide information on digitization technologies and workflows. Since 2005, we have written 21 digitization procedures guidelines on different themes (the full text of all of the 21 books can be downloaded from the “Taiwan Digital Archives Expansion Project” website under “Virtual Library: Digitization Books”), selecting exquisite digital objects, such as ceramics, paintings, calligraphy, and string-bound books, combining the experiences of different institutional projects, and supporting them with domestic and foreign theories and practice results.

Since 2008, we have continuously revised and expanded our “Digitization

Procedures Guideline” book series, hoping to expand distribution channels so that they may be provided to even more museums, libraries, institutions and individuals for reference. Our preparations are mainly divided into revising existing guidelines for “selected objects” and compiling new guidelines on “common principles”. The former refers to revising the existing 21 guidelines with a focus on introducing new digitization technologies and specifications, more practical software and hardware, and digital content protection mechanisms; we expect to revise seven books per year and complete all 21 books within three years. As for compiling guidelines on “common principles,” our emphasis will be on the introduction of key concepts, such as the “life cycle” of digital information and quality control, studying multiple types of objects instead of a single type of object, and adopting common principles as the guideline framework. The so called common principles refer to project planning, integrated workflow, audiovisual data, text data, color management, outsourcing management, and digital content protection and authorization. These eight common principles are topics of which we will investigate, study and write guidelines for; we expect to publish eight guidelines in three years.

Guidelines for selected objects and guidelines on common principles in fact complement each another. Guidelines on common principles emphasize on the analysis of important topics in digitization work, guiding readers to thoroughly consider the advantages and disadvantages of digitization. Guidelines on selected objects describe practices and techniques for digitizing specific objects, helping readers to select the most suitable, most effective digitization workflow. By publishing this “Digitization Procedures Guideline” book series, we believe that we are providing institutions and individuals with the intention to engage in digitization work with a series of practical guidelines that provide an overall view, while guiding them step by step through the digital workflow. Here we must stress that the theoretical foundation of this book series is the precious experiences of institutional and request-for-proposal project teams accumulated throughout the years. These experiences allow higher quality digital content to be produced, presented and maintained with less cost, further enriching our digital archives and e-learning content. As we continue to publish our “Digitization Procedures Guideline” book series, we must give special thanks to working partners who were interviewed and colleagues who were a part of writing the guidelines, and are grateful to the scholars and specialists that reviewed and provided their advice on the book series. Finally, we hope that readers will not be reluctant to correct any mistakes or make recommendations that will help us be even better.

Taiwan e-Learning and Digital Archives Program
Taiwan Digital Archives Expansion Project · Digital Archives Sub-project of Project
Integration



Project Director
February 10th, 2010

This book is the revised and enlarged edition of “Digitization Procedures Guideline: Philology Thematic Group” (2006). For the original book, Academician Chin-Chuan Cheng, convener of the Philology Thematic Group at the time, was responsible for overall planning, and assistant Chia-Min Lai was responsible for editing and compiling its 58 pages. The introduction of the original book briefly described projects of the Philology Thematic Group, and then shared experiences of project teams, including digitization procedures of Academia Sinica Balanced Corpus of Modern Chinese, Language Distribution GIS, TSL Online Dictionary, and Taiwan Child Language Corpus. In this revised and enlarged edition new chapters and an appendix were added, including Metadata and Related International Standards, Procedures for Building, Digital Learning, and Extended Issues. Furthermore, some new text was added to examples of the original book, and building procedures of the second phase language archives project “A Socio-phonetic Study of Spoken Taiwan Mandarin Project” were introduced. Samples provided in this book are not limited to projects of the Philology Thematic Group; the experience of the Chinese WordNet Group led by research fellow Chi-Jen Huang of the Academia Sinica Institute of History and Philology is also provided in this book.

This book is the joint achievement of all personnel and research teams that contributed to the first edition and this revised and enlarged edition. We are grateful to the project teams that provided information, articles, and took part in discussions. We would especially like to thank professors Hao-Yi Tai and Jane Tsay of National Chung Cheng University Graduate Institute of Linguistics for updating and reviewing introductions to the “TSL Online Dictionary” and “Taiwan Child Language Corpus”, the Chinese WordNet Group of Academia Sinica Institute of History and Philology for providing articles and sharing their experiences, Associate Research Fellow Shu-Chuan Tseng of Academia Sinica Institute of History and Philology for reviewing articles on “A Socio-phonetic Study of Spoken Taiwan Mandarin”, and Research Fellow Elizabeth Zeitoun and “Formosan Language Digital Archive” Research Assistant Walis Buya of Academia Sinica Institute of History and Philology for providing metadata, cross-database retrieval, and equipment information, as well as taking part in discussions on drawing the digitization flowchart. In addition, we would also like to thank colleagues of the “Taiwan Digital Archives Expansion Project” for providing suggestions to this book.

In this revised and enlarged edition, I was responsible for framing chapter outline, soliciting articles, writing articles, reviewing the book and text editing. Assistant Ching-Hsun Chan of the “Philology, Videography and Journalism Thematic Group” was responsible for writing most of the additions and arranging illustrations.

Finally, there was less than three months for editing before this book was finalized. Due to the lack of time and experience, there are sure to be sections that need revision or addition, please do not feel reluctant to offer your advice.

Academia Sinica Institute of History and Philology
Assistant Research Fellow
Su-Ying Hsiao
January 26th, 2010

2 Publisher's Preface

4 Editor's Preface

7 Introduction

I. What is a corpus?.....Su-Ying Hsiao, Pei-Ying Li 008

II. Contents of this book.....Su-Ying Hsiao 010

11 ONE. Metadata and Related International Standards

I. Dublin Core.....Ching-Hsun Chan 013

II. Corpus metadata and international standards.....
.....Su-Ying Hsiao, Ching-Hsun Chan 016

24 TWO. Procedures for Building.....Su-Ying Hsiao

30 THREE. Samples for Building

I. Digitization Procedures of Academia Sinica Balanced Corpus of Modern Chinese.....Chia-Min Lai, Chiu-Jung Lu, Chih-Ming Chiu 032

II. Lexical Knowledgebase Establishment Procedures.....
.....Chinese WordNet Group 041

III. Procedures for Building the Language Distribution GIS.....
.....Chia-Min Lai, Chiu-Jung Lu, Chu-Fang Huang, Yu-Tsen Kuo 049

IV. Digitization Procedures of TSL Online Dictionary.....
.....Hao-Yi Tai, Jane Tsay, Hsiu-Fen Su, Chia-Min Lai 058

V. Digitization Procedures of Taiwan Child Language Corpus.....
.....Pei-Yu Hsieh, Chia-Min Lai 068

VI. A socio-phonetic study of spoken Taiwan Mandarin.....
.....Shu-Chuan Tseng 079

83 FOUR. Digital Learning

- I. Digital Resources Center for Global Chinese Teaching and Learning.....
.....Ching-Hsun Chan, Pei-Ying Li 084
- II. NCKU Eagle Project and the CANDLE Project for Reading.....
.....Ching-Hsun Chan, Pei-Ying Li 087

90 FIVE. Extended Issues

- I. Digital Content Protection.....Ching-Hsun Chan 091
- II. Human Resource and Equipment Cost Analysis.....
.....Ching-Hsun Chan 095

118 SIX. Conclusions

120 References

122 Appendix

Introduction



I. What is a Corpus?

A corpus is a large and structural language database that can collect either only a single language or multiple languages; contents include text, sign language and speech audio files. Corpora are important results of linguistic research, and also a research tool for statistical analysis of language. To the typical user, corpora are a tool for learning languages.

Language is an important medium for human beings to express themselves and communicate; linguistics is a science that studies human languages. There are currently over three thousand known existing languages in the world, and as new languages are born, some languages wither. Building a corpus is a good option for preserving withering or developing languages; corpora are crystals of linguistic research combined with information technology. A corpus usually refers to samples of natural language that are collected for linguistic research, that are preserved in digital form, and that are used to express a specific language or language change. Corpora of suitable scale can be used to show and record the actual usage of languages. Observations of a corpus can find facts and patterns of a language, which are important resources to linguistic theory research, application research and linguistic engineering.

Depending on the target language(s), corpora can be divided into Monolingual, Bilingual and Multilingual. Corpora have diversified in recent years, developing from monolingual to multilingual, with some even combining audiovisual representation. These developments not only benefit linguistic research and analysis, but also greatly contribute to language learning.

Corpora are closely related to linguistic data processing. Before corpora were used, linguistic analysis of studies on natural language processing and machine translation was mainly based on rules, but some rules cannot be expressed and not all rules can be covered, making it difficult for computers to process. After the appearance of corpora, people could use corpora to investigate and compile statistics on natural languages, establish statistical models, and research natural language processing technology. From another aspect, studies on natural language processing technology have provided key technologies for corpora, including information search, text input, automatic segmentation and tagging, and corpus statistics and search.

Corpus functions mainly cover three aspects, corpus scale, distribution, and degree of processing. Scale affects the reliability of statistics, distribution is

related to the applicable scope of statistical results, and degree of processing determines what type of linguistic information the corpus can offer.

Corpora can be divided into the following four categories based on data gathering principles and methods:

- 1. Heterogeneous:** Linguistic data is collected non-specific to its type; a variety of linguistic data is widely collected and stored in its original form.
- 2. Homogeneous:** Only the same type of linguistic data is gathered.
- 3. Systematic:** Linguistic data is collected based on a pre-defined principle and ratio, creating a balanced and systematic corpus that represents language facts within a certain range. The Brown Corpus was established by Brown University in the 1960s, and was the world's first standard corpus with samples gathered based on systematic principles; the corpus had a scale of one million terms. The "Academia Sinica Balanced Corpus of Modern Chinese", simplified as "Sinica Corpus", is the world's first balanced modern Chinese corpus with part-of-speech tagging.
- 4. Specialized:** Only linguistic data for a specific purpose are gathered.

Linguistic data processing mainly refers to text format processing and text description; in text format processing, linguistic data is organized and converted into digital text of the same format, such as database format, XML format, etc.

Text description refers to the description of properties or characteristics of each linguistic data sample, including head and body descriptions. The head description describes metadata properties of the entire linguistic data sample, such as style, field of the content, author, publishing data, and publisher; the body description adds tags to the text, e.g. term segmentation tag, part-of-speech tag, grammatical feature tag, semantic role tag, dialogue tag...etc. Chinese corpus processing generally starts from segmentation, part-of-speech tagging, to syntax, semantic tagging; increased tags indicate increased degree of linguistic data processing.

Linguistic data that doesn't have a body description is called plain linguistic data. Plain linguistic data that is Chinese text can only be used for searches and statistics based on single words; linguistic data that have undergone segmentation can be used for searches, statistics and quantified analysis based on terms; if part-of-speech tags are added, then even more information can be

acquired. If tagging is done manually, correctness can be ensured, but it will be very slow. To large scale corpora, manual tagging is obviously too slow and can't meet requirements. Therefore, large scale corpora rely on automation technology to complete segmentation and part-of-speech tagging.

II. Contents of this Book

This book consists of eight chapters, including introduction, metadata and related international standards, procedures for building, samples for building, digital learning, extended issues, conclusion and appendix. “Metadata and related international standards” introduces the widely adopted metadata standard “Dublin Core (DC)”; the Open Language Archives Community (OLAC) used DC as a basis for developing the metadata standard OLACMS; OLAC adopts the cross-database harvesting protocol OAI-PMH and international standards for language codes. “Procedures for building” divides corpus building into three parts, linguistic data digitization, corpus system building, and metadata establishment. “Samples for building” collects experiences of building corpora with texts, speech audio files, videos, and linguistic GIS, providing them as reference for other institutions or projects to use when building a corpus. “Digital learning” introduces two examples of corpus value-added applications in teaching websites. “Extended issues” discusses digital content protection, and cost of human resources and equipment.

Written by: Su-Ying Hsiao, Pei-Ying Li



ONE. Metadata and Related International Standards

The most common definition of metadata in the field of digital archiving is “data about data.” Using photographs taken by a digital camera as an example, each photo is a digital file, and besides image data, each file will also include EXIF metadata, which records the date, time, place, aperture, shutter, focal length, lens and white balance settings.

According to the definition of the Taiwan e-Learning and Digital Archives Program (TELDAP) Metadata Architecture and Application Team (MAAT):¹

Metadata is a set of structural and standardized background information associated with objects that falls into three categories: descriptive, structural, and administrative, describing the contents and characteristics of each object in terms of semantics, syntax, and lexicology. Metadata allows digital collections to achieve optimal resource discovery performance in a digital environment or system, and effectively provides search, display, management, control and execution functions that facilitates digital resource interoperability and sharing, fulfilling its role as basic information for the permanent preservation of digital collections.

This tells us that in order to achieve permanent preservation of digital archives, each collection should have a set of metadata; metadata ensures the digitization results can be meaningfully and permanently preserved, as well as efficiently found and utilized. In terms of function, there are three types of metadata:²

- 1. Descriptive metadata:** Describes contents and relations of a document or resource so that resources can be found and identified, e.g. bibliographic records and the Dublin Core, which will be introduced in the following section.
- 2. Structural metadata:** Provides actual results of digital archives for browsing, search and representation, e.g. chapter outline of a book, electronic full text with page turning function, and links between text and related images.
- 3. Administrative metadata:** Records information for long-term management, usage and viewing, e.g. file format, resolution, and

¹Taiwan e-Learning and Digital Archives Program Metadata Architecture and Application Team website: <http://metadata.teldap.tw/introduction/introduction-frame.html>.

²Han-Tsung Shen “Digital Archives Technology Collection” electronic book, National Digital Archives Program, 2004, ch.9-1.

intellectual property rights.

Metadata can be adjusted according to project requirements, resulting in different sizes of metadata and cataloging rules that may vary along with object type. Metadata for corpora must give consideration to numerous aspects. In order to achieve integrity of data, metadata establishment will adopt multiple international standards, such as the Dublin Core, OLACMS, and ISO language codes. If a corpus were to allow cross-database retrieval or achieve data exchange with foreign corpora, then it will need to use the network protocol recommended by OLAC – OAI-PMH.

The purpose of linguistic research is to understand behavioral models of languages and to analyze different languages. When doing research work, the person speaking the language and the location where linguistic data is collected are both topics of discussion, so each language sample must come with detailed metadata if it is to be used as basic information for linguistic research. Therefore, metadata is relatively large when it is for a corpus.

Before collecting linguistic data, it is best to first complete metadata establishment according to project requirements. Carefully consider information that must be recorded when collecting linguistic data, this way waste of project funds can be avoided in case information is found to be insufficient after leaving the investigation site.

MAAT stresses that metadata planning and implementation are essential to digital archiving, whether or not search functions are practical and linguistic data is complete depends on the thoroughness of metadata. In the light of this, when the project team is establishing metadata, it should spend more time on making metadata complete and comprehensive. Below we will introduce a commonly used metadata element set “Dublin Core” and a number of international standards related to corpus metadata.

I. Dublin Core

The Dublin Core (DC) was a seminar jointly sponsored by the Online Computer Library Center (OCLC) and National Center for Supercomputing Applications (NCSA) in 1995. Its members included scholars and specialists in the field of libraries, computers, the internet and other professional fields, establishing a set of metadata elements to describe network resources. This set of elements was named after the venue of the seminar, which was in

Dublin, Ohio, and called the Dublin Core (DC). The Dublin Core today is an international standard that is managed by the Dublin Core Metadata Initiative.

Rules of the Dublin Core strive to be simple and effective, and it is why DC is widely used for digital objects. Each element of the Dublin Core is optional and can also be used repeatedly; most elements have a set of restrictive detailed options to choose from, allowing more complete expression of meaning. The elements can be arranged in any order, and cataloging rules can be established based on project requirements. This flexibility makes the Dublin Core easy to control and use, but may not be suitable for all objects, especially ones with complex meanings and concepts.

At present, the Dublin Core has two levels, the “simple” Dublin Core uses 15 elements to describe digital objects; the qualified Dublin Core uses qualifiers to extend or refine the 15 elements, and has three additional elements: audience, provenance and RightsHolder, making data easier to find.

In addition to, element refinement the Qualified Dublin Core includes a set of recommended encoding schemes, which are used based on three principles:

1. **One to one principle:** The Dublin Core only describes one object at a time. Replicas or different versions of an object will have different values in the Creator and Contributor columns.
2. **Simplification principle:** Elements are not required to use any qualifiers and can retain only their data values.
3. **Suitable data values:** Different objects have different qualifier contents; careful consideration can bring effects of metadata into full play.

Dublin Core is easy to establish, even allowing users without professional training to establish metadata, or even develop their own editor. Furthermore, the Dublin Core is highly flexible; its contents can be extended, are optional and can be repeated, which meets the requirements of most digital archives. Finally, the Dublin Core was developed based on English, which is a global language, and has therefore become a widely applied international metadata standard.

Table 1-1 Dublin Core element list³

Element	Definition	Qualifiers	
		Element Refinements	Element Encoding Schemes
Title	A name given to the resource	Alternative	
Creator	An entity primarily responsible for making the content of the resource		
Subject and Keywords	The topic of the content of the resource		LCSH MESH DDC LCC UDC
Description	An account of the content of the resource	Table of Contents Abstract	
Publisher	An entity responsible for making the resource available		
Contributor	An entity responsible for making contributions to the content of the resource		
Date	A date associated with an event in the life cycle of the resource	Created Valid Available Issued Modified	DCMI Period W3C-DTF
Resource Type	The nature or genre of the content of the resource		DCMI type vocabulary
Format	The physical or digital manifestation of the resource	Extent Medium	IMT

³ Dublin Core element list, Taiwan e-Learning and Digital Archives Program Metadata Architecture and Application Team website: <http://metadata.teldap.tw/standard/standard-frame.html>.

Element	Definition	Qualifiers	
		Element Refinements	Element Encoding Schemes
Source	Reference to a resource from which the present resource is derived		URI
Language	A language of the intellectual content of the resource		ISO 639-2 RFC 1766
Relation	A reference to a related resource	Is version of Has version Is replaced by Requires Is part of Has part Is referenced by References Is format of Has format	URI
Coverage	The extent or scope of the content of the resource	Spatial Temporal	DCMI point ISO 3166 DCMI box TGN DCMI Period W3C-DTF
Rights Management	Information about rights held in and over the resource	AccessRights License RightsHolder	

II. Corpus Metadata and International Standards

1. OLACMS

The Open Language Archives Community (OLAC) is an international partnership of institutions or individuals that was founded in December 2000. Its main coordinators are Steven Bird and Gary Simons, Academician Chin-Chuan Cheng of Academia Sinica is on its Advisory Board, and Research Fellow Chu-Ren Huang of Academia Sinica Institute of History and Philology is a council member.

Seeing that many organizations in the world need linguistic resources, e.g. linguists, engineers, archive administrators, software developers and publishers, and that most users hope to access resources using a single interface, including information that describe languages and language search tools, but couldn't

find the required resources in one search because resources were scattered all over the internet, OLAC set down two objectives⁴, the first was to develop a consistent practice guide for language archives, and the second was to develop an interoperable linguistic resource repository and service center.

To achieve these two objectives, the OLAC used the two standards established by the Dublin Core Metadata Initiative and Open Archives Initiative (OAI) as a basis to achieve data exchange with foreign databases and cross-database search.

In terms of metadata, the OLAC revised the Dublin Core's 15 elements and established a more detailed set of metadata elements – OLACMS, as shown in Table 1-2:

Table 1-2 OLACMS Elements

Element	Element
Contributor	Language
Coverage	Publisher
Creator	Relation
Date	Rights
Description	Source
Format	Subject
Format.cpu	Subject.language
Format.encoding	Title
Format.markup	Type
Format.os	Type.functionality
Format.sourcecode	Type.linguistic
Identifier	

OLACMS adopts four attributes for more detailed characteristic definitions, and an additional auxiliary attribute – langs.

1. refine: used to identify more detailed meanings and characteristics.
2. scheme: prescribes the standard names of element text content.
3. code: used to tag metadata; unique tagging system of OLAC.
4. lang: a required attribute for every element in OLACMS that specifies the language used for an element.
5. langs: prescribes the language used to read the metadata element.

⁴Ju-Ying Chang "Introduction to the Open Language Archives Community and Participation of the Linguistic Anchoring Project", Language Archives Division, National Digital Archives Program, http://www2.ndap.org.tw/newsletter06/news/read_news.php?nid=888.

2. Network Protocol for Cross-database Search

OLAC also provides a solution for corpus projects that intend to engage in cross-database searches. To facilitate searches between databases, OLAC adopts the network protocol established by the Open Archives Initiative (OAI)⁵ – OAI-PMH; contents of this protocol allow users to search for data on the internet, including metadata, without be limited by system, application, field and language.

OLAC uses OAI-PMH to retrieve data from Data Providers, or corpora, and then creates an index in OAI Service Providers. Once users search for data on the internet, they can rapidly find complete and rich index results. If the project team wants to conduct cross-database searches, there are two methods, one is to set up an OAI Data Provider server by themselves, allowing OAI Service Providers to retrieve data on a regular basis; the other is to follow the recommendations of the OLAC, convert corpus data into XML documents, and provide the documents to OAI Service Providers.

3. International Standards for Language Codes

ISO 639 is a set of language codes established by the International Organization for Standardization, and consists of six parts.⁶ ISO 639-1 is the first part that was published in 2002, and uses two letter codes to represent main languages of the world; the registration authority is Infoterm (International Information Center for Terminology)⁷. ISO 639-2 is the second part that was published 1998, and uses three letter codes to represent languages, macrolanguage, language family and collections of languages, in which a macrolanguage is the name of several closely related languages. Furthermore,

ISO 639-2 consists of four special codes (mis, mul, und, and zxx) and a reserved area for user definition (qaa~qtz); “mis” indicates an “uncoded language”, “mul” indicates multiple languages that are not individually labeled, “und” indicates an “Undetermined Language”, and “zxx” indicates “No Linguistic Content”, which means that the language must be labeled, but no linguistic information is contained. The registration authority of ISO 639-2 is the Library of Congress. ISO 639-3 is the international standard currently recommended by the OLAC, and was published in 2007. It is an extension of ISO 639-2, but doesn’t include language families, and collections of languages; its goal is to represent all languages, existing, extinct, historical, ancient and artificial, using three letter codes. SIL International⁸ in 2002 began participating in the establishment of ISO 639-3, and integrated SIL language codes into the new standard; Ethnologue 15th edition adopted the standard. SIL International is also the registration authority of ISO 639-3. ISO 639-5 was published in 2008 and extends collections of languages in ISO 639-2, using three letter codes to represent language families, language groups or homogeneous language collections (e.g. sign language, mixed language, and artificial language); the registration authority of 639-5 is also the Library of Congress. ISO 639-6 was published in November 2009, and attempts to describe all language variants in the world using a four letter code. Due to its recent publication, ISO 639-6 is only adopted by very few institutions, institutions that took part in its establishment, such as GeoLang Ltd⁹.

