

# PET: A Standalone Tool for Assessing Machine Translation through Post-editing

ASLIB 2012

**Lucia Specia** and Wilker Aziz

University of Sheffield  
University of Wolverhampton

November 29, 2012

# Outline

- 1 Introduction
- 2 Interface
- 3 File format
- 4 Examples of uses
- 5 Demo and Remarks

# Motivation

## Post-editing of Machine Translation (MT)

- Increase translation **productivity**: larger volumes, less time, less costs
- **Evaluate** translation quality (e.g. MT system comparison)
- **Diagnose** problems in MT systems
- **Collect data** for different purposes, e.g. quality estimation, paraphrases, etc.

### PET

A standalone tool for post-editing

# Motivation

## Post-editing of Machine Translation (MT)

- Increase translation **productivity**: larger volumes, less time, less costs
- **Evaluate** translation quality (e.g. MT system comparison)
- **Diagnose** problems in MT systems
- **Collect data** for different purposes, e.g. quality estimation, paraphrases, etc.

## PET

A standalone tool for post-editing

# Motivation

## Post-editing of Machine Translation (MT)

- Increase translation **productivity**: larger volumes, less time, less costs
- **Evaluate** translation quality (e.g. MT system comparison)
- **Diagnose** problems in MT systems
- **Collect data** for different purposes, e.g. quality estimation, paraphrases, etc.

## PET

A standalone tool for post-editing

# Motivation

## Post-editing of Machine Translation (MT)

- Increase translation **productivity**: larger volumes, less time, less costs
- **Evaluate** translation quality (e.g. MT system comparison)
- **Diagnose** problems in MT systems
- **Collect data** for different purposes, e.g. quality estimation, paraphrases, etc.

## PET

A standalone tool for post-editing

# Motivation

## Post-editing of Machine Translation (MT)

- Increase translation **productivity**: larger volumes, less time, less costs
- **Evaluate** translation quality (e.g. MT system comparison)
- **Diagnose** problems in MT systems
- **Collect data** for different purposes, e.g. quality estimation, paraphrases, etc.

PET

A standalone tool for post-editing

# Motivation

## Post-editing of Machine Translation (MT)

- Increase translation **productivity**: larger volumes, less time, less costs
- **Evaluate** translation quality (e.g. MT system comparison)
- **Diagnose** problems in MT systems
- **Collect data** for different purposes, e.g. quality estimation, paraphrases, etc.

## PET

A standalone tool for post-editing



# Goals of PET

- 1 Facilitate **post-editing** (and **translation**)
- 2 Be **simple** and **flexible**
  - Any MT system (or TM)
  - Any **evaluation purpose**
- 3 **Collect segment- and word-level information** from post-editing (and translation)
  - **quality** assessment
  - **diagnostic** evaluation
  - **productivity** assessment
  - etc.
- 4 **Open-source**: can be **customised** in many ways
- 5 **Free** of cost

# Goals of PET

- 1 Facilitate **post-editing** (and **translation**)
- 2 Be **simple** and **flexible**
  - Any MT system (or TM)
  - Any **evaluation purpose**
- 3 **Collect segment- and word-level information** from post-editing (and translation)
  - **quality** assessment
  - **diagnostic** evaluation
  - **productivity** assessment
  - etc.
- 4 **Open-source**: can be **customised** in many ways
- 5 **Free** of cost

# Goals of PET

- 1 Facilitate **post-editing** (and **translation**)
- 2 Be **simple** and **flexible**
  - Any **MT system** (or **TM**)
  - Any **evaluation purpose**
- 3 **Collect segment- and word-level information** from post-editing (and translation)
  - **quality** assessment
  - **diagnostic** evaluation
  - **productivity** assessment
  - etc.
- 4 **Open-source**: can be **customised** in many ways
- 5 **Free** of cost

# Goals of PET

- 1 Facilitate **post-editing** (and **translation**)
- 2 Be **simple** and **flexible**
  - Any **MT system** (or **TM**)
  - Any **evaluation purpose**
- 3 **Collect segment- and word-level information** from post-editing (and translation)
  - **quality** assessment
  - **diagnostic** evaluation
  - **productivity** assessment
  - etc.
- 4 **Open-source**: can be **customised** in many ways
- 5 **Free** of cost

