The Story of the Learner Corpus LINDSEI_CZ

Tomáš Gráf

ABSTRACT:
The article presents the recently completed Czech subcorpus of the multinational learner corpus of advanced spoken English LINDSEI and aims to draw attention to some of the methodological concerns the field of learner corpus linguistics faces. First, it describes the Louvain family of learner corpora, where this project originated, and provides a detailed description of LINDSEI, its history, design, structure, transcription system and metadata. It then outlines the nature of the Czech subcorpus LINDSEI_CZ, telling the story of its compilation and providing a quantitative description of the corpus size, task sizes and learner variables, as well as a description of the transcription process. The core part of this text discusses methodological concerns affecting learner corpus design and construction and deals with such issues as task design, recording instructions, the matter of learner-participant proficiency, and transcription system employed. It concludes with a consideration of various methodological suggestions and offers the possible view that, despite certain weaknesses, LINDSEI is an invaluable source of highly authentic learner data. The last section provides a thematic categorisation of existing studies on LINDSEI and concludes with descriptions of some future projects. The article calls for a thorough reconsideration of learner corpus design and practice and for the formulation of compilation and research standards which would lead to an increase in the reliability and exploitation potential of learner corpora.

KEY WORDS:
corpus methodology, learner corpora, learner corpus linguistics, LINDSEI, spoken corpora

1 INTRODUCTION

Learner corpora — "electronic collections of spoken or written texts produced by foreign language learners" (Granger, 2004, p. 124) — have long been a widely recognized part of the broader field of corpus linguistics. They offer direct evidence of the processes which are involved in the production of written and spoken texts in an L2,[1] and through the wealth of the data they contain they enable researchers to explore areas as diverse as second language acquisition, psycholinguistics and natural language processing and to utilize their research findings within L2 pedagogy. A recent survey of learner corpus research — *The Cambridge Handbook of Learner Corpus Research* (Granger — Gilquin — Meunier, 2015) — comprises no fewer than twenty-seven specialised chapters, which shows the enormous potential that learner data has for research. Somewhat sadly — if perhaps not surprisingly — the target language for most learner corpora is English. Other languages, however, are also appearing on the scene; of the 144 monolingual learner corpora listed in the survey of

1 See e.g. fluency studies by Götz (2013) and Gráf (2015; 2017) based on the LINDSEI data.

"Learner Corpora around the World",[2] 89 (62%) are for English, 12 (22%) for Spanish, 10 (18%) for German, 9 (16%), for French, 6 (11%) for Italian, 3 (6%) for Finnish, 2 (4%) for Arabic, and 2 (4%) for Persian, and there are also single instances for such languages as Czech, Dutch, Estonian, Gaelic, Hungarian, Chinese, Korean, Norwegian, Russian, Slovene, and Swedish. Besides these, there also exist 14 multilingual corpora with diverse target languages including — besides English and some of the above-mentioned languages — also Catalan, Portuguese and Romanian. The field is thus becoming truly international in two aspects: in the increasing variety of L2s for which learner corpora are being produced, as well as in terms of the growing abundance of learner corpora featuring language produced by learners of English with different mother-tongue backgrounds.[3]

When I was approached by Ondřej Tichý in 2012 and asked whether I might consider making a contribution to the international LINDSEI project by compiling its Czech subcorpus (i.e. learner English produced by speakers with Czech as their L1), I did not hesitate. As an English teacher of twenty years' standing and a teacher trainer at university level with a strong interest in learner language I already had high hopes that learner corpus research might have valid and perhaps even invaluable pedagogical implications. Furthermore, at that time there still existed no such corpus containing data produced by Czech learners of English.

In this article, written five years after the launching of the LINDSEI_CZ project and two years after its completion, I would like to provide not only the story of that subcorpus but also an evaluation of the LINDSEI design. In so doing I aim to provide an outline of how far we have come in our understanding of both pedagogical and learner-corpus-methodological implications.

1.1 THE LOUVAIN COLLECTION OF CORPORA

It was at the end of the 1980s that Sylviane Granger conceived the idea of the Centre for English Corpus Linguistics (CECL) at the Université catholique de Louvain, which set out to create learner and multinational corpora especially for pedagogical purposes. Since then the Centre has generated 14 corpora,[4] some of which are amongst the largest of their kind and include contributions from several countries. This extensive work has led to the creation of a new learner corpus methodology: Contrastive Interlanguage Analysis (Granger, 1996).