⁵“National Archives Administration 2008 Work Results – Collection of Division Knowledge”, from the NAA website http://wiki.archives.gov.tw/index.php?option=com_content&view=article&id=556&Itemid=107.

⁶Six parts of ISO 639: ISO 639-1:2002 Codes for the representation of names of languages -- Part 1: Alpha-2 code; ISO 639-2: 1998 Codes for the representation of names of languages -- Part 2: Alpha-3 code; ISO 639-3: 2007 Codes for the representation of names of languages -- Part 3: Alpha-3 code for comprehensive coverage of languages; ISO 639-4 Codes for the representation of names of languages -- Part 4: General principles of coding of the representation of names of languages and related entities, and application guidelines (not published yet); ISO 639-5: 2008 Codes for the representation of names of languages -- Part 5: Alpha-3 code for language families and groups; ISO 639-6: 2009 Codes for the representation of names of languages -- Part 6: Alpha-4 code for comprehensive coverage of language variants.

⁷Infoterm: <http://www.infoterm.info/>

⁸Main work items of the SIL include language development, academic research, language proficiency training, technology for language development, translation, and technical language development. The Ethnologue: Languages of the World published by SIL used language codes defined by SIL before the 14th edition.

⁹Geolang: <http://www.geolang.com>

Table 1-3 International standard codes for languages and collections of languages¹⁰
(Table made by Su-Ying Hsiao)

Name	ISO 639-5	ISO 639-3	ISO 639-2	ISO 639-1	Note
Altaic languages	tut		tut		Collection of language
Amis		ami			
Amis, Nataoran		ais			
Artificial languages	art		art		Collection of language
Atayal		tay			
Austro-Asiatic languages	aav				Collection of language
Austronesian languages	map		map		Collection of language
Babuza		bzg			
Basay		byq			
Bunun		bnn			
Buriat		bua	bua		Macrolanguage
Buriat, China		bxu			
Buriat, Mongolia		bxm			
Buriat, Russia		bxr			
Chinese		zho	zho/ chi	zh	Macrolanguage
Chinese, Gan		gan			
Chinese, Hakka		hak			
Chinese, Huizhou		czh			
Chinese, Jinyu		cjy			
Chinese, Late Middle		ltc			
Chinese, Literary		lzh			
Chinese, Mandarin		cmn			


¹⁰ Source: Languages of Taiwan, “Ethnologue: Languages of the World”, Ethnologue: Web, http://www.ethnologue.org/show_country.asp?name=TW, search date: January 21st, 2010; ISO 639 http://en.wikipedia.org/wiki/ISO_639, search data: January 21st, 2010; ISO 639 Code Tables <http://www.sil.org/iso639-3/codes.asp>, search date: January 21st, 2010, List of ISO 639-5 codes http://en.wikipedia.org/wiki/List_of_ISO_639-5_codes, search date: January 21st, 2010.

Name	ISO 639-5	ISO 639-3	ISO 639-2	ISO 639-1	Note
Chinese, Min Bei		mnp			
Chinese, Min Dong		cdo			
Chinese, Min Nan		nan			
Chinese, Min Zhong		czo			
Chinese, Old		och			
Chinese, Pu-Xian		cpx			
Chinese, Wu		wuu			
Chinese, Xiang		hsn			
Chinese, Yue		yue			
Creoles and pidgins	crp		crp		Collection of language
Daur		dta			
Dongxiang		sce			
English		eng	eng	en	
English, Middle (1100-1500)		enm	enm		
English, Old (ca. 450-1100)		ang	ang		
Esperanto		epo	epo	eo	
Formosan languages	fox				Collection of language; Hierarchical relationship: map:fox
German		deu	deu/ ger	de	
Germanic languages	gem		gem		Collection of language; Hierarchical relationship: ine:gem
Indo-European languages	ine		ine		Collection of language
Japanese		jpn	jpn	ja	
Jurchen		juc			
Kalmyk~Oirat		xal	xal		
Kanakanabu		xnb			

Name	ISO 639-5	ISO 639-3	ISO 639-2	ISO 639-1	Note
Kavalan		ckv			
Ketangalan		kae			
Kitan		zkt			
Korean		kor	kor	ko	
Kulon-Pazen		uun			
Manchu		mnc	mnc		
Mon-Khmer languages	mkh		mkh		Collection of language; Hierarchical relationship: aav:mkh
Mongolian		mon	mon	mn	Macrolanguage
Mongolian, Classical		cmg			
Mongolian, Halh		khk			
Mongolian, Middle		xng			
Mongolian, Peripheral		mvf			
Mongolian languages		xgn			Collection of language; Hierarchical relationship: tut:xgn
Oirat, Written		xwo			
Paiwan		pwn			
Papora-Hoanya		ppu			
Puyuma		pyu			
Qiang, Northern		cng			
Qiang, Southern		qxs			
Rukai		dru			
Saaroa		sxr			
Saisiyat		xsy			
Sign languages	sgn		sgn		Collection of language
Sino-Tibetan languages	sit		sit		Collection of language
Siraya		fos			
Taiwan Sign Language		tss			

Name	ISO 639-5	ISO 639-3	ISO 639-2	ISO 639-1	Note
Tangut		txg			
Taroko		trv			
Thao		ssf			
Tibetan		bod	bod/ tib	bo	
Tibetan, Amdo		adx			
Tibetan, Classical		xct			
Tibetan, Khams		khg			
Tibetan, Old		otb			
Tibeto-Burman languages	tbq				Collection of language; Hierarchical relationship: sit:tbq
Tsou Tungus languages	tuw	tsu			Collection of language; Hierarchical relationship: tut:tuw
Turkic languages	trk				Collection of language; Hierarchical relationship: tut:trk
Uighur		uig	uig	ug	
Uighur, Old		oui			
Yugur, East		yuy			
Yugur, West		ybe			

Written by: Su-Ying Hsiao and Ching-Hsun Chan; Special thanks to: Walis Buya



TWO. Procedures for Building

This chapter outlines procedures for building a linguistic corpus. The next chapter uses examples to introduce corpora of speech, text, and sign language. Corpus building procedures are divided into linguistic data digitization, system establishment, and metadata establishment, as shown in Fig 2-1.¹¹ Metadata was introduced in the previous chapter, and system establishment is closely related to characteristics of linguistic data and establishment purpose. Therefore, this chapter will mainly discuss linguistic data digitization procedures.

When planning a corpus, first decide the contents that will be collected based on the purpose, and then establish digitization specifications, equipment and tagging rules. Data collected for a corpus that will be for speech recognition and synthesis studies might be high quality audio files recorded in a studio, in which the tags added might be acoustic parameters for speech sound. Data collected for a corpus that will be for history and philology studies might be text files, in which tags added might include provenance, segmentation and part-of-speech.

Linguistic data collected by corpora might be in written form, or could require further investigation. Some written data might have electronic text files, while others might only be manuscripts or printed copies. After deciding on the content that will be collected, authorization issues also need to be considered. According the Copyright Act of the R.O.C., economic rights endure for the term of the author's life and fifty years after the author's death; economic rights in works authored by a juristic person endure for fifty years after the public release of the work; A creation adapted from one or more pre-existing works is a derivative work and shall be protected as an independent work; the rights of the plate maker shall subsist for ten years from the time the plate is completed.

Corpora that collect text data need to acquire authorization from the copyright owners. Although ancient books can be freely used by anyone, the version being used might still involve plate rights, and derivative works, e.g.

¹¹ Drawn by Ching-Hsun Chan; the metadata part referred to data provided by the “TELDAP MAAT” and “Taiwan Digital Archives Expansion Program – The Establishment and Integration of Digital Content”, procedures for collecting audiovisual data were drafted by research assistant Chiung-Yi Yu of Southern Min Archives after referring to “Planning and Procedures for Audiovisual File Digitization” (http://content.ndap.org.tw/index/?dl_id=76, downloaded on January 31st, 2009) and practical experiences, “Formosan Language Digital Archive” research assistant Walis Buya took part in discussions and provided many recommendations for system establishment and data backup.

textual criticism, remarks, might also have existing copyright. If the linguistic data is speech sound provided by a speaker, then the speaker's letter of authorization should be acquired in consideration of the speaker's copyright, right of publicity, and right to privacy.

Furthermore, according to the Copyright Act, where a work is completed by an employee within the scope of employment, the economic rights to such work shall be enjoyed by the employer¹²; where a work is completed by a person under commission, enjoyment of the economic rights to such work shall be assigned through contractual stipulation to either the commissioning party or the commissioned person; where no stipulation regarding the enjoyment of economic rights has been made, the economic rights shall be enjoyed by the commissioned person.¹³ Corpus building might involve "works" protected by the Copyright Act, including audiovisual works, sound recordings, photographic works, oral and literary works, translations, system designs, and computer programs; any person commissioned to complete works might have economic rights in accordance with the law; special attention must be paid during contract signing.

After recording audiovisual data, besides saving the file, the recording must be edited to eliminate parts that are not suitable or involve sensitive content, e.g. personal privacy¹⁴, or the parts might be processed that that areas are blurred or mute.¹⁵ After editing is complete, output the file in a permanent preservation

¹²Article 11 of the Copyright Act of the R.O.C.: "Where a work is completed by an employee within the scope of employment, such employee is the author of the work; provided, where an agreement stipulates that the employer is the author, such agreement shall govern. Where the employee is the author of a work pursuant to the provisions of the preceding paragraph, the economic rights to such work shall be enjoyed by the employer; provided, where an agreement stipulates that the economic rights shall be enjoyed by the employee, such agreement shall govern. The term "employee" in the preceding two paragraphs includes civil servants."

¹³Article 12 of the Copyright Act of the R.O.C.: "Where a work is completed by a person under commission, except in the circumstances set out in the preceding article, such commissioned person is the author of the work; provided, where an agreement stipulates that the commissioning party is the author, such agreement shall govern. Where the commissioned person is the author pursuant to the provisions of the preceding paragraph, enjoyment of the economic rights to such work shall be assigned through contractual stipulation to either the commissioning party or the commissioned person. Where no stipulation regarding the enjoyment of economic rights has been made, the economic rights shall be enjoyed by the commissioned person. Where the economic rights are enjoyed by the commissioned person pursuant to the provisions of the preceding paragraph, the commissioning party may exploit the work."

¹⁴To corpora that collect videos filmed based on scripts, such as a sign language corpus, the same vocabulary might be recorded numerous times before a suitable part is edited.

¹⁵Free dialogue corpora like the IFA Dialog Video corpus of Holland although avoid saying sensitive content, unsuitable content still need to be deleted before the corpus can be released.

format¹⁶, as well as files with lower standards¹⁷ so they can enter the text conversion stage. Paper data can be scanned first, and then converted by text recognition software into a digital file, or entered manually. For important collections, it is recommended to preserve scanned images along with the physical collection, so that users can link to the original image file when conducting searches.¹⁸

Text files need to be proofread twice, the first time is to check whether or not contents are consistent with the original; the second time is to check whether or not markup is correct, e.g. the article name, page. In practice, the first proofreading can be conducted by a computer program after two personnel input or convert a document; two people are required here because it reduces the chances of a mistake being made, and allows any mistake to be rapidly found and corrected.

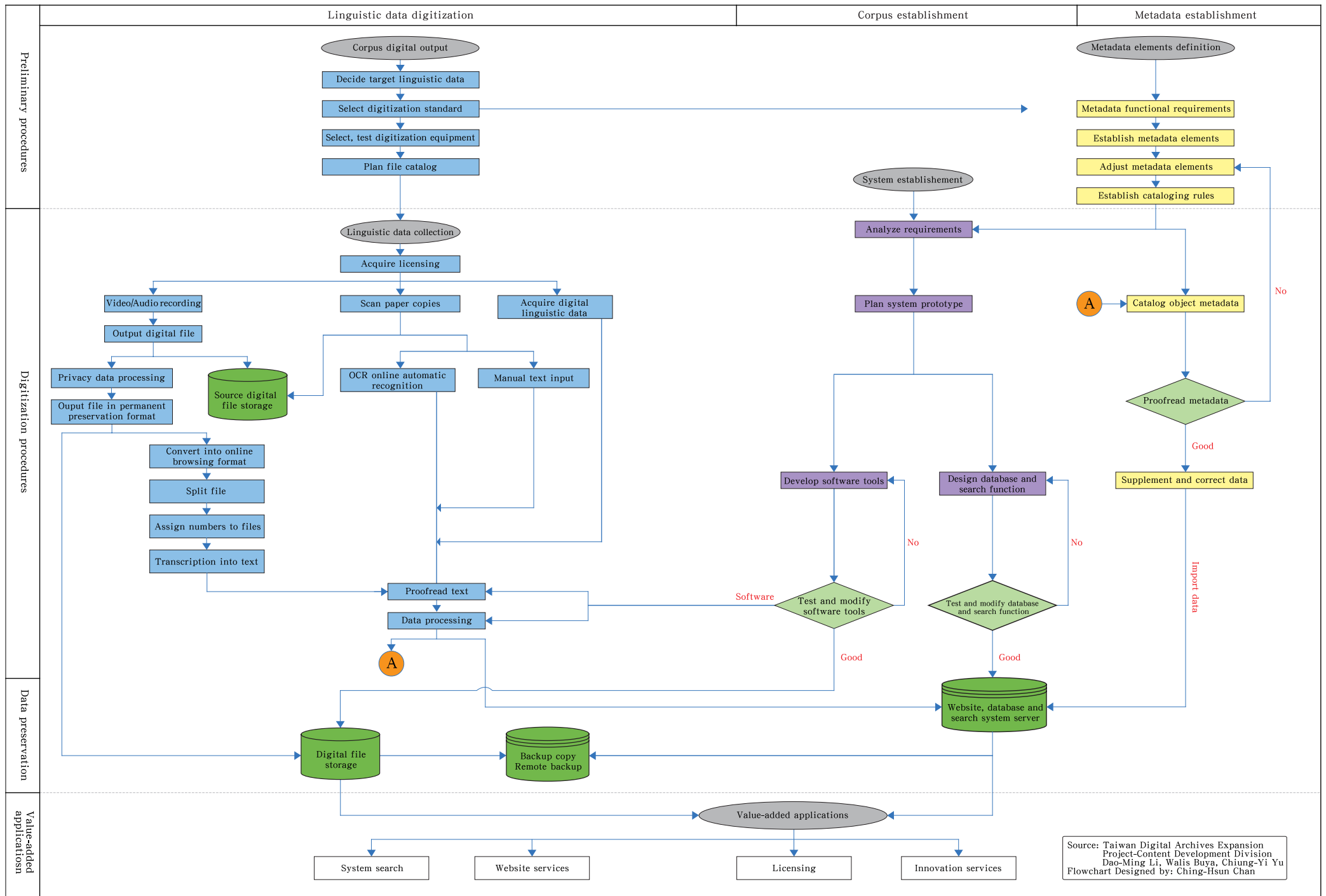
Files that have been proofread can begin the tagging process. Corpora for different purposes use different tag sets; for example, a speech corpus might use a rhythm tag set, a dialogue corpus might use a speech tag set, a typical corpus might use segmentation and part-of-speech...etc. Manually tagging linguistic data is not only time consuming and laborious, but also hard to maintain consistent quality; an interface for automatic processing and manual proofreading usually needs to be designed and developed.

Corpus system establishment includes establishing a database, designing a search system, developing tools and a working interface, and constructing a website. These tasks should be initiated during early stages of corpus planning, and completed in coordination of linguistic data processing. Finally, after a corpus is built, administration and maintenance is extremely important after it goes online. Therefore, file backup, remote backup and system security issues should also be planned and implemented.

¹⁶Permanent preservation format is the highest acceptable specification after considering current technology, compatibility, processing speed, storage space and cost. TELDAP currently recommends MPEG-2 with 8M/sec data rate for the permanent preservation of video files; WAVE with 44.1KHz sampling rate, 16-24bit for audio files.

¹⁷Public formats are usually compressed and could come in numerous versions for different bandwidths and platforms.

¹⁸Search results of the "Southern Min Archives: A Database of Historical Change and Language Distribution" (<http://southernmin.sinica.edu.tw>) "Min and Hakka Language Archives" (<http://minhakka.ling.sinica.edu.tw/>) provide links to images of the physical collection.





THREE. Samples for Building

Languages are the main content collected in corpora, but languages can have different representations, e.g. text, sign language and speech, therefore different digitization media and procedures are required.

Languages in the form of text usually appear in literature or on objects, and might involve photography, scanning, or text input for digitization. Some texts might already have electronic versions, speech sound is recorded, transcribed and then tagged, while sign language has to be expressed via images, and involves significantly different processing methods.

This chapter collects examples of text, speech, video and language distribution corpora, hoping to help users understanding building procedures for various corpora by sharing experiences of other projects.

Text corpus building examples include “Academia Sinica Balanced Corpus of Modern Chinese” and “Chinese WordNet”. The former introduces the world’s first balanced Chinese corpus with complete part-of-speech tagging; the later introduces a search system for WordNet, Hantology, and statistical word frequency distribution corpora.

For speech corpus building examples, “Linguistic GIS” not only consists of lexical field investigations, but also uses GIS to study language distribution; “Taiwan Child Language Corpus” makes recording on a regular basis and traces child language development; “A Socio-phonetic Study of Spoken Taiwan Mandarin Project” records speech via street intercept interviews.

In video corpora, “Taiwan Sign Language Online Dictionary” is a website that uses videos to teach, query and interpret sign language; different procedures are used for content establishment compared with text and speech corpora. Some speech corpora in Taiwan and overseas also collect videos, recording gestures and facial expressions along with sounds. The “Archives and Linguistic Representations of Spoken Taiwan Mandarin” of Associate Research Fellow Shu-Chuan Tseng of Academia Sinica Institute of History and Philology is one such corpus. However, due to limited time, this book wasn’t able to include building procedures of the project; users who are interested can visit the project website for themselves.¹⁹ In addition, typical audiovisual speech corpora normally don’t release image files due to authorization and

¹⁹ Archives and Linguistic Representations of Spoken Taiwan Mandarin: <http://mmc.sinica.edu.tw/>

privacy issues, and only release tag files; however, IFA Dialog Video Corpus²⁰ strives to establish public audiovisual linguistic resources, and releases all videos, audio files, tag files and related documents.

We hope that these examples will allow users to more clearly understand building procedures for different types of corpora, and be used as reference by institutions and projects that intend to build a corpus.

I. Digitization Procedures of Academia Sinica Balanced Corpus of Modern Chinese

Date of Creation: 2005/10/13

Update date: 2010/01/25

Corpus-based research has been a major development in linguistics and computational linguistics in recent years, and its influence has even extended its reach into computational studies on literature and sociology. In terms of studies on theoretical linguistics or natural language processing, the function of a corpus is to find a representative sample in a countless number of language facts. The sample can't be too large, or it will lose the meaning and advantages of sampling; it can't be too small, or it won't provide sufficient information, and won't provide adequate materials for statistical studies or testing. Therefore, the first issue of building a corpus is: How do we use limited linguistic data to represent an entire complex modern language?²¹

“Academia Sinica Balanced Corpus of Modern Chinese”, simplified as “Sinica Corpus”, is the world's first part-of-speech tagging balanced Chinese corpus. It was developed by the “Chinese Knowledge and Information Processing Group (CKIP)” of Academia Sinica, which was led by Research Fellow Ke-Chien Chen of Academia Sinica Institute of Information Science and Research Fellow Huang Chi-Ren of Academia Sinica Institute of History and Philology. The group began collecting linguistic data for the corpus in 1990,²² and had collected nearly 20 million words of modern Chinese text

and over 5 million words of ancient Chinese text.²³ Using their experience in processing Chinese text and electronic dictionary terms, the group in 1994 received a project from Academia Sinica for a cross-institute research on “Chinese information”, as well as subsidies from the National Science Council; the group thus began constructing Sinica Corpus. In order to give consideration to both ideal and practicality, the initial goal was two million words, which is twice the scale of typical small scale, and the final goal was five million words. The corpus was released in 1996, and by 1997, Sinica Corpus 3.0 had already reached the target scale of five million words. When the National Digital Archives Program was initiated in 2001, the group believed that they should continue to collect linguistic data on modern Chinese, so that corpus samples could fully represent Chinese used in Taiwan in the 21st century; the new goal was to collect five million additional words. Sinica Corpus is currently at version 4.0, and offers even more complete search functions.

Digitization Procedures:

The project's digitization procedures can roughly be divided into the following six stages: 1. Part-of-speech analysis, definition and verification; 2. Source selection; 3. Electronic linguistic data retrieval; 4. Automatic segmentation and part-of-speech tagging; 5. Manual part-of-speech inspection; 6. Import into the corpus; each stage is briefly described below:

1. Part-of-speech analysis, definition and verification:

The segmentation standard was established from two aspects, the general direction for establishing a segmentation standard was discussed by renowned scholars and specialists in Taiwan from their professional perspectives, while the Chinese Knowledge and Information Processing Group sorted out detailed segmentation rules by analyzing millions of words of linguistic data. In 1998, CKIP held a public hearing on the segmentation standard, and in 1999 the Chinese segmentation standard formally passed national standards; it was numbered CNS14366.²⁴ The group then engaged in part-of-speech analysis, definition and verification based on the segmentation standard.

²⁰ van Son, R., Wesseling, W., Sanders, E., and van den Heuvel, H. (2009). "Promoting free Dialog Video Corpora: The IFADV Corpus Example," in M. Kipp et al. (Eds.): *Multimodal Corpora*, LNAI 5509, pp. 18–37, 2009. Corpus website: <http://www.fon.hum.uva.nl/IFA-SpokenLanguageCorpora/IFADVcorpus/>

²¹ Chinese Knowledge Information Processing Group, 1995, “Contents and Description of Sinica Corpus”, Chinese Knowledge Information Processing Group Technical Report#95-02, Nangang, Academia Sinica.

²² Chu-Ren Huang and Keh-jiann Chen. 1992. A Chinese Corpus for Linguistics Research. In the Proceedings of the 1992 International Conference on Computational Linguistics (COLING-92). 1214-1217. Nantes, France.

