A Computational Theory of Writing Systems

This book breaks new ground by developing, for the first time, a formal computational theory of writing systems. It offers specific proposals about the nature of the linguistic objects that are represented by orthographic elements; what levels of linguistic representation are involved and how they may differ across writing systems; and what formal constraints hold of the mapping relation between linguistic and orthographic elements. Sproat demonstrates that this mapping relation is regular in the technical sense that it can be implemented computationally using the simplest class of computational devices, namely finite-state machines. He also argues that the level of linguistic representation reflected in written forms is always consistent within a given writing system. This level may be “shallow” (the language is “spelled as it sounds”) or “deep,” but in any case it extends over the entire vocabulary. Based on his insights as to what linguistic elements can be represented in writing, Sproat proposes a new taxonomy of writing systems. The treatment of theoretical linguistic issues and their computational implementation is complemented with discussion of empirical psycholinguistic work on reading and its relevance for the computational model developed here. Throughout, the model is illustrated with a number of detailed case studies of writing systems around the world.

This innovative book will be of interest to students and researchers in a variety of fields, including theoretical and computational linguistics, the psycholinguistics of reading and writing, and speech technology.

Richard Sproat has done published research in numerous areas of linguistics and computational linguistics. He is currently a Technology Consultant in the Human/Computer Interface Department at AT&T Labs – Research.
Studies in Natural Language Processing

Series Editor: Branimir Boguraev, IBM T. J. Watson Research

Editorial Advisory Board
Don Hindle, AT&T Labs – Research
Martin Kay, Xerox PARC
David McDonald, Content Technologies
Hans Uszkoreit, University of Saarbrücken
Yorick Wilks, Sheffield University

Also in the series
Douglas E. Appelt, Planning English Sentences
Madeleine Bates and Ralph M. Weischedel (eds.), Challenges in Natural Language Processing
Steven Bird, Computational Phonology
Peter Bosch and Rob van der Sandt (eds.), Focus
T. Briscoe, Ann Copestake, and Valeria Paiva (eds.), Inheritance, Defaults and the Lexicon
Ronald Cole, Joseph Mariani, Hans Uszkoreit, Giovanni Varile, Annie Zaenen, Antonio Zampolli, and Victor Zue (eds.), Survey of the State of the Art in Human Language Technology
David R. Dowty, Lauri Karttunen, and Arnold M. Zwicky (eds.), Natural Language Parsing
Ralph Grishman, Computational Linguistics
Graeme Hirst, Semantic Interpretation and the Resolution of Ambiguity
András Kornai, Extended Finite State Models of Language
Kathleen R. McKeown, Text Generation
Martha Stone Palmer, Semantic Processing for Finite Domains
Terry Patten, Systemic Text Generation as Problem Solving
Manny Rayner, David Carter, Pierreette Bouillon, Vassilis Digalakis, and Mats Wirén, The Spoken Language Translator
Ehud Reiter and Robert Dale, Building Natural Language Generation Systems
Michael Rosner and Roderick Johnson (eds.), Computational Linguistics and Formal Semantics
Patrick Saint-Dizier and Evelyn Viegas (eds.), Computational Lexical Semantics
A Computational Theory of Writing Systems

Richard Sproat

AT&T Labs – Research
For Lisa, who is learning to read
Contents

List of Figures                                      page xiii
List of Tables                                       xv
Preface                                              xvii

1 Reading Devices                                     1
1.1 Text-to-Speech Conversion: A Brief Introduction  2
1.2 The Task of Pronouncing Aloud: A Model           6
  1.2.1 A Simple Example from Russian                 6
  1.2.2 Formal Definitions                            11
    1.2.2.1 AVMs and Annotation Graphs                 11
    1.2.2.2 Definitions                                13
    1.2.2.3 Axioms                                    14
  1.2.3 Central Claims of the Theory                  15
    1.2.3.1 Regularity                                16
    1.2.3.2 Consistency                               19
  1.2.4 Further Issues                                20
    1.2.4.1 Why a Constrained Theory of Writing Systems?  21
    1.2.4.2 Orthography and the “Segmental” Assumption  23
1.3 Terminology and Conventions                      25
1.A Appendix: An Overview of Finite-State Automata and  29
  Transducers                                         29
   1.A.1 Regular Languages and Finite-State Automata  29
   1.A.2 Regular Relations and Finite-State Transducers  30

