Incorporating Alternate Translations into English Translation Treebank

Ann Bies, Justin Mott, Seth Kulick, Jennifer Garland, Colin Warner
{bies,jmott,skulick,garjen,colinw}@ldc.upenn.edu

Linguistic Data Consortium

Motivation: Request from MT system developers to include alternate translations of idiomatic expressions in the English translations of informal genres in order to facilitate better handling of idiomatic expressions

Example, English translated from Chinese:

Teacher salaries are very high, is that so? If there comes a day when everyone will become a teacher [by hook or by crook | sharpen their heads], as popular as the civil servants of today, that will mean that teachers are really attractive.

- **Original Chinese idiomatic expression**: 削尖脑袋 (xuè jiān nǎodai)
- **Fluent translation**: *by hook or by crook*
- **Literal translation**: *sharpen their heads*

New Annotation Guidelines

- Both fluent and literal translation alternates were annotated for word-level tokenization and part-of-speech (POS), as in BOLT/GALE
- Only the fluent translation alternates were annotated as part of the syntactic structure of the tree
- Literal translation alternates and markup left flat under META tree nodes
- An additional version of the trees was generated with the META nodes and their children removed from the trees, so only the fluent alternates appear in the resulting trees
- Both versions of the trees are published, with and without literal alternates

Adapting Annotation Pipeline Parsing Process

- Input annotation for parsing process = word-level tokenization and POS annotation, including both fluent and literal alternates
- META tree nodes inserted to indicate the scope of the literal alternates, mainly automatically with subsequent manual validation
- Literal alternates and markup (META nodes) removed prior to parsing
- Parser run on input without literal alternates to reduce parser confusion
- Literal alternates re-inserted in the parsed trees, tokens left flat under a META node
- Resulting trees used in the treebank annotation pipeline, as input to manual syntactic annotation and correction

FINAL TREES

Both translation alternates (with META nodes)

```
(S (SBAR-ADV If
        (S (NP-SBJ there)
           (VP comes
               (NP (NP a day)
                   (SBAR (WHADVP-1 when)
                         (S (NP-SBJ everyone)
                            (VP will
                                (VP become
                                    (NP-PRD (NP a teacher)
                                            (ADJP-2 *ICH*))
                                    (PP-MNR (META -LRB- )
                                            (PP by
                                                (NP hook))
                                           or
                                            (PP by
                                                (NP crook))
                                            (META | sharpen their heads -RRB- ))
                                    ,
                                    (ADJP-2 (ADJP as popular)
                                            (PP as
                                                (NP (NP the civil servants)
                                                    (PP of
                                                        (NP today)))))
                                    (ADVP-TMP-1 *T*)))))))))
   ,
   (NP-SBJ that)
   (VP will
       (VP mean
           (SBAR that
                 (S (NP-SBJ teachers)
                    (VP are
                        (ADJP-PRD really attractive))))))
   .)
```

If there comes a day when everyone will become a teacher [by hook or by crook | sharpen their heads], as popular as the civil servants of today, that will mean that teachers are really attractive.

Only fluent translation alternate (no META nodes)

```
(S (SBAR-ADV If
        (S (NP-SBJ there)
           (VP comes
               (NP (NP a day)
                   (SBAR (WHADVP-1 when)
                         (S (NP-SBJ everyone)
                            (VP will
                                (VP become
                                    (NP-PRD (NP a teacher)
                                            (ADJP-2 *ICH*))
                                    (PP-MNR (PP by
                                                (NP hook))
                                           or
                                            (PP by
                                                (NP crook)))
                                    ,
                                    (ADJP-2 (ADJP as popular)
                                            (PP as
                                                (NP (NP the civil servants)
                                                    (PP of
                                                        (NP today)))))
                                    (ADVP-TMP-1 *T*)))))))))
   ,
   (NP-SBJ that)
   (VP will
       (VP mean
           (SBAR that
                 (S (NP-SBJ teachers)
                    (VP  are
                        (ADJP-PRD really attractive))))))
   .)
```

If there comes a day when everyone will become a teacher by hook or by crook, as popular as the civil servants of today, that will mean that teachers are really attractive.

Distribution of Translation Alternates (META Nodes) in the Annotated Corpus

Alternate translation pairs in the corpus

- Total of 147,433 tokens/words (145,427 tokens without literal translation alternates), in a total of 5012 sentences
- 288 alternate translation pairs, affecting 248 sentences
  - NP (136 instances)
  - VP (88 instances)
  - S (23 instances)
  - Also ADJP, ADVP, FRAG, INTJ, NML, PP and UCP, each with less than 10 instances

Alternate translation pairs in the trees

- Alternates map exactly to the span of a single syntactic node
  - *[Anything | Divine horse] is possible.*
- Alternates correspond to a subspan of the node
  - *But the [Russians | Old Hairy Ones] occupied it …*
- Alternates are at different levels in the tree, so cannot be annotated as sister constituents
  - *Those who have bad luck will face [unqualified figures of | red lights all the way to] GDP*

Informal Genre Treebank Corpora with Translation Alternates

- **Completed:** English/Chinese Treebank web data (discussion forum (DF)). *Ann Bies, Justin Mott, Colin Warner, Seth Kulick. (2013). BOLT Phase 1 English Treebank DF – ECTB, Parts 6-7. Linguistic Data Consortium, Catalog numbers LDC2013E50, LDC2013E76*.
- **New translation treebank corpora in progress:** English/Chinese SMS/Chat, English/Chinese conversational telephone speech (CTS), English/Egyptian Arabic SMS/Chat, English/Egyptian Arabic CTS
- **Multiple annotations:** Parallel treebanks (Chinese, Egyptian Arabic) and word alignment also in progress → parallel aligned treebanks possible for DF, SMS/Chat, CTS

Acknowledgements: This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract Nos. HR0011-11-C-0145 and HR0011-12-C-0014. The content does not necessarily reflect the position or the policy\ of the Government, and no official endorsement should be inferred.