Appendix A - Tagset and Conversions

Below is a list of all tags used in the final simplified tagset for the models, and lists of what tags from the source materials were converted to that tag in the simplified tagset. For the OHG Referenzkorpus Altdeutsch, the simpler 'posLemma' tags were used as the basis for these conversions where possible, and the more fine-grain 'pos' tags were used elsewhere.

1. **VB (verb)**

OE, OS, and OIce - All tags indicating the verb 'be', the verb 'have', auxiliary verbs, modal verbs, and other verbs in the documentation. (BE, BEI, BEPH, BEPI, BEPS, BEP, BEDI, BEDS, BED, BAG, BEN, HV, HVI, HVPI, HVPS, HVP, HVDI, HVDS, HVD, HAG, HVN, AX, AXI, AXPI, AXPS, AXP, AXDI, AXDS, AXD, AXG, AXN, MD, MDI, MDPI, MDPS, MDP, MDDI, MDDS, MDD, TO, VB, VBI, VBPH, VBPI, VBPS, VBP, VBDI, VBDS, VBD, VAG, VBN)

GO - 'Verb' and 'Participle'

OHG - All tags indicating verbs (''VA', 'VV', 'VM' 'VVPP', 'VVPS', 'VVFIN')

2. **N (noun)**

OE, OS, and OIce - All nominal tags (N, NR, MAN)

GO - 'Noun'

OHG - All tags indicating nominals ('NA', 'NE')

3. **PRO (pronoun)**

OE, OS, and OIce - All pronominal tags (PRO, PRO$)

GO - 'Pronoun'

OHG - All tags indicating pronominals ('PPER', 'PI', 'PRF')

4. **AP (adposition)**

OE, OS, and OIce - All prepositions (P)

GO - ‘Preposition’

OHG - Tags for adposition (‘AP’)

5. **D (determiner)**

OE, OS, and OIce - All determiners (D)

GO - N/A

OHG - All tags indicating determiner forms ('DD', 'DPOS', 'DDS', 'DIS', 'DI')

6. **ADV (adverb)**

OE, OS, and OIce - all forms associated with adverbs (ADV, ADVR, ADVS)

GO - ‘Adverb’

OHG - adverbs (‘ADV’)

7. **CONJ (conjunction)**

OE, OS, and OIce - All conjunctions (CONJ)

GO - ‘Conjunction’

OHG - conjunctions (‘KO’)

8. **C (complementizer)**

OE, OS, and OIce - All complementizers (C)

GO - N/A

OHG - N/A

9. **ADJ (adjective)**

OE, OS, and OIce - all forms associated with adjectives (ADJ, ADJR, ADJS)

GO - ‘Adjective’

OHG - Adjectives (‘ADJ’)

10. Q (quantifier)

OE, OS, and OIce - All quantifiers (Q, QR, QS)

GO - N/A

OHG - N/A

11. NUM (numeral)

OE, OS, and OIce - All numerals (NUM)

GO - ‘Numeral’

OHG - Cardinal Numbers (‘CARD’)

12. NEG (negation marker)

OE, OS, and OIce - All negation (NEG)

GO - N/A

OHG - N/A

13. PART (other particle)

OE, OS, and OIce - All particles (FP, RP)

GO - ‘Particle’

OHG - Particles (‘PTK’)

14. WH (wh word)

OE, OS, and OIce - All WH-words (WPRO, WADV, WADJ, WQ)

GO - N/A

OHG - Question words (‘PW’, ‘PWG’, ‘DW’)

15. INTJ (interjection)

OE, OS, and OIce - All interjections (INTJ)

GO - ‘Interjection’

16. PUNCT (punctuation)

OE, OS, and OIce - All punctuation

GO - ‘Punctuation’

OHG - All tags indicating some form of punctuation (‘$,’, ‘$.’, ‘$(’, ‘$_’)

17. XX (problematic)

OE, OS, and OIce - All tags marked as FW (foreign word) or not appearing in the source documentation

GO - ‘nan’, ‘unassigned’, ‘foreign word’

OE, OS, and OIce - All tags marked as ‘nan’, empty tags, tags not occurring in the documentation.