The truly pioneering project — in both the local and the international context — was the International Corpus of Learner English (henceforth ICLE). This was started

2 For a comprehensive list of learner corpora around the world see <https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html> (last accessed on 17 August 2017).

3 The term "mother-tongue background" is preferred over "speakers with different L1s" as the former includes speakers who share a common L1 but come from different countries (e.g. the French of France and Belgium), thus making it possible for research to take into account not only linguistic but also educational and cultural factors.

4 See <https://uclouvain.be/en/research-institutes/ilc/cecl/corpora.html> for a complete list.

in the early 1990s and its first version was launched in 2002, with a second version appearing in 2009 and a third being planned for 2017. At the moment of writing it contains 16 national subcorpora[5] and is accompanied by a parallel native corpus, LOCNESS (the Louvain Corpus of Native English Essays), with 324,304 words. ICLE consists of argumentative essays written by high-intermediate and advanced learners of English as L2 and contains 3.7 million words.

As the years went by, CECL launched several other corpora. These include learner corpora such as FRIDA (the French Interlanguage Database), LINDSEI (the Louvain International Database of Spoken English Interlanguage), LONGDALE (the Longitudinal Database of Learner English) and VESPA (the Varieties of English for Specific Purposes Database); pedagogical corpora such as CoNNECT (the Corpus of Native and Non-Native EFL Classroom Teacher Talk) and TeMa (a corpus of textbook materials); translation corpora such as Label France, MUST (Multilingual Student Translation) and PLECI (Poitiers-Louvain Échange de Corpus Informatisé); and specialised corpora such as LOCRA (the Louvain Corpus of Research Articles), MULT-ED (the Multilingual Editorial Corpus) and NESSI (New Englishes Student Interviews).[6]

2 LINDSEI

From the start, ICLE was to be accompanied by a spoken counterpart. This project was commenced in 1995 under the name The Louvain International Database of Spoken English Interlanguage (LINDSEI). It was started by Sylviane Granger, who was later joined by Gaëtanelle Gilquin as the project's co-ordinator. The first version of LINDSEI was launched and released on CD-ROM in 2010; its second version is planned for release in 2018. LINDSEI is accompanied by a parallel native corpus, LOCNEC (the Louvain Corpus of Native English Conversation), containing almost 122,000 words in 50 interviews with native English students of linguistics, and following the same format as that of LINDSEI.

LINDSEI is a corpus of spontaneous spoken English produced by high-intermediate and advanced learners of English with various mother-tongue backgrounds. Version One was made up of 11 national subcorpora[7] which together contained approximately one million words, 554 interviews and 130 hours of recorded material. Version Two is to add a further nine subcorpora[8] to give a total of 20 subcorpora, 1,000 interviews, approximately 250 hours of recordings and almost two million words. At the time of writing, the corpus is distributed with only the orthographic transcriptions and not the actual recordings.

5 Bulgarian, Chinese, Czech, Dutch, Finnish, French, German, Italian, Japanese, Norwegian, Polish, Russian, Spanish, Swedish, Tswana and Turkish.

6 Detailed information on these corpora may be found at <https://www.uclouvain.be/en-258636.html>.

7 Bulgarian, Chinese, Dutch, French, German, Greek, Italian, Japanese, Polish, Spanish and Swedish.

8 Arabic, Brazilian, Basque, Czech, Finnish, Norwegian, Latvian, Turkish and Taiwanese.

2.1 LINDSEI — CORPUS STRUCTURE, TASK DESIGN, PROFICIENCY, AND TRANSCRIPTION SYSTEM

Each national subcorpus comprises a minimum of fifty transcriptions of approximately fifteen-minute recordings, all of which contain three tasks that are identical not only within the given national corpus but right across the whole LINDSEI project. The first is a monologue on one of a choice of three set topics (an experience which has affected you; a journey which has affected you; or a memorable film or play), all of which invite the use of the past tense and the present perfect. The speaker is given two to three minutes to choose a topic and think about what to say on it; in an ideal situation, the task itself takes a minimum of three minutes. The second task is a free conversation with the interviewer, topics here typically including the student's history of studying English, his or her experience of university life and studies, plans for the future, etc., and thus attempt to elicit a variety of grammatical tenses. The third task is a reconstruction of a story based on a set of four pictures. This is a spontaneous, improvisatory task which poses demands on the student's ability to construct coherent, logical text including linking devices and a variety of prepositions.

