

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-ter¹ / redʒɪstə \$ -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
 - the register(s) it comprises
 - the English language as a whole
 - 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 Meinel
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?



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



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 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...?

Samples

```
InterTab
<file = InterTab 1: "Help me prove my husband is not the father of my 2
kids!>
<speakers = Triaha (T), Calvis (C), Louisa (L)>
<source = The Triaha Show, UK, 04.11.2012,
https://www.youtube.com/watch?v=7M8G2j4E8: 14:12 - 15:12>
01 T: Oh, I'm interested in a few other things
02 L: How many times have you cheated?
03 C: How, how
04 L: (cheated maybe twice
05 (audience: ooh))
06 T: maybe
07 C: Yeah
08 T: Let me ask you about the Valentine's Day (T) basket?
09 L: (oh)
10 T: Did you (L) give that to somebody who had already given it
11 to you and you...?
12 C: No
13 T: (that's crazy
14 C: No
15 C: No
16 C: No
17 T: How you spent a lot of time at your mum's when you go -->
18 C: (no I don't)
19 C: Oh, I don't
20 I go over to her
21 I do whatever I gotta do for my Mum
22 and I leave
23 T: and where do you go?
24 C: I've no destination
25 There's no way possible I can tell you (L) where I'm going
26 cause I don't know
27 I just go
28 (audience applauding)
```

```
BBSCC
<file = BBSCC 1: Facewell1
<speakers = Crenrod>
<source = "Hard Shell 2 - cut time for 1 game", Twitch, Uploaded by Crenrod, 12 Dec.
2017, https://www.twitch.tv/videos/14168809, 03:14:48-03:16:00>
<transcripts>
Either way, I am done now. Thank you for watching. If you are new here, click the follow
button to follow along, check out all the links below the stream. We got Jinn's
we got subscribe, we got tips, we got YouTube, we got Instagram, Twitter, Jinn. If you
haven't checked out my latest YouTube video, I just got up the about my
characters and what's in their back. Bit of a nostalgia trip, check that one out.
Also, I guess I could also give him if I wanted to show my team value, which might
actually help a bit. Oh, I don't know. I think about it. Also thanks to everybody who
subscribed and rewatched, and donated, and dropped bits. I appreciate all of it.
And, yeah, thanks for watching. Hopefully I helped you fall asleep. Hopefully I
got a bit. Calmed your anxieties. Had some fun. Had some girlfriends.
Either way, thank you for watching. We are back tomorrow. Yeah. Oh! Oh. See ya!
</transcripts>
<chatroom>
Triaaha:chat2427 : goodnight
Mordstein : Bible7ump
MetaBil9 : crenrod/inspwr crenrodinsp
DerVothousand : Thanks for streaming, Crenrod! 1113
Wuzger : @D_VITAL_HBSE
MetaBil9 : finto
</chatroom>
</file>
```

```
FCC
<file = FCC 1:
<speakers = Devin Swearinge (SR), Terry Gibson (TG)
<source = Cup 04: Ray: FC Barcelona - Real Madrid 16-04-2014, YouTube,
http://youtu.be/8fCP895d10T-64030
ER: The headlines for the 2014 Copa del Rey Final are yet to be written. Whose name awaits the history
books? I And there was only one goal when they met here (I) in the final three years ago but it was a
cracking game (I) settled by a scorching Cristiano Ronaldo-beater in extra-time. (I) Barcelona have won
both league meetings this season. Terry (I) Does that have any part to play?
SG: No, it doesn't. Err and I think that we can -- for that part to play. This is a one-off. (I)
We've come to that stage of the season where you do get a proverbial cup final every game, but this is
the real thing. (I) The things going on between the two teams this season (I) count for nothing
though. Both teams will believe they're the best team, both teams will believe they will win this
game (I) And expect plenty of goals. (I)
ER: It's fair to say that neither (I) come in to it in the greatest of form (I). Real Madrid were very
fortunate (I) to advance to the semi-finals of the (I) Champions' League. After (I) an amazing
performance away to Borussia Dortmund. (I) And Barcelona (I) have the prospect (I), should they lose
none of possibly joining a potential finale (I) in the shape of a week (I) after their defeat. (I)
For the League title. After that defeat at Grenade. (I) Here's Madrid. (I) Now Depa (I) Di Maria, had
a sparkling season! (I) Stranded to Barcelona and now Jordi. Also, (I) Dieste (I) Here's Marc Bartra,
the (I) looks mention-back what (I) big evening it is for him.
```

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.