²³ Chu-Ren Huang, 1994. Corpus-based Studies of Mandarin Chinese: Foundational Issues and Preliminary Results. In Matthew Chen and Ovid Tzeng Eds. In Honor of William S-Y. Wang: *Interdisciplinary Studies on Language and Language Change*. Pp. 165-186. Taipei: Pyramid.

²⁴ “A Segmentation Standard for Chinese Information Processing” National Standard, File No.CNS14366.

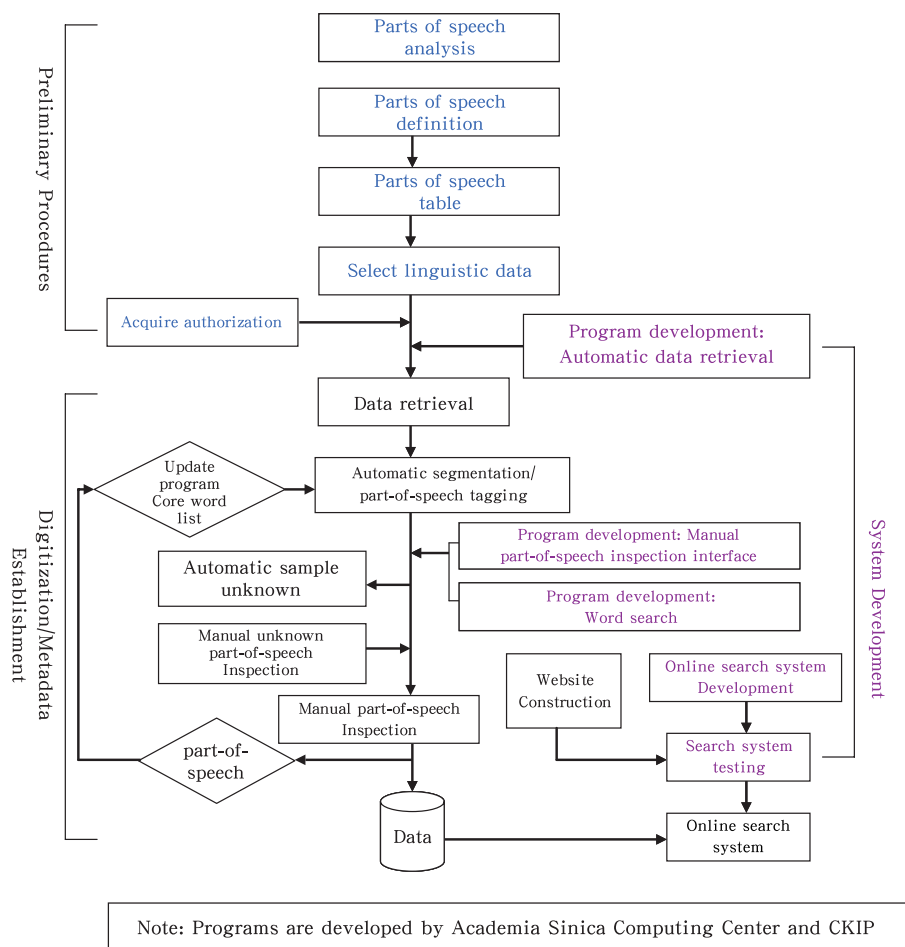


Fig 3-1-1 Workflow of the Balanced Corpus of Modern Chinese
Flowchart provided by Miss Chiu-Jung Lu, Academia Sinica Institute of History and Philology

“A Segmentation Standard for Chinese Information Processing” made the following two breakthroughs: (1) Proposes three levels of segmentation standard: faithful, truthful and graceful. In which “faithful” is the easiest to achieve and is set as the basic data exchange standard. The “truthful” level uses an automatic segmentation program to achieve natural language processing, e.g. automatic translation and data search. Graceful is the ultimate goal of computer

processing and understanding Chinese, but requires manual operation. (2) Separates the segmentation standard into a fixed core (definition of a segmentation unit and segmentation principles) and segmentation guidelines (auxiliary principles). After verifying the structure of the segmentation standard, the segmentation standard can be maintained by regularly updating the basic vocabulary or specialized glossary.²⁵

2. Source Selection

Linguistic data collected by CKIP (nearly 20 million words of modern Chinese text) was given priority for sampling, but different linguistic data of different literary styles were also acquired from a variety of channels. The different types of sources are listed below:

- (1) Linguistic data acquired from exchanges: Includes data exchanged from cooperative projects, e.g. China Times, Hong’s Foundation for Education and Culture, and National Taiwan Normal University Mandarin Training Center. Or exchanges within the Association for Computational Linguistics and Chinese Language Processing, e.g. with Behavior Design Corp. and National Taiwan University.
- (2) Directly acquired from the copyright owner: Organizations that have generously provided linguistic data that they own for academic research include: Commonwealth Magazine, Mandarin Daily News, Infopro Magazine, producer of “女人女人”, producer of “伴我成長”, producer of “我們一家都是人”, and many other units in Academia Sinica. In addition, Professor Yong-E Bi from San Francisco State University, Professor Sai-Hua Kuo from NTHU, Professor Mei-Chun Liu from NCTU, and Professor Cheng-Shu Yang from Fu Jen Catholic University also provided speech data that they transcribed.
- (3) Public data acquired from public spaces: Most data are from United Daily Network, China Times, BBS and Yam.

3. Electronic linguistic data retrieval

The computer program CKIP Corpus&Spider1.4.6a is used to retrieve online electronic linguistic data. When assistants use the program to retrieve

²⁵ A Segmentation Standard for Chinese Information Processing: Design Criteria and Content

electronic linguistic data, they must first select the source of the linguistic data, and then choose the article that is to be imported into the corpus. Articles need to be re-categorized before import because the source might have different categories for articles; the six categories in the balanced corpus are: literature, philosophy, art, science, sociology and life. (Fig 3-1-2~Fig 3-1-6)

The balanced corpus's content breakdown by theme is currently: literature 20%, philosophy 10%, art 5%, science 10%, sociology 35% and life 20%; linguistic data selections is based on these percentages.



Fig 3-1-2 Window for extracting digital linguistic data (Demonstration by Chih-Ming Chiu)

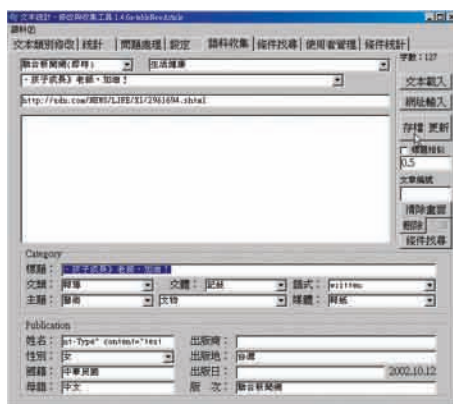


Fig 3-1-3 Linguistic data collection window

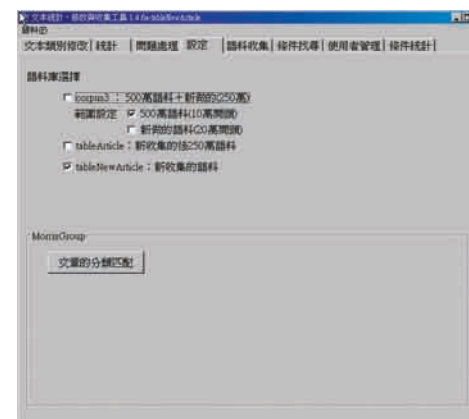


Fig 3-1-4 Verify the location of the corpus for import



Fig 3-1-5 Corpus text collection status



Fig 3-1-6 Data modification and categorization

4. Automatic segmentation and part-of-speech tagging

After selecting linguistic data, the next step is to tag parts of speech, but before doing so, the linguistic data has to be segmented. Part-of-speech tagging can only commence after each word is clearly segmented. The current accuracy of automatic segmentation done by machine is roughly 95%.

Basically, automatic segmentation procedures are based on the 80 thousand words in Academia Sinica's dictionary. Parts that are not listed in the dictionary are considered as single character words. Morphological rules are

then applied to strongly derivational affixes and numbers to combine words. The current segmentation standard adopted is “A Segmentation Standard for Chinese Information Processing”, which was formulated by the Association for Computational Linguistics and Chinese Language Processing under commission of the Bureau of Standards, Metrology and Inspections.

Automatic segmentation is completed by CKIP Tag Tool V1.8a. The program is a tool for assisting part-of-speech tag inspections; after entering the number of the text and executing automatic segmentation, the program will display the segmented text in the lower column (Fig 3-1-7, Fig 3-1-8).

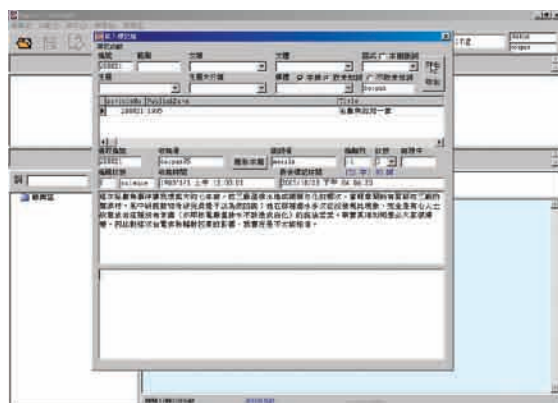


Fig 3-1-7 Select linguistic data for automatic segmentation

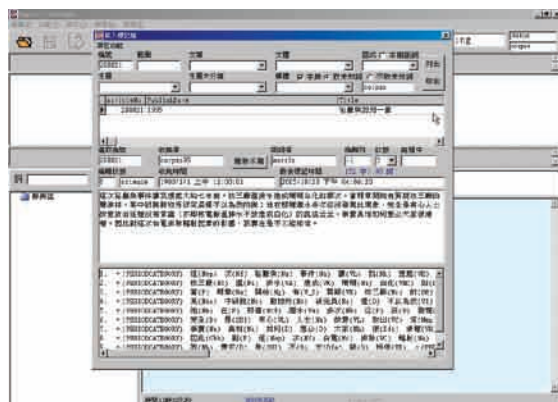


Fig 3-1-8 Automatic segmentation completed

5. Manual part-of-speech inspection

After executing automatic segmentation and part-of-speech tagging, a manual part-of-speech inspection is carried out to avoid segmentation being inconsistent with textual meaning.

The Chinese segmentation editing interface is used for manual verification. The system displays sentences before and after the sentence being inspected for reference. After verifying that segmentation for the sentence is correct, the up and down buttons can be used to continue the inspection (3-1-9).

If segmentation is found to be inappropriate, words can be modified by clicking on them (Fig 3-1-10~Fig 3-1-12).

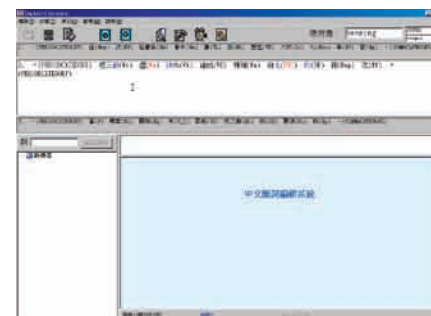


Fig 3-1-9 Chinese segmentation editing system

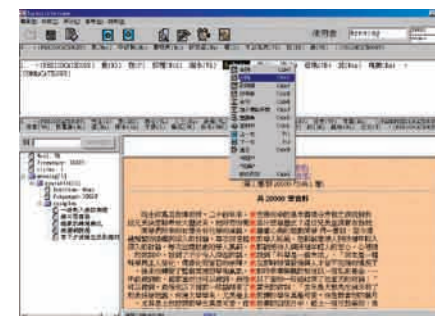


Fig 3-1-10 Segmentation Modification

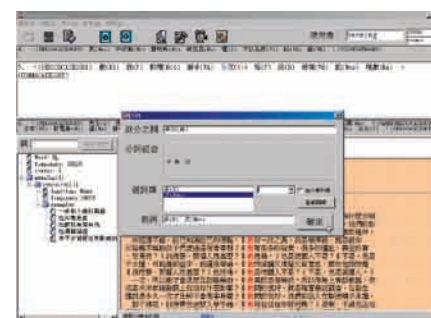


Fig 3-1-11 Input the word to correct and segmentation method

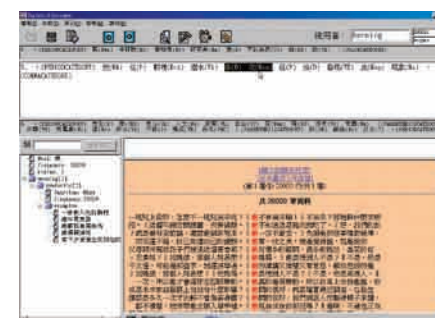


Fig 3-1-12 Segmentation correction completed

6. Import into the corpus

After completing manual part-of-speech inspection, the linguistic data is transferred to Academia Sinica Computing Center, and then imported into the Balanced Corpus of Modern Chinese.



Fig 3-1-13 Balanced Corpus of Modern Chinese



Fig 3-1-14 Search page of Sinica Corpus

Production: Content Development Division, National Digital Archives Program; Language Archives Project, Academia Sinica Institute of History and Philology
Written by: Philology Thematic Group Assistant Chia-Min Lai, Content Development Division, National Digital Archives Program; Language Archives Project Assistants Chiu-Jung Lu and Chih-Ming Chiu, Academia Sinica Institute of History and Philology

Images photographed by: Philology Thematic Group Assistant Chia-Min Lai and Shu-Hui Lin, Content Development Division, National Digital Archives Program
Images edited by: Philology Thematic Group Assistant Chia-Min Lai and Hsiu-Hua Chen, Content Development Division, National Digital Archives Program
Special thanks to: Academician Chin-Chuan Cheng, the director of the Language Archives Project of Academia Sinica Institute of Linguistics, joint director Chi-Jen Huang, and Ke-Chien Chen, Chiu-Jung Lu and Chih-Ming Chiu for their advice, assistance with photography, and providing data.

II. Lexical Knowledgebase System Establishment Procedures

Date of Creation: 2010/01/25

Chinese WordNet was developed by the Chinese WordNet Group of Academia Sinica Institute of History and Philology by combining detailed analysis of Chinese words and meanings with network technology, benefiting studies on Chinese words and meanings.

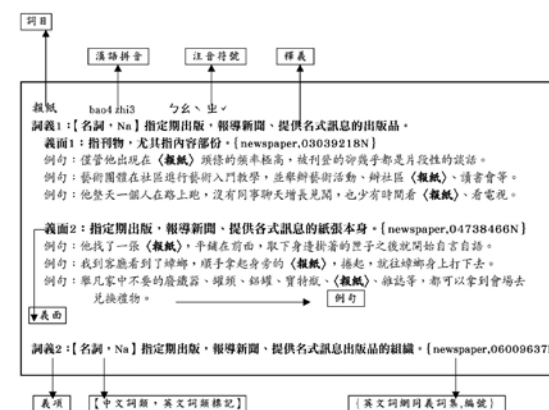


Fig 3-2-1 An example of contents of a Chinese word

We attempt to use word meanings analyzed by the Chinese WordNet Group as data, and utilize system structural analysis and design methods to construct a Chinese WordNet that meets requirements. In practical applications, a WordNet is the foundation of language processing research. Therefore, this corpus can be expected to become the base of Chinese processing and knowledge engineering.

Typical full text search systems can only search for words contained in target documents, and cannot search for word meanings or peripheral information; this type of search function apparently can't satisfy linguistic research requirements. From past studies we can see that linguistic research gives consideration to different types of word information. After analyzing studies on Chinese vocabulary, we listed the following types of information for Chinese words: headword, field of knowledge, explanation, semantic relation, English translation, sentence, and remarks.

Thorough analysis of word information not only allows word knowledge to be systematically preserved, but also satisfies the diverse requirements of linguistic research.

From the beginning of 2003 to the end of June 2009, research results of the Chinese WordNet Group had accumulated to 9,362 morphologies and 25,173 word meanings. For developing the Lexical Knowledgebase, we used the results above and followed structural system analysis and design methods to construct a platform on the internet, providing it for researchers to use. Besides sharing research results, we also hope that the system can be used as basis for Chinese Wordnet research.

System analysis and design of the Lexical Knowledgebase can be discussed in terms of two topics, functional model analysis and data model analysis.

1. Functional Model Analysis:

The Chinese Word Knowledge Search System provides a convenient environment on the internet for users to search for knowledge on Chinese words. The system consists of a database for storing data on Chinese words and an online user interface. Main functional models of the online user interface are divided into word search and word index. The word search function allows users to key in words to search for related knowledge. With consideration to word search flexibility, the system also includes a fuzzy search model. When users input key words but can't find an exact match, the system calls the adjustment mechanism to convert search criteria so that a match is found. This greatly increases search flexibility and provides users with a more convenient environment. Next, the system provides 44 accurate parts of speech categories for users to choose from; the system's dataflow is as shown in the data flowchart below (Fig 3-2-2):

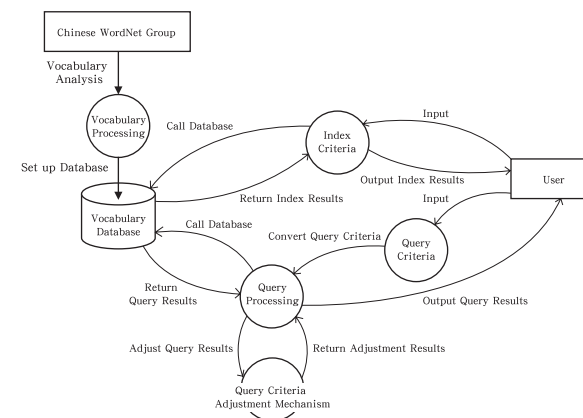


Fig 3-2-2 Data flowchart of the Lexical Knowledgebase

2. Data Model Analysis:

The word database includes 9,362 accurate analyses of Chinese words that can be presented via different search criteria. Fig 3-2-3 shows the entity relation model of the search function; the figure shows three entity types: search criteria, vocabulary and adjust criteria, and one relation: process query. From the entity relation diagram we know that query criteria has a many-to-many relationship with vocabulary, indicating that multiple entries can be searched for at the same time. Each word data entry can be accessed by multiple query criteria; therefore, the two entities search criteria and vocabulary is in a many-to-many relationship. However, when the query criteria cannot find an exact match, the system automatically adjusts the criteria, therefore the diagram also shows attributes that can be adjusted from fuzzy search.

The design of the Lexical Knowledgebase is as shown below:

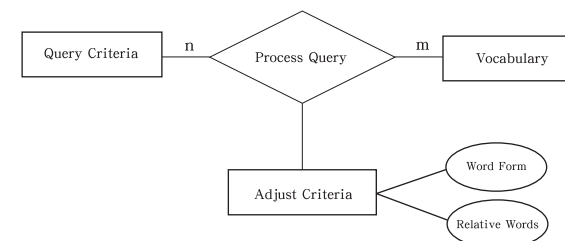


Fig 3-2-3 Entity relation diagram of the Lexical Knowledgebase

The operating environment for developing the Lexical Knowledgebase is Microsoft Windows Server, Microsoft Access and Active Server Page, which have the advantage of being easy-to-use and high compatibility, making the entire system development process smoother, while giving consideration to the feasibility of future maintenance.

During the design stage, the Lexical Knowledgebase's system architecture and operation procedures were established after considering user perspective and system function development; data structure and operation procedures within the system's scope are described in detail, especially the design of an integrated real-time search method, which provides system users with an integrated search interface for rapid searching and browsing information they are interested in, as shown in Fig 3-2-4. The scope of search allowed by the system includes: Chinese words, definition, English translation, fuzzy search, and phonetic systems to describe Chinese; users can choose a query method according to the type of information or their requirements. The main purpose is to achieve an effective extensive search of words and meanings, and to ensure the consistency and strength of explanations and semantic differential via comparison of search results. In terms of results presentation, data is arranged in the order: headword, meaning, field, explanation, semantic relation, English translation, sentences and notes.



Fig 3-2-4 Lexical Knowledgebase Query Interface



Fig 3-2-5 Search Results

The Lexical Knowledgebase developed according to structural system analysis and design methods described above can be effectively used as a source of lexical knowledge for related studies. Using “比得上” as an example, if “Chinese vocabulary” is selected as the scope of search, then search results will be as shown in Fig 3-2-5; Information on the word will be displayed on the interface. Likewise, if “比得上” is used as the key word for a search in “explanations”, any explanation of a word that has “比得上” will be displayed on the interface. The same process applies to other search functions.

In Fig 3-2-5, search results for “比得上” include phonetic symbols, explanation, semantic relation, English translation, and sentences, in which “semantic relation” not only indicates the relationship between two words, e.g. synonyms, and extended links to the individual character “比”; in “English translation” there is a link to SinicaBOW (<http://bow.sinica.edu.tw/>) access information on “比得上 (Compare)”. Furthermore, the sentences shown on the interface are from Sinica Corpus, and best represent word usage.

In the search process, if an exact match cannot be found for a key word, the system will automatically switch to fuzzy search, and on the results page it will display “The meaning of the word has not been analyzed, fuzzy search results are as shown below for your reference”, reminding users to browse through related word data. An example of a search on “一點鐘” is as shown in Fig 3-2-6.



Fig 3-2-6 Fuzzy Search Results

After completing system development, our attention shifted the update and management of content, which is required for sustainable operation and the continued provision of detailed lexical knowledge, and we developed a mechanism for regular data access between the system and data sources, for user interaction and exchanges, and for update and synchronization. Studies on natural language processing often need to conduct in-depth discussions on lexical semantics; thoroughly analyzed data is not only valuable in that it contains knowledge, but also because it can be provided for related research. This study uses lexical analysis research results that the Chinese WordNet Group accumulated in recent years as a basis, builds 9,362 entries of Synset data on to the Chinese WordNet, and displays an integrated user friendly interface via the internet. Besides allowing researchers and users who interested to conduct searches, the system hopes to be used as a basis for advanced research on lexical knowledge, and further achieve the purpose of research results sharing and academic exchanges.

Aside from the Chinese WordNet described above, the Chinese WordNet Group in recent years has also been devoted to studies on different wordnets, and set up databases that can convert knowledge from different backgrounds into interoperable information. The databases that the group has constructed are as follows:

(1) Chinese Wordnet:

Chinese WordNet is a lexical knowledgebase that provides complete Chinese sense analysis. Entries are limited to the Modern Chinese lexicon, and complete information is provided on headwords. This lexical knowledgebase can be used as basic data for studies on word sense theories and cognition, and can become a basic framework for practical applications in Chinese language processing and knowledge engineering.

