Title

Penn Korean Universal Dependency Treebank

Authors

Jinho D. Choi, Na-Rae Han, Jena D. Hwang and Hansaem Kim

Language

Korean

Recommended/Expected use of corpus

The Pen Korean Universal Dependency Treebank corpus brings the Korean Treebank Annotations Version 2.0 corpus (https://catalog.ldc.upenn.edu/LDC2006T09, 2006) up to the popular modern standards of the Universal Dependencies (UD, https://universaldependencies.org/). Recommended use include linguistic research focused on dependency syntax of the Korean language as well as training of statistical natural language processing models for Korean. Now in the Universal Dependencies framework known for its wide cross-linguistic compatibility, this dataset will be particularly useful to those seeking to develop a multilingual NLP application with a Korean component.

Collection Procedure - format, method, and timespan

The sentence annotations in this corpus were algorithmically converted from the source according to the procedure described in the paper: Chun, Jayeol, Na-Rae Han, Jena D. Hwang, and Jinho Choi "Building Universal Dependency Treebanks in Korean," Proceedings of LREC 2018 (http://www.lrec-conf.org/proceedings/lrec2018/pdf/378.pdf). The output was then quality-checked and error-corrected by Hansaem Kim's research group.

Directory Structure & File Format Specific Details

The corpus consists of 112 files in the CoNLL-U format which is the standard adopted by the Universal Dependencies project (https://universaldependencies.org/format.html). The files are numbered `302000.fid.conllu` through `320009.fid.conllu`, where the `XXXXXX.fid` portion refers to their file counterpart in the original Korean Treebank Annotations Ver 2.0. An example sentence annotation:

```
# sent_id = 302000.fid-s1
# text = 프랑스의 르노 자동차 그룹은 다음주 김대중 대통령의 프랑스 방문 중 한국 삼성자동차 인수를 공식 제의할지 모
른다고 르노사의 한 관계자가 1 일 밝혔다.
1       프랑스의        프랑스+의       PROPN   NPR+PAN _       4       nmod    _       _
2       르노    르노    PROPN   NPR     _       4       compound        _       _
3       자동차  자동차  NOUN    NNC     _       4       compound        _       _
4       그룹은  그룹+은 NOUN    NNC+PAU _       15      csubj   _       _
5       다음주  다음주  NOUN    NNC     _       15      obl     _       _
6       김대중  김대중  PROPN   NPR     _       7       compound        _       _
7       대통령의        대통령+의       NOUN    NNC+PAN _       10      nmod    _       _
8       프랑스  프랑스  PROPN   NPR     _       10      compound        _       _
9       방문    방문    NOUN    NNC     _       10      compound        _       _
10      중      중      NOUN    NNX     _       15      obl     _       _
11      한국    한국    PROPN   NPR     _       13      compound        _       _
12      삼성자동차      삼성+자동차     NOUN    NPR+NNC _       13      compound        _       _
13      인수를  인수+을 NOUN    NNC+PCA _       15      obj     _       _
14      공식    공식    NOUN    NNC     _       15      obl     _       _
15      제의할지        제의+하+을지    VERB    NNC+XSV+EFN     _       16      obj     _       _
16      모른다고        모르+는다+고    VERB    VV+EFN+PAD      _       22      ccomp   _       _
17      르노사의        르노+사+의      PROPN   NPR+XSF+PAN     _       19      nmod    _       _
18      한      한      NUM     NNU     _       19      nummod  _       _
19      관계자가        관계자+이       NOUN    NNC+PCA _       22      nsubj   _       _
20      1       1       NUM     NNU     _       21      nummod  _       _
21      일      일      NOUN    NNX     _       22      obl     _       _
22      밝혔다  밝히+었+다      VERB    VV+EPF+EFN      _       0       root    _       _
23      .       .       PUNCT   SFN     _       22      punct   _       _
```

One language-specific detail is the morphological analysis information, which is carried over from the 2006 source. Column 3 and column 5 make use of the + sign as the morphological boundary marker.

As with the original Korean Treebank Annotations Version 2.0 corpus, this resource contains a total of 5,010 sentences and 132,041 tokens.