Colours, clothing and food in CorTrad:
why corpus-based translation studies are revealing

Diana Santos (Linguateca & Univ. of Oslo)
Stella E. O. Tagnin (DLM-USP)
Elisa Duarte Teixeira (Projeto CoMET)

- ICAME 2011 (Oslo, 1-5 June 2011) -
Table of contents

• Short introduction to CorTrad

• Description of the current semantic information and its human revision
  – Size and distribution
  – Annotation rationale

• Some contrastive issues
Parallel corpora and translation


CorTrad in a nutshell

- Annotated, multiversion pt-en-pt parallel corpus
- Started in May 2008
- Partnership:
  - Linguateca (design, development & implementation of computational framework)
  - CoMET Project (design & text collection and edition)
  - NILC - Inter Institutional Center for Computational Linguistics (web hosting)
CorTrad in a nutshell

✓ **Innovations** compared to other parallel corpora:
  
  ✓ **Multiversion** format allows
    ✓ comparison of different translation stages
    ✓ study of revision process

  ✓ **Refined search system** – tailored especially for each genre and text type

  ✓ **Semantic information** – added and human-revised
CorTrad Parallel Subcorpora

Journalistic
Scientific (pt → en)
1,076 texts

Technical-Scientific
Cookbook (pt → en)
130,000 words

Literary
Australian Short Stories (en → pt)
28 texts

Legal
Mercosul Agreements (pt ↔ en)
Coming soon!

USP PhD abstracts (pt → en)
Coming soon!

Canadian Short Stories (en → pt)
Coming soon!

Alice in wonderland (en → pt)
Coming soon!
Journalistic (Science): Revista FAPESP

Original (Brazilian Portuguese)

Published translation (online publication)
Technical-Scientific: Cookbook

Original (Brazilian Portuguese)

Translators’ first version (English)

Revised text (by native speaker)

Published translation (not yet avail. online)
Literary:
Australian short stories
(*learner corpus)

Original (Australian English)
Student’s translation (Brazilian Portuguese)
Revised draft (after teacher’s suggestions)
Published translation
DISPARA (Santos 2002) – system to make parallel corpora available on the Web

✓ Corpus processing system
   → IMS-CWB (Christ et al. 1999), now Open CWB (Evert 2010)

✓ Underlying parser and tagger
   → Portuguese: PALAVRAS (Bick, 2000)
      http://visl.hum.sdu.dk/visl/pt/
   → English: CLAWS (Rayson & Garside 1998)
      http://www.comp.lancs.ac.uk/computing/research/ucrel/claws/
   → Semantic annotation: corte-e-costura (Santos & Mota 2010)

✓ Interface (graphic design by Patricia Tagnin)
CorTrad’s Semantic annotation:

some findings on colour and clothing
## Semantic information: colour

<table>
<thead>
<tr>
<th></th>
<th>Cooking</th>
<th>Scientific news</th>
<th>Short stories</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pure colour</td>
<td>574</td>
<td>372</td>
<td>344</td>
<td>1290</td>
</tr>
<tr>
<td>Conventional</td>
<td>310</td>
<td>153</td>
<td>2</td>
<td>465</td>
</tr>
<tr>
<td>Race</td>
<td>0</td>
<td>45</td>
<td>13</td>
<td>58</td>
</tr>
<tr>
<td>Human</td>
<td>0</td>
<td>7</td>
<td>39</td>
<td>46</td>
</tr>
<tr>
<td>Absence</td>
<td>7</td>
<td>21</td>
<td>22</td>
<td>50</td>
</tr>
<tr>
<td>Wine</td>
<td>87</td>
<td>1</td>
<td>6</td>
<td>94</td>
</tr>
<tr>
<td>Totals</td>
<td>985</td>
<td>599</td>
<td>428</td>
<td></td>
</tr>
<tr>
<td>Word count</td>
<td>134,093</td>
<td>776,284</td>
<td>121,253</td>
<td></td>
</tr>
</tbody>
</table>

Search expression: [sema="cor.*"]
Result type: semantic field
# 10 most recurring colour terms

<table>
<thead>
<tr>
<th>Scientific news (or)</th>
<th>Short stories (tr)</th>
<th>Cookbook (or)</th>
</tr>
</thead>
<tbody>
<tr>
<td>cor</td>
<td>branco</td>
<td>dourar</td>
</tr>
<tr>
<td>branco</td>
<td>azul</td>
<td>branco</td>
</tr>
<tr>
<td>vermelho</td>
<td>vermelho</td>
<td>vermelho</td>
</tr>
<tr>
<td>verde</td>
<td>cor</td>
<td>verde</td>
</tr>
<tr>
<td>negro</td>
<td>negro</td>
<td>preto</td>
</tr>
<tr>
<td><strong>buraco negro</strong></td>
<td>preto</td>
<td>cor</td>
</tr>
<tr>
<td><strong>ultravioleta</strong></td>
<td>verde</td>
<td>tinto</td>
</tr>
<tr>
<td><strong>infravermelho</strong></td>
<td>amarelo</td>
<td>dourado</td>
</tr>
<tr>
<td>azul</td>
<td>pálido</td>
<td>amarelo</td>
</tr>
<tr>
<td><strong>amarelinho</strong></td>
<td>cinza</td>
<td>roxo</td>
</tr>
</tbody>
</table>

