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 coordination conjunction
- CLO object clitic pronoun
- CLR reflexive clitic pronoun
- CLS subject clitic pronoun
- CS subordination 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

Additionally, the following combined annotations can occur, e.g. “P+D” for a preposition with a determiner like *aux*. The following list is ordered by the number of occurrences within the corpus:

- CLS+V
- CLS+CLO
- CS+CS
- CLS+CLO+V

- ADV+CLR+V+ADV
- DET+NC
- CLS+CLR
- CLS+CLR+V
- PRO+V
- P+NC
- CLR+V
- CLO+V
- DET+ADJ
- V+CLS
- CS+CLS
- P+PRO
- ADV+V
- DET+DET
- DET+PRO
- CLO+CLO
- P+VINF
- CLS+CLO+P
- P+ADJ
- CLS+VS
- CLS+CLO+CLO
- CLR+VINF
- CLS+NC
- CLS+DET
- PROWH+V+CLS+CS
- ADV+ADV+ADV
- NPP+V
- CLS+CLR+CLO
- DET+VPP
- ADV+ADV+CS
- ET+CLO+V
- ADV+VPP
- ADV+VINF
- CLS+P
- P+VPP
- CLS+VPP
- CLR+NC
- ET+CLO

Swiss German dialect

A small part of the Swiss German dialectal data has 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, i.e. "gloss".

- tt_lem: The lemma as assigned by TreeTagger

The tagset uses the following tags:

- ADJA attributive adjective (including participles used adjectivally) *das große Haus die versunkene Glocke*
- ADJD predicate adjective; adjective used adverbially *der Vogel ist blau er fährt schnell*
- ADV adverb (never used as attributive adjective) *sie kommt bald*
- APPR preposition left hand part of double preposition *auf dem Tisch an der Straße entlang*
- APPRART preposition with fused article *am Tag*
- APPO postposition *meiner Meinung nach*
- APZR right hand part of double preposition *an der Straße entlang*
- ART article (definite or indefinite) *die Tante; eine Tante*
- CARD cardinal number (words or figures); also declined *zwei; 526; dreier*
- FM foreign words (actual part of speech in original language may be appended, e.g. FMADV/ FM-NN) *semper fidem*
- ITJ interjection *Ach!*
- KON co-ordinating conjunction *oder ich bezahle nicht*
- KOKOM comparative conjunction or particle *er arbeitet als Straßenfeger, so gut wie du*
- KOUI preposition used to introduce infinitive clause *um den König zu töten*
- KOUS subordinating conjunction *weil er sie gesehen hat*
- NA adjective used as noun *der Gesandte*
- NE names and other proper nouns *Moskau*
- NN noun (but not adjectives used as nouns) *der Abend*
- PAV [PROAV] pronominal adverb *sie spielt damit*
- PAVREL pronominal adverb used as relative *die Puppe, damit sie spielt*
- PDAT demonstrative determiner *dieser Mann war schlecht*
- PDS demonstrative pronoun *dieser war schlecht*
- PIAT indefinite determiner (whether occurring on its own or in conjunction with another determiner) *einige Wochen, viele solche Bemerkungen*
- PIS indefinite pronoun *sie hat viele gesehen*
- PPER personal pronoun *sie liebt mich*
- PRF reflexive pronoun *ich wasche mich, sie wäscht sich*
- PPOSS possessive pronoun *das ist meins*
- PPOSAT possessive determiner *mein Buch, das ist der meine/meinige*
- PRELAT relative depending on a noun *der Mann, dessen Lied ich singe […], welchen Begriff ich nicht verstehe*
- PRELS relative pronoun (i.e. forms of der or welcher) *der Herr, der gerade kommt; der Herr, welcher nun kommt*
- PTKA particle with adjective or adverb *am besten, zu schnell, aufs herzlichste*
- PTKANT answer particle *ja, nein*
- PTKNEG negative particle *nicht*
- PTKREL indeclinable relative particle *so*
- PTKVZ separable prefix *sie kommt an*
- PTKZU infinitive particle zu
- PWS interrogative pronoun *wer kommt?*
- PWAT interrogative determiner *welche Farbe?*
- PWAV interrogative adverb *wann kommst du?*
- PWAVREL interrogative adverb used as relative *der Zaun, worüber sie springt*
- PWREL interrogative pronoun used as relative *etwas, was er sieht*
- TRUNC truncated form of compound *Vor- und Nachteile*

- VAFIN finite auxiliary verb *sie ist gekommen*
- VAIMP imperative of auxiliary *sei still!*
- VAINF infinitive of auxiliary *er wird es gesehen haben*
- VAPP past participle of auxiliary *sie ist es gewesen*
- VMFIN finite modal verb *sie will kommen*
- VMINF infinitive of modal *er hat es sehen müssen*
- VMPP past participle of auxiliary *sie hat es gekonnt*
- VVFIN finite full verb *sie ist gekommen*
- VVIMP imperative of full verb *bleibt da!*
- VVINF infinitive of full verb *er wird es sehen*
- VVIZU infinitive with incorporated zu *sie versprach aufzuhören*
- VVPP past participle of full verb *sie ist gekommen*

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