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).

re·gis·ter[1] /ˈredʒɪstə $ -ər/ **noun**

1 Definition by Biber & Conrad (2009: 6) [**countable**] a variety associated with a particular situation of use (including particular communicative purposes)

EXAMPLES *dinner-table conversations, academic research articles, hip-hop lyrics* (Kreyer 2016), *crossword puzzles* (Pham 2016)

- **Each corpus opens up new opportunities** to learn more about
  a) the register(s) it comprises
  b) the English language as a whole
  c) linguistic variation.
- Even small corpora contribute to more **comprehensive linguistic description** and can be clustered or compared.
  ➢make existing resources available

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:

Corpus	Acronym	Compiler
Tabloid Talk Show Interruption Corpus	InterTab	Petra Krammer
Football Commentary Corpus	FCC	Daniel Meinl
Late-night Show Corpus	LnSC	Luise Neumüller
Blood Bowl Stream Commentary Corpus (= Live Video Game Commentary)	BBSCC	Heiko Ermer

- **Downloadable material**:
  - Corpus (txt)
  - Standardised information sheet (pdf):

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/date of texts • collection method/source • compilation and annotation conventions (e.g. tokenisation, lemmatisation, 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/version number

3 Where?

Erlingo Linguistik Erlanger online

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**
  ➢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. double spaces, *I'm so happy witeh who I am*
- Standardised coding of metalinguistic information, e.g. <filename>, [cut]
- Development of guidelines for compilers of new corpora

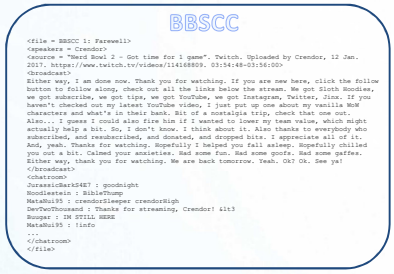
The future

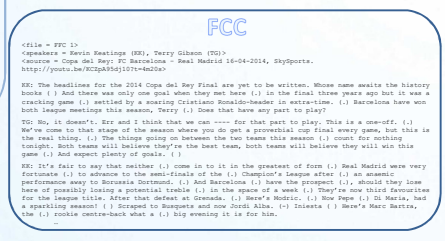
- All supervisors of linguistic research projects at the FAU (Friedrich-Alexander-Universität Erlangen-Nürnberg) can suggest corpora for inclusion in ESCC.

➲ Creation of new student corpus collections...?

5 Samples


InterTab


BBSCC


FCC

6 References

Biber, Douglas & Susan Conrad. 2009. *Register, genre, and style*. Cambridge: CUP.

Kreyer, Rolf. 2016. "Now niggas talk a lotta Bad Boy shit": The register hip-hop from a corpus-linguistic perspective. In Christoph Schubert & Christina Sanchez-Stockhammer (eds.), *Variational text linguistics: Revisiting register in English*. Berlin: de Gruyter Mouton. 87-109.

Pham, Teresa. The register of English crossword puzzles: Studies in intertextuality. In Christoph Schubert & Christina Sanchez-Stockhammer (eds.), *Variational text linguistics: Revisiting register in English*. Berlin: de Gruyter Mouton. 111-136.

Schubert, Christoph & Christina Sanchez-Stockhammer. 2016. *Variational text linguistics: Revisiting register in English*. Berlin: de Gruyter Mouton.

Poster template design: https://www.sciencposters.co.uk/otherservices.htm, 02.05.2017

Christina Sanchez-Stockhammer
christina.sanchez@lmu.de

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Combining corpus linguistic research and teaching may benefit both the linguistic description of the English language and the development of students' research skills.