Search expression: [sema="cor.*"]

Result type: lemma distribution
“white” collocates in ≠ genres

<table>
<thead>
<tr>
<th>Science news 45 (31)</th>
<th>Short stories 54 (27)</th>
<th>Cookbook 126 (24)</th>
</tr>
</thead>
<tbody>
<tr>
<td>globo</td>
<td>4</td>
<td>mano</td>
</tr>
<tr>
<td>cabelo</td>
<td>4</td>
<td>homem</td>
</tr>
<tr>
<td>cubo</td>
<td>3</td>
<td>pena</td>
</tr>
<tr>
<td>mancha</td>
<td>3</td>
<td>cabelo</td>
</tr>
<tr>
<td>pelagem</td>
<td>2</td>
<td>látex</td>
</tr>
<tr>
<td>luz</td>
<td>2</td>
<td>galão</td>
</tr>
<tr>
<td>população</td>
<td>2</td>
<td>blusa</td>
</tr>
<tr>
<td>célula</td>
<td>2</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Search expr.: [pos="N.*"] ([lema="branco"])[[grupo=“Branco”])  Lemma distrib.
Simple search: “dour.*” is not “golden.*”
Result: “dour.*” ≠ “golden.*”

Meanwhile, lightly toast pine nuts in a dry large frying pan and set aside.

Meanwhile, lightly toast pine nuts in a dry large frying pan and set aside.

Bake for about 25 minutes to heat thoroughly and brown the top (if you want, you can assemble lasagna a day ahead and refrigerate, unbaked).

Drizzle with olive oil and bake for about 15 minutes, until dough is dark brown and risen on the edges.

Drizzle with olive oil and bake for about 15 minutes, until dough is dark brown and risen on the edges.

Heat a drizzle of olive oil in a large ovenproof casserole dish, add half the meat and sear, over a high heat to prevent juices from escaping, until cubes are browned.

Heat a drizzle of olive oil in a large pot that can go into the oven, add half the meat and sear, over high heat to prevent juices from escaping, until cubes are browned.

Transfer browned chunks to a plate, heat another drizzle of olive oil, sear remaining meat and transfer to the same plate.

Transfer browned chunks to a plate, heat another drizzle of olive oil, sear remaining meat and transfer to the same plate.

Transfer browned chunks to a plate, heat another drizzle of olive oil, sear remaining meat and transfer to the same plate.

Transfer browned chunks to a plate, heat another drizzle of olive oil, sear remaining meat and transfer to the same plate.

Heat a little more olive oil, add bacon and, when it begins to brown, stir in diced onion, carrot, celery, and a pinch of salt.

Heat a little more olive oil, add bacon and, when it begins to brown, stir in diced onion, carrot, celery, and a pinch of salt.

Brush roll with egg yolk and bake for 25 minutes or so, until it begins to brown.

Brush cylinder with egg yolk and bake for 25 minutes or so, until it begins to brown.

Spread slices out on baking sheet and return to oven for another 20 minutes, turning once halfway through, so they brown evenly.

in) thick slices. Spread slices out on baking sheet and return to oven for another 20 minutes, turning once halfway through, so they brown evenly.

To start with, as the atmosphere exudes relaxation, there are the medium to large-sized shrimps, coated with sesame seeds, browned in the oven, and served with the aid of toothpicks for dipping in a sweet-n sour pineapple sauce.

To start with, as the atmosphere exudes relaxation, there are the medium to large-sized shrimps, coated with sesame seeds, browned in the oven, and served with the aid of toothpicks for dipping in a sweet-n sour pineapple sauce.

Transfer salmon to preheated oven and bake for 15 minutes or so, just enough to crisp the outer crust and to tenderize the opaque pink flesh in the center.

Transfer salmon to preheated oven and bake for 15 minutes or so, just enough to crisp the outer crust and to tenderize the opaque pink flesh in the center.

The second takes green salad leaves, mango and browned and spicy strips of chicken breasts.

The second takes salad greens, mango and browned and spicy strips of chicken breasts.

