Parallel corpora in translation and contrastive studies

Lucie Chlumská
Faculty of Arts, Charles University in Prague
OUTLINE

1. corpus classification and terminology in TS/CS

2. parallel corpora: objectives and issues

3. InterCorp 9: corpus design

4. languages in contrast based on the parallel corpus
Corpora in TS/CS: terminology

PARALLEL CORPORA
Objectives and issues

- to include **originals** and their **translations**
  - segment/sentence alignment, word-to-word alignment?
- to provide a basis for research in TS/CS
- main resource of data for machine translation

- **representativeness** – genres/text types **matter**
  - obvious issue in CL: what texts to include? > what translations to include...?
- **directionality** – small languages vs. big languages
  - different amount of texts translated and available
  - highbrow literature and classics vs. virtually anything is available
PCA: fiction vs. non-fiction
Bidirectional parallel corpus

- same size in **both directions** > „reciprocal“ (Zanettin 2011)
- both a parallel and comparable corpus (e.g. ENPC) > perfect for the analysis of translation universals (s-universals, t-universals)
Directionality matters

- usually, there is no symmetry in translation equivalence
- ALWAYS DEPENDS ON THE CONTEXT

example:

EN shout > CS křičet > EN scream, shout, yell (EN scream > CS křičet, řvát, ječet)
EN come > CS jít > EN go, come
CS hned > DE gleich > CS stejný, hned, stejně
Basic information

- multilingual parallel corpus focused on Czech (pivot)
- Czech as pivot, sentence/segment alignment
  - word-to-word alignment > used in Treq (treq.korpus.cz)

<table>
<thead>
<tr>
<th>Name</th>
<th>Czech – core</th>
<th>Czech – collections</th>
<th>other – core</th>
<th>other – collections</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positions</td>
<td>120,443,181</td>
<td>117,981,673</td>
<td>278,445,878</td>
<td>1,556,840,965</td>
</tr>
<tr>
<td>Number of word forms</td>
<td>96,956,714</td>
<td>89,645,545</td>
<td>231,501,606</td>
<td>1,228,896,294</td>
</tr>
<tr>
<td>Structural attributes</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of documents</td>
<td>1,430</td>
<td>5</td>
<td>2,934</td>
<td>89</td>
</tr>
<tr>
<td>Number of div</td>
<td>1,430</td>
<td>111,263</td>
<td>2,934</td>
<td>1,849,184</td>
</tr>
<tr>
<td>Number of sentences</td>
<td>8,308,814</td>
<td>13,588,082</td>
<td>17,210,601</td>
<td>143,478,514</td>
</tr>
<tr>
<td>Further information</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>reference</td>
<td>YES</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>representative</td>
<td>NO</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>publication date</td>
<td>2016</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>foreign languages</td>
<td>39</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>tagged languages</td>
<td>23</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>lemmatized languages</td>
<td>20</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
InterCorp 9: design

- currently 39 languages
  - in different proportions, not all are lemmatized and/or tagged

- design: core and collections (incl. subtitles)
  - fiction, manual alignment
  - journalism:
    - Project Syndicate: http://www.project-syndicate.org/
    - PressEurop: http://www.presseurop.eu
  - legal texts in the EU languages:
    - Europarl: http://www.statmt.org/europarl/
  - Open Subtitles
    - www.opensubtitles.org
Collections

