1.	Record Nr.	UNINA9910770258703321
	Titolo	Chinese Language Resources : Data Collection, Linguistic Analysis, Annotation and Language Processing / / Chu-Ren Huang, Shu-Kai Hsieh, and Peng Jin, editors
	Pubbl/distr/stampa	Cham, Switzerland : , : Springer, , [2023] ©2023
	ISBN	3-031-38913-1
	Edizione	[First edition.]
	Descrizione fisica	1 online resource (0 pages)
	Collana	Text, Speech and Language Technology Series ; ; Volume 49
	Disciplina	495.1072
	Soggetti	Chinese language - Data processing Computational linguistics
	Lingua di pubblicazione	Inglese
	Formato	Materiale a stampa
	Livello bibliografico	Monografia
	Nota di bibliografia	Includes bibliographical references.
	Nota di contenuto	Intro -- Biography of Prof. Shiwen Yu -- A Chronological Biography of Professor Shiwen Yu -- Acknowledgments -- Contents -- Editors and Contributors -- Part I: Overview -- Chapter 1: Chinese Language Resources Through One-Third of a Century -- 1.1 Headwater -- 1.2 Vision: Of Peaks and Giants -- 1.3 From the Great Mountains Long Streams Flow -- 1.4 The Versatility of Language Resources -- 1.5 Giving Shape to Water -- 1.6 Deriving Sharable and Versatile Knowledge -- 1.7 The Power of Language Data as Water -- 1.8 Conclusion and Dedication -- References -- Chapter 2: Chinese Comprehensive Language Knowledge Base -- 2.1 Why Was the Chinese Language Knowledge Base Constructed? -- 2.2 Cornerstone of the CLKB: Grammatical Knowledge Base of Contemporary Chinese -- 2.3 Profile of the CLKB -- 2.3.1 PSKB -- 2.3.2 BPTC -- 2.4 What Was Learned from the Development of the CLKB? -- 2.4.1 Fundamental Research and Application Research -- 2.4.2 Theoretical Research and Engineering Practices -- 2.4.3 Development Goals and Process Monitoring -- 2.4.4 Balance of Scale and Quality -- 2.5 Conclusion -- References -- Chapter 3: Introduction to CKIP´s Language Resources and Their Applications -- 3.1 Background -- 3.2 Language Resources -- 3.2.1 Chinese Writing System Resources -- Database of Component Parts of Chinese Characters -- Hantology -- 3.2.2 Lexical Databases

and Grammar -- CKIP Lexical Knowledge Base -- Information-Based Case Grammar -- 3.2.3 Corpora -- Sinica Chinese Corpus -- Sinica Ancient Chinese Corpus -- Sinica Treebank -- Language Resources Derived from the Sinica Corpus -- 3.2.4 WordNet and Ontologies -- Bilingual Ontological WordNet -- Chinese WordNet -- Extended-HowNet -- 3.2.5 Integrated Resources -- Chinese Sketch Engine -- 3.3 Core Tools in Chinese Language Processing -- 3.3.1 Word Segmentation and POS Tagging.

