Advantages and Challenges of Using Corpora in Translation Practice

Ana Frankenberg-Garcia

Translators use language that is not their own

Translators use language that is not their own

Translators use language that is not their own

Translators use language that is not their own

Why should translators use corpora?

- Instant access to combined intuitions of dozens (or hundreds) of language users
- Without informants having to think
- Without translators having to bother informants
- Answers to questions that have not been documented in dictionaries, termbases and other edited resources
  - Equivalence
  - Terminology
  - Phraseology
  - Translation decisions


I am looking forward to hearing from you...

I look forward to hearing from you...
Why should translators use corpora?

Business Letter Corpus
A collection of around 1 million words of U.S. and U.K. business letters compiled into a corpus by Professor Yasumasa Someya for his MA at the University of Tokyo

Free online access
http://www.someya-net.com/concordancer

Interpreting corpus data

- Empirical evidence
- But no clear explanations and rules
- It’s up to users to interpret the data
- And results are only as good as the corpus
Interpreting corpus data

- Interpretation is a matter of common sense
- Corpora are made of natural language
- They contain
  - Mistakes
  - Non-standard usage
- What is interesting is what is conventional

Concordances

Portuguese > Casada com um francês
Literal English translation > *Married with a Frenchman*

Is this right?

BNC concordance query: married sort right

<table>
<thead>
<tr>
<th>English</th>
<th>Portuguese literal translation</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Married with a Frenchman</em></td>
<td><em>Casada com um francês</em></td>
</tr>
</tbody>
</table>

Concordances

BNC concordance query: married sort right

<table>
<thead>
<tr>
<th>English</th>
<th>Portuguese literal translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>They are very flexible nowadays.</td>
<td>Eles são muito flexíveis hoje em dia</td>
</tr>
</tbody>
</table>

Concordances

BNC concordance query: married sort right

<table>
<thead>
<tr>
<th>English</th>
<th>Portuguese literal translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Married with a Frenchman</td>
<td>Casada com um francês</td>
</tr>
</tbody>
</table>

English > They are very flexible nowadays.
Portuguese literal translation > Eles são muito flexíveis hoje em dia

Does this sound good?
Tuesday, 07 July 2015

**Concordances**

**BNC concordance query (sentence view): nowadays**

Nowadays, they do this nowadays.

**Concordances**

**ptTenTen concordance query (sentence view): hoje em dia**

Hoje em dia, eles têm mais acesso a esse tipo de conhecimento científico. Dali, há mais acessibilidade.

**Concordances**

**EuroParl corpus concordance query: in my opinion, align with French**

Dans mon opinion, je pense que le français est plus familier.

*Which is better?* peixe assado

**Frequencies**

**BNC frequency query, distribution per text type: preoccupied, worried**

<table>
<thead>
<tr>
<th>Text Type</th>
<th>Preoccupied</th>
<th>Worried</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preoccupied</td>
<td>43.0%</td>
<td>21.0%</td>
</tr>
<tr>
<td>Worried</td>
<td>45.0%</td>
<td>22.0%</td>
</tr>
</tbody>
</table>
Collocations

enTenTen collocation query: opinion

- What verbs can I use before opinion?
- What adjectives can I use before opinion?

Collocations

ptTenTen collocation query: opinião

- How about opinião in Portuguese?

Collocations: subtle differences

Portuguese > elétrico
English translation > electric or electrical

Collocation: subtle differences

Electric or electrical?

Collocation: looking for equivalence

EN - tooth
PT - dente

- escovar 0.06
- pailar 204.85
- arreganhar 153.75
- clarear 199.75
- cerrar 172.37
- lavar 912.67
- branquear 94.67
- ranger 129.62
- trincar 87.66
- ajar 121.63
- desascar 128.66
- cravar 140.60
Collocation: looking for equivalence II

Kilgarriff et al. (2013)