Speakers are required and expected to be advanced learners of English. Proficiency is not, however, checked prior to the interview (i.e. the speakers are not expected to produce any proof of their proficiency), but is defined institutionally; the speakers are to be students in their 3rd or 4th year of study of English philology, therefore at or near the end of their BA studies, as it is assumed that such speakers ought to possess the requisite level of English.[9] As I will later show, this selection method is far from ideal.

The recordings are orthographically transcribed, and anonymized. There is no punctuation, and full stops are used to mark unfilled pauses. The transcription of filled pauses differentiates between short, long and nasalised pauses, and overlaps are marked using a tag. Non-standard forms (such as *cos*, *dunno*, *kinda* etc.) are maintained, and so are contracted forms. Truncated words are transcribed using the equals sign (e.g. *I reme= remember*). A certain number of phonetic features are also recorded, including syllable or vowel lengthening (using colons, e.g. *I went to: an interesting place*) and stressed articles (e.g. a[ei] and the[i:]). Attention is also paid to prosodic features (whispering, laughing etc.) and non-verbal vocal sounds (e.g. coughing, lip smacking). The speakers' turns are marked using the tags <A>, </A> for the interviewer and <B>, </B> for the learner. Example 1 shows a short extract from one of the transcriptions.

<B> (er) she: she was an economist but now she (er) takes care of a farm with horses </B>
<A> that's a change <overlap /> wow </A>
<B> <overlap /> <starts laughing> yeah <stops laughing> </B>
<A> and she married and Irishman <overlap /> didn't she </A>

9 In the case of the Czech subcorpus, students completing their BA studies are expected to have attained C1 level of the Common European Framework of Reference for Languages.

<B> <overlap /> yeah . yeah </B>
<B> so: . I usually go to Dublin first and visit my friends over there and then . I move to her place and spend . (er) really a . leisure time over there </B>

EXAMPLE 1: A brief extract from one of the LINDSEI_CZ transcriptions.

2.2 LINDSEI — METADATA

Prior to the recording, the speakers are asked to sign an informed consent form and complete a questionnaire. The purpose of the latter is to collect variables which are believed to play a role in the acquisition process; these include social and language-acquisition-related variables such as name, age, gender, nationality, language background (parents' L1s, language(s) spoken at home, other languages spoken by the student), length of study of English at various levels of education, and lengths of stays in English-speaking countries. These are later transferred to a database along with information regarding choice of set topic, duration of the interview and length in tokens, as well as with basic information on the interviewer and his/her degree of familiarity with the student.

3 LINDSEI_CZ — THE CZECH SUBCORPUS OF LINDSEI

The compilation of the Czech subcorpus of LINDSEI was started in 2012 and completed in 2015 at the Department of English Linguistics and ELT Methodology, the Faculty of Arts, Charles University, Prague. The project received financial support from the Institute of the Czech National Corpus, on whose web site it is currently hosted within the KonText interface.[10] The requisite fifty interviewees were recruited from among 3rd- and 4th-year students of English philology at the aforementioned English department; they were interviewed by two of their teachers,[11] whose prime concern was to maintain as natural a flow of communication as possible by providing encouraging feedback and asking open questions whilst keeping their own oral input to a minimum. Most of the recordings were made at the recording studio of the same faculty's Institute of Phonetics.

3.1 LINDSEI_CZ — DATA DESCRIPTION

LINDSEI_CZ contains 123,761 tokens.[12] These include filled pauses and truncated words. Contracted forms with an apostrophe are counted as one word. 77.5% of the total number of words are uttered by the students.

10 For more information see <http://wiki.korpus.cz/doku.php/cnk:lindsei_cz>.
11 Sarah Peters Gráfová and Tomáš Gráf.
12 The total count of positions including all special characters and tags is 135,366.

	A & B turns	Mean	B turns only	Mean	Min. (B turns)	Max. (B turns)
Length in tokens	123,761	2,475 (SD = 386)	95,904	1,918 (SD = 407)	920	3,045
Duration (hh:mm:ss)	12:52:25	15:27 (SD = 2:14)	10:37:42	00:12:45 (SD = 2:24)	0:06:51	0:17:21

TABLE 1: LINDSEI_CZ — corpus size. Length and duration of interviews.