3.3.2 Part-of-Speech Tagging -- 3.3.3 Parsing -- 3.3.4 Automatic Semantic Role Assignment -- 3.4 Applications of CKIP´s Resources -- 3.4.1 Word Segmentation and Part-of-Speech Tagging Using YamCha and CRF++ -- 3.4.2 Viterbi PCFG Parser, Syntactic Complexity, and Chinese Readability -- 3.4.3 Chinese Dependency Parser -- 3.4.4 Chinese Dependency Relations Database and Lexicology -- 3.4.5 Selectional Preferences -- 3.4.6 Unsupervised and Minimally Supervised Approaches to Word Sense Disambiguation -- 3.4.7 Vector Semantics and Deep Neural Net -- 3.5 Conclusion: Interdisciplinary Impact and Future Research -- References -- Part II: Language Resources: Annotation and Processing -- Chapter 4: Practical and Robust Chinese Word Segmentation and PoS Tagging -- 4.1 Introduction -- 4.2 Word Boundary Detection Model Robust Word Segmentation -- 4.2.1 From Word Identification to Boundary Decision -- 4.2.2 Word Boundary Decision (WBD) -- 4.3 From Online Learning to Active Learning -- 4.3.1 Online Semi-supervised Learning with Labeled Data -- 4.3.2 Online Semi-supervised Learning with Unlabeled Data -- 4.3.3 Performances of WBD -- 4.3.4 Results of Online Learning with Unlabeled Data -- 4.3.5 Active Learning Approach for CWS: Meeting the Three Challenges -- 4.4 Robustness of PoS Tagging and Quality Assurance: A Two-Tagset Model -- 4.4.1 Corpus-Based POS Tag Mapping -- 4.5 Linguistic Ramification -- 4.6 Conclusion: The Convergence of Linguistic and Stochastic Modeling -- References -- Chapter 5: Describing the Grammatical Knowledge of Chinese Words for Natural Language Processing -- 5.1 Introduction -- 5.2 Overall Design of the Knowledge Base -- 5.2.1 Databases -- 5.2.2 Selection of Words -- 5.3 Classification of Words -- 5.3.1 Basic Word Classes -- 5.3.2 Purpose of Word Classification -- 5.3.3 Word Class Definitions by Grammatical Functions.

5.3.4 Multi-class Words, Homographs, and Homonyms -- 5.4 Description of Grammatical Properties -- 5.4.1 Selection of Grammatical Attributes -- Morphological Attributes -- Syntactic Attributes -- Semantic Attributes -- Collocation -- 5.4.2 Data Redundancy -- 5.4.3 Value Types -- 5.5 Semantic Considerations in the GKB -- 5.5.1 Word Entries Distinguished by Their Meanings -- 5.5.2 Semantic Properties Described for Word Entries -- 5.5.3 Grammatical Properties Distinguished Based on Semantic Clues -- 5.6 Conclusion -- References -- Chapter 6: DeepLEX -- 6.1 Introduction -- 6.2 Current Approaches and Issues -- 6.2.1 Chinese Lexical Resources -- 6.3 DeepLEX -- 6.3.1 Modules -- 6.3.2 Fluid Annotation -- 6.4 Conclusion -- References -- Chapter 7: The Chinese Generalized Function Word Usage Knowledge Base and Its Applications -- 7.1 Introduction -- 7.2 Chinese Function Word Usage Knowledge Base -- 7.2.1 Framework and Construction Process of the CFKB -- 7.2.2 Function Word Usage Dictionary -- 7.2.3 Function Word Usage Rule Base -- 7.2.4 Function Word Usage Corpus -- 7.3 Automatic Identification of Function Word Usages -- 7.3.1 Rule-Based Method -- 7.3.2 Statistics-Based Method -- 7.3.3 Combined Rule-Based/Statistics-Based Method -- 7.4 Applications Based on the CFKB -- 7.4.1 Syntactic Analysis -- 7.4.2 Grammar Error Analysis -- 7.4.3

Information Extraction -- 7.4.4 Chinese Deep Semantic Understanding -- 7.5 Conclusion -- References -- Chapter 8: A Generic Study of Linguistic Information Based on the Chinese Idiom Knowledge Base and Its Expansion -- 8.1 Introduction -- 8.2 Construction, Structure, and Properties of the Knowledge Base for Chinese Idiomatic Expressions -- 8.3 Nature, Living, History, Culture, and Idiomatic Expressions -- 8.4 Conclusion -- References -- Chapter 9: Lexical Knowledge Representation and Semantic Composition of E-HowNet.