# Goals of PET

- 1 Facilitate **post-editing** (and **translation**)
- 2 Be **simple** and **flexible**
  - Any **MT system** (or **TM**)
  - Any **evaluation purpose**
- 3 **Collect segment- and word-level information** from post-editing (and translation)
  - **quality** assessment
  - **diagnostic** evaluation
  - **productivity** assessment
  - etc.
- 4 **Open-source**: can be **customised** in many ways
- 5 **Free** of cost

# Goals of PET

- 1 Facilitate **post-editing** (and **translation**)
- 2 Be **simple** and **flexible**
  - Any **MT system** (or **TM**)
  - Any **evaluation purpose**
- 3 **Collect segment- and word-level information** from post-editing (and translation)
  - **quality** assessment
  - **diagnostic** evaluation
  - **productivity** assessment
  - etc.
- 4 **Open-source**: can be **customised** in many ways
- 5 **Free** of cost

# Goals of PET

- 1 Facilitate **post-editing** (and **translation**)
- 2 Be **simple** and **flexible**
  - Any **MT system** (or **TM**)
  - Any **evaluation purpose**
- 3 **Collect segment- and word-level information** from post-editing (and translation)
  - **quality** assessment
  - **diagnostic** evaluation
  - **productivity** assessment
  - etc.
- 4 **Open-source**: can be **customised** in many ways
- 5 **Free** of cost

# PET is not...

- A complete post-editing (or translation) environment, such as those provided by Trados, Systran, etc.
  - 1 Intuitive interfaces
  - 2 Large number of translation/editing functionalities
  - 3 Tighter integration with specific MT or TM systems



# PET is not...

- A complete post-editing (or translation) environment, such as those provided by Trados, Systran, etc.
  - 1 Intuitive interfaces
  - 2 Large number of translation/editing functionalities
  - 3 Tighter integration with specific MT or TM systems

# Focus

## Translation **quality** evaluation via post-editing

- ① *Quality* can be defined in different ways
- ② Some post-editing facilities (e.g. intuitive interface, shortcuts, dictionaries)
- ③ Controlled environment: **logging** of explicit (scores, etc.) and implicit (time, keystrokes, edits) quality indicators

# Focus

## Translation **quality** evaluation via post-editing

- 1 *Quality* can be defined in different ways
- 2 Some post-editing facilities (e.g. intuitive interface, shortcuts, dictionaries)
- 3 Controlled environment: **logging** of explicit (scores, etc.) and implicit (time, keystrokes, edits) quality indicators

# Focus

## Translation **quality** evaluation via post-editing

- 1 *Quality* can be defined in different ways
- 2 Some post-editing facilities (e.g. intuitive interface, shortcuts, dictionaries)
- 3 Controlled environment: **logging** of explicit (scores, etc.) and implicit (time, keystrokes, edits) quality indicators

# Focus

## Translation **quality** evaluation via post-editing

- 1 *Quality* can be defined in different ways
- 2 Some post-editing facilities (e.g. intuitive interface, shortcuts, dictionaries)
- 3 Controlled environment: **logging** of explicit (scores, etc.) and implicit (time, keystrokes, edits) quality indicators

# Outline

- 1 Introduction
- 2 Interface**
- 3 File format
- 4 Examples of uses
- 5 Demo and Remarks

# Editing/translating

editing...		partial: 8s	revisions: 0	total: 0s
vais voltar para Índia ?				1/10
My brother-- he's got a big crush on Bernadette.	meu irmão tem uma queda por Bernadette .			
You're moving back to India?	vais voltar para Índia ?			
It's-it's not what it looks like.				
It's not what it looks like.				
It's not what it looks like.				
It's not what it looks like.	não é o que parece .			
				1 saved 04:37:34

# Editing/translating (ctd)

The screenshot displays the PET interface during an editing session. At the top, a status bar shows "editing...", "partial: 5s", "revisions: 0", and "total: 0s". The main workspace is a table with two columns: source text on the left and target text on the right. The rows are color-coded: yellow for the current row being edited, green for completed rows, and red for rows that are not yet translated. The source text in the yellow row is "não é o que parece ." and the target text is "meu irmão tem uma quedinha por Bernadette .". The source text in the green row is "My brother-- he's got a big crush on Bernadette." and the target text is "você vai voltar para Índia?". The source text in the yellow row below is "It's-It's not what it looks like." and the target text is "não é o que parece .". The sidebar on the right contains icons for undo, redo, search, and other navigation functions.