(Website: <http://cwn.ling.sinica.edu.tw/> open access)

(For internal use: <http://140.109.150.20/>)

(2) Sinica BOW:

This is the world's first ontological wordnet and is based on the Princeton WordNet and language usage experiences in Taiwan. Information that is provided includes English-Chinese bilingual information conversion, links between linguistic information and

concept structure (ontology), sense distinction and semantic relation links and domain; for using language and word data, the system provides an infrastructure for knowledge logistics, allowing content from different sources to be converted into interoperable information.

(Website: <http://bow.sinica.edu.tw/> open access)

(3) Chinese Word Sketch:

This is a syntactic knowledge generation system that is combined with a large number of corpora. Besides typical key word and language environment searches, this system also provides automatically generated syntactic knowledge, such as word sketches, syntactic relation, and synonym analysis. After "Chinese Word Sketch" was combined with the LDC Chinese Gigaword, which has 1.4 billion words, it provides descriptions of most actual usage patterns of Chinese, and can be applied to dictionary compilation, Chinese language teaching, linguistic research and natural language processing.

(Website: <http://bow.sinica.edu.tw/> open access; account required)

(4) Hantology:

The Chinese character writing system crosses over three millenniums. Its knowledge representation system is the most stable; it is an ontology that represents the richest knowledge. Knowledge represented by Hantology includes structure and form of Chinese characters, signifier, phonetic element, ancient sound, medieval sound, modern sound, meaning, character variants, and derivatives, including form, sound, and meaning changes and relations during different periods; construction of Hantology achieves systematic representation of knowledge on Chinese characters.

(Website: <http://hantology.sinica.edu.tw/> not open for access)

(5) Taiwan Presidential Corpus:

This corpus collects speeches given by the four presidents between 1955 and 2007 on Taiwan's National Day and New Years. Segmentation and part-of-speech tagging was completed for every sentence in the corpus. This corpus was specially designed for Chinese political language analysis, and is a representative sample of Taiwan's political language.

(Website: <http://140.109.19.114/president/>, currently only accessible by IP addresses within Academia Sinica)

(6) Chinese word frequency distribution system:

This system is based on Chinese GigaWord, allows queries on word frequency differences, can be used to discuss word usage, and can be used as a research tool for comparing word usage across the Taiwan Strait and other Chinese speaking areas.

(Website: <http://140.109.150.156/sinica/cwordfreq/>, currently only accessible by IP addresses within Academia Sinica)

Written by: Chinese WordNet Group, Academia Sinica Institute of History and Philology

III. Procedures for Building the Language Distribution GIS

Date of Creation: 2005/12/05

Update date: 2010/01/25

This section introduces procedures for building the Language Distribution GIS of the “Southern Min Archives – A Database of Historical Change and Language Distribution”, which was a phase one sub-project of the Language Archives Project of Academia Sinica Institute of History and Philology. Min and Hakka are major dialects of the Han language, and important language assets that are distributed through Southern Fukien, Guangdong, Taiwan and Southeast Asia. However, due to the influence of school education and mass media using Mandarin, fewer and fewer people are able to speak Min and Hakka, making them relatively minor languages that require research and preservation.

Taiwan has a large and frequent population flow, and growingly frequent language contact; the language ecology has undergone great change, and “regional variation” and “social variation” of dialects are rapidly changing. Only in recent years did scholars begin to actively investigate and draw a map of languages spoken in Taiwan. However, drawing electronic language maps are still in preliminary stages, and maps that show language distribution and changes are still rare.

This project is one of the five subprojects under the “Language Archives – Chinese Branch Project” of Academia Sinica, and targets plays and songbooks of popular literature for building a Min and Hakka corpus. The project investigates Hsinfeng Township, Hsinchu County, where the population speaks Min and Hakka, and studies how the language spoken by residents is affected by the other language. From the perspective of historical language and language distribution, the project combines literature language with spoken language, and establishes a Min and Hakka corpus and language distribution GIS, providing powerful research tools for academia.

Hsinchu County Hsinfeng Township is the home of Min and Hakka people, so the project conducted a study on language distribution in the area to develop the language distribution GIS.

Digitization Procedures:

The project's digitization procedures and roughly be divided into the following six stages: 1. Interview and audio recording; 2. Road positioning

and digital map; 3. Questionnaire input; 4. Audio recording backup; 5. Layer production; 6. Import into database. Brief descriptions are as follows:

1. Interview and audio recording

We conducted a survey study on language distribution with Hsinchu County Hsinfeng Township as the target area; surveys are conducted in the form of interviews, language used by interviewees in daily life was recorded; picture cards and word cards were shown to interviewees and their speech was recorded. However, it was hard to persuade residents to let us interview them, so we visited the head of the community first, and asked the head of the community to accompany us to gain the trust of residents (Fig 3-3-1~Fig 3-3-4).



Fig 3-3-1 Interview and audio recording (demonstration by Chi-Jung Lu, Yueh-Hsia Cheng and Yu-Tsen Kuo)



Fig 3-3-2 Examples of picture cards used for interviews (photographed by Su-Ying Hsiao)

寫字	關門
查某	迫迫
血	桃園

Fig 3-3-3 Examples of Min word cards used for interviews

朋友	講話
胃	肥
時	四

Fig 3-3-4 Examples of Hakka word cards used for interviews

The questionnaire asks differences in how the subject would speak to different family members, e.g. differences when speaking with elders and juniors. The questionnaire uses one “household” as a unit for investigating language distribution, and the language most commonly used in the family as the family’s language (Fig 3-3-5 and Fig 3-3-6).

Hsinfeng Township Min and Hakka Dialect GIS Data Sheet

Time of Interview: 2005 Year ___ Month ___ Day ___ Hr ___ Min GIS:			
Address: Fl., No. , Alley , Lane , Sec. , Rd./St., Neighborhood, Village			
Name:		Date of Birth: Year Month Day Sex: <input type="checkbox"/> 1. Male; <input type="checkbox"/> 2. Female	
Birth place: <input type="checkbox"/> 1. Hsinfeng Township; <input type="checkbox"/> 2. _____			
Occupation: <input type="checkbox"/> 1. Agriculture/Forestry/Fishery/Animal Husbandry; <input type="checkbox"/> 2. Industry; <input type="checkbox"/> 3. Commercial; <input type="checkbox"/> 4. Education; <input type="checkbox"/> 5. Public; <input type="checkbox"/> 6. Other; <input type="checkbox"/> 7. None of the above (retired)			
Education Level: <input type="checkbox"/> 1. None; <input type="checkbox"/> 2. Elementary school; <input type="checkbox"/> 3. Junior high; <input type="checkbox"/> 4. High (Vocational) school; <input type="checkbox"/> 5. Junior College; <input type="checkbox"/> 6. College and above			
Ethnic group: <input type="checkbox"/> 1. Hakka; <input type="checkbox"/> 2. Min; <input type="checkbox"/> 3. Other provinces of China; <input type="checkbox"/> 4. Indigenous; <input type="checkbox"/> 5. Other _____			
1. What is the name of the place you live in?			
2. How long have you lived here?			
3. Have you lived in other places before?			
4. What languages do you speak?			
		<input type="checkbox"/> 1. Hakka	<input type="checkbox"/> 2. Min
		<input type="checkbox"/> 3. Mandarin	<input type="checkbox"/> 4. _____
Fluency: <input type="checkbox"/> Fluent			
		<input type="checkbox"/> Fluent	<input type="checkbox"/> Fluent
		<input type="checkbox"/> Moderate	<input type="checkbox"/> Moderate
		<input type="checkbox"/> Moderate	<input type="checkbox"/> Moderate
		<input type="checkbox"/> Understand	<input type="checkbox"/> Understand
		<input type="checkbox"/> Understand	<input type="checkbox"/> Understand
		<input type="checkbox"/> but can't say	<input type="checkbox"/> but can't say
		<input type="checkbox"/> but can't say	<input type="checkbox"/> but can't say
5. What language do you use when you are at work? (Mark 1, 2, 3, 4 according to frequency if the answer is multiple choice)			
		<input type="checkbox"/> 1. Hakka	<input type="checkbox"/> 2. Min
		<input type="checkbox"/> 3. Mandarin	<input type="checkbox"/> 4. _____
6. What language do you use when you are worshipping ancestors? (Mark 1, 2, 3, 4 according to frequency if the answer is multiple choice)			
		<input type="checkbox"/> 1. Hakka	<input type="checkbox"/> 2. Min
		<input type="checkbox"/> 3. Mandarin	<input type="checkbox"/> 4. _____
7. What language do you use when you are talking to elders? (Mark 1, 2, 3, 4 according to frequency if the answer is multiple choice)			
		Father: <input type="checkbox"/> 1. Hakka	<input type="checkbox"/> 2. Min
		<input type="checkbox"/> 3. Mandarin	<input type="checkbox"/> 4. _____
		Mother: <input type="checkbox"/> 1. Hakka	<input type="checkbox"/> 2. Min
		<input type="checkbox"/> 3. Mandarin	<input type="checkbox"/> 4. _____
		Grandfather (Father's side): <input type="checkbox"/> 1. Hakka	<input type="checkbox"/> 2. Min
		<input type="checkbox"/> 3. Mandarin	<input type="checkbox"/> 4. _____
		Grandmother (Father's side): <input type="checkbox"/> 1. Hakka	<input type="checkbox"/> 2. Min
		<input type="checkbox"/> 3. Mandarin	<input type="checkbox"/> 4. _____
		Grandfather (Mother's side): <input type="checkbox"/> 1. Hakka	<input type="checkbox"/> 2. Min
		<input type="checkbox"/> 3. Mandarin	<input type="checkbox"/> 4. _____
		Grandmother (Mother's side): <input type="checkbox"/> 1. Hakka	<input type="checkbox"/> 2. Min
		<input type="checkbox"/> 3. Mandarin	<input type="checkbox"/> 4. _____

Fig 3-3-5 Questionnaire sample - page one (questionnaire design: Su-Ying Hsiao)

<p>8. What language do you use when you are talking to your spouse? (Mark 1, 2, 3, 4 according to frequency if the answer is multiple choice)</p> <p><input type="checkbox"/>1. Hakka <input type="checkbox"/>2. Min <input type="checkbox"/>3. Mandarin <input type="checkbox"/>4. _____</p> <p>Your spouse is: <input type="checkbox"/>1. Hakka; <input type="checkbox"/>2. Min; <input type="checkbox"/>3. Other provinces of China; <input type="checkbox"/>4. Indigenous; <input type="checkbox"/>5. Other _____</p>
<p>9. What language do you use when you are talking to your siblings? (Mark 1, 2, 3, 4 according to frequency if the answer is multiple choice)</p> <p>Older brother: <input type="checkbox"/>1. Hakka <input type="checkbox"/>2. Min <input type="checkbox"/>3. Mandarin <input type="checkbox"/>4. _____</p> <p>Older sister: <input type="checkbox"/>1. Hakka <input type="checkbox"/>2. Min <input type="checkbox"/>3. Mandarin <input type="checkbox"/>4. _____</p> <p>Younger brother: <input type="checkbox"/>1. Hakka <input type="checkbox"/>2. Min <input type="checkbox"/>3. Mandarin <input type="checkbox"/>4. _____</p> <p>Younger sister: <input type="checkbox"/>1. Hakka <input type="checkbox"/>2. Min <input type="checkbox"/>3. Mandarin <input type="checkbox"/>4. _____</p>
<p>10. What language do you use when you are talking to your children? (Mark 1, 2, 3, 4 according to frequency if the answer is multiple choice)</p> <p><input type="checkbox"/>1. Hakka <input type="checkbox"/>2. Min <input type="checkbox"/>3. Mandarin <input type="checkbox"/>4. _____</p>
<p>11-1. What language do you use when you are talking to your daughter-in-law? (Mark 1, 2, 3, 4 according to frequency if the answer is multiple choice)</p> <p><input type="checkbox"/>1. Hakka <input type="checkbox"/>2. Min <input type="checkbox"/>3. Mandarin <input type="checkbox"/>4. _____</p> <p>Your daughter-in-law is: <input type="checkbox"/>1. Hakka; <input type="checkbox"/>2. Min; <input type="checkbox"/>3. Other provinces of China; <input type="checkbox"/>4. Indigenous; <input type="checkbox"/>5. Other _____</p>
<p>11-2. What language do you use when you are talking to your son-in-law? (Mark 1, 2, 3, 4 according to frequency if the answer is multiple choice)</p> <p><input type="checkbox"/>1. Hakka <input type="checkbox"/>2. Min <input type="checkbox"/>3. Mandarin <input type="checkbox"/>4. _____</p> <p>Your son-in-law is: <input type="checkbox"/>1. Hakka; <input type="checkbox"/>2. Min; <input type="checkbox"/>3. Other provinces of China; <input type="checkbox"/>4. Indigenous; <input type="checkbox"/>5. Other _____</p>
<p>12-1. What language do you use when you are talking to your grandchildren (son's children)? (Mark 1, 2, 3, 4 according to frequency if the answer is multiple choice)</p> <p><input type="checkbox"/>1. Hakka <input type="checkbox"/>2. Min <input type="checkbox"/>3. Mandarin <input type="checkbox"/>4. _____</p>
<p>12-2. What language do you use when you are talking to your grandchildren (daughter's children)? (Mark 1, 2, 3, 4 according to frequency if the answer is multiple choice)</p> <p><input type="checkbox"/>1. Hakka <input type="checkbox"/>2. Min <input type="checkbox"/>3. Mandarin <input type="checkbox"/>4. _____</p>
<p>Thank you! Now I would like to ask you to say some simple words out loud. We will record your voice for analysis.</p>
<p>1. Please count from 1 to 10</p> <p>2. Please read the picture card or number card</p>

Fig 3-3-6 Questionnaire sample - page two (questionnaire design: Su-Ying Hsiao)

2. Road positioning and digital map

The geographical coordinates of each interview location is required to produce the language distribution map; there are two methods for acquiring files required by the GIS.

The first method is automatic satellite positioning. We originally used GARMIN eTrex Vista, but switched to more advanced Trimble ProXH for its higher precision; the error margin of Trimble ProXH is less than one meter. X and Y coordinates are transferred via Bluetooth to a notebook computer, which is then directly added to the address by a computer program. However, satellite information is harder to acquire where buildings are densely located (Fig 3-3-7 and Fig 3-3-8).



Fig 3-3-7 Satellite positioning device (left: device currently used)



Fig 3-3-8 Positioning in the field (demonstration by Chin-Chuan Cheng, Chih-Chieh Chang; photographer: Chu-Fang Huang)

The second method is to use a vectorized aerial map to find coordinates. The task is completed by two assistants; one assistant uses ArcView to open the vectorized aerial map and finds the position of the interviewee's home on the map; X and Y coordinates are found by clicking on the position. The first assistant then reads the coordinates to the other assistant, and the other assistant inputs the information into the computer (Fig 3-3-9 and 3-3-10).

The second computer shows the Microsoft Excel file created by the household registration office that X and Y coordinates are input into. Afterwards, the Microsoft Excel file is converted into a DBF and opened with ArcView; the file is used to establish the language distribution GIS (Fig 3-3-11).



Fig 3-3-9 Verify the interviewee's doorplate address and geographical position (Demonstration by: Chien-Hui Lin and Yu-Tsen Kuo; Photographer: Chu-Fang Huang)

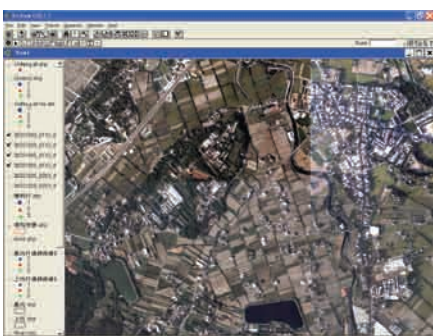


Fig 3-3-10 Aerial image opened using ArcView

門牌號碼	村名	路名	房屋座落	X	Y
170	三德興村	11 建興路二段	759	1	24792.05
171	三德興村	11 建興路二段	758	2	24791.82
227	三德興村	11 建興路二段	758	4	24791.88
230	三德興村	11 建興路二段	759	3	24793.94
279	三德興村	11 建興路二段	770	1	
710	三德興村	11 建興路二段	772	2	24793.38
711	三德興村	11 建興路二段	774	3	24792.77
712	三德興村	11 建興路二段	775	2	
713	三德興村	11 建興路二段	776	2	24793.34
714	三德興村	11 建興路二段	778	2	24793.81
715	三德興村	11 建興路二段	778	2	25094.45
716	三德興村	11 建興路二段	778	2	
717	三德興村	11 建興路二段	780	1	24793.34
718	三德興村	11 建興路二段	782	2	24793.95
719	三德興村	11 建興路二段	784	1	24794.15

Fig 3-3-11 Information keyed into a Microsoft Excel file

3. Questionnaire input

Questionnaire results are input into the computer, so that users can understand the language distribution in Hsinchu County Hsinfeng Township by clicking on entries in the Language Distribution GIS (Fig 3-3-12 and Fig 3-3-13).

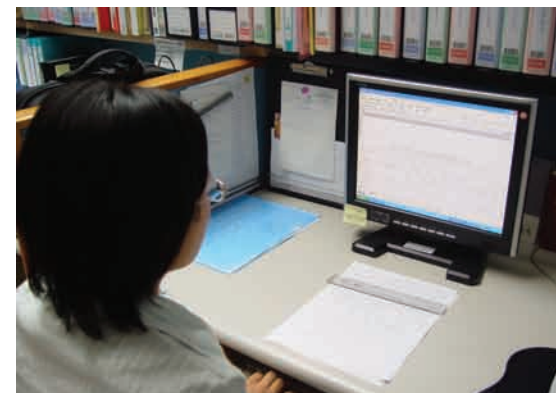


Fig 3-3-12 Work condition of questionnaire input (demonstration by: Chu-Fang Huang)

序號	戶籍村名	門牌	年齡	性別	語言別
1	三德興	311	28	男	客語
2	三德興	40	34	男	客語
3	三德興	49	34	男	客語
4	三德興	76	34	男	客語
5	三德興	82	34	男	客語
6	三德興	87	34	男	客語
7	三德興	88	34	男	客語
8	三德興	89	34	男	客語
9	三德興	90	34	男	客語
10	三德興	91	34	男	客語
11	三德興	92	34	男	客語
12	三德興	93	34	男	客語
13	三德興	94	34	男	客語
14	三德興	95	34	男	客語
15	三德興	96	34	男	客語
16	三德興	97	34	男	客語
17	三德興	98	34	男	客語
18	三德興	99	34	男	客語
19	三德興	100	34	男	客語
20	三德興	101	34	男	客語
21	三德興	102	34	男	客語
22	三德興	103	34	男	客語
23	三德興	104	34	男	客語
24	三德興	105	34	男	客語
25	三德興	106	34	男	客語
26	三德興	107	34	男	客語
27	三德興	108	34	男	客語
28	三德興	109	34	男	客語
29	三德興	110	34	男	客語
30	三德興	111	34	男	客語
31	三德興	112	34	男	客語
32	三德興	113	34	男	客語
33	三德興	114	34	男	客語
34	三德興	115	34	男	客語
35	三德興	116	34	男	客語
36	三德興	117	34	男	客語
37	三德興	118	34	男	客語
38	三德興	119	34	男	客語
39	三德興	120	34	男	客語
40	三德興	121	34	男	客語
41	三德興	122	34	男	客語
42	三德興	123	34	男	客語
43	三德興	124	34	男	客語
44	三德興	125	34	男	客語
45	三德興	126	34	男	客語
46	三德興	127	34	男	客語
47	三德興	128	34	男	客語
48	三德興	129	34	男	客語
49	三德興	130	34	男	客語
50	三德興	131	34	男	客語
51	三德興	132	34	男	客語
52	三德興	133	34	男	客語
53	三德興	134	34	男	客語
54	三德興	135	34	男	客語
55	三德興	136	34	男	客語
56	三德興	137	34	男	客語
57	三德興	138	34	男	客語
58	三德興	139	34	男	客語
59	三德興	140	34	男	客語
60	三德興	141	34	男	客語
61	三德興	142	34	男	客語
62	三德興	143	34	男	客語
63	三德興	144	34	男	客語
64	三德興	145	34	男	客語
65	三德興	146	34	男	客語
66	三德興	147	34	男	客語
67	三德興	148	34	男	客語
68	三德興	149	34	男	客語
69	三德興	150	34	男	客語
70	三德興	151	34	男	客語
71	三德興	152	34	男	客語
72	三德興	153	34	男	客語
73	三德興	154	34	男	客語
74	三德興	155	34	男	客語
75	三德興	156	34	男	客語
76	三德興	157	34	男	客語
77	三德興	158	34	男	客語
78	三德興	159	34	男	客語
79	三德興	160	34	男	客語
80	三德興	161	34	男	客語
81	三德興	162	34	男	客語
82	三德興	163	34	男	客語
83	三德興	164	34	男	客語
84	三德興	165	34	男	客語
85	三德興	166	34	男	客語
86	三德興	167	34	男	客語
87	三德興	168	34	男	客語
88	三德興	169	34	男	客語
89	三德興	170	34	男	客語
90	三德興	171	34	男	客語
91	三德興	172	34	男	客語
92	三德興	173	34	男	客語
93	三德興	174	34	男	客語
94	三德興	175	34	男	客語
95	三德興	176	34	男	客語
96	三德興	177	34	男	客語
97	三德興	178	34	男	客語
98	三德興	179	34	男	客語
99	三德興	180	34	男	客語
100	三德興	181	34	男	客語

Fig 3-3-13 Questionnaire data of the Language Distribution GIS (WebGIS format)

4. Audio file backup

Backup copies of audio files recorded in the first stage are made for future conversion into different formats. The audio files will be available in the Language Distribution GIS for users to hear the unique accents of Min and Hakka in Hsinchu County Hsinfeng Township.

5. Layer production

The Microsoft Excel file of the second stage is imported into ArcView. Different layers are then produced using ArcView for research purposes.