9.1 Introduction -- 9.2 Overview of E-HowNet -- 9.2.1 Ontology of Concepts -- 9.3 Lexical Knowledge Representation and Semantic Composition -- 9.3.1 Principles of Sense Definitions -- 9.3.2 Uniform Representation of Content Words and Function Words for Semantic Composition -- 9.3.3 Basic Composition Process -- 9.4 Semantic Role Labeling -- 9.4.1 Establishing a Reasonable Set of Semantic Roles -- 9.4.2 Guidelines for Pursuing Role Assignment -- 9.4.3 Difficulties and Solutions -- 9.5 Conclusion and Future Work -- References -- Chapter 10: Sense Tagging Unknown Chinese Words with Word Embedding -- 10.1 Introduction -- 10.2 Related Work -- 10.3 Linguistic Features of Unknown Words -- 10.3.1 Paradigmatic Features of Words -- 10.3.2 Features of Chinese Word Formation -- 10.4 Introduction of the Semantic Resource -- 10.5 Model Construction -- 10.5.1 Model Based on Word Embedding -- 10.5.2 Combined Model Based on Word Embedding and POS Filtering -- 10.5.3 Model Based on Word Embedding, POS Filtering, and Suffix Filtering -- 10.6 Experiments -- 10.6.1 Evaluation Metrics -- 10.6.2 Experimental Setting -- 10.6.3 Experiments and Analysis -- Setting the Size of Related Words´´ K Values -- Results and Analysis -- 10.7 Multimodel Cascade -- 10.8 Conclusion -- References -- Chapter 11: PKUSenseCor: A Large-Scale Word Sense Annotated Chinese Corpus -- 11.1 Introduction -- 11.2 Corpus and Knowledge Base Selection -- 11.2.1 Corpus -- 11.2.2 Sense Inventory: The Grammatical Knowledge Base of Contemporary Chinese -- 11.3 Corpus Annotation -- 11.3.1 The Annotation Process -- 11.4 Inter-annotator Agreement -- 11.5 Conclusion -- References -- Chapter 12: Semantic Annotation and Mandarin VerbNet -- 12.1 Introduction -- 12.2 Issues in the Annotation of Chinese Verbs -- 12.3 Frame-Based Constructional Approach to Semantic Annotation -- 12.3.1 Emotion Archi-frames.

12.3.2 Motion Archi-frames -- 12.4 Advantages of the Approach -- 12.5 Conclusion -- References -- Chapter 13: The Construction of a Chinese Semantic Dependency Graph Bank -- 13.1 Introduction -- 13.2 Annotation Scheme of the Semantic Dependency Graph -- 13.2.1 Graph Structure of Semantic Dependency -- 13.2.2 Semantic Relation Set -- 13.2.3 Special Situations -- 13.3 Corpus -- 13.3.1 Corpus Origin -- 13.3.2 Annotation Tool -- 13.4 Evaluation of the Corpus -- 13.5 Corpus Statistics -- 13.6 Conclusion -- References -- Chapter 14: A Chinese Dialogue Corpus Annotated with Dialogue Act -- 14.1 Introduction -- 14.2 Related Work -- 14.2.1 Existing Datasets -- 14.2.2 Traditional Methods -- 14.2.3 Deep Learning Models -- 14.3 Annotation of a Group Chat Corpus -- 14.3.1 Data Collection and Preprocessing -- 14.3.2 Annotation Specification -- Dimension 1 (D1): Semantic Information -- Dimension 2 (D2): Reaction to Context -- Dimension 3 (D3): Effect of Turn-Taking on Topic -- 14.3.3 Annotated Example -- 14.3.4 Consistency Check -- 14.3.5 Dataset Statistics -- 14.3.6 Confusion Matrix -- 14.4 Baseline Methods -- 14.4.1 Dataset -- 14.4.2 Evaluation Metrics -- 14.4.3 Conditional Random Field (CRF) Model -- 14.4.4 Recurrent Neural Networks (RNN) -- 14.5 Experimental Results -- 14.5.1 Human Performance -- 14.5.2 Model Performance -- Random Selection -- CRF Model -- RNN Model -- 14.6

Conclusions -- References -- Chapter 15: Automatic Construction of Parallel Dialogue Corpora with Rich Information -- 15.1 Introduction -- 15.2 Related Work -- 15.3 Building a Parallel Dialogue Corpus -- 15.3.1 Script and Subtitle -- 15.3.2 Movie Alignment -- 15.3.3 Sentence Alignment -- 15.4 Experiments and Results -- 15.4.1 Parallel Dialogue Corpus Construction -- 15.4.2 Improved Translation with Speaker Information -- 15.5 Conclusion and Future Work -- References.

Chapter 16: A Chinese Event-Based Emotion Corpus: Emotion Cause Detection.