Part of speech tagging

Wikipedia defines PoS tagging as follows: "In corpus linguistics, part-of-speech tagging (POS tagging or POST), also called grammatical tagging or word-category disambiguation, is the process of marking up a word in a text (corpus) as corresponding to a particular part of speech, based on both its definition, as well as its context, i.e. relationship with adjacent and related words in a phrase, sentence, or paragraph. " In this corpus, we applied PoS tagging to the German, French and Italian parts using Helmut Schmid's TreeTagger For both varieties of German (i.e. dialectal and non dialectal), there is also a sub-corpus available that was annotated with the RFTagger. For Romansh, unfortunately, there is no parameter file available for TreeTagger and there are in fact no other tools available for this language, either.

German (both dialectal and non-dialectal)

TreeTagger

- PoS tagging was applied to the normalised level of each SMS, and each SMS was tagged as one unit.
- TreeTagger was used, applying a tailor-made German parameter file (courtesy of Helmut Schmid)
- The STTS tagset was used.
- The tagger's lexicon was systematically supplemented with borrowings and proper nouns (thanks to Andrea Suter).
- The tag PTKINF was added for infinitive particle (go, goge etc.) for the german dialect.
- The resulting sub-corora are: deu-tagged and gsw-tagged

RFTagger

- The same varieties of German were also tagged with the RFTagger, resulting in the sub-corpora deu-rftagged and gsw-rftagged

French

1. PoS tagging was applied to the normalised level of each SMS, and SMS was tagged as one unit.
2. TreeTagger was used out of the box.
3. Achim Stein's TagSet for French was used.
4. The tags DET:DEM and DET:IND were added.

Italian

- PoS tagging was applied to the normalised level of each SMS, and SMS was tagged as one unit.
- TreeTagger was used out of the box.
- Achim Stein's TagSet for Italian was used.

- The tag ADJ:poss was added.

Precision

Our test gave the following precision for the respective sub-corpora:

- gsw: 2'734 tokens checked: 96.3% correct
- deu: 2'922 tokens checked: 93.1% correct
- fra: 3'133 tokens checked: 94.6% correct
- ita: 2527 tokens checked: 90.5% correct