2 Regularity                                          34
2.1 Planar Regular Languages and Planar Regular      35
  Relations                                           35
2.2 The Locality Hypothesis                           41
2.3 Planar Arrangements: Examples                     42
  2.3.1 Korean Hankul                                  43
  2.3.2 Devanagari                                     45
## Contents

2.3.3 Pahawh Hmong 47  
2.3.4 Chinese 48  
2.3.5 A Counterexample from Ancient Egyptian 54  
2.4 Cross-Writing-System Variation in the SLU 55  
2.5 Macroscopic Catenation: Text Direction 59  
2.6 Sample Chinese Characters and Their Analyses 62  

### 3 ORL Depth and Consistency

3.1 Russian and Belarusian Orthography: A Case Study 67  
3.1.1 Vowel Reduction 68  
3.1.2 Regressive Palatalization 73  
3.1.3 Lexical Marking in Russian and Other Issues 76  
3.1.4 Summary of Russian and Belarusian 79  
3.2 English 79  
3.3 The Orthographic Representation of Serbo-Croatian Consonant Devoicing 89  
3.3.1 Methods and Materials 91  
3.3.2 Results 92  
3.4 Cyclicity in Orthography 95  
3.5 Surface Orthographic Constraints 96  
3.6 English Deep and Shallow ORLs 99  
3.6.1 Lexical Representations 99  
3.6.2 Rules for the Deep ORL 127  
3.6.3 Rules for the Shallow ORL 129  

### 4 Linguistic Elements

4.1 Taxonomies of Writing Systems: A Brief Overview 132  
4.1.1 Gelb 132  
4.1.2 Sampson 133  
4.1.3 DeFrancis 134  
4.1.3.1 No Full Writing System Is Semasiographic 134  
4.1.3.2 All Full Writing Is Phonographic 135  
4.1.3.3 Hankul Is Not Featural 135  
4.1.4 A New Proposal 139  
4.1.5 Summary 143  
4.2 Chinese Writing 144  
4.3 Japanese Writing 154  
4.4 Some Further Examples 160  
4.4.1 Syriac Syame 160  
4.4.2 Reduplication Markers 161  
4.4.3 Cancellation Signs 162
Contents

5 Psycholinguistic Evidence 163
  5.1 Multiple Routes and the Orthographic Depth Hypothesis 165
    5.1.1 Evidence for the Orthographic Depth Hypothesis 168
    5.1.2 Evidence against the Orthographic Depth Hypothesis 170
  5.2 “Shallow” Processing in “Deep” Orthographies 171
    5.2.1 Phonological Access in Chinese 171
    5.2.2 Phonological Access in Japanese 173
    5.2.3 Evidence for the Function of Phonetic Components in Chinese
        175
    5.2.4 Summary 176
  5.3 Connectionist Models: The Seidenberg-McClelland Model 178
    5.3.1 Outline of the Model 179
    5.3.2 What Is Wrong with the Model? 181
  5.4 Summary 184

6 Further Issues 185
  6.1 Adaptation of Writing Systems: The Case of Manx Gaelic 185
  6.2 Orthographic Reforms: The Case of Dutch 192
    6.2.1 The 1954 Spelling Rules 193
    6.2.2 The 1995 Spelling Rules 195
  6.3 Other Forms of Notation: Numerical Notation and Its Relation to Number Names 197
  6.4 Abbreviatory Devices 204
  6.5 Non-Bloomfieldian Views on Writing 209
  6.6 Postscript 212

Bibliography 217
Index 227
List of Figures

1.1 A partial linguistic representation for the sentence in (1.1). Shown are a phonetic transcription, a prosodic analysis into two intonational phrases (i) and one utterance (U), accent assignment (*), a set of part-of-speech tags, and a simple phrase-structure analysis. Phonetic symbols are IPA. Note that 'MP' means 'measure phrase'.

1.2 A simple FST implementing the rewrite rules in (1.5). In this example the machine has a single state (0), which is both an initial and a final state. The labels on the individual arcs consist of an input label (to the left of the colon) and an output label (to the right). Here, capital Roman letters are used to represent the equivalent Cyrillic letters.

1.3 An acceptor for ab*. The heavy-circled state (0) is (conventionally) the initial state, and the double-circled state is the final state.

1.4 An FST that accepts a:b (b:a)*.

1.5 Three transducers, where T3 = T1 o T2.

2.1 \( \gamma(\alpha) \triangleright \gamma(\beta) \triangleright \gamma(\zeta) \triangleright \gamma(\delta) \).