<table>
<thead>
<tr>
<th>Character</th>
<th>Fictional</th>
<th>Cartoon</th>
<th>Favorite</th>
<th>Male</th>
<th>Female</th>
<th>Lead</th>
<th>Moral</th>
<th>Chinese</th>
<th>Disney</th>
<th>Central</th>
<th>Comic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Main</td>
<td>43241</td>
<td>8.79</td>
<td>11580</td>
<td>8.32</td>
<td>7293</td>
<td>7.95</td>
<td>11268</td>
<td>7.58</td>
<td>8897</td>
<td>7.5</td>
<td>4154</td>
</tr>
<tr>
<td>Fictional</td>
<td>2732</td>
<td>7.02</td>
<td>1735</td>
<td>6.38</td>
<td>511</td>
<td>7.09</td>
<td>1109</td>
<td>7.29</td>
<td>11668</td>
<td>7.58</td>
<td>8996</td>
</tr>
</tbody>
</table>

Collocation: looking for equivalence III

Navigating through bilingual word sketches (Kilgarriff et al. 2013)

Verb collocates of goal in general English corpus (enTenTen)

- Accomplish
- Attain
- Reach
- Pursue
- Achieve
- Score
- Fulfill

Word lists

Becoming acquainted with new terminology

- mg
- vial
- ribavirin
- olanzapine
- excipients
- ritonavir
- expiry
- pharmacist
- administration
- ml
- insulin
- dose
- syringe
- warning
- dosage

Phraseology

- mg
- vial
- ribavirin
- olanzapine
- excipients
- ritonavir
- expiry
- pharmacist
- administration
- ml
- insulin
- dose
- syringe
- warning
- dosage

Word lists

Becoming acquainted with new terminology

oral use
marketing authorisation
for injection
medicinal products
authorisation holder
film-coated tablets
special warning
batch number
expiry date
injection site
active substance
package leaflet
pre-filled syringe
special precautions
adverse reactions
hepatic impairment

From terminology to phraseology

Phraseology?

- mg
- vial
- ribavirin
- olanzapine
- excipients
- ritonavir
- expiry
- pharmacist
- administration
- ml
- insulin
- dose
- syringe
- warning
- dosage
### From terminology to phraseology

**Dose?**

Less than 2% of the dose is excreted in the urine. No problem to administer this dose at a different time than the same day.

The maximum tolerated dose (MTD) of fenopirrol for LP is 4 mg. To avoid doses as appropriate.

Anorexia is taken divided doses at meal times. You will usually inject two doses of 0.5 ml in a day.

The recommended starting dose is 15 mg or 30 mg once daily after Day 4, reduce next dose by 25%.

If you accidentally miss a daily dose, just take the next dose as normal. Take the full daily dose prescribed by your doctor.

You can do your dose one unit at a time.

**Increase the dose gradually.**

**EMEA English corpus**

### Consistency

**IPCC report 2013 – volume 1 – The Physical Science Basis**

- Extract ST terminology
  - Aerosol
  - Tropospheric
  - Stratospheric
  - Cryosphere
  - Paleoclimate
  - Interannual variability
  - North Atlantic Oscillation (NAO)
  - Global mean surface temperature
  - Coupled Model Intercomparison Project Phase 5 (CMIP 5)
- Agree on TT terminology beforehand to ensure consistency among different translators
- Feed it to a terminology management system like SDL MultiTerm

### Interpreting

**Preparing for IPCC interpreting job**

- Extract ST terminology
  - Aerosol
  - Tropospheric
  - Stratospheric
  - Cryosphere
  - Paleoclimate
  - Interannual variability
  - North Atlantic Oscillation (NAO)
  - Global mean surface temperature
  - Coupled Model Intercomparison Project Phase 5 (CMIP 5)
- Research TT equivalent terms
- Sight translate concordances with terms to prepare and practice (Xu 2015)

### More

**Populating translation memories with existing parallel corpora**

**OPUS** – the open parallel corpus

- Books - A collection of translated literature (DOGC 2014-07-17.tar.gz - 236 MB)
- DGT - A collection of EU Translation Memories provided by the JRC
- DOGC - Documents from the Catalan Government (DOGC 2014-07-17.tar.gz - 762 MB)
- ECB - European Central Bank
- EMEA - European Medicines Agency
- The EU bookshop corpus (EU)
- EUconstr - The European construction
- EUROPALIER v7 - European Parliament GB
- EUROPALIER - European Parliament GB
- GNOME - GNOME localization
- The Croatian - English WoC collection
- JRC - Agro- legislative EU terms

**CLIMATE CHANGE 2013**

**The Physical Science Basis**

- 375 MB, 1552 pages
- Two-week deadline
Do translators use corpora?

