PET: a Tool for Post-editing and Assessing Machine Translation

Wilker Aziz†, *Sheila C. M. de Sousa*† *and Lucia Specia*§

† Research Group in Computational Linguistics
University of Wolverhampton
{w.aziz, sheila.castilhomonteirodesousa}@wlv.ac.uk

§ Department of Computer Science
University of Sheffield
l.specia@sheffield.ac.uk

Post-editing of MT

- Larger volumes of translations, less time and costs
- Help understand problems in such translations
- Way of evaluating the quality of translations

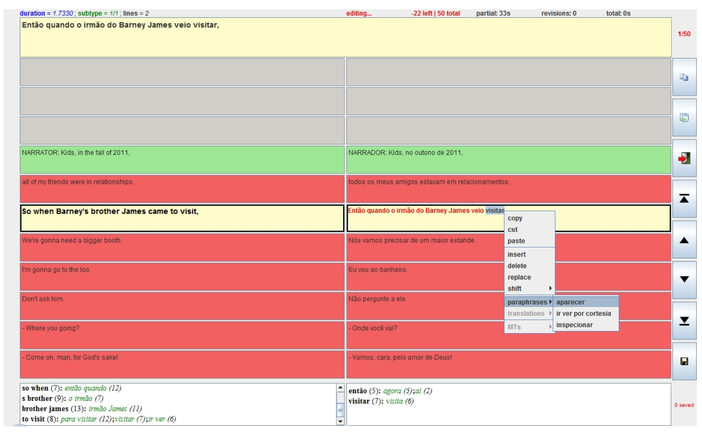
Goals

- Facilitate the post-editing of translations from **any MT system** so that they reach publishable quality
- **Collect sentence-level and word-level information** from post-editing (and translation)

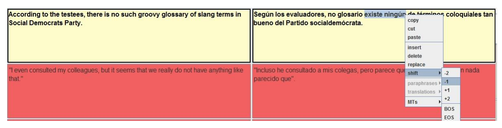
PET: Post-Editing Tool

- **Java** tool - works on any platform
- Standalone tool - **multiple MT systems**

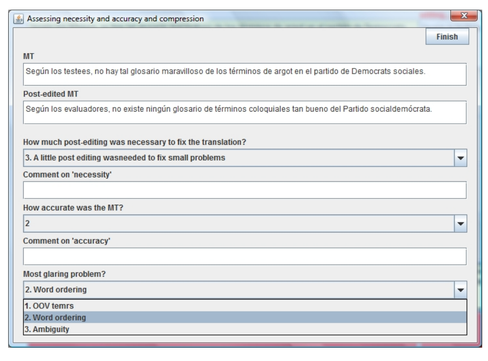
- **Unit**: sentence, paragraph, phrase, etc.
- Monolingual and bilingual **dictionaries**



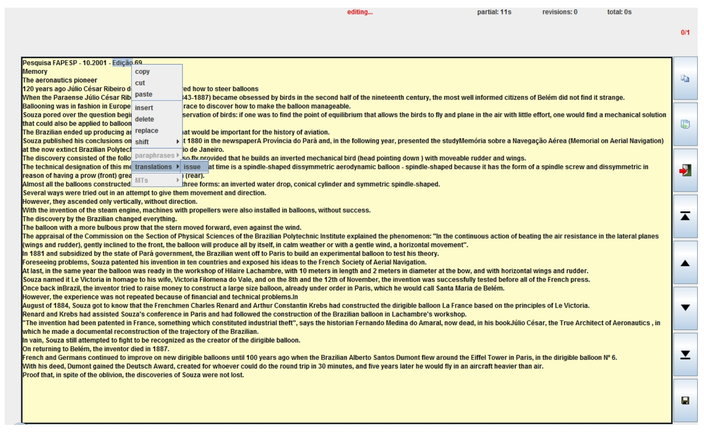
- For the active unit:
  - **Top box** displays the original translation, an alternative translation, or a reference translation
  - **Top bar** displays attributes that can be made visible, e.g.: the "producer" of the translation
  - **Bottom boxes** display additional **sub-sentential** information for source / translation
  - **Edit operations**