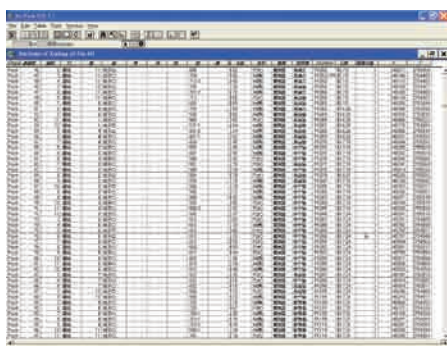


Fig 3-3-14 Import the Microsoft Excel file into ArcView (personal information omitted)

The language distribution is produced into a spotted map, different color spots are used to represent Hakka – Sze Hsien Accent, Hakka – Hai Lu Accent, Min, and other. The languages are produced into different layers along with other data, including: locations that have already been investigated, buildings, roads, rivers, land usage, boundaries and aerial images. Different layers are produced to allow users to select layers according to their requirements. Users can download ArcRead free of charge and use it to select different layers (Fig 3-3-15 and Fig 3-3-16).

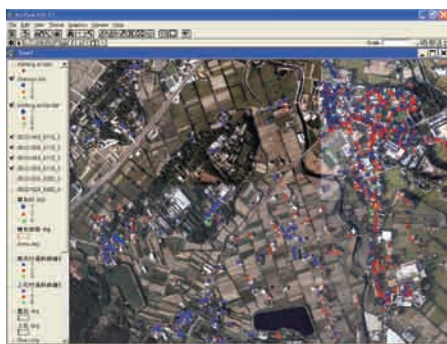


Fig 3-3-15 Layer production using ArcView
(Blue: Min; Red: Hakka – Hai Lu Accent; Green: Other)

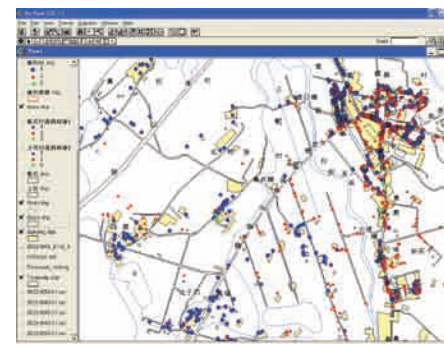


Fig 3-3-16 Language distribution displayed on a road map

Road positioning and digital maps of the second stage and Microsoft Excel files that contain questionnaire data of the third stage are handed to Academia Sinica Computing Center for layer production and file conversion. Files are converted into WebGIS format, allowing users to link to the language distribution GIS on WebGIS (Fig 3-3-17).



Fig 3-3-17 Online user interface of the Language Distribution GIS

6. Import into database

Finally, the “Southern Min Archives – A Database of Historical Change and Language Distribution” of Academia Sinica Institute of History and Philology is linked together with the WebGIS – “Language Distribution GIS” of Academia Sinica Computing center, allowing users to use the Language Distribution GIS while browsing through the “Southern Min Archives – A Database of Historical Change and Language Distribution”.

Production: Content Development Division, National Digital Archives Program; Language Archives Project, Academia Sinica Institute of History and Philology

Written by: Philology Thematic Group Assistant Chia-Min Lai, Content Development Division, National Digital Archives Program; Philology Thematic Group Assistants Chiu-Jung Lu, Chu-Fang Huang and Yu-Tsen Kuo, Academia Sinica Institute of History and Philology

Images photographed by: Philology Thematic Group Assistants Chia-Min Lai and Shu-Huei Lin, Content Development Division, National Digital Archives Program

Edited by: Philology Thematic Group Assistants Chia-Min Lai and Hsiu-Hua Chen, Content Development Division, National Digital Archives Program

Special thanks to: Academician Chin-Chuan Cheng, the director of the Language Archives Project of Academia Sinica Institute of Linguistics, assistants Chiu-Jung Lu, Chu-Fang Huang and Yu-Tsen Kuo for their advice, assistance with photography, and providing data.

IV. Digitization Procedures of Taiwan Sign Language Online Dictionary

Date of Creation: 2005/11/16

Update Date: 2009/12/03

The purpose of the “A Study of Taiwan Sign Language: Phonology, Morphology, Syntax and Digital Graphic Dictionary” Project of National Chung Cheng University Graduate Institute of Linguistics is to make a complete description and analysis of Taiwan Sign Language, which includes the compilation of a reference book on syntax that possesses both academic and practical value, and the construction of a digital graphic dictionary on the internet. The term of this National Science Council project is between August 1st, 2001 and December 31st, 2005.

Professor Hao-Yi Tai was responsible for the overall planning, execution and supervision of the project. Professors Hao-Yi Tai and Jung-Hsing Chang were responsible for the words and syntax in the reference book, and professors Jane Tsay and James Myers were responsible for the phonology and morphology portions of the book. Professor Jane Tsay was responsible for compiling the online dictionary, and the first edition of TSL Online Dictionary was published online in July 2008, in which website construction

was completed by graduate student Chia-Hsiung Lu under the supervision of Professor Tzu-Chiang Chen of the Department of Electrical Engineering. The second edition was published in September 2009, and graduate student Chiung-Yi Yu of the Graduate Institute of Linguistics was put in charge of website maintenance.

This digital graphic dictionary currently collects 3,000 words (including Chinese and English search interfaces and explanations), and will continue to expand. Procedures for compiling the digital graphic dictionary and constructing the website are described below.

Digitization Procedures:

Digitization procedures of TSL Online Dictionary can roughly be divided into the follow eight stages: 1. Collect sign language words; 2. Prepare video recording materials; 3. Video recording; 4. File conversion; 5. Video editing; 6. Text descriptions of videos; 7. Database establishment; 8. Website construction. Each stage is briefly described below.

1. Collect sign language words

We collected sign language words in the “Taipei City Department of Labor Sign Language Interpreter Training Manual Volume One”, “手能生橋” volumes one and two, and “A Reference Grammar of Taiwan Sign Language” (compiled by National Chung Cheng University Graduate Institute of Linguistics), including words and compound words, and compiled them into a list for recording the digital graphic dictionary (Fig 3-4-1).



Fig 3-4-1 Sign language lexicon reference books and images collected as linguistic data

2. Prepare video recording materials

The sign language words we collected are typed into a Microsoft Word file, translated into English to create a bilingual word list, and added with a description of the sign language gesture (Fig 3-4-2).

The word list is then converted into a Microsoft PowerPoint file. Each word is made into one page for the sign language consultant to demonstrate words one by one; white characters and black background makes it easier for the sign language consultant to recognize words during video recording (Fig 3-4-3).



Page	Chinese	English	Description
1	太太	太太	太太
2	結婚	結婚	結婚
3	女	女	女
4	太太 (結婚+女)	太太 (結婚+女)	太太 (結婚+女)

Fig 3-4-2 Word list

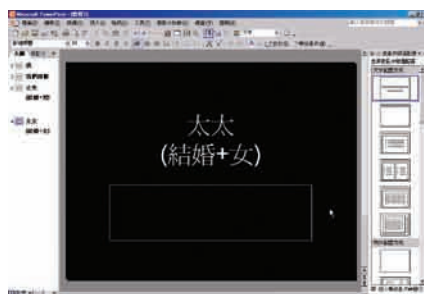


Fig 3-4-3 PPT file

3. Video recording

First, prepare the computer in which the PowerPoint file is saved, and then set up recording equipment (note: a digital video camera is used to record videos). A fixed black cloth is used as the background; the sign language consultant wears a red T-shirt for larger contrast. Before recording, the assistant discusses contents with the sign language consultant to confirm words that will be recorded (Fig 3-4-4, Fig 3-4-5).

Fig 3-4-4 Background and clothes worn by sign language consultant (Demonstration: Yu-Shan Ku)



Fig 3-4-5 Personnel discussing contents that will be recorded with the sign language consultant (Demonstration: Hsiu-Fen Su, Yu-Shan Ku)

Due the fact that each word may have many sign language gestures, the research assistant will discuss differences with the sign language consultant (for example, a specific gesture belongs to northern Taiwan dialect, southern Taiwan dialect, or synonyms). In principle, the multiple forms of a word are all recorded.

The video recording process usually requires two personnel, one sign language consultant and one sign language translator. One person plays the powerpoint file and determines the video recording progress; the other person ensures the correctness of sign language words, and controls the video recording quality. During the recording process, the sign language translator assists with communication. Due to the strict requirements on image quality and clarity, and the need to constantly check with the sign language consultant, the same word often needs to be repeatedly recorded to achieve correctness and the best image quality (Fig 3-4-6, Fig 3-4-7).



Fig 3-4-6 Filming site (Demonstration: from the left are Hsiu-Fen Su, Yu-Shan Ku and Pei-Lan Wu)



Fig 3-4-7 Sign language translator assisting with discussion between personnel and sign language consultant (Demonstration: from the left are Yueh-Hsia Hsiao, Yu-Shan Ku and Hsiu-Fen Su)

4. File conversion

We use CyberLink PowerVCR to convert the video file on the DV tape into a MPEG file, and then use Quick Time Player to convert the MPEG file into a smaller MOV file, which is stored in the computer server; this reduces the time users have to wait for videos to load (Fig 3-4-8).



Fig 3-4-8 File conversion interface

5. Video editing

We use the “editing function” of CyberLink PowerVCR to separate individual words or compound words in video files. Since time cannot be one hundred percent controlled during the recording process, sometimes images of two words are too close together and cannot be completely separated; if this occurs, then the words need to be recorded again. In addition, to achieve the best image quality, separated videos need to be repeatedly examined; the length of a video clip needs to be modified several times to gain the most suitable

length. If video clips are too short, then words cannot be completely presented; if they are too long, not only will they take up more storage, but also make users wait longer (Fig 3-4-9).



Fig 3-4-9 Video editing (Demonstration: Hsiu-Fen Su)

6. Text descriptions of videos

Each sign language word has a text description to describe the gesture. The Chinese description is also translated into English; both are listed in Microsoft Excel files (Fig 3-4-10, Fig 3-4-11).

序號	中文名稱	中文描述	英文描述
1	中文名稱	中文動作說明	
2	7 (L) AAD	一手握拳與面中偏中處，另一手虎口對準面中下下，掌心朝向內	
3	7 (L) BUD	一手握拳與面中偏中處，另一手虎口對準面中下下，掌心朝向內	
4	轉表筆	一手握拳與面中偏中處，另一手虎口對準面中下下，掌心朝向內	
5	轉表筆	一手握拳與面中偏中處，另一手虎口對準面中下下，掌心朝向內	
6	轉表筆	一手握拳與面中偏中處，另一手虎口對準面中下下，掌心朝向內	
7	打表筆	一手掌心朝上，另一手虎口對準面中下下，掌心朝向內	
8	打表筆	一手掌心朝上，另一手虎口對準面中下下，掌心朝向內	
9	打表筆	一手掌心朝上，另一手虎口對準面中下下，掌心朝向內	
10	打表筆	一手掌心朝上，另一手虎口對準面中下下，掌心朝向內	
11	打表筆	一手掌心朝上，另一手虎口對準面中下下，掌心朝向內	
12	打表筆	一手掌心朝上，另一手虎口對準面中下下，掌心朝向內	
13	打表筆	一手掌心朝上，另一手虎口對準面中下下，掌心朝向內	
14	打表筆	一手掌心朝上，另一手虎口對準面中下下，掌心朝向內	
15	打表筆	一手掌心朝上，另一手虎口對準面中下下，掌心朝向內	
16	打表筆	一手掌心朝上，另一手虎口對準面中下下，掌心朝向內	
17	打表筆	一手掌心朝上，另一手虎口對準面中下下，掌心朝向內	
18	打表筆	一手掌心朝上，另一手虎口對準面中下下，掌心朝向內	
19	打表筆	一手掌心朝上，另一手虎口對準面中下下，掌心朝向內	
20	打表筆	一手掌心朝上，另一手虎口對準面中下下，掌心朝向內	

Fig 3-4-10 Chinese description

序號	英文名稱	中文名稱	英文描述
1	ADREE BY C	害怕	Two hands touch then separate, each hand claps flat as if looking at
2	ADREE ON	害怕	The shaky finger of fist on the bottom claps onto the palm of the fist
3	ADREE-A	懼	AFRAID-THINK/THE GAME
4	ADREE-B	懼	AFRAID-SUBJECT
5	ADREE-C	懼	AFRAID-SUBJECT
6	ADREE-D	懼	AFRAID-SUBJECT
7	ADREE-E	懼	AFRAID-SUBJECT
8	ADREE-F	懼	AFRAID-SUBJECT
9	ADREE-G	懼	AFRAID-SUBJECT
10	ADREE-H	懼	AFRAID-SUBJECT
11	ADREE-I	懼	AFRAID-SUBJECT
12	ADREE-J	懼	AFRAID-SUBJECT
13	ADREE-K	懼	AFRAID-SUBJECT
14	ADREE-L	懼	AFRAID-SUBJECT
15	ADREE-M	懼	AFRAID-SUBJECT
16	ADREE-N	懼	AFRAID-SUBJECT
17	ADREE-O	懼	AFRAID-SUBJECT
18	ADREE-P	懼	AFRAID-SUBJECT
19	ADREE-Q	懼	AFRAID-SUBJECT
20	ADREE-R	懼	AFRAID-SUBJECT

Fig 3-4-11 English description

To complete the most appropriate text description, text descriptions for sign language words are compared by different staff members, and deliberated on by the project director, joint directors and research assistants when necessary. English translations are verified by joint directors or English experts.

Besides describing the hand gestures, descriptions also include the direction or position of contact with the body, swinging methods and facial expressions. Furthermore, the region where the sign language form is used (Northern Taiwan or Southern Taiwan) or its origin is also described when necessary.

7. Database establishment

The online dictionary database is managed via PhpMyAdmin, in which the list of Chinese words completed in the previous procedure is added with Hanyu Pinyin, and number of strokes of the first character. Each word consists of a video file, Chinese description, English description, and links between video and text. The database interface is as shown below (Fig 3-4-12, Fig 3-4-13).



Fig 3-4-12 Chinese Version Database



Fig 3-4-13 English Version Database

8. Website construction

All image files, text descriptions and links from each word are completed in the previous procedure. Data entries are proofread in the order of number of strokes for Chinese characters and in alphabetical order for English.

This dictionary has two interfaces, Chinese and English (Fig 3-4-14). In the Chinese interface searches can be based on the number of strokes (1~19) of the first character in a word, Hanyu pinyin, or key words; in the English interfaces searches can be based on the first letter (A~Z) or key words (Fig 3-4-15).



Fig 3-4-14 Homepage of the TSL Online Dictionary



Fig 3-4-15 Chinese version search function

Using the word “交通” as an example, the first character “交” is written in six strokes, so under the six stroke category you will find the word “交通”, click on the word to see the sign language gesture for “交通” with Chinese descriptions on the right (Fig 3-4-16).



Fig 3-4-16 Search results for “交通” in the TSL Online Dictionary

In the English interface, you can enter “TRAFFIC” (note: uppercase and lowercase letters are considered the same) to find the sign language gestures for “TRAFFIC” with English descriptions on the right (Fig 3-4-17).



Fig 3-4-17 Search results for “TRAFFIC” in the TSL Online Dictionary

The digital workflow of the TSL Online Dictionary is tightly linked together; each individual step can affect the end result. Therefore, careful consideration is required when planning digitization procedures and countless discussions must be completed to correct and improve results. Since the first edition of the TSL Online Dictionary was formally published online in July 2008, the suggestions and feedback from different sectors have given us great encouragement. We are grateful to the strenuous efforts of our researchers, and thank the NSC for its support and assistance, without such support we wouldn't have been able to yield the fruitful results of “TSL Online Dictionary.” In the future, we will continue expand the TSL Online Dictionary so that it keeps pace with the times. We hope that Taiwan Sign Language will gain international respect in the field of linguistics, and that it will provide substantial aid to teaching people with hearing impairments in Taiwan.

Producer: Content Development Division, National Digital Archives Program; Graduate Institute of Linguistics “A Study on Taiwan Sign Language: Phonology, Morphology, Syntax and Digital Graphic Dictionary,” National Chung Cheng University

Written by: Philology Thematic Group assistant Chia-Min Lai, Content Development Division, National Digital Archives Program; Graduate Institute of Linguistics Project Director Hao-Yi Tai, National Chung Cheng University; Graduate Institute of Linguistics Joint Director Jane Tsay, National Chung Cheng University; Graduate Institute of Linguistics Project Assistants Hsiu-Fen Su and Hsin-Hui Chen, National Chung Cheng University

Image photographed by: Philology Thematic Group assistants Chia-Min Lai, Shu-Huei Lin, and Hsiu-Hua Chen, Content Development Division, National Digital Archives Program

Edited by: Philology Thematic Group assistants Chia-Min Lai, Mei-Chih Chen, and Hsiu-Hua Chen, Content Development Division, National Digital Archives Program

Special thanks to Professor Hao-Yi Tai, Project Director of Graduate Institute of Linguistics “A Study on Taiwan Sign Language” National Chung Cheng University, Joint Director Jane Tsay, and Assistant Hsiu-Fen Su for their advice, assistance with photography, and providing data, and sign language translator Yueh-Hsia Hsiao.

V. Digitization Procedures of Taiwan Child Language Corpus

Date of Creation: 2005/12/14

Update Date: 2010/01/25

The Taiwan Child Language Corpus (TAICORP) of National Chung Cheng University Graduate Institute of Linguistics is a corpus that collects audio recordings of Taiwan child language and was established according to the Child Language Data Exchange System (CHILDES; MacWhinney and Snow 1985, MacWhinney 1995) format. Its main purpose is to (1) provide domestic and foreign scholars with the convenience of linguistic data sharing and a tool for linguistic data analysis; (2) make collection of Taiwan child language more systematic and efficient by setting a standard format, allowing all languages in Taiwan to be rapidly covered. The corpus will eventually be established into a website for domestic and foreign scholars to use.

Under the prevalence of Mandarin spoken by the next generation, linguistic data of child Min language acquisition in Taiwan is exceptionally precious. This corpus can be provided for different aspects of linguistic and child language acquisition research, including phonetics, phonology, morphology, syntax, semantics and pragmatics, as well as research and applications of speech engineering. This project is directed by Professor Jane Tsay of National Chung Cheng University Graduate Institute of Linguistics, and began audio recordings in October 1997, spending nearly nine years on transcription, tagging and formatting. A total of 431 person-times were recorded, total length of recordings reaching 330 hours. Text files consist of roughly 500 thousand sentences and over 1.6 million words.

Digitization Procedures:

Digitization procedures of the Taiwan Child Language Corpus can be divided into the following six procedures: 1. Audio recording. 2. Transcribe audio files into text files. 3. Corpus establishment. 4. Automated system establishment. 5. Automated system applications. 6. Website construction and maintenance. The six procedures are further divided into 23 steps, and briefly described below.

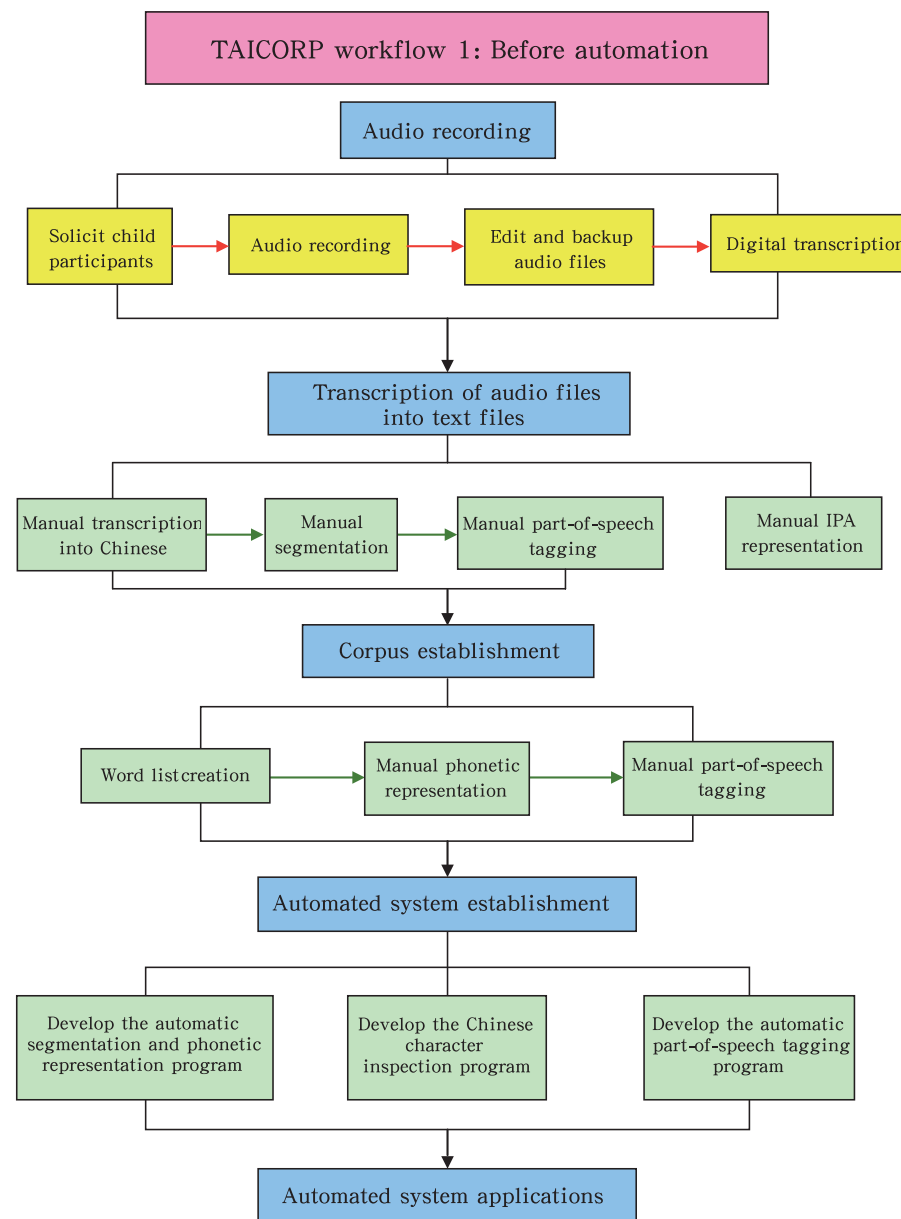


Fig 3-5-1 Workflow of the Child Language Corpus before Automation

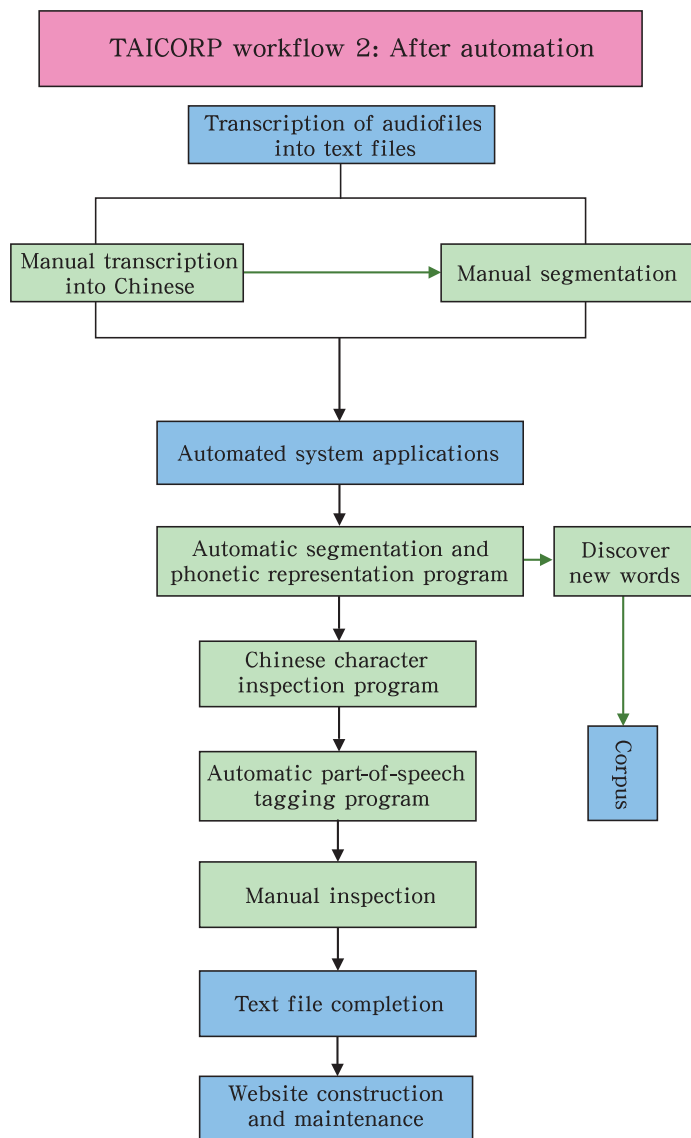


Fig 3-5-2 Workflow of the Child Language Corpus after Automation
Date of Creation: 2005/11/17

Flowcharts provided by Professor Jane Tsay of National Chung Cheng University Graduate Institute of Linguistics

1. Audio recording

“Audio recording” is divided into five steps, “research assistant training,” “solicit child participants,” “audio recording,” “edit and backup audio files,” and “digital transcription.”

