CLEG13

version 07-19-2013

CLEG13 stands for ***C***orpus of ***LE***arner ***G***erman and is a publically available resource for written German learner corpus research. The number ‘13’ refers to the corpus’ public release year, 2013, and also serves to distinguish CLEG13 from an earlier version of the corpus, CLEG, that contained a smaller data set (which is now contained within CLEG13) and was used in several studies before the release of CLEG13.

Details of the corpus

The learners

The learners contributing to CLEG13 were students of German in the Department of European Languages and Cultures at Lancaster University, UK, in the period between 2003 and 2007. Learner profiles were collected from students in all three year-groups of the Lancaster University undergraduate degree course (in CLEG13 these are called Year A, Year B and Year C) so that meta-data could be recorded for all participating students containing:

- Age
- Sex
- Father’s L1, Mother’s L1
- Language spoken at home
- Years of German tuition in school
- Number/length of stays in Germany
- Other foreign languages in acquisition order

All learners in the corpus were part of the ‘post-A level’ strand, i.e. they had passed A-levels in German before starting their university course at Lancaster. This equates to between five and seven years of German tuition at secondary school. On the basis of the learner profiles those participants were selected that had British English as their mother tongue (spoken by both parents) and were not mature students (i.e. they were between 18 and 19 years of age at the start of their degree). This was done in order to achieve a high level of homogeneity across the learner groups. Most students in the first two years (Year A and Year B) had spent a few weeks in Germany on vacation or as part of a school exchange. The students in the final year (Year C) had all spent between six and twelve months in a German-speaking country as part of their ‘year abroad’, which is a compulsory part of the degree scheme for all language majors in the third year of study. Different levels of proficiency were determined by the external criteria of “year of study”. In accordance with the guidelines by the DfES (Department for Education and Skills) achieving an A-level in a Modern Foreign Language is equivalent to level B1/B2 in the Common European Framework of References (CEFR) for Languages. The CEFR defines levels of attainment in different aspects of its descriptive scheme with illustrative descriptors scale. On the CEFR scale, levels A1 and A2 are classed as “basic users”, B1 and B2 as “independent users” and C1 and C2 as “proficient users”. During their university course at Lancaster University, students are expected to work at levels B1/B2 and B2 in Years A and B respectively and at level C1 in Year C, after the year abroad.

For practical and confidentiality reasons, it was not possible to verify these levels for individual learners or to record marks and degree classes. Although it is recognised that this is not an ideal solution, their passing the overall assessment each year and moving into the next year of study (or obtaining their overall degrees at the end) was considered their qualification for the advancing proficiency levels.

2.2 The texts

CLEG13 consists of written learner texts. The texts chosen for the corpus can be classified as "expository-argumentative". This is defined in a purely operational way as texts where the task instructions imply "the presentation and weighing up of arguments, writer's criticism or systematic outlines of abstract concepts" (Lorenz 1999:12). Incidentally, expository-argumentative texts are also the kind of texts that learners are asked to produce most frequently throughout their study course at Lancaster University. This means that this collection criterion yielded the largest amount of reasonably homogeneous texts from all year groups. The only slight deviance to this occurs in the second year of study (Year B), due to different foci in the course design in that year. Firstly, in preparation for their year abroad, the students have to work on an 'intercultural project' where they explore differences in student life and culture between British and foreign students by means of a questionnaire. Secondly, many students take a module on the German press and write a critical analysis of either two German news articles or a German and English news article or a critical summary of the history of the German press after WWII. They also write a report on a collaborative task with students at the University of Graz, Austria. This means that these three projects (which add up to about 72% of Year B tokens) contain large sections of descriptive writing, although all of them also contain some elements of argumentative writing. The Year B subcorpus therefore contains overall longer but fewer individual texts.

All texts are free compositions, which can be broadly divided into text-based tasks and essays (*Aufsätze*). Text-based tasks are *kritische Zusammenfassung* (critical summary) and *kritischer Kommentar* (critical commentary). The *kritische Zusammenfassung* is used more in the first two years of study, where students are asked to first summarise a German text in their own words and then take a critical stance towards the arguments presented in the text. The *kritischer Kommentar* is a task mainly for the final year of study, where students have to develop their own arguments on a topical question, but they receive a suitable text as background information. *Aufsätze*, on the other hand, are independent argumentative texts on general topics such as violence in the media, tuition fees, the death penalty etc. This distinction is important to keep in mind when carrying out lexical studies, as the lexis in a text-based composition may be influenced by its source text.