Source Text	Target Text
não é o que parece .	meu irmão tem uma quedinha por Bernadette .
My brother-- he's got a big crush on Bernadette.	você vai voltar para Índia?
It's-It's not what it looks like.	não é o que parece .



# Editing/translating (ctd)

editing... partial: 2s revisions: 0 total: 0s

0/1

0 saved

<p>Pesquisa FAPESP Edição 69 Outubro 2001</p> <p><b>Memória</b></p> <p>O pioneiro da aeronáutica</p> <p>Há 120 anos, Júlio César Ribeiro de Souza descobria como dirigir balões</p> <p>Quando o <b>paraense</b> Júlio César Ribeiro de Souza (1843-1887) tornou-se obcecado por pássaros, na segunda metade do século 19, as pessoas mais bem informadas de <b>Belém</b> não estranharam. O <b>balonismo</b> estava na moda na Europa e havia uma corrida para descobrir como tornar os balões dirigíveis.</p> <p>Souza mergulhou na questão a partir da observação das aves: se achasse o ponto de equilíbrio que permite aos pássaros voar e planar no ar com pouco esforço, encontraria uma solução mecânica que poderia ser aplicada também aos balões.</p> <p>O brasileiro acabou por produzir um estudo original que seria importante para a história da aviação.</p> <p>Souza publicou suas conclusões em 1º de agosto de 1880 no jornal <b>A Província do Pará</b> e, no ano seguinte, apresentou o estudo <b>Memória sobre a Navegação Aérea</b> no já extinto Instituto Politécnico Brasileiro, do <b>Rio de Janeiro</b>.</p>	<p>Pesquisa FAPESP - 10.2001 - Edição 69</p> <p><b>Memória</b></p> <p>The aeronautics pioneer</p> <p>120 years ago Júlio César Ribeiro de Souza discovered how to steer balloons</p> <p>When the <b>Paraense</b> Júlio César Ribeiro de Souza (1843-1887) became obsessed by birds in the second half of the nineteenth century, the most well informed citizens of <b>Belém</b> did not find it strange.</p> <p>The <b>balonismo</b> was in fashion in Europe and there was a race to discover how to make the balloon manageable.</p> <p>Souza pored over the question beginning with the observation of birds: if one was to find the point of equilibrium that allows the birds to fly and plane in the air with little effort, one would find a mechanical solution that could also be applied to balloons.</p> <p>The Brazilian ended up producing an original study that would be important for the history of aviation.</p> <p>Souza published his conclusions on the 1st of August 1880 in the <b>newspaper A Província do Pará</b> and, in the following year, presented the study <b>Memoria sobre a Navegação Aérea</b> (Memorial on Aerial Navigation) at the now extinct Brazilian Polytechnic Institute of <b>Rio de Janeiro</b>.</p>
---	--

# Editing/translating (ctd)

**errors** = 2,4450; **subtypes** = 1/1; **lines** = 2      **editing...**      **-9 left | 48 total**      **partial: 9s**      **revisions: 0**      **total: 0s**

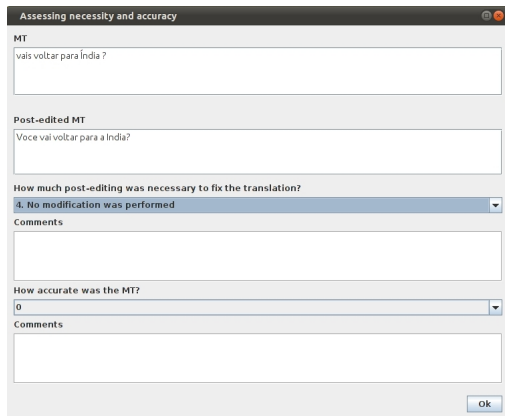
É somente quando adultos que se tornam difíceis.		7/50 7 saved 03:42:58
It's a boy.	É um menino.	
How is he? Is he okay?	Como ele está? Ele está bem?	
Another, you were tired.	Outro, estavam cansados.	
He's perfect.	Ele é perfeito.	
It's only as adults that we become difficult.	É somente quando adultos que se tornam difíceis.	
<b>difficult</b> (9): <i>difícil</i> (7) <b>it's</b> (5): <i>é</i> (1) <b>adults that</b> (11): <i>adultos que</i> (11)	<b>somente</b> (7): <i>só</i> (2) <b>quando</b> (8): <i>em que ocasião</i> (14); <i>em que tempo</i> (12); <i>em que época</i> (12)	