- (1) Research assistant training: Research assistants are trained by the project director; there are three core research assistants. Research assistants are required to have a master’s degree in linguistics and speak Min as a native language. The project director spends three to six hours each week to train research assistants, help them understand Min phonetics and writing systems, Min lexicon, syntax, semantics, part of speech tagging, and literature on CHILDES and child language acquisition; research assistants also become familiar with IPA transcription.
- (2) Solicit children from families that speak Southern: Our subjects are infants between ages 1 and 3 from Southern Min speaking families that are in CCU’s subsidiary nursery, kindergarten and nearby townships; we selected 14 children in total.
 - i. We posted ads on posters and the internet and use parents day of our kindergarten to solicit children from families that speak Southern Min.
 - ii. Arranging audio recording time: We contacted parents to set the time schedule for audio recordings.
- (3) Audio recording
 - i. Prepare audio recording equipment: We selected convenient and portable recording equipment with large storage space that benefit the long-term preservation of linguistic data (Fig 3-5-3).



Fig 3-5-3 Audio recording equipment, from the left is mini-discs, professional earphones, professional microphone, and portable mini-disc recorder.

ii. Recording interviews: We visit children in their homes and record our interviews. The interviews are periodical and continue throughout summer and winter vacation. We visit children under the age of 2 once every week; children between 2 and 3 years old once every two weeks; and children between 3 and 4 years old once every two to three weeks. Each interview varies between 1 to 2 hours, actual recording time ranges between 40 to 60 minutes. We recorded interviews between October 1997 and May 2000, for a total of 431 person times and roughly 330 hours. Our interview method was to let children speak as they normally do when accompanied by a parent or babysitter. Besides natural language, we also used illustrated books, story books, toys, dolls, paper cutting, origami, or other games to cause children to talk.

(4) Edit and backup audio files

i. Editing audio files: The assistant deletes irrelevant sounds or silent parts and cuts the audio file into smaller segments, which are labeled on the mini-disc; dates and file names are keyed into the disc. Each hour recorded takes 1.5 hours to edit, meaning that the total editing time was: 1.5×330 hours = 495 hours (Fig 3-5-4).

ii. Audio file backup: We used the mini-disc recorder stand and mini-disc recorder for making a backup mini-disc (Fig 3-5-5).



Fig 3-5-4 Editing audio files (Demonstration: Pei-Yu Hsieh)



Fig 3-5-5 Audio file backup (Demonstration: Pei-Yu Hsieh)

(5) Digital transcription: We convert the audio file in the mini-disc into MP3 format for storage convenience. The file can be converted into other formats for speech analysis (e.g. *.wav). The transcription software we used was GoldWave Digital Audio Editor (developed by GoldWave Inc., see Fig 3-5-6).

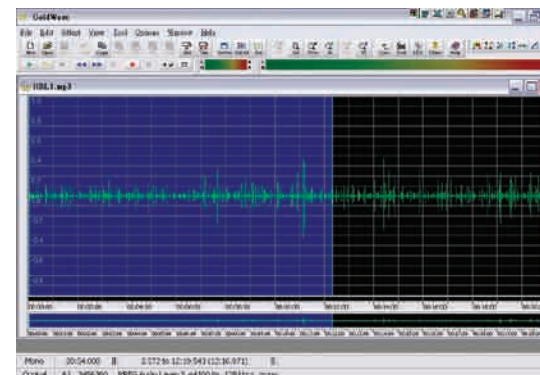


Fig 3-5-3 Audio recording equipment, from the left is mini-discs, professional earphones, professional microphone, and portable mini-disc recorder.

2. Transcribe audio file into text file

“Transcription” consists of four steps: “Manual transcription into Chinese characters,” “manual segmentation,” “manual part-of-speech tagging,” and “manual IPA representation”.

(1) Manual transcription into Chinese characters: There currently isn’t an established Chinese character writing system for Southern Min, while some original words cannot be verified and some words only have sounds but no character. Therefore, it is necessary to establish a text transcription principle. Before implementing text transcription, the first task is to establish the Southern Min writing system, the four dictionaries we used according to their priority are: “臺灣閩南語辭典 (Taiwan Southern Min Dictionary),” “台灣話大辭典 (Taiwanese Dictionary),” “廈門方言詞典 (Xiamen Dialect Dictionary),” and “閩南語詞彙 (Southern Min Dictionary),” as shown from left to right in Fig 3-5-7. The transcription platform we used was CHILDES. Every hour of audio recording requires some 10 hours to transcribe into text; total work time: $330 \text{ hours} \times 10 = 3,300$ hours.

(6) Website construction and maintenance

- i. Website structure and contents: The project director and research assistants discuss the website's contents and the interface. Website contents include an introduction, database, user manual, related programs, and links to related websites.
- ii. Website construction and maintenance: We constructed a dedicated website to provide our corpus to researchers around the world. After completing final tests, the website will be opened to the public (Fig 3-5-14).



Fig 3-5-14 Website Homepage

Producer: Content Development Division, National Digital Archives Program; Graduate Institute of Linguistic “Taiwan Child Language Corpus Project,” National Chung Cheng University

Written by: Assistant Pei-Yu Hsieh of the Taiwan Child Language Corpus Project; Philology Thematic Group assistant Chia-Min Lai, Content Development Division, National Digital Archives Program

Images photographed by: Philology Thematic Group assistants Chia-Min Lai, Shu-Huei Lin, and Hsiu-Hua Chen, Content Development Division, National Digital Archives Program

Edited by: Philology Thematic Group assistants Chia-Min Lai, Mei-Chih Chen, and Hsiu-Hua Chen, Content Development Division, National Digital Archives Program

Special thanks to Professor Jane Tsay, Project Director of the “Taiwan Child Language Corpus” National Chung Cheng University, former assistants Ting-Yu Huang and Hui-Chuan Liu, and current assistant Pei-Yu Hsieh for their advice, assistance with photography, and providing data.

VI. A Socio-phonetic Study of Spoken Taiwan Mandarin

Date of Creation: 2010/01/25

The director of the Language Archives phase two subproject “A Socio-phonetic Study of Spoken Taiwan Mandarin” of Academia Sinica Institute of History and Philology is Associate Professor Shu-Chuan Tseng. The project's purpose is to conduct a socio-phonetic study based on social linguistics and supported with acoustic phonetic tools. Database establishment provides structural data that facilitates systematic research. Social linguistic research methods are used to arrange social, economic, educational and linguistic background data. Digital audio recording technology systematically processes audio contents and tags them. Data processing methods of computational linguistics allow phonetic and social linguistic data to be effectively integrated and analyzed. We hope to establish a speech corpus for Spoken Taiwan Mandarin, using it to record the social, regional and linguistic characteristics of speech. Taiwan Mandarin is affected by an environment with multiple languages, and is unique in terms of lexicon and tones compared with other areas that use Modern Chinese. We use social, economic and internet usage indicators in coordination with acoustic analysis of speech in audio recordings. This subproject collects natural language from main prefectures in Taiwan; on one hand recording the connection between language usage and social transition, on the other hand recreating original speech sounds.

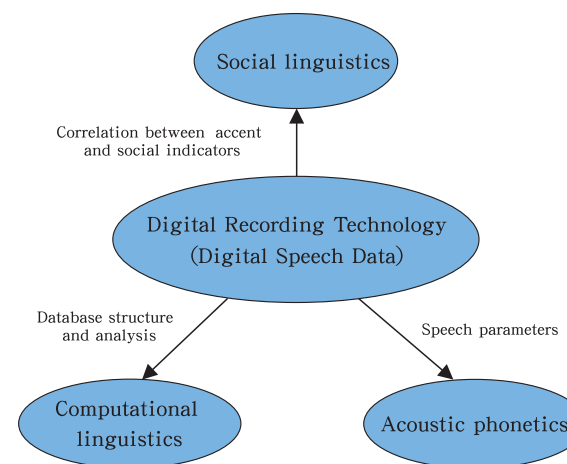


Fig 3-6-1 Relationship of the socio-phonetic corpus

Digitization procedures:

Digitization procedures of “A Socio-phonetic Study of Spoken Taiwan Mandarin” are divided into the follow seven procedures: 1. Decide on the sampling location and subjects; 2. Audio recording equipment; 3. Submit application to county/city government; 4. Questionnaire survey; 5. Transcription of audio recording contents into text; 6. Establish questionnaire content database; 7. Website construction and maintenance.

1. Decide on the sampling location and subjects

- (1) Sampling location: Main prefectures in Taiwan; locations where there are non-student crowds and relatively low noise, such as post offices, large parks, cultural centers and libraries.
- (2) Time of single sample: Three days.
- (3) Sampling personnel: Four groups in total; 2 people in 1 group.
- (4) Sampling subjects: Each group samples 30 subjects; four groups sample 120 people, at least 100 effective samples per location.
- (5) Operation method: One person is responsible for asking questions and recording, the other person is responsible for recording related data: Name, Gender, Location, and Profession.
- (6) Age of subjects: 20~40 years old.
- (7) Gender of subjects: Not limited.

2. Audio recording equipment



Fig 3-6-2 Audio recorder, microphone

- (1) Digital recording system:

Recorder: SONY Hi-MD MZ-RH1

Microphone: SONY ECM MS907

- (2) Audio format:

The audio input is Hi-SP stereo, audio output format is wave, mono sampling rate 44.1KHz, 16 bits.

3. Submit application to county/city government

Before setting out, we ask administrative staff of Academia Sinica Institute of History and Philology to submit an application to the county/city government to ensure that project personnel are aided and safe.

4. Questionnaire survey

- (1) Language background: Language usage of the subject and the subject's family members.
- (2) Social and economic background: Academic background from elementary school to the highest academic degree, any work experience over six months, and current salary. The area of residency is recorded as the township.
- (3) Internet usage and international perspective: Whether or not the subject often uses the internet, purpose and behavior using the internet, language of websites visited, type of news browsed, and whether or not the subject ever went abroad.
- (4) Language self-assessment: We ask subjects to self-assess whether or not their Mandarin is affected by the local accent based on their own experiences and habits when speaking, as well as which sounds are incomplete or easy to get confused with.



Fig 3-6-3 Outdoor interview

5. Transcription of audio recording contents into text

All digital speech data is transcribed using PRAAT, audio files are segmented according to answers and then aligned.

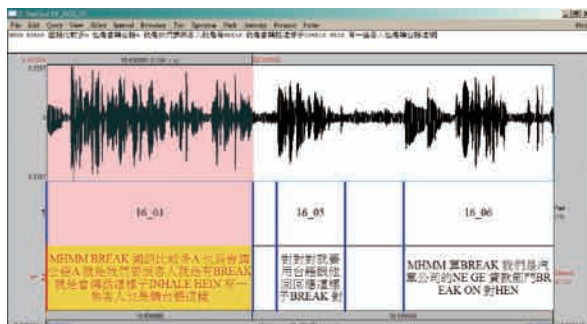


Fig 3-6-4 Transcription of audio recording contents into text

6. Establish questionnaire content database

An entry is created for each subject's questionnaire; answers are input in the order on the questionnaire.

ID	姓 名	性別	年齡	學歷	職業	工作 地點	工作 時間	工作 內容	工作 時間	工作 內容	工作 時間	工作 內容
FW_001	許	男	40	高中	工程師	台中市	8.0	台中縣	8.0	科技業	調	調
FW_002	羅	女	35	高中	司機	台中市	8.0	台中市	8.0	汽車公司	調	調
FW_003	李	男	45	高中	司機	台中市	8.0	台中市	8.0	新通車	調	調
FW_004	鄭	男	40	高中	司機	台中市	8.0	台中市	8.0	新通車	調	調
FW_005	張	男	40	高中	司機	台中市	8.0	台中市	8.0	新通車	調	調
FW_006	林	男	40	高中	司機	台中市	8.0	台中市	8.0	新通車	調	調
FW_007	陳	男	40	高中	司機	台中市	8.0	台中市	8.0	新通車	調	調

Fig 3-6-5 Questionnaire Database

7. Website construction and maintenance

Archives and Linguistic Representations of Spoken Taiwan Mandarin consist of corpora established by associate research fellow Shu-Chuan Tseng, and not limited to the digital archives project. In which collection of linguistic data for the Socio-phonetic Study of Spoken Taiwan Mandarin is still in progress, so it has not been launched online yet. However, contents of other corpora already cover most matters that require special attention in corpus establishment.



FOUR. Digital Learning

The earliest learner’s corpus was the Longman Learners’ Corpus established in the late 1980s. In the mid 1990s, Sylvaine Granger of the Catholic University of Louvain Centre for English Corpus Linguistics established the International Corpus of Learner English (ICLE). The corpus is an extensive international cooperation project that currently collects over two million words and data on English learners from 14 different native language backgrounds.

Corpus applications in learning is a growing trend, below we will use the “Digital Resources Center for Global Chinese Teaching and Learning” of Academia Sinica Institute of Linguistics and the NCKU “Eagle Project” as examples of how academic resources of corpora can be applied to teaching.

I. Digital Resources Center for Global Chinese Teaching and Learning

The “Digital Resources Center for Global Chinese Teaching and Learning” is a project under the National Science and Technology Program for e-Learning that is directed by Academician Chin-Chuan Cheng of Academia Sinica Institute of Linguistics.

Academia Sinica has implemented digital archives projects for numerous years and accumulated rich results in corpora. However, these resources were mainly developed in response to academic research requirements, making it hard for typical Chinese teachers and students to use. The purpose of the “Digital Resources Center for Global Chinese Teaching and Learning” is to integrate these corpora and extended resources, and provide an easy to use learning tool and teaching resource, constructing and digital resources center that serves both teaching and research functions.



Fig 4-1 Homepage of the Digital Resources Center for Global Chinese Teaching and Learning

This project has two main objectives, one is to use the theory of “word-focused extensive reading” as a basis to help students more rapidly learn how to use words; the other is to provide Chinese teachers with the language information they need to compile teaching materials.

The “word-focused extensive reading” learning model is the core concept of the “Digital Resources Center for Global Chinese Teaching and Learning,” using immense databases to help users find related sentences when they search for a word. This helps users understand the linguistic environment in which words appear, as well as combinations with different words, allowing them to better understand how to use specific words, which further accelerates language learning speed. This type of “word-focused extensive reading” is especially effective to adult learners.

Corpora used by the “Digital Resources Center for Global Chinese Teaching and Learning” include the “Classical Chinese Corpus,” “Academia Sinica Tagged Corpus of Early Mandarin Chinese,” and “Academia Sinica Balanced Corpus of Modern Chinese” of Academia Sinica Institute of Linguistics, the “NICT Elementary School Chinese Textbook Corpus,” and the “300 Tang Poems Corpus” established in collaboration with Yuan Ze University. Furthermore, Academia Sinica acquired licensing to use the British Nation Corpus, so the “word-focused extensive reading” learning model also provides English word usage for English learners.

Contents of the abovementioned corpora are as follows:

1. **“Classical Chinese Corpus”**: “The Analects of Confucius,” “Mencius,” “The Great Learning,” “The Book of Chuang Tzu,” and “Lao tzu.”
2. **“Academia Sinica Tagged Corpus of Early Mandarin Chinese”**: “Dream of the Red Chamber,” “Journey to the West,” “Outlaws of the Marsh,” and “The Unofficial History of the Scholars.”
3. **“Academia Sinica Balanced Corpus of Modern Chinese”**: Modern Chinese text of various topics with a total of 5 million words (over 200 thousand sentences, roughly 140 thousand entries).
4. **“NICT Elementary School Chinese Textbook Corpus”**: Over 50 thousand words.
5. **“300 Tang Poems Corpus”**: Roughly 7 thousand entries.
6. **“British National Corpus”**: A corpus that contains 100 million English words.

All pages on the website provide Chinese or English links for the convenience of foreign language learners; the website is divided into two main areas, “Learning Area” and “Teaching Resources.” The “Learning Area” allows users to use Chinese and English “word-focused extensive reading” functions; the “Teaching Resources” area helps teachers find the materials they need.

To provide a model that benefits learning, in the “Learning Area” the difficulty of a sentence is calculated based on the length of the sentence, frequency of words and semantics; search results can either be arranged in the mode “Read for simple to complex,” which lists readings from simplest to the most complex, allowing learners to choose readings of the difficulty they desire, or “Read randomly,” for which the system randomly displays search results. Contents of the “Near Synonyms” mode are retrieved from the “同義詞詞林(Thesaurus of Chinese Synonyms)”, listing words with similar meanings according to their degree of similarity.



Fig 4-2 Word-focused extensive reading search interface

The “Teaching Resources” area provides word frequency statistics of the Academia Sinica Balanced Corpus of Modern Chinese, Academia Sinica Tagged Corpus of Early Mandarin Chinese, Classical Chinese Corpus, 300 Tang Poems Corpus, and 300 Songci Corpus, as well as tagged text reading. Users can search for word frequencies of an individual corpus, frequency of an individual word, and accumulated word frequencies; teachers can learn the number and frequency words are used, and determine which words to teach first, while the tagged text reading provides parts of speech tags for words.

The “Digital Resources Center for Global Chinese Teaching and Learning” integrates rich corpora resources, offers a user interface that helps learners effectively learn word usage based on learning theories, arranges readings according to their difficulty, and provides objective statistics to Chinese teachers, offering numerous benefits to both learners and teachers.

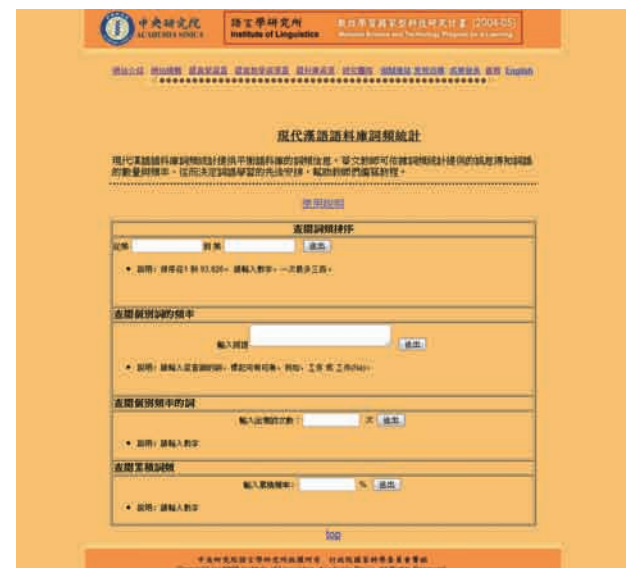


Fig 4-3 Language teaching resource search interface

II. NCKU Eagle Project and the CANDLE Project for Reading

The NCKU Eagle Project was planned and implemented by the Department of Foreign Languages and Literature of NCKU under commission by the Office of Academic Affairs. Starting in 2006, to enhance the English proficiency of undergraduate students, the project purchased an online English teaching platform, constructed an online English proficiency testing system, and established online multimedia interactive English learning courses. This project hopes to encourage English teachers to improve their ability to utilize information technologies, and to use online English teaching materials, which helps students enhance their English proficiencies and computer skills at the same time during class.

Project Contents Include:

1. Online English proficiency testing system

- (1) Establish software/hardware equipment for the online testing system.
- (2) Complete different levels of difficulty of English questions.
- (3) Test the online English proficiency testing system, conduct evaluations, and make improvements.
- (4) Offer tests for students (free of charge) and others (charged) to take.

2. Multi-purpose English resources classroom

- (1) Plan classroom functions and purchase software/hardware audiovisual equipment.
- (2) Complete classroom installations to provide services to students.
- (3) Plan integration of the resources classroom and courses.
- (4) Professional growth of all teachers in audiovisual multimedia teaching.

