Contents

1 Introduction **1**

1.1 Research Questions 3

1.2 Overview 4

2 Literature review **7**

2.1 Learner corpora 7

2.2 Error annotation 12

2.2.1 Defining *error* 13

2.2.2 Describing errors 15

2.2.3 Error-tagging systems 18

2.2.4 Inter annotator agreement 21

2.2.5 Implicit error annotation 23

2.3 Linguistic annotation 24

2.3.1 PoS tagging, lemmatisation and morphological feature annotation 24

2.3.2 Syntactic annotation 28

2.4 Learner corpora and NLP tasks 33

2.4.1 NLI 34

2.4.2 GED and GEC 36

3 VALICO-UD design **39**

3.1 Data description 39

3.1.1 Gold standard 45

3.1.2 Silver standard 48

3.2 Target Hypothesis writing 49

3.2.1 Guiding principles 53

4 Error Annotation **57**

4.1 Methodology 57

4.2 Tagset description 64

4.2.1 Spelling errors 66

4.2.2 Derivation errors . . . 69
4.2.3 Form errors . . . 70
4.2.4 Inflection errors . . . 72
4.2.5 Unnecessary, missing and replace word errors . . . 74
4.2.6 Word order errors . . . 78
4.2.7 Complex errors . . . 79
4.3 Error distribution per macro-categories and L1s . . . 80
4.4 Inter Annotator Agreement . . . 85
4.4.1 Methodology . . . 86
4.4.2 Experiment 1 and 2: error identification and normalisation . . . 87
4.4.3 Experiment 3: error coding system evaluation . . . 100

5 Linguistic annotation **105**
5.1 Treebanking VALICO-UD . . . 105
5.1.1 Universal Dependencies formalism . . . 105
5.1.2 Annotation scheme . . . 109
5.2 Inter Annotator Agreement . . . 129
5.3 Treebank statistics . . . 130
5.3.1 Distribution of PoS tags . . . 130
5.3.2 Distribution of dependency relations . . . 131
5.3.3 Parallel reading of the treebank . . . 134
5.4 Incremental parsing evaluation . . . 136
5.4.1 Error analysis . . . 138

6 Quantitative data exploration **147**
6.1 Quantifying the parser performance . . . 147
6.2 Quantifying learners' proficiency via MT metrics . . . 149
6.2.1 Quantifying string distance . . . 150
6.2.2 Quantifying tree distance . . . 152

7 Conclusions **155**
7.1 Future work . . . 158

Bibliography **161**

Acknowledgements **179**