For each text, a text profile was created detailing:

- Text type (*Aufsatz, Zusammenfassung, kritischer Kommentar*)
- Topic
- Number of words
- Timed/untimed (homework or exam)
- Available reference materials (grammar books, dictionaries, online resources etc.)
- Date of production

- Year of collection

Size and make-up

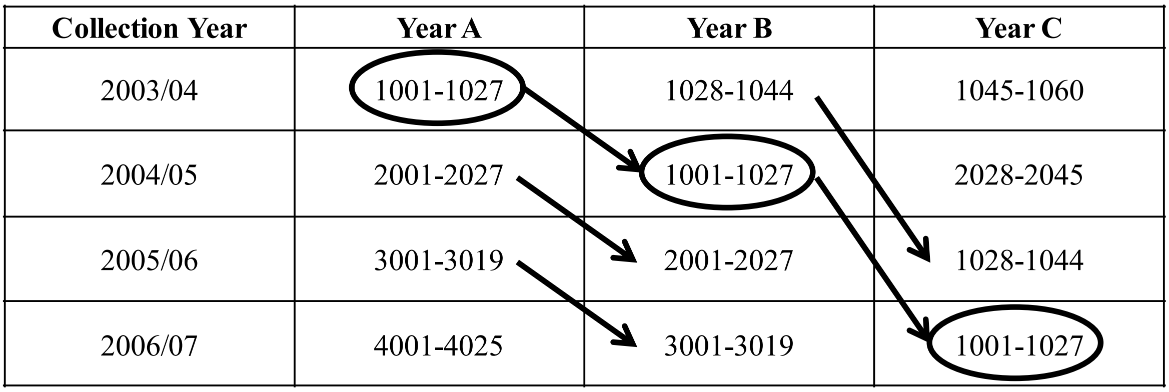
Overall, 149 learners contributed to the corpus, some of them over the whole period of data collection, which was four years (see 2.4 for details of the truly longitudinal part of CLEG13). Learner profiles and text profiles are stored in a Microsoft EXCEL file, where a unique text ID was created for each text. This is linked to all the meta-data on the text as well as the corresponding learner who created the text. This allows the researcher to create subcorpora according to different learner or text specifications if desired. The text ID is devised in the following format that gives information on year group and point of collection at a glance: [(year group)(learner ID)_(running number of text produced by learner)].
To give an example, [a1030_04] means the text was produced in the first year of study ('a') by learner '1030' and it is the fourth text produced by that learner ('_04').
Incidentally, one piece of information contained in the meta-data file is the year of collection for each text. It is therefore possible to recreate the subcorpus CLEG, referred to in several publications before the availability of CLEG13, e.g. Maden-Weinberger 2009; Maden-Weinberger 2013; Hirschmann et al. 2013. CLEG simply refers to the first two years of data collection (2003/04 and 2004/05). The corresponding text IDs can easily be identified from the meta-data file.
Table 1 gives an overview of the overall size of CLEG13 and the size of the different year group subcorpora:

Year group	Texts	Tokens	No. of different topics
Year A	271	73,537	18
Year B	145	99,339	14
Year C	315	146,545	24
Total	**731**	**319,421**	**56**

Table 1: Size of CLEG13

Truly and quasi-longitudinal data in CLEG13

Overall, data is available from three progressive proficiency levels of German as a foreign language, meaning developmental aspects of learner language can be investigated with this quasi-longitudinal corpus. This gives CLEG13 an additional dimension that cannot be found in many learner corpora. However, CLEG13 provides another unique characteristic that is invaluable for investigations into developmental aspects: a truly longitudinal core. This means one cohort of students contributed texts throughout their whole degree course, from first year to final exams. As the students spent one year abroad between the second and the final year of studies at Lancaster, it was necessary to collect data over four years in order to capture this one cohort's progression. However, as texts were always collected in all three year groups A, B and C throughout the four years, there are several more groups of students that provided texts across two years. Table 2 below indicates the truly longitudinal data that is available from different student groups across the years.



Collection Year	Year A	Year B	Year C
2003/04	1001-1027	1028-1044	1045-1060
2004/05	2001-2027	1001-1027	2028-2045
2005/06	3001-3019	2001-2027	1028-1044
2006/07	4001-4025	3001-3019	1001-1027

Table 2: Learner IDs moving through the four collection years

The table shows the truly longitudinal cohort of 27 students (Learner ID 1001-1027), as they move from Year A to B and, after the year abroad, Year C. Some students decide not to continue with German studies after their first year, some decide on German studies as a minor degree option, which means that in Year C, out of the original 27 students, 15 students remain. Their contributions are distributed across the corpus as follows:

Year group	Texts	Tokens
Year A	52	10,818
Year B	37	31,429
Year C	89	37,246
Total	**178**	**79,493**

Table 3: Truly longitudinal section of CLEG13 – 15 learners