3. Online English courses

- (1) Plan online English course contents and implementation methods; complete software/hardware equipment establishment.
- (2) Provide students with a variety of online courses to choose from, allowing students to learn online without being limited by time and space.
- (3) Online course implementation assessment and revision; students can select suitable courses based on assessment results.
- (4) Increase the number of courses for students (free of charge) and others (charged) to take.
- (5) Courses integrate training of the four language skills, listening, speaking, reading, and writing, topics cover food, clothing, living, transportation, education, and entertainment. Popular elements and topics can be included to provide rich and diverse course contents for students to connect and compare Chinese and Western cultures via language learning.
- (6) In response to the policy the government planned to implement in the 2006 academic year, we recognize credits of online learning courses, for which students can gain a degree; acquiring credits and a degree via online learning is the trend of the new era.

The CANDLE (Corpus and NLP for Digital Learning of English) system, which is categorized as a digital English teaching material, was developed by the research team of Professor Hsien-Chin Liu of National Tsing Hua University between 2003 and 2006 under the National Science and Technology Program for e-Learning.



Fig 4-4 CANDLE Homepage

The project uses advanced corpora and natural language processing tools to set up learning support on the internet, and establishes a learning center – CANDLE to aid their English learning. According to students' English proficiency, the learning center provides suitable listening, speaking, reading, writing, cultural and translation teaching materials, as well as suitable practice tests to better their English skills. Besides typical English corpora, CANDLE also includes the bilingual “Taiwan Panorama Magazine” corpus; its contents mainly reports information on various aspects of modern Taiwan. Bilingual corpora are an extremely forward-looking research topic in computational linguistics; this system adopted a bilingual corpus to let NCKU students learn English using strengths and background knowledge of their native language, this is “computerized” learning support for students both mentally and systematically. The current stage CANDLE system provides listening, speaking, reading and writing practices, as well as translation and search functions.

Although the “Eagle Project” is not a teaching website entirely based on corpus applications, the integration of a corpus with a teaching project for teaching is also a type of value-added application of corpus resources. Therefore, this project is still a valuable reference.



FIVE. Extended Issues

I. Digital Content Protection

After dedicating years of effort, projects of the “Taiwan e-Learning and Digital Archives Program” have output great digital results, and continue to accumulate even more results. To projects with substantial achievements, besides developing new value-added applications and creating new value, protecting existing resources is also extremely important. If a result that took years of effort were stolen, it would significantly impact morale and have long-term effects on the digital archives environment. In recent years, digital content protection mechanisms have made great progress, besides protecting digital output, the entire digital archiving process can also be protected, whether it may be digitization procedures, data transfer, or follow-up on usage state.

At present, the complete concept of digital content protection is what we call Digital Rights Management (DRM), combining hardware and software for content protection. In terms of software, DRM limits access rights and the number of times a digital content can be accessed; in terms of hardware, DRM limits the storage media on which digital content can be stored. The combination of software and hardware limitations defines a life cycle in which users can access digital content; during the life cycle, DRM traces and limits the number of times the digital content is accessed, duplicated, and how it is used. Digital content will become inaccessible after its life cycle ends.

Reasons for the rise and development of DRM are as follows:

1. To protect intellectual property rights

Many technologies of DRM originated from antipiracy concepts, meaning that they were developed to protect intellectual property right. In a digital age, many digital contents have become intangible assets, protecting these intellectual property rights from being misused benefits holding institutions by allowing the contents to be used for other value-added applications.

2. To protect privacy and classified content

Each time data is transferred it is at risk of being stolen. Protection technologies, such as data encryption and applying restrictions on the software and hardware that can be used to access data, were developed to prevent information from being intercepted. Many sensitive units use these technologies to protect their classified information.

3. To create new market opportunities

The establishment of DRM mechanisms also indicates the development of a digital content usage model. This model can be used for commercial applications, digital content with usage restrictions can be transformed into a product, providing digital information and services to clients.

4. To unify standards

From a commercial perspective, the more digital content there is, the more important DRM will be. Many firms saw this development potential, and became engaged in technology developments to get a share of the market. A unified standard benefits both technology developments and usage, which attracts firms and consumers, and the rise of DRM was the result.

After seeing the reason for the rise of DRM, it is not hard to understand that the purpose of DRM is to protect intellectual property rights, to prevent the unlimited dissemination of digital content without licensing, and even when digital content is licensed, to keep track of digital content usage, guaranteeing that digital content is not pirated.²⁶ Effective protection of property benefits digital content output, allowing the continued progress towards future visions, at the same time developing value-added applications.

Common technologies of DRM include digital watermarks, public keys and digital rights expression language; each technology is briefly described below:

1. Digital watermark

“Digital watermark” technology refers to embedding symbols or totems into images or videos. Digital watermarks can be a basis for establishing copyright; in the event there is a dispute over copyright, digital watermarks can serve as evidence for the rightful owner. For this reason embedding digital watermarks in digital content can prevent piracy to a certain extent; digital watermarks can be considered a declaration of copyright.

Digital watermarks are suitable for photos, audio files, and image files.

²⁶ Hsin Yu Chen, Cheng-Hung Li, Yi-Hang Chiu and Wei-Ling Lin, “Digital Rights Management Mechanism Implementation – Using Digital Archive Management Systems as an Example”, “Proceedings of the fourth workshop on digital archives technology”, September 2005, pages 93~100.

Depending on its visibility, digital watermarks can be divided into visible and invisible watermarks. The former refers to watermarks directly visible on digital files, so it intimidates people and prevents them from illegally using the image; the later refers to watermarks not visible to the naked eye, but provides protection to copyright. Watermarks mentioned in digital archive projects typically refer to the later.

The design of digital watermarks must give consideration to the following factors:²⁷

- (1) Transparency: Watermarks must not affect image or audio quality.
- (2) Robustness: Watermarks must remain in digital content even if attacked.
- (3) Security: Watermarks should be undetectable, and even when the watermark structure is known, the corresponding key is still required to remove the watermark.
- (4) Capacity: The number of watermarks that can be added; this usually contradicts transparency requirements.
- (5) Complexity: The time and difficulty of embedding and removing watermarks, and whether or not the extraction of watermarks requires source data or related data for comparison (watermark blindness).
- (6) Invertibility: Can the source data be retrieved by removing the watermark.
- (7) Unambiguous: The watermark must clearly indicate the copyright owner.

Whether digital archive projects develop their own watermark technology or purchase commercialized watermark technology, they should use the factors above as a standard for evaluating antipiracy technology. However, following the development of new technologies, the antipiracy function of digital watermarks will gradually weaken until it serves only to declare copyright; watermarks no longer guarantee that digital content will not be pirated, and are considered a relatively passive prevention measure.

²⁷ Jen-Hao Hsiao, Hsin-Hui Lin, Jin-Lung Lin and Li-Hung Lin, “Development status of digital watermark technology: Using digital archive projects as an example”, “Proceedings of the third workshop on digital archives technology”, August 2004, pages 163~169; Steinebach, M., J. Dittmann & E. Neuhold “Digital Watermarking – Common watermarking techniques, Important Parameters, Applied mechanisms, Applications, Invertible watermarking, Content-fragile watermarking,” <http://encyclopedia.jrank.org/articles/pages/6725/Digital-Watermarking.html>, downloaded on January 27th, 2010.

2. Keys

Keys are a type of encryption technology that embeds cryptography technology into digital content to limit file access and duplication. There are two different designs:

(1) Symmetric encryption

This is the conventional encryption method, in which both ends have a private key. If one end wishes to transfer a file, then both ends must generate a private key to remove restrictions. For transfers between groups, this encryption method lacks efficiency, but provides better file security.

(2) Asymmetric encryption

This is the public key encryption method; both ends are required to have two keys, a private key and a public key. Encryption requires a pair of matching keys to complete; after a file is encrypted using the public key, it is decrypted using private keys; the purpose of this method is to use the irreversibility of keys to prevent anyone from calculating the encryption algorithm to steal files.

3. Digital rights expression language

Digital rights expression language refers to a language that describes the scope of rights and obligations associated with a digital content between the creator and user. At present, XrML (eXtensible Rights Markup Language) is the most common type of digital rights expression language; XrML is the standard digital rights expression language of the ISO, and provides DRM, metadata management, content management, and content transfer services. Furthermore, XrML can be used as the standard language for DRM of various media, e.g. electronic books, digital publications, broadcasts and music, and has been adopted by numerous firms. DRM can add a digital signature to digital content to control its circulation and duplication. Besides XrML, other digital rights expression languages include ODRL (Open Digital Rights Language), EBX (Electronic Book Exchange) and MPEG (Moving Picture Experts Group).

Following the growing diversity of corpus resources, corpora now face different issues of digital rights. Data of ancient literature corpora have no copyright restrictions and can be used by anyone, but the search interface and database system were built by the strenuous efforts of project units; when

most corpora are open for searches, they can only call on users to give a clear indication of the source, and not completely neglect the contribution of the project unit. Some projects don't provide full text as a strategy, showing only partial text in search results, which not only achieves the purpose of the corpus, but also protects the copyright owner of data or the project unit's efforts.²⁸

The appearance of multimedia corpora further showed the importance of digital content protection technology. In the future, when value-added applications of corpora are being developed, the necessity of DRM concepts will further grow to ensure digital content is properly protected.

II. Human Resource and Equipment Cost Analysis

Generally speaking, digitization procedures of corpora are unlike digitization procedures for antiquities or artworks, which require extremely expensive equipment. Therefore, equipment costs of corpora only represent a small portion of overall expenses. However, due to the extensive time required for corpus digitization procedures, human resource expenses will take up most of the project's funds. In addition, transportation, accommodations, food, miscellaneous, and personnel expenses of field investigations, which are required to collect linguistic data, also total to a considerable amount, which is why the project director should pay special attention to budget planning when planning project details and writing the proposal.

However, not all corpora need to conduct field investigation. Therefore, this book does not elaborate on budget planning for field investigations, and only introduces human resources and equipment required for corpus digitization.

1. Human resource cost analysis

Professor Shu-Chuan Tsai of the National Chung Cheng University Graduate Institute of Linguistics and director of the Taiwan Child Language Corpus kept a record of the human resources and work hours required for each operation procedure, as shown in Table 5-1. In this section, we will explain human resource cost using Table 5 as an example.

The term of the "Taiwan Child Language Corpus" establishment project

²⁸ Search results of Sinica Corpus do not provide the source or complete text; one reason is that the corpus did not acquire sufficient licensing from the copyright owner.

was three years (August 1st, 2000 to July 31st, 2003); the source of linguistic data was the NSC project “A Study on Developments of Taiwanese Tone Acquisition” (project term was from August 1st, 1998 to July 31st, 2000) implemented by Professor Shu-Chuan Tsai, in which the free conversations of 14 children between ages 1 and 3 in the Chiayi area when playing games or reading books were recorded; the total time of recordings was 329 hours and 14 minutes, and transcribed into roughly 2.3 million words; the entire process from data collection to completion of corpus establishment took six years.

Telling from the records of the Taiwan Child Language Corpus, the project’s field investigations made a total of 431 person-times recordings, recording time reaching 330 hours. For every hour that is recorded, it takes 1.5 hours to edit, meaning that 495 hours must be spent on editing alone; calculating based on 8 hours of work each day, editing will take roughly 62 days, meaning that each assistant needs to spend 20 days on editing. Editing is extremely time-consuming, 1.5 hours spent on editing 1 hour of audio recordings by the Taiwan Child Language Corpus is an extremely ideal state; generally speaking, editing might take a significantly longer amount of time.

When the project reached the stages “manual transcription into Chinese characters” and “manual IPA representation”, they took 10 times and 4.5 times the amount of audio recording time respectively, which is a total of 4,785 hours. Calculating based on 8 hours of work a day, 5 days of work a week, and 3 assistants sharing the work, these two stages still require 10 months to complete. This is just the time noted in the operation procedures table, time spent on other procedures, e.g. manual segmentation, manual part-of-speech tagging, and manual phonetic representation, were not included.

Associate Research Fellow Hsin-Min Wang of Academia Sinica Institute of Information Science once spent three years establishing a corpus of Mandarin news reports, collecting 250 hours of Mandarin news programs broadcasted by Taiwan Public Television between November 2001 and June 2003. Even after using a computer program to automatically compare and extract corresponding text data on the Taiwan Public Television Service Foundation website, with two full-time assistants he was only able to complete transcription and part-of-speech tagging for 198 hours in three years.²⁹ This shows the extremely

²⁹Hsin-Min Wang, Berlin Chen, Jen-Wei Kuo and Shih-Sian Cheng. 2005. "MATBN: A Mandarin Chinese Broadcast News Corpus," *Computational Linguistics and Chinese Language Processing* 10.2, pp.219-236.

long amount of time that must be spent if the project director intends to build a corpus with detailed contents, and doesn’t have computer programs to automatically perform segmentation and part-of-speech tagging.

Up to now, corpus digitization procedures remain a task that require massive amounts of human resources, excluding cost of human resources on general affairs, expenses on project implementation personnel will make up the majority of expenses. Therefore, the project director must fully understand project requirements and carefully calculate human resource costs when planning the project budget.

2. Equipment cost analysis

The most frequently used equipment in corpus building is the audio recorder. The ultimate objective of corpus building is to establish a database server that offers search functions, so servers also need to be purchased. In the following section we will briefly introduce these two devices.

Audio recorders can be divided into analog recorders, e.g. Cassette Recorders, and digital recorders, e.g. Minidisc (MD) Recorder, Digital Audio Tape (DAT) Recorders, Solid-State Recorder, Hard-disk Recorder and computers. There is currently a wide variety of pen recorders in the market, even typical MP3 players come with recording functions. However, in order for the quality of linguistic data to satisfy acoustic speech analysis or digital archiving requirements, we recommend using high-end digital recorders with external microphones instead of typical cassette recorders or pen recorders.

(1) Linear PCM recorder

Some PCM recorders are equipped with a highly sensitive condenser microphone; besides high reception sensitivity, it also gathers excellent stereo sound quality. In terms of audio files, PCM recorders have a sampling rate up to 96kHz, in contrast to the 44.1kHz sampling rate of typical MP3 files, showing



Fig 5-1 PCM Recorder³⁰

³⁰Image provided by Roland Taiwan Enterprise Co., Ltd.

the better recording performance of PCM recorders. PCM recorders can directly store sound as uncompressed high quality files (e.g. WAV format), which benefits following procedures, e.g. monitoring, reading and post processing. Due to its better performance, PCM recorders are relatively expensive and vary between NT\$10~20 thousand.

(2) Digital Recording Studio

Digital recording studios are even more expensive, and cost almost the same as a high-end notebook computer. The digital recording studio mentioned here does not refer to the device used for recording music, and is a relatively small device suitable for collecting linguistic data. Digital recording studios often need to be used together with a notebook computer; following technology advancements, the size of digital recording studios has significantly shrunk to weighing around 1Kg, making it easier to carrier for field investigators.

Digital recording studios can be connected to a notebook computer via the USB interface; data that is collected can be directly saved in the computer. Some models have memory card slots and don't need to be connected to a computer; audio files are directly stored in the memory card. The main advantage of digital recording studios is that it provides multiple channels for input, and can collect the speech sound of several people after connecting it with microphones. Furthermore, the audio quality of digital recording studios can match that of linear PCM recorders, files can be stored in 16-bit or 24-bit, and sampling rate can reach 192kHz, which is even better than PCM recorders.



Fig 5-2 Digital recording station ³¹

structure, Dynamic Microphones and Condenser Microphones.³² Dynamic microphones have coils, diaphragm, and permanent magnet; the microphone you hold in a KTV belongs to this category; advantages include low manufacturing cost and warmer sounds, disadvantages include larger size, low sensitivity, and poor performance at high and low frequencies.

Condenser microphones record audio signals via voltage changes on a capacitive diaphragm; advantages include smaller size and higher sensitivity, making it suitable for high sensitivity recording. However, condenser microphones require a stable voltage source, and some products require additional batteries. The high sensitivity of condenser microphones makes it suitable for collecting linguistic data, its size and weight makes it suitable for use outdoors, which is why we recommend condenser microphones for corpus projects.

Based on different sensitivity to sounds, microphones can be divided into Omnidirectional, Cardioid and Bi-directional. Omnidirectional microphones gather sounds from all around, bi-directional microphones collect sounds to the front and back, and cardioid microphones collect sounds from only one direction. When corpus projects collect speech sounds, one microphone only focuses on the sound of one speaker, surrounding noise should be minimized for future procedures. Therefore, cardioid microphones are more suitable for collecting linguistic data.

When selecting microphones, also notice the format of the connector and whether or not it is compatible with the pen recorder or digital recording studio. Depending on its response frequency, sensitivity, and resistance, microphone prices may vary widely; mini-microphones and lavalier microphones connected with a PCM pen recorder using a 3.5mm connector can range between NT\$2,000~4,000; table or handheld microphones, which are larger but have better performance, can range between NT\$4,000~6,000; while the best microphones can have prices of over NT\$10 thousand.

If you are using a PCM pen recorder, we recommend using a lavalier microphone for better convenience; if you are using a digital recording studio, then a table or handheld microphone would be your best option.

³¹ Same as 30.

³² Microphone, Wikipedia: <http://zh.wikipedia.org/zh-tw/>.



Fig 5-3 Table and handheld condenser microphone³³

(4) Server³⁴

In terms of hardware, a server refers to computer hardware dedicated to storing digital resources; in terms of software, a server refers to a computer program that manages digital resources and provides services, e.g. file server, database server and applications server. In this section we will introduce servers for storing digital resources.

There are numerous differences between a server and typical desktop computer; servers are used by the owner of digital resources, and provide digital resources via the internet to users with typical desktop computers. Server hardware was designed to endure 24 hour non-stop operation to satisfy user requirements. Servers also have better processing ability than typical desktop computers. Generally speaking, common servers fall into three categories according to their size, form and performance, including Tower Servers, Rack-Mount Servers and Blade Servers.

Tower servers are basic servers suitable for small companies; they are often confused with desktop computers due to their similar appearance. A tower server takes up similar space as a typical computer, but is significantly more stable, and has better CPU and memory to sustain long periods of non-stop operation. Due to space limitations, the number of hard drives that can be expanded for a tower server is similar

to a desktop computer. Most small digital archive projects use tower servers; basic tower servers cost roughly NT\$50,000~60,000, while high-end tower servers are still within NT\$100,000.



Fig 5-4 Tower Server

When using tower servers, it is necessary to purchase a UPS to ensure the stability of data storage and transfer. There are currently three types of UPS in the market: Off-Line, On-Line, and Line Interactive UPS; these three types of greatly varying prices. For servers we recommend using an On-Line UPS, which is the most stable type of UPS and costs approximately NT\$10,000.

When server requirements reach ten tower servers or higher, it will take an extremely large amount of space, in this case we recommend using rack-mount servers to save space. Rack-mount servers are flat; calculated in units of the smallest rack-mount server 1U,³⁵ a rack-mount server can range from 1U to 5U in size. For effective management and space utilization, rack-mount servers must be installed in a server rack, which has roughly 42U of space.

The advantage of rack-mount servers is its extreme expandability, while providing better performance than tower servers. However, maintenance is a somewhat daunting task; due to the number of servers installed in the server rack, heat dispersion becomes a major issue. Rooms that hold such servers

³³ Image provided by Roland Taiwan Enterprise Co., Ltd.

³⁴ Image provided by IBM.

³⁵ Rack Unit, Wikipedia: http://zh.wikipedia.org/zh-tw/Rack_unit. The rack unit was established by the EIA to describe the height of equipment, such as servers; height: 44.45mm, width 482.6mm.

must be air conditioned at all times, and its position, maintenance methods, and personnel all require special arrangements.



Fig 5-5 1U Rack-Mount Server



Fig 5-6 Rack-mount servers installed in a server rack

The back of server racks is usually riddled with cables, and may heat up to astonishing temperatures even with air conditioning. Besides heat dispersion, the cables also cause difficulties with management and maintenance, which led to the development of blade servers.

Blade servers need to be used with a blade base, which provides power, fans and internet functions; slots in the base are for connecting blade servers. The signal lines replace cables while providing heat dispersion functions, making it easier to manage and maintain; under the same number of servers, blade servers offer better heat dispersion.



Fig 5-7 Blade Server

Rack-mount servers have good performance, but are much more expensive than tower servers. A full height server rack costs around NT\$60,000 (includes the KVM monitor), one rack-mount server costs at least NT\$80,000 (hard drive not included), adding the UPS and peripheral equipment, the cost of a set of equipment can easily exceed NT\$200,000. Blade servers are even more expensive, and involve considerable cost on future maintenance, making it even harder to affordable by typical projects.

In general, only SMEs or large organizations, e.g. schools or culture and education foundations, will use rack-mount servers; too small scale project institutions, this type of server will prove too expensive to maintain. Therefore, besides purchasing your own tower server, renting servers from large institutions is also a feasible option.

Written by: Ching-Hsun Chan and Pei-Ying Li

Table 5-1 Procedures of Taiwan Child Language Corpus
 Unit: National Chung Cheng University Graduate Institute of Linguistics
 Digital Object Name: Linguistic data on child language Subproject Name: Taiwan Child Language Corpus
 Director (Person in Charge) (E-mail, Tel): Professor Jane Tsay lngfsay@ccu.edu.tw 05-2720411*31502
 Contact Person (E-mail, Tel): Pei-Yu Hsieh astpnh@ccu.edu.tw 05-2720411*21509

Procedure	Work Content	Operators (Number, Professional Competency Requirements)	Hardware (Name, Version, Price)	Software (Name, Version, Price)	Standards (Technical standards, product specifications, quality requirements, etc.)	Time	Conclusion (Difficulties, Flaws, Features, etc.)	Cost Estimation
1	Research Assistant Training (Understand Southern Min phonology, writing system, lexicon, syntax, semantics and part of speech tagging system; understand CHILDES; understand IPA transcription)	Project Director 3 Research Assistants (bachelor or master's degree in philology; native language is Southern Min)	Audio Recorder (NT8,000/Set) Audio Tape (NT150/Tape)		“Southern Min Vocabulary” volumes 1 and 2, Yang Hsiu-Fang, National Languages Committee, 1998. “Taiwan Southern Min Syntax Manuscript”, Yang Hsiu-Fang, Daan Publishing House, 1995. “Brief Record of Taiwan Southern Min Dialects”, Chang Chen-Hsing, Wenshizhe Publishing House, 1993. Handbook of the International Phonetic Association, (1999) The CHILDES Project, Brian MacWhinney (1995)	3-6 hours a week for discussion or transcription practice		1 Full time assistant (master's degree) NT34,000/ Month 2 Part time assistants (master's degree) NT6000*2/ Month

Procedure	Work Content	Operators (Number, Professional Competency Requirements)	Hardware (Name, Version, Price)	Software (Name, Version, Price)	Standards (Technical standards, product specifications, quality requirements, etc.)	Time	Conclusion (Difficulties, Flaws, Features, etc.)	Cost Estimation
2	Solicit children from Southern Min speaking families (Post ads on posters and the internet; use parents day of our kindergarten)	3 Research Assistants (Familiar with computer network applications; basic art and poster design)	3 Desktop Computers (NT50,000 each)	(1) Microsoft OS: 98/2000/XP Microsoft Office: Word/Excel			Our subjects are infants between ages 1 and 3 from Southern Min speaking families that are in CCU's subsidiary nursery, kindergarten and nearby townships; we selected 14 children in total.	
3	Arrange audio recording time (contact parents to set the time schedule)	3 Research Assistants	Desktop Computer (NT50,000 each)					

Procedure	Work Content	Operators (Number, Professional Competency Requirements)	Hardware (Name, Version, Price)	Software (Name, Version, Price)	Standards (Technical standards, product specifications, quality requirements, etc.)	Time	Conclusion (Difficulties, Flaws, Features, etc.)	Cost Estimation
4	Prepare audio recording equipment	3 Research Assistants	Minidisc Portable Audio Recorder (NT12,000 each) Professional microphone (NT8,000 each) Minidisc (NT800/15 discs) Professional earphone (NT3,000 each)			Two weeks	We selected convenient and portable recording equipment with large storage space that benefit the long-term preservation of linguistic data.	