# Editing/translating

- **Unit of text**: phrase, sentence, text of any length
- Units grouped into **jobs**: mixture of units to translate or post-edit from one or more “systems”
- Top text box for **alternative translations**: other systems, past revisions, or reference translation
- Bottom text boxes for **external information** e.g. translation options that match words or phrases in the active unit, from monolingual and bilingual dictionaries
- Optional **attributes**, possibly with certain “behaviour”, e.g. number of characters used in the post-edited unit; block edits above a certain length
- Most widgets are **configurable**

# Assessing

Once a unit is completed, optional **assessment window(s)**:



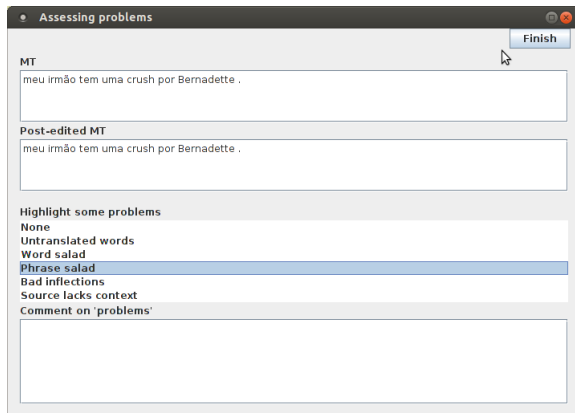
The screenshot shows a dialog box titled "Assessing necessity and accuracy" with a standard window control bar (minimize, maximize, close). The dialog is divided into several sections:

- MT:** A text box containing the original machine translation: "vais voltar para Índia ?".
- Post-edited MT:** A text box containing the human-edited translation: "Voce vai voltar para a Índia?".
- How much post-editing was necessary to fix the translation?** A dropdown menu with the selected option "4. No modification was performed".
- Comments:** An empty text box for providing feedback on the post-editing level.
- How accurate was the MT?** A dropdown menu with the selected option "0".
- Comments:** An empty text box for providing feedback on the overall accuracy of the machine translation.

An "Ok" button is located at the bottom right of the dialog.

# Assessing (ctd)

Once a unit is completed, optional **assessment window(s)**:



# Logging

Built-in implicit assessment indicators:

- **Time** spent **editing** a unit
- **Time** spent **assessing** a unit
- **Counts of groups of keys**, such as white keys, non-printable keys and non-white/printable keys
- **Timestamped edits** (deletion, insertion, substitution, shift), i.e., words or phrases edited and how much **time** each edit required
- **Edit distance** between translation and its post-edited version
- Support for **revisions**: every time box is edited is recorded separately

# Outline

- 1 Introduction
- 2 Interface
- 3 File format**
- 4 Examples of uses
- 5 Demo and Remarks

# Logging

Input:

```
<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>
```

Output:

```
-<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>
```

Tools available to produce input and process output files



# Logging

Input:

```
<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>
```

Output:

```
-<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>
```

Tools available to produce input and process output files

# Availability

## Where:

- Java build + documentation + examples:

`http://pers-www.wlv.ac.uk/~in1676/pet`

- Source code + Java docs:

`https://github.com/wilkeraziz/PET`

**License:** LGPL

# Availability

## Where:

- Java build + documentation + examples:

<http://pers-www.wlv.ac.uk/~in1676/pet>

- Source code + Java docs:

<https://github.com/wilkeraziz/PET>

License: LGPL

# Availability

## Where:

- Java build + documentation + examples:

`http://pers-www.wlv.ac.uk/~in1676/pet`

- Source code + Java docs:

`https://github.com/wilkeraziz/PET`

**License:** LGPL

# Outline

- 1 Introduction
- 2 Interface
- 3 File format
- 4 Examples of uses**
- 5 Demo and Remarks

# Comparing translation systems

## RANLP-11: [de Sousa et al., 2011]

- PET to compare 3 MT vs 1 TM systems vs translation from scratch
- Sitcom and movie subtitles:
  - Translating from scratch can be **73% slower than post-editing** a draft translation
  - SMT systems (Google and Moses) performed the best