Table 1 provides basic descriptive data regarding the size of the corpus. As is clear from the standard deviations and the large ranges of the values, there is much variability in the data, which can be explained by the fact that the recorded students included both very reticent and rather talkative ones. The data, however, has a normal distribution.

As was mentioned above, each interview is divided into three tasks. As is apparent from Table 2 and Chart 1, responses to Tasks One and Two are similar in terms of number of tokens and overall duration. Task One comprises 42% of the whole corpus, Task Two 47% and Task Three 11%.

	Task One	Task Two	Task Three
Length in tokens	40,584	42,850	12,535
Mean token count	812 (SD = 329)	857 (SD = 284)	251 (SD = 85)
Duration	4 hours 26 minutes	4 hours 38 minutes	1 hour 32 minutes
Mean duration (mm:ss)	5:20	5:56	1:51

TABLE 2: LINDSEI_CZ — task sizes.

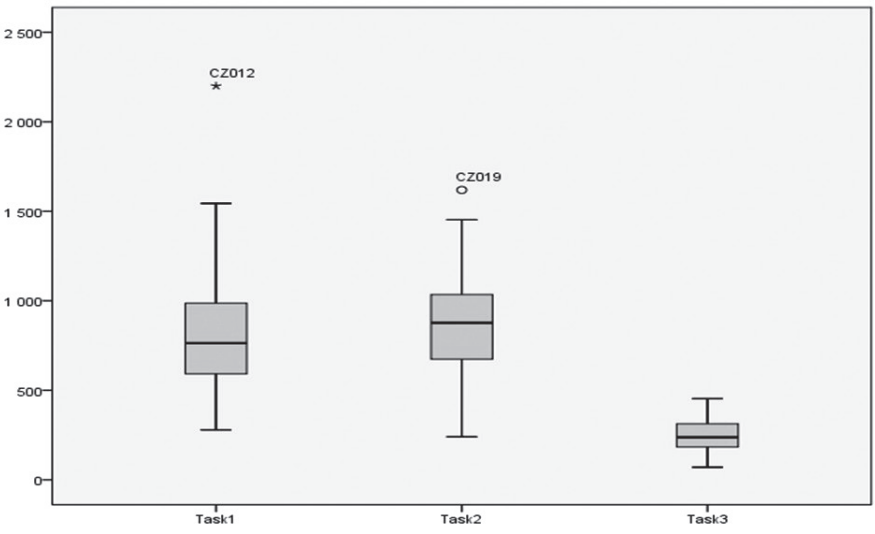

CHART 1: LINDSEI_CZ — comparison of task sizes (Y-scale = number of tokens).

3.2 LINDSEI_CZ — LEARNER VARIABLES

Whilst we strove to achieve balance in the structure of the data, the fact that the majority of our students are women meant that we were unable to recruit the same number of men and women, the final ratio of females to males being 43:7. The average age of the speakers is 22.5 years (SD = 1.6), and prior to university studies they studied English for an average of 9.9 years (SD = 2.6). At the time of being interviewed they had spent 3.4 years (SD = 0.9) studying English at university and an average of 1.2 months in English-speaking countries. This shows that the students learnt their English largely in institutional settings. As for their knowledge of other foreign languages, 25 students mentioned German, 14 French, 7 Spanish, and 4 other languages including Russian, Italian and Dutch. In all of the cases, the home language was Czech.

3.3 LINDSEI_CZ — THE TRANSCRIPTION PROCESS

The laborious process of transcribing the recordings was performed by the speakers themselves as part of their credit requirements for courses in SLA and ELT methodology, both of which included seminars on the specifics of spoken language. Whilst the pedagogical dimension of this experiment proved to be meaningful and — as became clear in subsequent seminar discussions — led the students towards a deeper understanding of the complexity and structure of spoken discourse, the resulting transcriptions varied in quality. The coordinator of the corpus was obliged to spend an average of three hours checking each transcription to guarantee its accuracy and consistency. Features which were especially taxing from both the transcribers' and the checker's view were unfilled pauses (esp. their length), filled pauses (nasalisation and length) and the marking of overlaps. The students reported having needed on average of five hours to transcribe each recording. The total number of hours for transcribing and checking was thus about 400. It would perhaps have been more time-efficient to have had all the recordings transcribed by one person.