Procedure	Work Content	Operators (Number, Professional Competency Requirements)	Hardware (Name, Version, Price)	Software (Name, Version, Price)	Standards (Technical standards, product specifications, quality requirements, etc.)	Time	Conclusion (Difficulties, Flaws, Features, etc.)	Cost Estimation
5	Recording interviews (We visit children in their homes and record our interviews. The interviews are periodical and continue throughout summer and winter vacation. We visit children under the age of 2 once every week; 2 and 3 years old once every two weeks; and 3 and 4 years old once every two to three weeks.)	3 Research Assistants (Familiar with minidisc audio recorder operations; patient, likes to interact with children)	Minidisc Portable Audio Recorder (NT12,000 each) Professional microphone (NT8,000 each) Minidisc (NT800/15 discs)			Each interview takes 1~2 hours; recording time 40~60 minutes. Date of recordings: October, 1997 to May 2000. 431 person-times, rought 330 hours.	During interviews, we let children speak as they normally do when accompanied by a parent or babysitter. Besides natural language, we also used illustrated books, story books, toys, dolls, origami, or other games to cause children to talk.	Field investigation fee of NT200/person-times Interview fee NT200/person-times

Procedure	Work Content	Operators (Number, Professional Competency Requirements)	Hardware (Name, Version, Price)	Software (Name, Version, Price)	Standards (Technical standards, product specifications, quality requirements, etc.)	Time	Conclusion (Difficulties, Flaws, Features, etc.)	Cost Estimation
6	Editing audio files	3 Research Assistants (Familiar with minidisc audio recorder operations)	Minidisc Portable Audio Recorder (NT12,000 each) Professional earphone (NT3,000 each) Minidisc (NT800/15 discs)			Each hour recorded takes 1.5 hours to edit. Total editing time: 1.5×330 hours = 495 hours.	Dates and filenames are entered into the minidisc. Irrelevant sounds or silent parts are deleted. The audio file is split into smaller parts and assigned numbers.	
7	Audio file backup (backup audio files on a minidisc)	3 Research Assistants (Familiar with minidisc audio recorder operations)	Minidisc Portable Audio Recorder (NT12,000 each) Minidisc Audio Recorder (NT35,000 each) Minidisc (NT800/15 discs)	GoldWave Digital Audio Editor (Developed by GoldWave Inc.)		Each minidisc takes roughly 2.5 hours. Total work time: 2.5*330 (hr) = 825 hrs		

Procedure	Work Content	Operators (Number, Professional Competency Requirements)	Hardware (Name, Version, Price)	Software (Name, Version, Price)	Standards (Technical standards, product specifications, quality requirements, etc.)	Time	Conclusion (Difficulties, Flaws, Features, etc.)	Cost Estimation
8	Digital transcription	Multiple Research Assistants (Familiar with minidisc audio recorder operations)	Desktop computer (NT50,000 each) Minidisc Portable Audio Recorder (NT12,000 each)				The audio file in the mini-disc is converted into MP3 format for storage convenience. The file can be converted into other formats for speech analysis (e.g. *.wav).	

Procedure	Work Content	Operators (Number, Professional Competency Requirements)	Hardware (Name, Version, Price)	Software (Name, Version, Price)	Standards (Technical standards, product specifications, quality requirements, etc.)	Time	Conclusion (Difficulties, Flaws, Features, etc.)	Cost Estimation
9	Establish the Southern Min writing system (There currently isn't an established Chinese character writing system for Southern Min; some original words cannot be verified and some words only have sounds but no character. Therefore, it is necessary to establish a text transcription principle.)	3 Research assistants (Familiar with computer word processing; basic knowledge of Southern Min writing system)	Desktop computer (NT50,000 each) Mimidis Portable Audio Recorder (NT8,000 each) Professional earphone (NT3,000 each)	(1) Microsoft OS:98/2000/XP (2) Microsoft Office: Word/Excel	The four dictionaries we used according to their priority are: "Taiwan Southern Min Dictionary", Tung Chung-Ssu, Wu-Nan Books Inc., 2001 "Taiwanese Dictionary," Chen Hsiu, Yuan-Liou Publishing Co., 1998 "Xiamen Dialect Dictionary," Li Jung, Jiang Su Education Press, 1998 and "Southern Min Dictionary", Yang Hsiu-Fang, National Languages Committee, 1998			

Procedure	Work Content	Operators (Number, Professional Competency Requirements)	Hardware (Name, Version, Price)	Software (Name, Version, Price)	Standards (Technical standards, product specifications, quality requirements, etc.)	Time	Conclusion (Difficulties, Flaws, Features, etc.)	Cost Estimation
10	Manual transcription into Chinese characters (Transcription of audio files into text files)	Multiple Research assistants (Familiar with computer word processing; basic knowledge of Southern Min writing system)	Desktop computer (NT50,000 each)	CHAT transcription platform of CHILDES	CHILDES (Child Language Data Exchange System; MacWhinney and Snow 1985, MacWhinney 1995)	Each hour recorded requires 10 hours to transcribe into text; total work time: 330 hours×10 = 3,300 hours		
11	Manual segmentation: natural speech into sentences with individual meanings)	Multiple research assistants (possess linguistic background knowledge, e.g. syntax, semantics, etc.)	Desktop computer (NT50,000 each)	(1) Microsoft OS:98/2000/XP (2) Microsoft Office: Word/Excel	Segmentation standards of CHILDES		This is a spoken corpus, so the segmentation principles for speech are used as reference.	

Procedure	Work Content	Operators (Number, Professional Competency Requirements)	Hardware (Name, Version, Price)	Software (Name, Version, Price)	Standards (Technical standards, product specifications, quality requirements, etc.)	Time	Conclusion (Difficulties, Flaws, Features, etc.)	Cost Estimation
12	Manual part-of-speech tagging (separating sentences into terms with individual meaning and serve specific syntactic functions)	Multiple research assistants (possess background knowledge of linguistics)	Desktop computer (NT50,000 each)	(1) Microsoft OS: 98/2000/XP (2) Microsoft Office: Word/Excel	According to "A Segmentation Standard for Chinese Information Processing" of the Association for Computational Linguistics and Chinese Language Processing.			
13	Manual IPA representation (Adopts phonetic transcription. Segments are represented using Unicode IPA. Five degree phonetic symbols are used to represent tones.	Multiple Research assistants (Familiar with computer word processing; familiar with IPA and phonetics)	Desktop computer (NT50,000 each) Minidisc Portable Audio Recorder (NT12,000 each) Professional earphone (NT3,000 each)	(1) CHAT transcription platform of CHILDES (2) Microsoft OS: 98/2000/XP (3) Microsoft Office: Word/Excel (4) Unicode IPA font software	CHILDES Handbook of the International Phonetic Association (1999)	Each hour requires roughly 4.5 hours to transcribe. Total time: 330 hours×4.5 = 1,485 hours.		

Procedure	Work Content	Operators (Number, Professional Competency Requirements)	Hardware (Name, Version, Price)	Software (Name, Version, Price)	Standards (Technical standards, product specifications, quality requirements, etc.)	Time	Conclusion (Difficulties, Flaws, Features, etc.)	Cost Estimation
14	Create word list (A list of all words in the transcribed text file is created, and then manually checked whether the Chinese characters are consistent with dictionaries.)	Multiple research assistants (possess background knowledge of linguistics)	Desktop computer (NT50,000 each)	(1) Microsoft OS: 98/2000/XP (2) Microsoft Office: Word/Excel				
15	Manual phonetic representation	Multiple research assistants (possess background knowledge of linguistics)	Desktop computer (NT50,000 each)	(1) Microsoft OS: 98/2000/XP (2) Microsoft Office: Word/Excel	Phonetic symbols used to represent Chinese characters are letters from the Roman alphabet promulgated by the Ministry of Education in 1998			

Procedure	Work Content	Operators (Number, Professional Competency Requirements)	Hardware (Name, Version, Price)	Software (Name, Version, Price)	Standards (Technical standards, product specifications, quality requirements, etc.)	Time	Conclusion (Difficulties, Flaws, Features, etc.)	Cost Estimation
16	Establish the part-of-speech tagging system	Multiple assistants (possess basic knowledge of syntax and semantics)	Desktop computer (NT50,000 each)	Microsoft OS: 98/2000/XP Microsoft Office: Word/ Excel "Southern Min Corpus"	CKIP "Principles for Part-of-speech Tagging" CANCORP: The Hong Kong Cantonese Child Language Corpus, Lee and Wong (1998). A Study on Part-of-speech Tagging for Taiwan Southern Min" Tsao Feng-Fu (1996).		We adopted parts of speech tags of CKIP, but only 46 simplified tags. This is to avoid subjectively forced categorization when categories for parts of speech are too detailed.	
17	Manual part-of-speech tagging	Multiple assistants (possess basic knowledge of syntax and semantics)	Desktop computer (NT50,000 each)	Microsoft OS: 98/2000/XP Microsoft Office: Word/ Excel (3) "Southern Min Corpus"				

Procedure	Work Content	Operators (Number, Professional Competency Requirements)	Hardware (Name, Version, Price)	Software (Name, Version, Price)	Standards (Technical standards, product specifications, quality requirements, etc.)	Time	Conclusion (Difficulties, Flaws, Features, etc.)	Cost Estimation
18	Develop the automatic segmentation and phonetic representation program	1 programmer (familiar with programming languages; basic knowledge of linguistics)	Desktop computer (NT50,000 each)	"Southern Min Corpus" Linux Operating System Visual C	Self developed		Besides segmentation and phonetic representation, this program also adds new words to the corpus.	Programming fee (Case payment)
19	Develop the Chinese character inspection program	1 programmer (familiar with programming languages; basic knowledge of linguistics)	Desktop computer (NT50,000 each)	(1) "Southern Min Corpus" (2) Linux Operating System (3) Visual C	Self developed			Programming fee (Case payment)
20	Develop the automatic part-of-speech tagging program	1 programmer (familiar with programming languages; basic knowledge of linguistics)	Desktop computer (NT50,000 each)	(1) Linux Operating System (2) Visual C	Self developed		Based on the part of speech tag in the "Southern Min Corpus"	Programming fee (Case payment)

Procedure	Work Content	Operators (Number, Professional Competency Requirements)	Hardware (Name, Version, Price)	Software (Name, Version, Price)	Standards (Technical standards, product specifications, quality requirements, etc.)	Time	Conclusion (Difficulties, Flaws, Features, etc.)	Cost Estimation
21	Manual Inspection (Check if automatic part-of-speech tagging is correct)	Several research assistants (familiar with word processing; segmentation and part-of-speech tagging)	Desktop Computer (NT50,000 each)	(1) Part-of-speech tagging program (2) Min Corpus (3) Microsoft OS: (4) 98/2000/XP (5) Microsoft Office: Word/ Excel CHAT transcription platform of CHILDES			Manual inspection – Checks if automatic part-of-speech tagging is correct; requires immense manpower.	

Procedure	Work Content	Operators (Number, Professional Competency Requirements)	Hardware (Name, Version, Price)	Software (Name, Version, Price)	Standards (Technical standards, product specifications, quality requirements, etc.)	Time	Conclusion (Difficulties, Flaws, Features, etc.)	Cost Estimation
22	Website Structure and Content Compilation (Website contents and interface discussed with research assistants)	Project Director Research Assistants	Desktop Computer (NT50,000 each)	(1) Microsoft FrontPage (2) Microsoft OS: 98/2000/XP (3) Microsoft Office: Word/ Excel	(1) CHILDES (2) CANCORP		Website contents include corpus introduction, database, user manual, applications and links to related websites.	
23	Website construction and maintenance (Build a dedicated website for the corpus so that it can be used for research by scholars around the world)	1-2 Computer technicians (computer and information related background and programming capabilities)	Desktop Computer – Server (NT50,000 each)				The website will be released to the public once final tests are complete.	1 Computer technician (compensation depends contents of work)



SIX. Conclusions

In terms of corpus applications, linguists are concerned about presenting the original appearance of a language, while computer scientists hope to organize and structuralize corpus data, and then use database technology to satisfy needs of different users. Therefore, the preservation of languages in digital form must overcome technical issues to rid itself of its physical form, paper and pen, and must rely on relational database technology to be achieved.

The development of a corpus in written form into a relational database indicates increased complexity, but is rewarded by more effective data utilization and higher operability. At a closer look, the increased complexity is not real, different data sheets that are linked together can be considered as the relations between linguistic knowledge. A database focuses on designing data fields, records, and data sheets, which is like linguistics, presenting words, sentences and compositions, and creating effective links between them.

Corpora introduced in this book utilize modern data storage and extraction technology, and convert source corpora into databases with computer data structures. In which relational database structure and specifications are used to not only provide more exact definitions of corpus data, but also make links between different data entries clearer.



References

Books

沈漢聰，〈數位典藏技術彙編〉電子書，數位典藏國家型科技計畫，2004年，初版。

van Son, R., Wesseling, W., Sanders, E., and van den Heuvel, H. (2009). "Promoting free Dialog Video Corpora: The IFADV Corpus Example," in M. Kipp et al. (Eds.): *Multimodal Corpora*, LNAI 5509, pp. 18–37, 2009.

Journal Papers

陳心渝、李政宏、邱一航、林韋伶，〈數位版權管理機制實作—以數位典藏管理系統為例〉，《第四屆典藏技術研討會論文集》，2005年9月，頁93~100。

蕭人豪、林欣慧、林金龍、林麗虹，〈數位浮水印技術發展現況：以典藏計畫為例〉，《第三屆數位典藏技術研討會》2004年8月，頁163~169。

詞庫小組，〈研究院語料庫的內容及說明〉，中文詞知識庫小組技術報告 #95-02，南港，中央研究院，1995年。

Chu-Ren Huang and Keh-jiann Chen. A Chinese Corpus for Linguistics Research. In the Proceedings of the 1992 International Conference on Computational Linguistics (COLING-92). 1214-1217. Nantes, France. 1992.

Chu-Ren Huang Corpus-based Studies of Mandarin Chinese: Foundational Issues and Preliminary Results. In Matthew Chen and Ovid Tzeng Eds. In Honor of William S-Y. Wang: *Interdisciplinary Studies on Language and Language Change*. Pp. 165-186. Taipei: Pyramid. 1994.

Hsin-Min Wang, Berlin Chen, Jen-Wei Kuo and Shih-Sian Cheng. "MATBN: A Mandarin Chinese Broadcast News Corpus," *Computational Linguistics and Chinese Language Processing* 10.2, pp.219-236. 2005.

Online Resources

李道明，〈影音檔案數位化之規劃與流程〉，拓展台灣數位典藏計畫網站，2009年1月31日下載，<http://content.ndap.org.tw/index/?p=843>。

Steinebach, M., J. Dittmann & E. Neuhold "Digital Watermarking - Common watermarking techniques, Important Parameters, Applied mechanisms, Applications, Invertible watermarking, Content-fragile watermarking," <http://encyclopedia.jrank.org/articles/pages/6725/Digital-Watermarking.html>，2010年1月27日下載。



Appendix

Online Resources for Linguistic Corpora Building:

1. Chinese Wordnet : <http://cwn.ling.sinica.edu.tw/> ◦
2. Chinese Word Sketch : <http://bow.sinica.edu.tw/> ◦
3. Balanced Corpus of Modern Chinese : <http://dbo.sinica.edu.tw/SinicaCorpus/> ◦
4. Linguistics Anchoring : <http://linganchor.sinica.edu.tw/> ◦
5. Sinica BOW : <http://bow.sinica.edu.tw/> ◦
6. TSL Research Group : <http://tsl.ccu.edu.tw/web/index.php> ◦
7. Formosan Language Archive : <http://formosan.sinica.edu.tw/> ◦
8. Roland Taiwan Enterprise Co., Ltd. : <http://www.rolandtaiwan.com.tw/roland/index.php> ◦
9. Digital Resources Center for Global Chinese Teaching and Learning : <http://elearning.ling.sinica.edu.tw/> ◦
10. The CANDLE Project for Reading : <http://elearning.eng.ntnu.edu.tw/CANDLE/> ◦
11. NCKU Eagle Project : <http://english.ncku.edu.tw/> ◦
12. Archives and Linguistic Representations of Spoken Taiwan Mandarin : <http://mmc.sinica.edu.tw/> ◦
13. Southern Min Archives: A Database of Historical Changes and Language Distribution : <http://southernmin.sinica.edu.tw/> ◦
14. Min and Hakka Language Archives : <http://minhakka.ling.sinica.edu.tw/bkg/index.php> ◦
15. Metadata Architecture and Application Team : <http://metadata.teldap.tw/index.html> ◦
16. E-books of Technical Assembly of Digital Archives : <http://www2.ndap.org.tw/eBook/showContent.php> ◦
17. AHDS Literature, Languages and Linguistics : <http://www.ahds.ac.uk/litlangling/index.htm> ◦
18. Brown Corups Manual : <http://icame.uib.no/brown/bcm.html> ◦
19. CGN--The Spoken Dutch Corpus project : http://www.tst.inl.nl/cgndocs/doc_English/topics/index.htm ◦
20. DoBES: Documentation of Endangered Languages : <http://www.mpi.nl/DOBES/> ◦
21. Dublin Core Metadata Initiative : <http://dublincore.org/> ◦
22. ELAN : <http://www.lat-mpi.eu/tools/elan/> ◦

23. GNU Operating System : <http://www.gnu.org/> ◦
24. GeoLang: The prime sponsor of the World Language Documentation Cengre : <http://www.geolang.net/> ◦
25. HCRC Map Task Corpus : <http://www.hcrc.ed.ac.uk/maptask/> ◦
26. IBM Taiwan : <http://www.ibm.com/tw/zh/> ◦
27. IFA Dialog Corpus : <http://www.fon.hum.uva.nl/IFA-SpokenLanguageCorpora/IFADVcorpus/> ◦
28. IMDI: ISLE Meta Data Initiative : <http://www.mpi.nl/IMDI/> ◦
29. ISO-1 Registration Authority : http://www.infoterm.info/standardization/iso_639_1_2002.php ◦
30. ISO 639-2 Registration Authority : <http://www.loc.gov/standards/iso639-2/> ◦
31. ISO 639-3 Registration Authority : <http://www.sil.org/iso639-3/> ◦
32. ISO 639-5 Registration Authority : <http://www.loc.gov:8081/standards/iso639-5/> ◦
33. Open Language Archives Community : <http://www.language-archives.org> ◦
34. OLAC Metadata Set : <http://www.language-archives.org/OLAC/olacms.html#Metadata elements> ◦
35. SIL International : <http://www.sil.org/> ◦
36. SMIL: W3C Synchronized Multimedia Integration Language : <http://www.w3.org/AudioVideo/> ◦
37. TEI: Text Encoding Initiative : <http://www.tei-c.org/index.xml> ◦
38. The Odrl Initiative : <http://odrl.net/> ◦
39. Unicode Standard : <http://www.unicode.org/standard/standard.html> ◦
40. Wynne, M (editor). 2005. Developing Linguistic Corpora: a Guide to Good Practice. Oxford: Oxbow Books. Available online from <http://ahds.ac.uk/linguistic-corpora/> [2010-01-08 Downloaded] ◦
41. XML: W3c Extensible Markup Language : <http://www.w3.org/TR/2006/REC-xml11-20060816/> ◦
42. XrML-The Digital Rights Language for Trusted Content and Services : <http://www.xrml.org/about.asp> ◦

A Starter's Guide to Linguistic Corpora Building

Advisory Unit: National Science Council, Executive Yuan

Issuer: Fu-Shih Lin

Editor-in-Chief: Peng-Sheng Chiu

Executive Editors: Yen-Hung Lin, Ting-Li Lin, Fang-Chih Lin, Lang-Hsuan Kao

Authors: Pei-Ying Li, Chih-Ming Chiu, Huo-Tsen Kuo, Shu-Chuan Tseng, Chu-Fang Huang, Ching-Hsun Chan, Su-Chuan Tsai, Chiu-Jung Lu, Su-Ying Hsiao, Chia-Min Lai, Hao-Yi Tai, Pei-Yu Hsieh, Hsiu-Fang Su, Chinese WordNet Group

Reviewer: Assistant Research Fellow Su-Ying Hsiao, Institute of Linguistics, Academia Sinica

Translator: International Collaboration and Promotion of Taiwan e-learning and Digital Archives Project

Publisher: Taiwan e-Learning and Digital Archives Program Taiwan Digital Archives Expansion Project

Address: No.128, Sec.2, Academia Rd., Nankang District, Taipei City, 115 Institute of History & Philology, Academia Sinica

Tel: +886-2-2782-9555 extn.288

Fax: +886-2-2786-8834

Website: <http://content.teldap.tw>

Email: content@gate.sinica.edu.tw

Typesetting: Yu-Mu Hsiao

Printing: Evergreen International Corporation

Publish Date: 1st Edition May 2011

ISBN: 978-986-02-7743-2

All Rights Reserved, Not for Sale

本書譯自拓展臺灣數位典藏計畫出版

數位化工作流程指南：語料庫建置入門