Access to corpora today

Corpus compilation today

Removing barriers is not enough

Awareness of corpora

Do translators want training?

- Survey by Bernardini (2006) — 623 responses, mostly UK and Europe
  - 42% never heard of corpora
- Survey by Gough (2013) — 540 responses, mostly EU
  - 62% Least used technology (compared with term bases, translation memories, glossaries, web searches)
- Survey by Gallego-Hemández (2015) — 526 responses, Spain
  - 50% never or almost never used corpora

- No pressure from the industry
  - Many translation jobs demand use of TM systems
  - Software developers/agencies/corporate clients profit
  - Corpora not mentioned
    - No immediate gains for big stakeholders
- Why bother?
  - Translations don’t get cheaper 😞
  - No obvious productivity gains 😞
  - Quality 😊
  - Reassurance 😊
  - Flexibility and autonomy 😊
Corpus linguistics modules often available, but…
- Focus on linguistics (not specific to translation)
- Corpora & translation mostly translation studies research (especially after Baker 1993)

Not simple to teach about corpora for translation practice
- Especially in UK
  - Students translate into and out of many languages
- Corpus instructors vs translation instructors

General imbalance regarding corpora of different languages
- Even when available (e.g. BNC, CREA, DeReKo), interfaces and query languages differ

Is it worth the trouble?
- Optional corpus module for translation practice at Surrey

Focus not on research, but on everyday translation
- 11 weeks, 22 hours
- 13 students
- Complicated multilingual setting

French > English
Russian > English
German > English
Spanish > English
English > Chinese
English > Greek
English > Italian

Task-based, consciousness-raising activities (Frankenberg-Garcia 2012)

Guided tasks with English corpora
Students try out similar queries using corpora of other languages
Main corpora used
- BYU: COCA & BNC (Davies 2004, 2008)
- OPUS collection (Tiedemann 2012)
- EMEA, ECB, EuroParl, OpenSubtitles
- Sketch Engine (Kilgarriff et al. 2004, 2014)
  - Via the same interface and using same query language, access to
  - BNC, OPUS collection
  - Huge webcrawled corpora in 60 languages (Jakubíček et al. 2013)
- DIY corpora
  - Pre-selected texts
  - Crawl the web via WebBootCaT, (Baroni et al. 2006)
  - TMX (parallel)

End-of-semester anonymous questionnaires
End-of-semester written assignments

I had never heard of corpora before my MA.
12 true
1 false

I had already used a corpus hands-on before I started my MA.
1 true
12 false

I understand the strengths and limitations of different types of corpora.
3 strongly agree
6 agree
4 neither agree nor disagree
0 disagree
0 strongly disagree
I can carry out simple word queries to retrieve KWIC concordances. 4 strongly agree 9 agree 0 neither agree nor disagree 0 disagree 0 strongly disagree

I can carry out queries involving more than one word. 2 strongly agree 11 agree 0 neither agree nor disagree 0 disagree 0 strongly disagree

I understand the difference between looking up lemmas and looking up plain words. 4 strongly agree 8 agree 1 neither agree nor disagree 0 disagree 0 strongly disagree

I can use part-of-speech tags in my queries. 2 strongly agree 8 agree 1 neither agree nor disagree 2 disagree 0 strongly disagree

I can use corpora to retrieve information about collocation. 5 strongly agree 8 agree 0 neither agree nor disagree 0 disagree 0 strongly disagree

I am able to compare the frequencies of different words or combinations of words within a corpus. 4 strongly agree 8 agree 1 neither agree nor disagree 0 disagree 0 strongly disagree
### Self-assessment after completion of the module

**Questionnaire**

**I am able to use normalized frequencies to compare words across different corpora or sub-corpora.**

<table>
<thead>
<tr>
<th>Strongly Agree</th>
<th>Agree</th>
<th>Neither Agree nor Disagree</th>
<th>Disagree</th>
<th>Strongly Disagree</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

**Self-assessment after completion of the module**

**I am able to build a simple corpus on my own.**

<table>
<thead>
<tr>
<th>Strongly Agree</th>
<th>Agree</th>
<th>Neither Agree nor Disagree</th>
<th>Disagree</th>
<th>Strongly Disagree</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