2.2 Another figure described by \( \gamma(\alpha) \triangleright \gamma(\beta) \triangleright \gamma(\zeta) \triangleright \gamma(\delta) \).

2.3 The five planar concatenation operations: (a) \( \gamma(\alpha) \triangleright \gamma(\beta) \), (b) \( \gamma(\alpha) \triangleright \gamma(\beta) \), (c) \( \gamma(\alpha) \triangleright \gamma(\beta) \), (d) \( \gamma(\alpha) \triangleright \gamma(\beta) \); (e) \( \gamma(\alpha) \triangleright \gamma(\beta) \).

2.4 A 2FSA that recognizes Figure 2.2. The labels “R” and “D” on the arcs denote reading direction; “left” on state 0 (the initial state) denotes the position at which scanning begins.

2.5 A 2FST that maps the expression \( \alpha \beta \zeta \delta \) to \( \gamma(\alpha) \triangleright \gamma(\beta) \triangleright \gamma(\zeta) \triangleright \gamma(\delta) \).

2.6 The syllables <mos/> /mo/ ‘cannot’ and <cal/> /čal/ ‘well’ in Hankul.

2.7 A two-word Mayan glyph representing the adjective sequence yax ch’ul ‘green sacred’, after Macri (1996, ...
List of Figures

page 179), with the arrangement Y AX \( \rightarrow \) [ch'uv'it]. Per convention, the capitalized gloss represents a logographic element, and the lower case glosses phonographic elements.

2.8 Three methods of wrapping the virtual tape: (a) standard non-boustrophedon; (b) boustrophedon; (c) inverted boustrophedon.

3.1 Waveform and voicing profile for one utterance of gradski 'urban'. The closure for the /d/ is labeled as "dcl", the voicing offset is labeled as "v", and the start of the /s/ is labeled as "s". The voicing profile is the third plot from the top in the second window, labeled as prob_voice.

3.2 Barplot showing the proportions of voicing for all samples of underlying /d/ (black bars) versus /t/ (shaded bars).

4.1 The taxonomy of Gelb (1963), along with examples of writing systems that belong to each case.

4.2 The taxonomy of Sampson (1985).

4.3 Featural representation of Korean Hankul, from Sampson (1985, page 124, Figure 19). (Presented with permission of Geoffrey Sampson.)

4.4 DeFrancis’s classification of writing systems.

4.5 A nonarboreal classification of writing systems. On Perso-Aramaic, see Section 6.1.

5.1 A model of reading a word aloud, simplified from Besner and Smith (1992).

5.2 The Seidenberg and McClelland model of lexical processing (Seidenberg and McClelland, 1989, page 526, Figure 1). Used with permission of the American Psychological Association, Inc.

5.3 The implemented portion of Seidenberg and McClelland’s model of lexical processing (Seidenberg and McClelland, 1989, page 527, Figure 2). Used with permission of American Psychological Association, Inc.

5.4 Replication of the Seidenberg, McRae, and Jared (1988) experiment, from Seidenberg and McClelland (1989, page 545, Figure 19). Used with permission of American Psychological Association, Inc.

6.1 A numeral factorization transducer for numbers up to 999.

6.2 Two-dimensional layout in an e-mail signature.
List of Tables

2.1 Chinese characters illustrating the five modes of combination of semantic (underlined) and phonetic components. 35
2.2 Full and diacritic forms for Devanagari vowels, classified by the position of expression of the diacritic forms. Thus “after” means that the diacritic occurs after the consonant cluster, “below,” below it, and so forth. 46
2.3 Distribution of placement of semantic component (Kang-Xi Radical) among 12,728 characters from the Taiwan Big5 character set. 51
2.4 Distribution of placement of semantic component (Kang-Xi Radical) among 2,588 characters from the Taiwan Big5 character set. 52
2.5 Linguistic units corresponding to Mayan complex glyphs, from Macri (1996); glossing conventions follow those of Macri. In some phrasal cases it is unclear whether the unit in question is really a constituent: Such cases are indicated with a question mark. 56
3.1 The ten most frequent lexical markings for the shallow ORL in the English fragment. 87
3.2 The ten most frequent lexical markings for the deep ORL in the English fragment. 87
4.1 Disyllabic morphemes collected from the ROCLING corpus (10 million characters) and 10 million characters of the United Informatics corpus. This set consists of pairs of characters occurring at least twice, and where each member of the pair only cooccurs with the other. 149
4.2 Further disyllabic morphemes collected from the ROCLING corpus (10 million characters) and 10 million characters of the United Informatics corpus. This set consists of other pairs of characters that do not exclusively occur with each other, but where there is nonetheless a high mutual information for the pair. Note that lV (i) indicates that the phonetic component in question occurs 9 xv
List of Tables

out of 38 times in characters pronounced with initial /l/ followed by some vowel.