# Comparing translation systems

## RANLP-11: [de Sousa et al., 2011]

- PET to compare 3 MT vs 1 TM systems vs translation from scratch
- Sitcom and movie subtitles:
  - Translating from scratch can be **73% slower than post-editing** a draft translation
  - SMT systems (Google and Moses) performed the best

**How often** a system produced an output that was more quickly post-edited than other systems:

<b>System</b>	Google	Moses	Systran	Trados
Google	-	139	161	187
Moses	69	-	122	164
Systran	69	106	-	145
Trados	48	67	89	-

# Comparing translation systems

## RANLP-11: [de Sousa et al., 2011]

- PET to compare 3 MT vs 1 TM systems vs translation from scratch
- Sitcom and movie subtitles:
  - Translating from scratch can be **73% slower than post-editing** a draft translation
  - SMT systems (Google and Moses) performed the best

**How often** post-editing a system output was faster than translating from scratch:

<b>System</b>	Faster than human translation
Google	94%
Moses	86.8%
Systran	81.20%
Trados	72.40%



# Collecting data for Quality Estimation

## EAMT-11: [Specia, 2011]

- QE systems aim at minimising post-editing time and human frustration
- PET to collect quality indicators (time, scores, edit distance) to learn QE models
- PET to assess QE models: models learnt from **time** reliably rank translations by their PE effort

**Time to post-edit** subset of sentences predicted as “low PE effort” **vs** time to post-edit random subset of sentences

Language	no QE	QE
fr-en	0.75 words/sec	<b>1.09</b> words/sec
en-es	0.32 words/sec	<b>0.57</b> words/sec

# Collecting data for Quality Estimation

## EAMT-11: [Specia, 2011]

- QE systems aim at minimising post-editing time and human frustration
- PET to collect quality indicators (time, scores, edit distance) to learn QE models
- PET to assess QE models: models learnt from **time** reliably rank translations by their PE effort

**Time to post-edit** subset of sentences predicted as “low PE effort” **vs** time to post-edit random subset of sentences

Language	no QE	QE
fr-en	0.75 words/sec	<b>1.09</b> words/sec
en-es	0.32 words/sec	<b>0.57</b> words/sec

# Post-editing subtitles with space constraints

## EAMT-12: [Aziz et al., 2012]

- Add compression constraints to an MT system to generate length compliant subtitles
- PET to **guide** post-editing according to length and time requirements for every unit
  - Show space limitation (attribute)
  - Change colour of translation if too long
  - Offer shorter paraphrases
- PET to evaluate edit distance and length

System	Dexter		How I Met..		Terra Nova	
	TER ↓	LENGTH	TER ↓	LENGTH	TER ↓	LENGTH
Moses <sub>t</sub>	30.3	116.0	20.0	108.5	33.8	120.2
Google	63.6	156.5	52.8	144.3	63.1	152.1
Moses <sub>LP2</sub>	29.5	115.5	21.0	109.1	33.4	119.3
Moses <sub>LP1</sub>	28.3	115.8	20.7	110.0	34.8	119.8

# Post-editing subtitles with space constraints

## EAMT-12: [Aziz et al., 2012]

- Add compression constraints to an MT system to generate length compliant subtitles
- PET to **guide** post-editing according to length and time requirements for every unit
  - Show space limitation (attribute)
  - Change colour of translation if too long
  - Offer shorter paraphrases
- PET to evaluate edit distance and length

System	Dexter		How I Met..		Terra Nova	
	TER ↓	LENGTH	TER ↓	LENGTH	TER ↓	LENGTH
Moses <sub>t</sub>	30.3	116.0	20.0	108.5	33.8	120.2
Google	<b>63.6</b>	<b>156.5</b>	<b>52.8</b>	<b>144.3</b>	<b>63.1</b>	<b>152.1</b>
Moses <sub>LP2</sub>	29.5	<b>115.5</b>	21.0	109.1	<b>33.4</b>	<b>119.3</b>
Moses <sub>LP1</sub>	<b>28.3</b>	115.8	20.7	110.0	34.8	119.8

# PE to assess domain adaptation

**Canadian AI:** NLP Tech and University of Montreal  
[Sankaran et al., 2012]

- Assess **domain adaptation** techniques
- PET to show a significant reduction in PE time: one second per word
- The reduction would save 3 hours/day in a production environment with a translation capacity of 10000 words/day