### Opinion about corpus output

**Questionnaire**

**I find concordances helpful.**

<table>
<thead>
<tr>
<th>Strongly Agree</th>
<th>Agree</th>
<th>Neither Agree nor Disagree</th>
<th>Disagree</th>
<th>Strongly Disagree</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

**I find word lists & frequencies helpful.**

<table>
<thead>
<tr>
<th>Strongly Agree</th>
<th>Agree</th>
<th>Neither Agree nor Disagree</th>
<th>Disagree</th>
<th>Strongly Disagree</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>4</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

**I find collocation queries helpful.**

<table>
<thead>
<tr>
<th>Strongly Agree</th>
<th>Agree</th>
<th>Neither Agree nor Disagree</th>
<th>Disagree</th>
<th>Strongly Disagree</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

### Present uses of corpora outside module

**Questionnaire**

**I use corpora to help me when I am writing in my native language.**

<table>
<thead>
<tr>
<th>Very Often</th>
<th>Often</th>
<th>Sometimes</th>
<th>Rarely</th>
<th>Never</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>3</td>
<td>5</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>

**I use corpora to help me when I am writing in a language that is not my native language.**

<table>
<thead>
<tr>
<th>Very Often</th>
<th>Often</th>
<th>Sometimes</th>
<th>Rarely</th>
<th>Never</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>5</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

**I use corpora to help me with my translation assignments.**

<table>
<thead>
<tr>
<th>Very Often</th>
<th>Often</th>
<th>Sometimes</th>
<th>Rarely</th>
<th>Never</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5</td>
<td>6</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

**I use corpora for other purposes.**

<table>
<thead>
<tr>
<th>Very Often</th>
<th>Often</th>
<th>Sometimes</th>
<th>Rarely</th>
<th>Never</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>3</td>
<td>4</td>
<td>4</td>
</tr>
</tbody>
</table>

### Future uses of corpora

**Questionnaire**

**I am likely to look things up in corpora during my translation exams or for writing my MA dissertation.**

<table>
<thead>
<tr>
<th>Strongly Agree</th>
<th>Agree</th>
<th>Neither Agree nor Disagree</th>
<th>Disagree</th>
<th>Strongly Disagree</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>7</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

**I am likely to carry on using corpora in the future in my work as a translator.**

<table>
<thead>
<tr>
<th>Strongly Agree</th>
<th>Agree</th>
<th>Neither Agree nor Disagree</th>
<th>Disagree</th>
<th>Strongly Disagree</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>6</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Student assignments

- Graded piece of assessment about student uses of corpora in translation with examples
- 3000 word limit (excluding references)
- Corpus of 47,123 running words
  - Overall picture
- Essays read from beginning to end
  - Detailed analysis

Student assignments: corpus analysis

<table>
<thead>
<tr>
<th>Search terms [lemmas]</th>
<th>Concept mentioned</th>
<th>Concept used</th>
</tr>
</thead>
<tbody>
<tr>
<td>frequency, hit, occurrence, token</td>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td>concordance</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>lemma</td>
<td>5</td>
<td>13</td>
</tr>
<tr>
<td>collocation, collocate, word sketch</td>
<td>12</td>
<td>7</td>
</tr>
<tr>
<td>word/frequency list</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>keyword list, keys</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>part-of-speech, part of speech, pos, tag</td>
<td>8</td>
<td>3</td>
</tr>
<tr>
<td>relative/normalized frequency</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

Student assignments: more details

Translation decisions

- English > Chinese technical translation
- Student not sure whether translation of mRNA should keep the English form mRNA or use the Chinese form 信使 RNA
- Looked up frequencies zhTenTen corpus
  - mRNA - 1674
  - 信使 RNA – 106
- Corpus helped her decide to use English form

Student assignments: more details

Translation decisions

- French > English translation of victoire total [total victory]
- Looked up collocates of victory in the BNC
- Unhelpful results
  - Labour/Conservative/great victory
- Reformulated query so as to look up synonyms of total in the context of victory
  - final/outright/complete/conclusive victory
- Immediately spotted what she considered to be best option in context of translation: outright victory

Student assignments: more details

Collocation

- Spanish > English business translation
- Student not sure how to translate cuadro macroeconómico into English
- Tried out several concordance queries in enTenTen
  - Macroeconomic picture – 67
  - Macroeconomic projection – 35
  - Macroeconomic prediction – 9
- Chose to use literal translation but not too happy

Student assignments: more details

Collocation

Had student tried a collocation query to identify which nouns follow macroeconomic...