4.3 A sample of Japanese kokuji (second column), with their componential analysis (third column). The first column is the entry number in Alexander's list (Lehman and Faust, 1951). The fourth column lists the phonetic, if any. The fifth column lists the kun pronunciation, and the sixth column the on pronunciation, if any: In one case – goza – there is no kun pronunciation. The last three kokuji shown are formed as semantic–phonetic constructs, with the last two being based on the kun pronunciation: Note that the phonetic component of masa also means 'straight', so it is possible that this one is also a semantic–semantic construct.

6.1 Total counts in Creggen's dictionary of words spelled with initial <Ch>, where C denotes a consonant, and the number and percentages of those words that are minimal pairs with homophonic or close-homophonic words spelled without the <h>.
Preface

Most general books on writing systems are written or edited by scholars who are specialists in a small subset of the writing systems that they cover and who have developed their views on writing in general based on their own experience in their particular specialized area.

This book is different: I cannot claim to be an expert on any particular writing system. My interest in writing systems stems in part from my interest in text-to-speech synthesis systems, and in particular the problem of converting from written text into a linguistic representation that specifies how that text would be read. Given that problem, it is natural to inquire about the formal nature of the relationship between the written form and the linguistic representation that the written form encodes: What linguistic elements do written symbols encode? Do writing systems differ in the abstractness of the linguistic representation encoded by orthography, and if so how? What are the formal constraints on the mapping between linguistic representation and writing? Some of these issues have, of course, been addressed elsewhere, though usually in an informal fashion. This book is an attempt to answer these questions in the context of a formal, computational theory of writing systems.

One point that needs to be made at the outset is that this book is not intended as an introduction to the topic of writing systems. There are many excellent books that serve that purpose, including Sampson (1985), Coulmas (1989), and DeFrancis (1989). Special mention must be given to the superb collection in Daniels and Bright (1996), without which the present book would not have been possible. Thus, while I do discuss aspects of several writing systems in some amount of detail, there are also a number of writing systems that are discussed in less detail. The reader unfamiliar with the general properties of the writing systems discussed here is urged to consult one of the many general introductions to the topic, such as those cited above.

In preparing this work, I have benefited greatly from discussions with and comments from a number of colleagues, listed here in alphabetical order: Harald Baayen, Alan Black, Wayles Browne, Roy Harris, Leonard Katz, George Kiraz, Martin Jansche, Kazuaki Maeda, Mark-Jan Nederhof, Anneke Neijt, Elena Pavlova, Geoffrey Sampson, Chiling Shih, Brian
Preface

Stowell, Robert Thomson, J. Marshall Unger, and Jennifer Venditti. I would especially like to thank Steven Bird, who read through two whole drafts of this work and gave me extensive comments on both. I also acknowledge an anonymous reviewer for Cambridge University Press. Portions of this work were presented at the University of Arizona and at Charles University in Prague, and I thank audiences there for useful comments and questions. I also thank Juergen Schroeter for help in using his recording setup for the experiment reported in Section 3.3.

The technical production of this book depended heavily upon several free or public domain resources including databases and software. I am indebted to Rick Harbaugh (developer of www.zhongwen.com) for kindly allowing me access to his data on Chinese character structure. Several of the more detailed analyses in this book, including the treatment of English in Section 3.2 and of Chinese in Section 2.3.4, were implemented, and these implementations depended upon the fsm library developed by my colleagues at AT&T Labs, Michael Riley, Fernando Pereira, and Mehryar Mohri. Chinese characters were incorporated into \TeX using Stephen Simpson’s PMC package; for Devanagari I used Frans Velthuis’s devtex package; Visible Speech fonts are due to Mark Shoulson. Editing of figures and graphics were done using Vectaport Inc.’s idraw, John Bradley’s xo, and Davor Matic’s bitmap.

Finally I would like to thank my editor at Cambridge University Press, Christine Bartels, for her support for this project.