Then, discard chicken marinade, heat a drizzle of oil in a large pan, and add zucchini and red onion.
Figurative expr. and terminology

- just declared itself **out to the blue**
- It was to be **black tie**.
- She never refused to go to Melbourne, but it was her hoodoo city, **a black jinx**.
- The dog knew they were coming, and barked **blue murder**.
- would quarrel with her till **the white hours**
- knowing about **brown rice**
- **blackfellow**
- **thin white sliced bread**
- **red wine**
- **red cabbage**

- surgira de repente, **do nada**
- O traje é **a rigor**.
- Nunca se negava a ir a Melbourne, mas era uma cidade de azar, **mau agouro**.
- O cachorro sabia que eles estavam vindo e latiu **desesperadamente**.
- de discutir com ela até **o amanhecer**
- Eu entendia sobre **arroz integral**
- **arborígene**
- **pão de forma**
- **vinho tinto**
- **repolho roxo**
Deemed unnecessary … or different associations

- She rushed to the back of the house and hauled the drowsy **black** pup out of the kennel.
- She began to prowl between the desks, waving the **white** letter like a flag.
- When the torrent of **white** water subsided
- The **brown** smell of his cigar
- **white** and blue enamel bowls
- She had the **tinted** view of the Irish
- **golden** summer

- Ela correu para os fundos da casa e puxou o sonolento filhote para fora da casinha.
- Então começou a rondar as carteiras, balançando a carta como se fosse uma bandeira.
- Quando a corrente de águas espumantes cessou,
- O aroma característico do charuto
- tigelas azuis esmaltadas
- Ela tinha a visão **cor-de-rosa** dos irlandeses
Concluding remarks on colour

- Totally different translation patterns for
  - Figurative language (most cases do not preserve colour)
  - Skin/race/culture colour (more differentiation in English)
  - Real colour
- Scientific news: a lot of (unexpected) colour in scientific terminology: disease names, stars, etc.
- Short stories: high correlation of clothing and colour
# 10 most recurring clothing words

<table>
<thead>
<tr>
<th>Scientific News (or)</th>
<th>Short Stories (tr)</th>
<th>Cookbook (or)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total oc.</strong></td>
<td>58 (20)</td>
<td>380 (72)</td>
</tr>
<tr>
<td><strong>roupa</strong></td>
<td>17</td>
<td>37</td>
</tr>
<tr>
<td><strong>sapato</strong></td>
<td>9</td>
<td>30</td>
</tr>
<tr>
<td><strong>luva</strong></td>
<td>5</td>
<td>27</td>
</tr>
<tr>
<td><strong>calçado</strong></td>
<td>4</td>
<td>26</td>
</tr>
<tr>
<td><strong>camiseta</strong></td>
<td>4</td>
<td>24</td>
</tr>
<tr>
<td><strong>vestir</strong></td>
<td>2</td>
<td>22</td>
</tr>
<tr>
<td><strong>bota</strong></td>
<td>2</td>
<td>13</td>
</tr>
<tr>
<td><strong>anel</strong></td>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td><strong>vestido</strong></td>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td><strong>...</strong></td>
<td>1</td>
<td>10</td>
</tr>
</tbody>
</table>

**Search expression:** [sema="roupa.*"]

**Result type:** lemma distribution
## Semantic information: clothing

<table>
<thead>
<tr>
<th></th>
<th>Cooking</th>
<th>Scientific news</th>
<th>Short stories</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pure clothing</td>
<td>21</td>
<td>50</td>
<td>378</td>
<td>449</td>
</tr>
<tr>
<td>Figurative</td>
<td>0</td>
<td>6</td>
<td>2</td>
<td>8</td>
</tr>
<tr>
<td>Totals</td>
<td>21</td>
<td>56</td>
<td>380</td>
<td></td>
</tr>
<tr>
<td>Word count</td>
<td>134,093</td>
<td>776,284</td>
<td>121,253</td>
<td></td>
</tr>
</tbody>
</table>

Search expression: `[sema="roupa.*"]`  
Result type: semantic field
Figurative expressions and non-clothing-referring clothes

- The arrangement **suited** Henry extremely well.
- He says the government will set an example in `belt-tightening`.
- desafio de **se pôr nos sapatos** de um arqueólogo
- do tamanho de uma caixa de sapatos,
- pense na **roupa**
- O esquema **caiu como uma luva** para Henry.
- Ele diz que o governo estabelecerá um exemplo de `apertar os cintos`.
- task of **wearing the hat** of an archeologist
- the size of a **shoebox**
- think of what you will wear
Concluding remarks on clothing

• Also different translation patterns for
  – Figurative language
  – The issue of wear vs. vestir/calçar and usar

• Scientific news: expressions related to clothing but no clothes (Shoebox, washing machine)

• Short stories: how culture dependent are some descriptions, and what was the stance taken by the translators?
Corpus-based TS are revealing because

• you discover what were the challenges for the translator
• you discover vagueness in both languages
• you find problems you had never thought about before
• you find creative solutions
Acknowledgements

• Thanks to Eckhard Bick and Paul Rayson, for the use of PALAVRAS and CLAWS, respectively.
• Thanks to Sandra Aluísio and Arnaldo Candido Júnior at NILC for hosting and corresponding technical support.
• Thanks to Research Computing Services at Univ. Oslo
• This work was partially funded by the Portuguese government, UMIC, FCCN and the European Union (FEDER and FSE), under grant POSC/339/1.3/C/NAC (Linguateca)