CoNLL Table

The CoNLL table is produced in output (in csv format) by the Stanford CoreNLP parsers (Klein and Manning 2003). The input to CoreNLP is a txt-formatted single file containing the text corpus to be analyzed. The CoNLL table provides the basis of all computational linguistics analyses.

The CoNLL csv file is the basis of a variety of statistical analyses and visualizations on the text corpus, namely frequency distribution of words (particularly lemmatized words but also by different types of syntactical categories, e.g., verbs or nouns, average sentence length, n-grams, co-occurrences).

The CoNLL file has a typical standard format (CoNLL format).

CoNLL-U & CoNLL-X

There are many different CoNLL formats since CoNLL has had many different updates over the years. In general, in a CoNLL file, each line represents a single word with a series of tab-separated fields (columns). _ (underscore) characters indicate empty values. CoNLL-U is becoming the new standard, based on an extension of CoNLL-X. In CoNLL-U annotations are encoded in plain text files (UTF-8) with three types of lines:

- Word lines containing the annotation of a word/token in **10 fields** separated by single tab characters (see list of fields below; in **bold** the 7 fields available from CoNLL-X).
- Blank lines marking sentence boundaries.
- Comment lines starting with hash (#).

Sentences consist of one or more word lines, and word lines contain the following fields from CoNLL-U (in **bold** the 7 fields from CoNLL-X):

1. **ID**: Word index, integer starting at 1 for each new sentence; may be a range for tokens with multiple words.
2. **FORM**: Word form or punctuation symbol (the very word found in the input document).
3. **LEMMA**: Lemma or stem of word form.
4. CPOSTAG: Universal part-of-speech tag drawn from CoreNLP revised version of the Google universal POS tags.
5. **POSTAG**: Language-specific part-of-speech tag; _ (underscore) if not available. See the specific TIPS file ***Part of Speech Tags (POSTAG)***.
6. **NER**: Named-Entity Recognition (a standard set of pre-defined categories such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc.); not available in CoNLL-U.

   For English, the Stanford CoreNLP, by default through the NERClassifierCombiner annotator, recognizes the following NER values: named (PERSON, LOCATION, ORGANIZATION, MISC), numerical (MONEY, NUMBER, ORDINAL, PERCENT), and temporal (DATE, TIME, DURATION, SET) entities (12 classes). Adding the regexner annotator and using the supplied RegexNER pattern files adds support for the fine-grained and additional entity classes EMAIL, URL, CITY, STATE_OR_PROVINCE, COUNTRY, NATIONALITY, RELIGION, (job) TITLE, IDEOLOGY, CRIMINAL_CHARGE, CAUSE_OF_DEATH (11 classes) for a total of 23 classes.

7. FEATS: List of morphological features from the universal feature inventory or from a defined language-specific extension; _ (underscore) if not available.
8. **HEAD**: Head of the current token, which is either a value of ID or zero (0).
   When the value of HEAD is 0, the DEPREL value is set to ROOT.
   There is ONLY one ROOT for each sentence.
9. **DEPREL**: Universal Stanford dependency relation to the HEAD (root iff HEAD = 0) or a defined language-specific subtype of one. See the specific TIPS file ***Stanford Dependency Relations (DEPREL)***
   When the value of HEAD is 0, the DEPREL value is set to ROOT.
   There is ONLY one ROOT for each sentence.
10. DEPS: List of secondary dependencies (head-deprel pairs).
11. MISC: Any other annotation (e.g., comments).

The fields must additionally meet the following constraints:

- Fields must not be empty.
- Fields must not contain space characters.

References

Klein, Dan and Christopher D. Manning. 2003. “Accurate Unlexicalized Parsing.” In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pp. 423–430. Association for Computational Linguistics.