# ISO 24611:2012-11 (E)

## Language resource management - Morpho-syntactic annotation framework (MAF)

## Contents

Page


Foreword ..... v

Introduction ..... vi

1 Scope ..... 1

2 Normative references ..... 1

3 Terms and definitions ..... 1

4 The MAF meta-model ..... 4
4.1 Overview ..... 4
4.2 MAF Meta-model ..... 4

5 Segmenting with tokens ..... 6
5.1 General ..... 6
5.2 Formal description: <token> ..... 7
5.3 Embedding notation ..... 7
5.4 Alternate representation for TEI based documents ..... 8
5.5 Stand-off notation ..... 9
5.6 Informative attributes ..... 9
5.7 Completing the inline token notation ..... 10
5.7.1 Joining tokens in embedded mode ..... 10
5.7.2 Overlapping tokens ..... 11

6 Word-forms as linguistic units ..... 11
6.1 Formal description: <wordForm> ..... 12
6.2 Token attachment ..... 12
6.2.1 One token; one word-form ..... 12
6.2.2 Several contiguous tokens; one word-form ..... 12
6.2.3 Several discontinuous tokens; one word-form ..... 13
6.2.4 Zero token; one word-form ..... 13
6.2.5 One token; several word-forms ..... 14
6.3 Referring to lexical entries ..... 14
6.4 Compound word-forms ..... 15
6.5 Identification of word-forms within a TEI-compliant document ..... 15

7 Morpho-syntactic content ..... 18
7.1 General ..... 18
7.2 Using feature structures ..... 18
7.3 Compact morpho-syntactic tags ..... 18
7.4 FSR libraries ..... 19
7.5 Designing tagsets ..... 20
7.6 Formal description: <tagset> ..... 22

8 Handling ambiguities ..... 22
8.1 Word-form content ambiguities ..... 22
8.2 Lexical Ambiguities ..... 23
8.3 Structural ambiguities ..... 23
8.3.1 Structural ambiguities with word-forms ..... 23
8.3.2 Structural ambiguities with tokens ..... 24
8.4 Simplified structuring variants ..... 24

8.4.1 Non-ambiguous linear representation ..... 24
8.4.2 Mixed linear and lattice representation ..... 25
8.5 Expanding the simplified variants ..... 26
8.5.1 Separating tokens and word-forms ..... 26
8.5.2 Wrapping into local lattices ..... 26
8.5.3 Merging local lattices ..... 27
8.5.4 Removing <wfAlt> ..... 28
8.6 Formal description: <wfAlt> and <fsm> ..... 29

Annex A (informative) Encoded example using the MAF serialization ..... 30

Annex B (normative) MAF specification ..... 33

B.1 Elements ..... 33
B.1.1 <dcs/> ..... 33
B.1.2 <fsm> ..... 34
B.1.3 <maf> ..... 34
B.1.4 <tagset> ..... 35
B.1.5 <token> ..... 35
B.1.6 <transition> ..... 36
B.1.7 <wfAlt> ..... 36
B.1.8 <wordForm> ..... 37
B.2 Model classes ..... 38
B.3 Attribute classes ..... 38
B.3.1 att.token.information ..... 38
B.3.2 att.token.join ..... 39
B.3.3 att.token.span ..... 39
B.3.4 att.wordForm.content ..... 39
B.3.5 att.wordForm.tokens ..... 40
B.4 Macros ..... 40
B.4.1 data.certainty ..... 40
B.4.2 data.code ..... 40
B.4.3 data.count ..... 40
B.4.4 data.duration.w3c ..... 41
B.4.5 data.enumerated ..... 41
B.4.6 data.key ..... 41
B.4.7 data.language ..... 42
B.4.8 data.name ..... 43
B.4.9 data.numeric ..... 43
B.4.10 data.pointer ..... 43
B.4.11 data.probability ..... 44
B.4.12 data.temporal.w3c ..... 44
B.4.13 data.truthValue ..... 44
B.4.14 data.word ..... 45
B.4.15 data.xTruthValue ..... 45

Annex C (normative) Morpho-syntactic data categories ..... 46

Bibliography ..... 58