

Contents

1	A Gentle Introduction	1
1.1	The Beginning	1
1.2	Some Unavoidable Terminology	1
1.3	A Few Intuitive Examples	8
1.4	Sharing Structure	16
1.5	Some Background in Set Theory	17
1.6	Composition and Rule Application	31
1.7	Lexicon and Rules	32
1.8	The Big Picture	36
1.9	Finite-State Networks are Good Things	37
1.10	Exercises	37
2	A Systematic Introduction	43
2.1	Where We Are	43
2.2	Basic Concepts	43
2.3	Simple Regular Expressions	45
2.4	Complex Regular Expressions	63
2.5	Properties of Finite-State Networks	74
2.6	Exercises	78
3	The <code>xfst</code> Interface	81
3.1	Introduction	81
3.2	Compiling Regular Expressions	84
3.3	Interacting with the Operating System	109
3.4	Incremental Computation of Networks	121
3.5	Rule-Like Notations	130
3.6	Examining Networks	182
3.7	Miscellaneous Operations on Networks	190
3.8	Advanced <code>xfst</code>	196
3.9	Operator Precedence	202
3.10	Conclusion	202

4	The <code>lexc</code> Language	203
4.1	Introduction	203
4.2	Defining Simple Automata	204
4.3	Defining Lexical Transducers	226
4.4	Useful <code>lexc</code> Strategies	240
4.5	Integration and Testing	259
4.6	<code>lexc</code> Summary and Usage Notes	270
5	Planning and Managing Finite-State Projects	279
5.1	Infrastructure	279
5.2	Linguistic Planning	283
5.3	Composition is Our Friend	293
5.4	Priority Union for Handling Irregular Forms	300
5.5	Conclusions	307
6	Testing and Debugging	311
6.1	The Challenge of Testing and Debugging	311
6.2	Testing on Real Corpora	311
6.3	Checking the Alphabet	319
6.4	Testing with Subtraction	327
7	Flag Diacritics	339
7.1	Features in Finite-State Systems	339
7.2	What Are Flag Diacritics?	339
7.3	Using Flag Diacritics	341
7.4	Flag Diacritics and Finite-State Algorithms	356
7.5	Examples	364
8	Non-Concatenative Morphotactics	375
8.1	Introduction	375
8.2	Formal Morphotactic Description	378
8.3	Practical Non-Concatenative Examples	383
8.4	Usage Notes for <code>compile-replace</code>	412
8.5	Debugging Tips for <code>compile-replace</code>	417
8.6	The Formal Power of Morphotactics	419
8.7	Conclusion	420
9	Finite-State Linguistic Applications	421
9.1	Introduction	421
9.2	The <code>tokenize</code> Utility	422
9.3	The <code>lookup</code> Utility	431
9.4	Using <code>xfst</code> in Batch Mode	439
9.5	Transducers for Morphological Analysis	439
9.6	Spelling	451

9.7	Beyond Tokenization and Morphological Analysis	453
9.8	Application Summary	455

A	Graphing Quiz	457
----------	----------------------	------------

B	Solutions to Exercises	459
----------	-------------------------------	------------

B.1	Graphing Quiz	459
B.2	The Better Cola Machine Diagram	466
B.3	The Better Cola Machine Network	466
B.4	Brazilian Portuguese Pronunciation	470
B.5	The Monish Language	473
B.6	Esperanto Nouns	476
B.7	Esperanto Adjectives	477
B.8	Esperanto Nouns and Adjectives	477
B.9	Esperanto Nouns and Adjectives with Upper-Side Tags	479
B.10	Esperanto Nouns, Adjectives and Verbs	480