Contents

Preface **xi**

Acronyms **xxi**

1 Introduction **1**

1.1 Machine translation 2

1.1.1 Taxonomy 3

1.2 Historical documents processing 5

1.2.1 Language modernization 5

1.2.2 Spelling normalization 7

2 Machine Translation **9**

2.1 Statistical machine translation 10

2.1.1 Phrased-based statistical machine translation 11

2.2 Neural machine translation 16

2.2.1 Architectures 17

2.2.2 Subword NMT 24

2.2.3 Synthetic data generation 25

3 Interactive Machine Translation **27**

3.1 State of the art . . . 28
3.2 Prefix-based IMT . . . 29
3.2.1 Implementation . . . 30
3.3 Segment-based IMT . . . 30
3.3.1 Implementation . . . 32
3.4 Segment-based IMT with active prediction . . . 40
3.5 Experimental framework . . . 40
3.5.1 Corpora . . . 41
3.5.2 Systems . . . 41
3.5.3 Metrics . . . 42
3.5.4 User simulation . . . 43
3.6 Evaluation . . . 46
3.6.1 Main approaches . . . 46
3.6.2 Segment-based IMT with active prediction . . . 49
3.7 Qualitative analysis . . . 50
3.8 Interactive neural machine translation . . . 53
3.8.1 Prefix-based INMT . . . 54
3.8.2 Segment-based INMT . . . 54
3.8.3 Comparison of INMT against IMT . . . 55
3.9 Conclusions . . . 56
3.10 Publications . . . 57

4 Language Modernization **59**

4.1 State of the art . . . 60
4.2 Approaches . . . 61
4.2.1 Statistical machine translation . . . 61
4.2.2 Neural machine translation . . . 61
4.2.3 NMT enriched with modern documents . . . 61

4.3 Experimental framework . . . 62
4.3.1 Corpora . . . 62
4.3.2 Systems . . . 64
4.3.3 Metrics . . . 65
4.4 Evaluation . . . 66
4.4.1 Automatic metrics . . . 66
4.4.2 Scholars . . . 67
4.4.3 Non-experts . . . 69
4.5 Qualitative analysis . . . 71
4.5.1 Initial evaluation . . . 72
4.5.2 User perception . . . 74
4.6 Conclusions . . . 77
4.7 Publications . . . 78

5 Spelling Normalization 79

5.1 State of the art . . . 80
5.2 Approaches . . . 81
5.2.1 Statistical dictionary . . . 81
5.2.2 Initial approaches . . . 81
5.2.3 Character-based MT . . . 82
5.2.4 CBNMT enriched with modern documents . . . 83
5.3 Experimental framework . . . 84
5.3.1 Corpora . . . 84
5.3.2 Systems . . . 85
5.3.3 Metrics . . . 86
5.4 Evaluation . . . 87
5.4.1 Main approaches . . . 87
5.4.2 Additional CBNMT approaches . . . 88
5.4.3 Generation of modern transcripts . . . 89
5.5 Qualitative analysis . . . 91

5.5.1 Main approaches . . . 91
5.5.2 Additional CBNMT approaches . . . 94
5.6 Conclusions . . . 96
5.7 Publications . . . 97

6 Interactive Machine Translation for the Processing of Historical Documents 99

6.1 Language modernization . . . 100
6.1.1 Experimental framework . . . 100
6.1.2 Evaluation . . . 101
6.1.3 Qualitative analysis . . . 102
6.2 Spelling normalization . . . 104
6.2.1 Evaluation . . . 105
6.2.2 Qualitative analysis . . . 106
6.3 Online demonstrator . . . 107
6.4 Conclusions . . . 109
6.5 Publications . . . 109

7 Conclusions 111

7.1 Scientific contributions . . . 112
7.1.1 Interactive machine translation . . . 112
7.1.2 Machine translation applications to historical documents . . . 113
7.2 Publications derived from the thesis . . . 114
7.3 Software contributions . . . 116
7.4 Other contributions . . . 117
7.4.1 Tokenization . . . 117
7.4.2 Incremental learning . . . 118
7.5 Future works . . . 119
7.5.1 Interactive machine translation . . . 119
7.5.2 Language modernization . . . 119

7.5.3 Spelling normalization 119

A Tuning the Number of Merge Operations for the BPE Algorithm **121**

B Language Modernization: Questionnaire from the User Study **123**

C Post-editing the abstracts **125**

List of Figures **129**

List of Tables **131**

List of Equations **134**

List of Examples **136**

Bibliography **149**