

# Urdu English Parallel Corpus



**By**  
**Atif Gulzar**

**National University of Computer and  
Emerging Sciences, Pakistan**

# Presentation Outline



- Parallel Corpus
- Levels of Alignment
- Usage of Parallel Corpus
- Urdu English Parallel Corpus
- General Translation Guidelines
- Translation Quality and Assurance

# Parallel Corpus



- Parallel Text
  - A set of texts, in which each is a translation into different language from single source text.
- Parallel Corpus
  - A large collection of parallel texts are called parallel corpus

# Levels of Alignment



- Sentence level
- Phrase level
- Word level
- He eats an apple.  

وہ ایک سیب کھاتا ہے۔

# Usage of Parallel Corpus



- Parallel corpora are essential tools to perform different kind of statistical lexical analysis in which two or more languages are involved
- Statistical machine translation systems based on statistical analysis of parallel corpora are currently used versus traditional rule based systems
- Parallel corpora can also be used for generating bilingual dictionaries

# Urdu English Parallel Corpus



- CRULP has produced Urdu English parallel corpus that is a translation of 150,000 English words (6500 sentences) into Urdu
- The source (English text) is taken from Penn Treebank corpus (Appendix A) for the following reasons
  - The corpus is available through a reliable consortium to everybody by LDC
  - It is in English, which is a widely understood language
  - It is completely tagged with POS, syntactic and grammatical structure and with other annotations. Thus, using this corpus also makes these annotations available for use, at least from English side

# General Translation Guidelines (cont.)



- There should be consistency in the translated text
- Try not to introduce extra words in translation and avoid idiomatic translation
- Each word of source text should have a translation in target language
- However idioms are explained in local language instead of literal word to word translation
  - "Break the ice" → برف کو توڑنا vs. جمود ختم کرنا

# General Translation Guidelines (cont.)



- The domain knowledge of source and target languages is essential while translating
  - Downtown → اندرون شہر (central part of city) vs. نشیبی علاقہ (Lower altitude region)
- Urdu has adopted many foreign words, it is encouraged to translate these in local language if the Urdu variation is available. E.g.
  - Museum → میوزیم vs. عجائب گھر
  - Government → گورنمنٹ vs. حکومت

# General Translation Guidelines (cont.)



- Word order

- English follows strict word order (subject-verb-object) rules
- Urdu is a free word order language but typically it uses subject-object-verb as natural word order

From natural to  
free word order



- وہ (subject) سیب (object) کھاتا ہے (verb) -
- وہ (subject) کھاتا ہے (verb) سیب (object) -
- کھاتا ہے (verb) وہ (subject) سیب (object) -

- It is tried to follow the natural word order of Urdu as much as possible

# General Translation Guidelines



- Whether to translate or transliterate proper nouns such as:
  - “New England Journal of Medicine”  
○ ینو انگلینڈ رسالہء طب
  - What about “Apple Computers”?  
○ ایپل کمپیوٹرز VS. سیب کمپیوٹر
- If the result of translation does not lose the meaning of proper noun then it is translated otherwise it is transliterated

