1.2.6 Part of Speech Tagging

Some sub-corpora have been annotated with Part Of Speech annotations. This concerns WUS_DIALOG_GSW, WUS_FRA, WUS_FRA_DEMOG, WUS_ITA, WUS_ITA_DEMOG.

French

The whole French corpus has been annotated with MElt (Modified French TreeBank) using the tag set CC Tagset. Available annotations are "mftb_pos" (for part of speech) and "mftb_lem" (for the lemma). The following tags are used:

- ADJ adjective
- ADJWH interrogative adjective
- ADV adverb
- ADVWH interrogative adverb
- CC coordinating conjunction
- CLO object clitic pronoun
- CLR reflexive clitic pronoun
- CLS subject clitic pronoun
- CS subordinating conjunction
- DET determiner
- DETWH interrogative determiner
- ET foreign word
- I interjection
- NC common noun
- NPP proper noun
- P preposition
- P+D preposition+determiner amalgam
- P+PRO prepositon+pronoun amalgam
- PONCT punctuation mark
- PREF prefix
- PRO full pronoun
- PROREL relative pronoun
- PROWH interrogative pronoun
- V indicative or conditional verb form
- VIMP imperative verb form
- VINF infinitive verb form
- VPP past participle
- VPR present participle
- VS subjunctive verb form

Swiss German dialect

Five chats of the Swiss German dialectal data (34,683 tokens) have been manually normalized and annotated for Part of Speech. The according corpus is called WUS_DIALOG_GSW. Three annotations have been added to each token:

- gloss: The manual normalization
- tt_pos: Part of Speech annotation with TreeTagger based on the manually normalized tokens.
- tt_lem: The lemma as assigned by TreeTagger

The tagset uses the following tags:

- ADJA attributive adjective (including participles used adjectivally)
- ADJD predicate adjective; adjective used adverbially
- ADV adverb (never used as attributive adjective)
- APPR preposition left hand part of double preposition
- APPRART preposition with fused article
- APPO postposition
- APZR right hand part of double preposition
- ART article (definite or indefinite)
- CARD cardinal number (words or figures); also declined
- FM foreign words (actual part of speech in original language may be appended, e.g. FMADV/ FM-NN)
- ITJ interjection
- KON co-ordinating conjunction
- KOKOM comparative conjunction or particle
- KOUI preposition used to introduce infinitive clause
- KOUS subordinating conjunction
- NA adjective used as noun
- NE names and other proper nouns
- NN noun (but not adjectives used as nouns)
- PAV [PROAV] pronominal adverb
- PAVREL pronominal adverb used as relative
- PDAT demonstrative determiner
- PDS demonstrative pronoun
- PIAT indefinite determiner (whether occurring on its own or in conjunction with another determiner)
- PIS indefinite pronoun
- PPER personal pronoun
- PRF reflexive pronoun
- PPOSS possessive pronoun
- PPOSAT possessive determiner
- PRELAT relative depending on a noun
- PRELS relative pronoun (i.e. forms of der or welcher)
- PTKA particle with adjective or adverb
- PTKANT answer particle
- PTKNEG negative particle
- PTKREL indeclinable relative particle
- PTKVZ separable prefix
- PTKZU infinitive particle zu
- PWS interrogative pronoun
- PWAT interrogative determiner
- PWAV interrogative adverb
- PWAVREL interrogative adverb used as relative
- PWREL interrogative pronoun used as relative
- TRUNC truncated form of compound
- VAFIN finite auxiliary verb
- VAIMP imperative of auxiliary
- VAINF infinitive of auxiliary
- VAPP past participle of auxiliary

- VMFIN finite modal verb
- VMINF infinitive of modal
- VMPP past participle of auxiliary
- VVFIN finite full verb
- VVIMP imperative of full verb
- VVINF infinitive of full verb
- VVIZU infinitive with incorporated zu
- VVPP past participle of full verb

As in the French corpus, there are also combined tags such as *VAFIN+PPER* when a personal pronoun is agglutinated to a verb (*hätti* for 'hätte ich').

Italian

The Italian corpus is annotated with the TreeTagger, too, but based on the original tokens, i.e. not manually normalized. In this sub-corpus, however, only some parts were manually normalized resulting in the following three annotations:

- gloss: The manual normalization (often _UNGLOSSED_)
- tt_pos: Part of Speech annotation with TreeTagger
- tt_lem: The lemma as assigned by TreeTagger

The following PoS tagset was used:

- ABR abbreviation
- ADJ adjective
- ADV adverb
- CON conjunction
- DET:def definite article
- DET:indef indefinite article
- FW foreign word
- INT interjection
- LS list symbol
- NOM noun
- NPR name
- NUM numeral
- PON punctuation
- PRE preposition
- PRE:det preposition+article
- PRO pronoun
- PRO:demo demonstrative pronoun
- PRO:indef indefinite pronoun
- PRO:inter interrogative pronoun
- PRO:pers personal pronoun
- PRO:poss possessive pronoun
- PRO:refl reflexive pronoun
- PRO:rela relative pronoun
- SENT sentence marker
- SYM symbol
- VER:cimp verb conjunctive imperfect
- VER:cond verb conditional

- VER:cpre verb conjunctive present
- VER:futu verb future tense
- VER:geru verb gerund
- VER:impe verb imperative
- VER:impf verb imperfect
- VER:infi verb infinitive
- VER:pper verb participle perfect
- VER:ppre verb participle present
- VER:pres verb present
- VER:refl:infi verb reflexive infinitive
- VER:remo verb simple past