Automatically aligned texts

Syndicate
Presseurop
Acquis
Europarl
Subtitles
# Tags in different languages

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>bg</td>
<td>✓</td>
<td>✓</td>
<td>TT</td>
<td>R Pde-os-n Ansi Ncnsi</td>
</tr>
<tr>
<td>cs</td>
<td>✓</td>
<td>✓</td>
<td>Morče</td>
<td>RR-6 PDXP6 AAFP6---3A NNFP6---A</td>
</tr>
<tr>
<td>de</td>
<td>✓</td>
<td>✓</td>
<td>TT</td>
<td>APPR ART ADJA NN</td>
</tr>
<tr>
<td>en</td>
<td>✓</td>
<td>✓</td>
<td>TT</td>
<td>IN DT JJS NNS</td>
</tr>
<tr>
<td>es</td>
<td>✓</td>
<td>✓</td>
<td>TT</td>
<td>PREP ART NC ADJ</td>
</tr>
<tr>
<td>et</td>
<td>✓</td>
<td>✓</td>
<td>TT</td>
<td>P--s3 A--p--s3 Nc--s3</td>
</tr>
<tr>
<td>fr</td>
<td>✓</td>
<td>✓</td>
<td>TT</td>
<td>PRP DET:ART ADJ NOM</td>
</tr>
<tr>
<td>hu</td>
<td>✓</td>
<td>✓</td>
<td>HunPos</td>
<td>ART ADJ ADJ NOUN (CAS (ILL))</td>
</tr>
<tr>
<td>it</td>
<td>✓</td>
<td>✓</td>
<td>TT</td>
<td>PRE PRO:demo NOM ADJ</td>
</tr>
<tr>
<td>lt</td>
<td>✓</td>
<td>✓</td>
<td>V.D.</td>
<td>prln jvrd bdvr dktv</td>
</tr>
<tr>
<td>nl</td>
<td>✓</td>
<td>✓</td>
<td>TT</td>
<td>600 370 103 000</td>
</tr>
<tr>
<td>no</td>
<td>✓</td>
<td>✓</td>
<td>OB</td>
<td>prep det adj subst</td>
</tr>
<tr>
<td>pl</td>
<td>✓</td>
<td>✓</td>
<td>TaKlPI</td>
<td>prep:loc:nwok adj:sg:loc:m3:pos</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>adj:sg:loc:m3:pos subst:sg:loc:m3</td>
</tr>
<tr>
<td>pt</td>
<td>✓</td>
<td>✓</td>
<td>TT</td>
<td>SPS DA0 NCFS AQ0</td>
</tr>
<tr>
<td>ru</td>
<td>✓</td>
<td>✓</td>
<td>TT</td>
<td>Sp-1 P--p1 Afp-plf Ncmlpln</td>
</tr>
<tr>
<td>sk</td>
<td>✓</td>
<td>✓</td>
<td>Morče</td>
<td>Eu6 PFfs6 AAsfs6x SSfs6</td>
</tr>
<tr>
<td>sl</td>
<td>✓</td>
<td>✓</td>
<td>totale</td>
<td>Sl Pd-nsq Agpfsq Ncnsl</td>
</tr>
</tbody>
</table>
Where to find the tagset description?

- In the Wiki: http://bit.ly/1bv3ll4

- In the KonText interface:
LANGUAGES IN CONTRAST
Examples of use

word-formation

1. EN: *-ridden, -laden*
   > meaning? combinations? text types? translations?

2. EN: *Hey, ain't you that demon-fighting-son-of-a-bitch?*
   *stared up at it with a the-bigger-they-are-the-harder-they-fall expression*
   > length? translations?

3. CS: deminutives ending in *–eček, -ička*
   > translations? possible equivalents in analytical languages?
Examples of use

**grammar**

4. EN: present perfect and its counterparts in other languages

   *he has never given me a present before vs. he’s got(ta), I’ve been divorced...*

   *have/has/’s/’ve + any word (been) + past participle (been, got(ta))*

   > tense? > aspect? > markers?

5. EN: *-ing clauses* – clauses with participle constructions

   *Having published a draft of this Regulation, ...*

   > transgressives? finite clauses?

6. EN: syntactical feature – *disjunct*

   *Sadly, he came late. Honestly, I didn’t do it.*
Examples of use

pragmatics

7. EN: ...and stuff, sort of..., kind of...
8. CS: vole vs. EN: man? dude? you?
   > use? translations? combinations?

lexicon and phraseology

9. proverbs and sayings in different languages
   EN: light as a feather > in other languages? (ADJ as NOUN)

stylistics / norms of translation

10. verba dicendi
   EN: ..., says Peter/Peter says. > CS? FI? FR?
Thank you for your attention!

Questions 😊 ?

lucie.chlumska@korpus.cz
Bibliography


NON-TYPICAL PATTERNS AND COLLOCATIONS
N-grams: extraction

• 3-grams & 4-grams (strings of 3-4 words, excl. punctuation)

1. automatically generated list of n-grams from Jerome

2. comparison of relative frequencies in T and N

3. selection of the most different ones (occurring in one of the subcorpus only, outliers etc.)

4. manual sorting out of irrelevant results (personal names, text-related phrases etc.)
N-gramy: typical in translations

• 3-grams:
  – Co to sakra, Děláš si legraci, to tak líto, je mi líto, mi to líto, ani v nejmenším, Zkrátka a dobře...

• 4-grams:
  – o čem to sakra, To je v pořádku, je to v pořádku, že je v pořádku, Všechno bude v pořádku, Moc mě to mrzí, až do morku kostí, co do činění s, Pokud jde o mě, Podle mě je to, Pro všechno na světě...

interference from EN (v pořádku, líto, mrzí...)
N-gramy: typical in non-translations

• 3-grams:
  – *jen a jen, další a další, v neposlední řadě, v té době...*

• 4-grams:
  – *stále nové a nové, čím dál tím méně, čím dál tím více, mezi nebem a zemí, a tak není divu, jako jeden z mála, od rána do noci...*

  repetitions, different phrasemes...