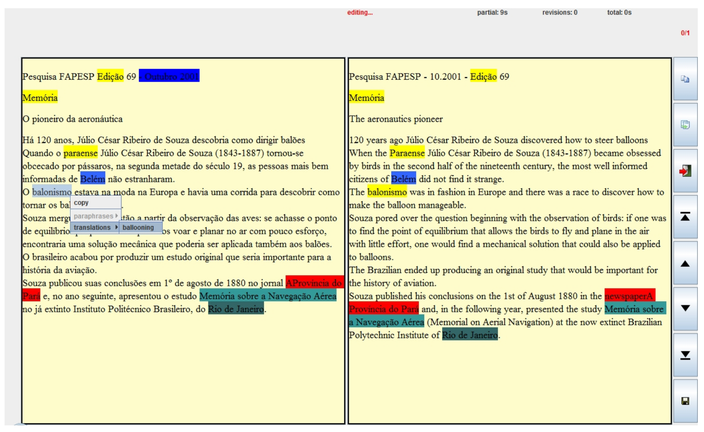
- API to add **new attributes** and **behavior**
- A job can be **paused**, **interrupted** and **re-stared** after a unit is completed
- After each unit, customizable **assessment window(s)** to collect **explicit** indicators:



- Suitable to **various tasks**: HT, (monolingual) PE



- Renders **HTML**



Input files

- The input format for PET is **XML**:

```
<task type="pe" id="3">
  <S producer="xfiles.en">Excuse me.</S>
  <R producer="xfiles.pt">- Com licença,</R>
  <MT producer="google">Desculpe-me.</MT>
</task>
```

- In a **config file**, PET allows the customization of the following, among other features:
  - how many units are displayed at a time
  - which attributes are displayed
  - whether explicit assessments, and which assessments, are requested in the assessment window
  - whether a unit should be hidden before its editing time starts to be recorded

Output files

- **XML** file: one annotation object per unit with the final translation, implicit and explicit effort indicators:

```
-<task id="3" status="FINISHED" type="pe">
   <S producer="xfiles.en">Excuse me.</S>
   <R producer="xfiles.pt">- Com licença,</R>
   <MT producer="google">Desculpe-me.</MT>
  -<annotations revisions="1">
   -<annotation r="1">
      <PE producer="pet">Desculpe-me.      </PE>
      <indicator id="editing">3s</indicator>
      <indicator id="assessing">0s</indicator>
      <comment/>
    </annotation>
  </annotations>
</task>
```

- Built-in **implicit effort indicators** (others can be added via the PET's API):
  - Editing time
  - Assessing time
  - Keystrokes: counts of specific groups of keys
  - Edits: types of edits (deletion, insertion, substitution), amount of time spent on each edit, position and offset of the edited segment and resulting text
  - HTER: edit distance between original translation and its PE version
  - Revisions: one log for each revision
- **Scripts** to create input files and parse output files are provided

PE vs translation

- Sousa et al. (2011): objective way of measuring translation quality in terms of PE time
- **Goals**
  - Check whether **post-editing is quicker than translation**
  - **Rank systems** by the amount of time required to post-edit their output
- 11 translators post-edited English-Portuguese TV series subtitles translated using 4 systems. They also translated such sentences
- Assessment: PE effort in [1-4]
- Results:
  - **PE is 40% faster than human translation**
  - PE vs translation in terms of time:

System	Faster than HT
Google	**94%**
Moses	86.8%
Systran	81.2%
Trados	72.4%

  - MT eval metrics - different references:

Metric	Ref	Google	Moses	Systran	Trados
TER	R_0	0.79	**0.75**	0.88	1.01
	P_i	**0.06**	0.21	0.22	0.66
	R_{0-17}	**0.06**	0.19	0.21	0.62
BLEU	R_0	21.51	**22.28**	13.90	09.22
	R_{0-17}	**92.24**	72.04	70.23	28.36

  - Correlation between PE time and HTER or human assessment:

Post-editing time vs	HTER	Assessments
Spearman's ρ	0.72±0.1	**-0.76±0.1**
Pearson's	0.46±0.1	**-0.53±0.1**

PE for quality estimation

- Specia (2011): PET to obtain training data for QE
- **Goal**: assess QE models built using different annotation types in a *task-based* evaluation
- Datasets: *news* **fr-en, en-es**
- **Annotations**: HTER, [1-4] scores and PE time (avg. seconds/word in sentence)
- **3 QE models built** for each language pair
- 4 non-overlapping subsets of 600 unseen translations randomly selected:
  - Quality predictions generated for 3 subsets – **translations ranked best-first**
  - Translations in the 4th subset **not ranked**
- Translators **post-edited as many sentences as possible** in 4 "tasks": **1-hour per task**
- **Results**: QE models built based on time allow post-editing more words in 1 hour

Dataset		Words/second
fr-en	HTER	0.96
	[1-4]	0.91
	time	**1.09**
	unsorted	0.75
en-es	HTER	0.41
	[1-4]	0.43
	time	**0.57**
	unsorted	0.32

Download

Free: http://pers-www.wlv.ac.uk/~in1676/pet/ or http://www.dcs.shef.ac.uk/~lucia/resources/