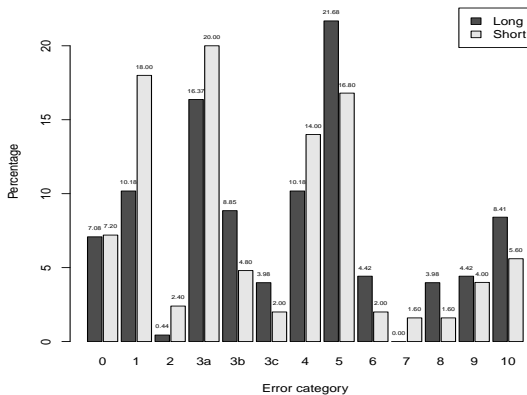
# PE time and cognitive effort

**AMTA-12 WPTP**: [Koponen et al., 2012]

- PET to analyse **post-editing process**:
  - TIME: post-editing time of a sentence
  - SPW: seconds per word
  - KEYS: number of keystrokes
  - HTER: edit distance between MT and PE

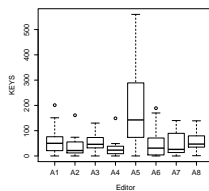
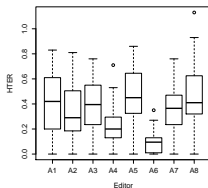
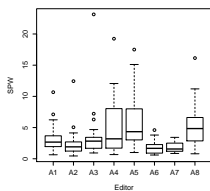
# PE time and cognitive effort (ctd)

Take sentences with **long/short PE times** and **similar # edits** and perform an error analysis to determine if errors were **easy** or **difficult**:



# PE time and cognitive effort (ctd)

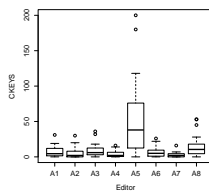
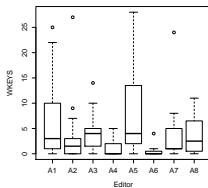
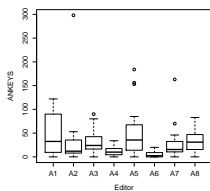
Analyse **human variability** in post-editing:



Seconds per word

HTER

Keystrokes



Alphanumeric keys

White keys

Control keys



# Other uses

- Teaching post-editing (@Gent)
- Assessing controlled languages (@Wolverhampton)
- Assessing multiple choice question generation systems (@Wolverhampton)
- ...

Interested in using?

[l.specia@sheffield.ac.uk](mailto:l.specia@sheffield.ac.uk)

# Outline

- 1 Introduction
- 2 Interface
- 3 File format
- 4 Examples of uses
- 5 Demo and Remarks**

# Future work

- Add more post-editing functionalities
- Add alignment information (source-target phrases)
- Different input formats
- More detailed logging
- Support for formatting tags

# PET: A Standalone Tool for Assessing Machine Translation through Post-editing

ASLIB 2012

**Lucia Specia** and Wilker Aziz

University of Sheffield  
University of Wolverhampton

November 29, 2012



Aziz, W., de Sousa, S. C. M., and Specia, L. (2012).  
Cross-lingual sentence compression for subtitles.  
*In The 16th Annual Conference of the European Association for Machine Translation, EAMT '12*, pages 103–110, Trento, Italy.



de Sousa, S. C. M., Aziz, W., and Specia, L. (2011).  
Assessing the post-editing effort for automatic and semi-automatic translations of DVD subtitles.  
*In Conference Recent Advances in Natural Language Processing 2011*, pages 97–103, Hissar, Bulgaria.



Koponen, M., Aziz, W., Ramos, L., and Specia, L. (2012).  
Post-editing time as a measure of cognitive effort .  
*In AMTA 2012 Workshop on Post-Editing Technology and Practice (WPTP 2012)*, pages 11–20, San Diego, USA.



Sankaran, B., Razmara, M., Farzindar, A., Khreich, W., Popowich, F., and Sarkar, A. (2012).  
Domain adaptation techniques for machine translation and their evaluation in a real-world setting.  
*In Proceedings of the 25th Canadian Conference on Artificial Intelligence*.



Specia, L. (2011).

Exploiting objective annotations for measuring translation post-editing effort.

In *Proceedings of the 15th Conference of the European Association for Machine Translation*, pages 73–80, Leuven.