Source: enTenTen corpus
• German > English business translation
• Compiled a small corpus of different types of companies to become more familiar with terminology in this area
• Used her DIY corpus to look up terms with liability and came up with
  • joint liability
  • non-current liabilities
  • interest-bearing liabilities, etc.
• Added terms to her glossary of business terminology

“Of all the types of corpora available, parallel are undoubtedly the easiest for translators to draw conclusions from because the necessary information can be accessed immediately and terms can be directly compared to their equivalents in another language”

“The parallel corpus often produced few results.”

Only a very small part of what people in general say or write ever gets to be translated, which seriously limits the number and types of texts available for the compilation of parallel corpora. Indeed, this is one of the main reasons why parallel corpora are usually much smaller in scale than monolingual corpora. Frankenberg-García (2009: 60)

“Although my corpus was put together in only a matter of minutes, it still allowed me to study terminology and phraseology related to astronomy in a reasonable amount of depth”

“I find that compiling corpora is more suitable for researchers, linguists and teachers, rather than translators and interpreters.”

“One thing that can make using corpora time-consuming is that once concordances are begun, in my experience, I can find myself looking further and often find interesting things out that I wasn’t looking for in the first place, which isn’t necessarily a negative observation.”

“Compared to dictionaries, they [corpora] offer translators with extensive genuine examples in various contexts, thus can be a powerful complementary tool for understanding the usage of language. However, it is also noted that translators should be careful with their own interpretations for data presented by corpora and examine the reliability of some examples in corpora before making further analysis.”

“Although corpus is highly informative, it is no substitute for other authoritative resources like dictionaries. A better solution would be to combine them both and utilise the advantages of both.”
“my comparable [DIY] corpora saved me time and effort.”

“unexpected insights on the native language […] a precious resource especially in regards with working into a non-native language, in this case English, during the writing of essays.”

“Producing an authentic-sounding TT is, however, especially difficult when you are working out of your native language and I therefore found corpora to be especially useful when translating a text about an Aztec artefact from English into my non-native language, German.”


<table>
<thead>
<tr>
<th>Advantages and Challenges of Using Corpora in Translation Practice</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thank you for your attention and cooperation with these audits.</td>
</tr>
<tr>
<td>Thank you for your attention and for being part of the GilFish.</td>
</tr>
<tr>
<td>Thank you for your attention. I share your view that this is</td>
</tr>
<tr>
<td>standard for parsing. Thank you for your attention.</td>
</tr>
<tr>
<td>I would like to thank you for your attention to this urgent</td>
</tr>
<tr>
<td>matter and it really needs attention. Thank you for your</td>
</tr>
<tr>
<td>attention to this vital issue. Joan, Nathalie and I</td>
</tr>
<tr>
<td>would like to thank you for your attention to this critical</td>
</tr>
<tr>
<td>issue and welcome your</td>
</tr>
<tr>
<td>Thank you for your attention to detail and great writing style.</td>
</tr>
<tr>
<td>Your</td>
</tr>
<tr>
<td>Thank you for your attention to the customers' innovation</td>
</tr>
<tr>
<td>Thank you for your attention to customer's interview.</td>
</tr>
<tr>
<td>Thank you for your attention, respectfully. Let's get ... the</td>
</tr>
<tr>
<td>President is feeling good. I thank you for your attention. You</td>
</tr>
<tr>
<td>had better not take any more from me. Thank you for your</td>
</tr>
<tr>
<td>attention to these issues. I look forward to your</td>
</tr>
<tr>
<td>thank you for your attention and interest. The next few years</td>
</tr>
<tr>
<td>should be in the preparation. Thank you for your attention.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>To Adam Kilgarriff</th>
</tr>
</thead>
<tbody>
<tr>
<td>12 Feb 1960—16 May 2015</td>
</tr>
</tbody>
</table>