# General Translation Guidelines



- Abbreviations

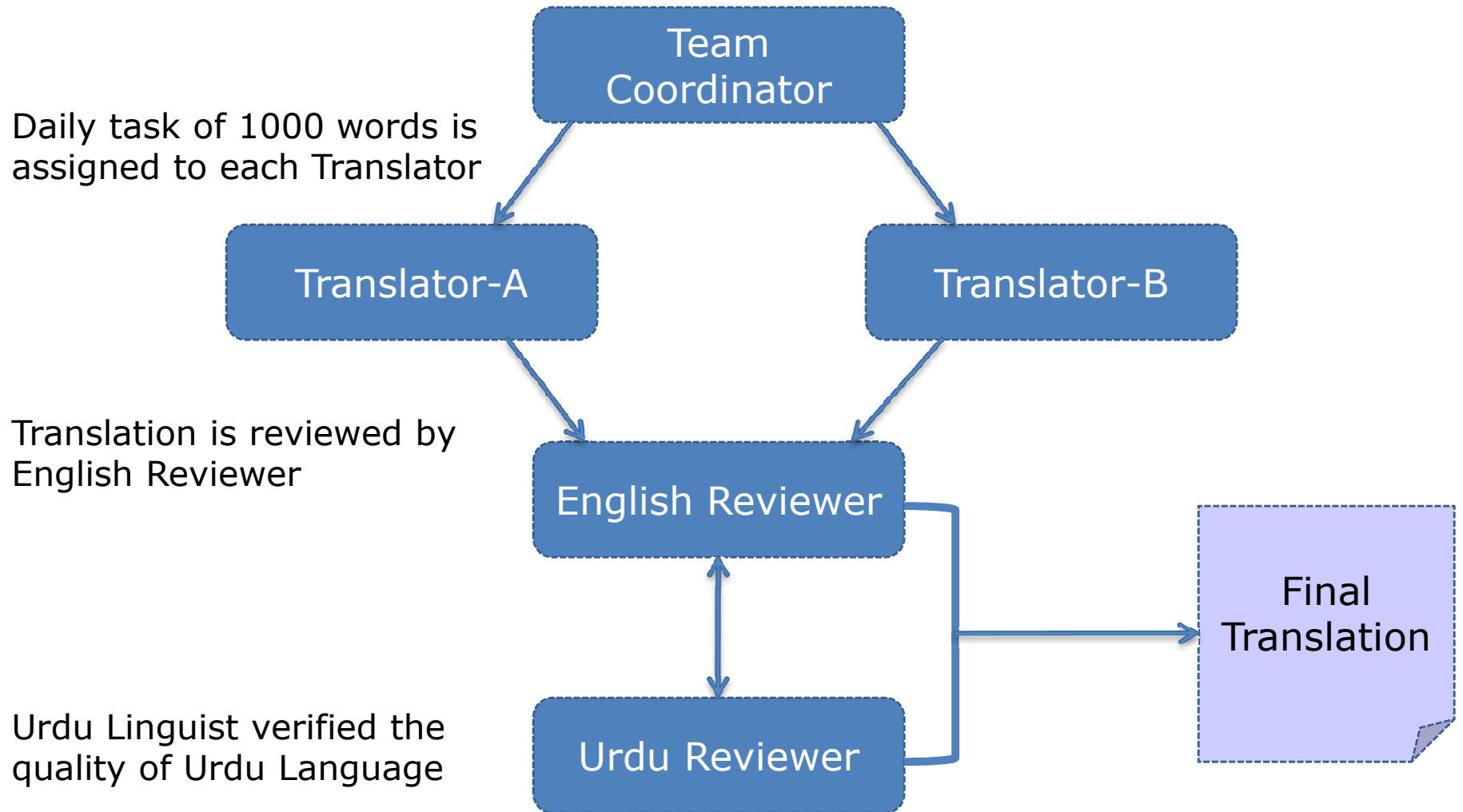
- PIA → پی آئی اے vs. پی آئی اے
- U.S. → یو-ایس
- Dr. → ڈاکٹر "Doctor"
- Sep. → ستمبر "September"

## Translation Quality and Assurance (cont.)



- A sample corpus of about 15,000 source words are translated to investigate the potential difficulties in translation and to acquire the domain knowledge
- Translation guidelines are documented to ensure the consistency in translation and to assist the translators

# Translation Quality and Assurance



# Thanks



The corpus is available at  
[http://crulp.org/software/ling\\_resources/UrduNepaliEnglishParallelCorpus.htm](http://crulp.org/software/ling_resources/UrduNepaliEnglishParallelCorpus.htm)

# Appendix A



- Penn Treebank source text
  - `.\Penn Tree\Treebank-3\PARSED\MRG\WSJ\00\`
  - `.\Penn Tree\Treebank-3\PARSED\MRG\WSJ\01\`
  - `.\Penn Tree\Treebank-3\PARSED\MRG\WSJ\02\`
  - `.\Penn Tree\Treebank-3\PARSED\MRG\WSJ\01\`