Creating a mosaic of English language usage with student-compiled micro-corpora

1 Why?

Because it’s a win-win situation!

a) The research perspective
  • Some English registers are still underresearched (cf. Schubert & Sanchez-Stockhammer 2016).

b) The pedagogical perspective
  • Numerous linguistics students compile corpora for their research (e.g. coursework, BA theses) as experts on specialised topics (e.g. video game commentary).
  • Most of these corpora fall into oblivion in spite of the time and effort their compilation required.
  • Offering students the opportunity to make such corpora available to other researchers may increase students’ motivation (product-orientation, recognition)
  • Accuracy (relevance).

2 What?

ESCC (Erlangen Student Corpus Collection)

• An open-ended collection of student-compiled micro-corpora (started in 2014)
• Corpora currently under construction:

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Acronym</th>
<th>Compiler</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table Talk Show Interruption Corpus</td>
<td>InterTab</td>
<td>Petra Krämer</td>
</tr>
<tr>
<td>Football Commentary Corpus</td>
<td>FCC</td>
<td>Daniel Moelt</td>
</tr>
<tr>
<td>Late-night Show Corpus</td>
<td>LnSC</td>
<td>Luise Neumüller</td>
</tr>
<tr>
<td>Blood Bowl Stream Commentary Corpus</td>
<td>BBSCC</td>
<td>Heiko Eimer</td>
</tr>
</tbody>
</table>

• Downloadable material:
  o Corpus (txt)
    o brief description of corpus • name of compiler • type of texts (e.g. genre, language) • corpus size (in orthographic words) • number of texts/speakers • length range of texts • collection period/dates of texts • collection method/source • compilation and annotation conventions (e.g. tokenisation, lexicon, tagging, syntactic parsing) • sociolinguistic description of speakers (age, sex, native language, education etc.) • research project for which corpus was compiled • name of supervisor • additional information (if relevant) • contact person/contact details • release date/revision number

3 Where?

Erlingo

http://www.erlanger-linguistik-online.uni-erlangen.de/projekte/escc.shtml

4 How?

The compilers
  • Students were informed about ESCC in class.
  • Compilers of potentially relevant high-quality coursework were asked if they would like to participate in the project.

The copyright filter
  • Close collaboration with the legal department
    o no texts or films of artistic/literary value
  • Relatively short extracts from certain types of broadcast spontaneous speech
  • Request permissions from speakers recorded for the corpus

Editing
  • Correction of obvious orthographic mistakes, e.g. ‘I’m so happy with who I am’
  • Standardised coding of metalinguistic information, e.g. ‘<filename>’ [cut]
  • Development of guidelines for compilers of new corpora

The future
  • All supervisors of linguistic research projects at the FAU (Friedrich-Alexander-Universität Erlangen-Nürnberg) can suggest corpora for inclusion in ESCC.
  • Creation of new student corpus collections...?

5 Samples

BBSCC

PCC

6 References