The transcriptions followed the rules outlined in the Louvain transcription manual,[13], which was made available to all LINDSEI coordinators, and some of whose details are described in section 2.1 above.

4 LINDSEI — METHODOLOGICAL CONCERNS AND CAVEATS

The experience of compiling LINDSEI_CZ revealed a number of methodological problem areas, some of which might — and perhaps rather ought to — have been considered prior to the project. Adolphs and Knight (2012) recommend that all phases of spoken corpus construction be considered at the design stage. It is perhaps due to the fact that LINDSEI was planned in the early days of learner corpus linguistics that some of these issues did not receive as much attention as they deserved. In the follow-

13 Its full version is available at <https://www.uclouvain.be/en-307849.html>.

ing paragraphs I will attempt to explore some of these "weaknesses" whilst remaining constantly aware of the fact that LINDSEI was and is a truly pioneering project in the field of spoken learner corpora.

4.1 CONCERNS REGARDING THE GENERAL PURPOSE OF THE CORPUS AND THE DESIGN OF THE TASKS

To this day any researcher interested in compiling a LINDSEI subcorpus receives what subsequent work on the given subcorpus reveals to be to be rather scanty information regarding the project. The idea stated is that LINDSEI is a spoken corpus of advanced learner English and a spoken counterpart to ICLE, but it is not specified whether any concrete research questions or interests are intended. How the interviews are actually carried out in terms of communicative content is thus left very much to the coordinator's own experience or initiative. The positive side is that the resulting data is more spontaneous and authentic in being less controlled, but on the other hand a number of questions arise and remain unanswered. Are the interlocutors to elicit particular grammatical forms and, if so, how is this to be achieved? What is the ideal proportion of the individual tasks in terms of duration? How conversationally active should the interlocutor become? These considerations can be illustrated by taking a brief look at task design.

As described above, Task One consists in a short speech on one of three set topics. These all seem to be designed to elicit statements in which the speakers need to contrast the present perfect and the past tense. But was this really the intention? Later analysis of the speech of those who chose the third topic (a film or play which has impressed you) reveals that some speakers opt for the historical present. But might this be the case because it is actually the speakers' avoidance strategy regarding the more complex tenses? As regards speech continuity, what is the interlocutor to do if the speaker is very brief? This was indeed the case with some of the more reticent speakers, and the interlocutor was obliged to intervene by asking supplementary questions and thus turning what had presumably been designed as a monologue into a dialogue. Could not this situation have been easily avoided if the interlocutor had been specifically instructed to inform the student that he would be expected to talk without any interruptions for a number of minutes and to say more than just a few sentences? As a result, in some of the interviews Task One is more dialogical than in others and hence the formal (genre) distinction between Task One and Task Two, which could be significant for research, is negligible or even lost. The dialogue form may be limiting also in the sense that if the interlocutor asks specific (leading) questions his/her own use of grammar might prompt the speaker to produce grammatical constructions which the latter otherwise would not have used or might not even be capable of using.

Task Two is intended to be a dialogue spurred by the interlocutor's questions. Here the instructions provide a brief list of possible topics (namely life at university, hobbies, and plans for after university) but can there be a hidden agenda here? Is the interlocutor to attempt to elicit different temporal references to the past, present and future? Indeed, this might have provided highly useful research data and would have

been quite easy to achieve if a list of suitable questions had been presented to the interlocutor with the explicit requirement that questions were to refer to the past, present and future. In this way more guidance would also have been provided as to the expected or ideal length of this part of the interview.

Task Three is a narrative based on a set of four pictures. According to the instructions, the speakers are to make up a short story, but no other guidance is given. To what extent is the interlocutor to intervene here, for example, if the speaker is very brief? Are questions allowed? Are the speakers to be prompted to produce more language, or is the task to be essentially monological? Could the instructions for the speakers regarding this part be more detailed and actually specify that, besides the retelling of the story, the pictures are also to be described? I have randomly checked this task in the various national subcorpora and found a variety of speaker solutions to this issue ranging from very short, one-sentence descriptions to lengthy dialogues about the pictures and even about the meaning of art. Whilst such data is authentic to the core and very valuable, comparisons across the subcorpora as well as of different speakers within one subcorpus may be problematic.

4.2 CONCERNS REGARDING SPEAKER PROFICIENCY

Carlsen (2012) describes proficiency in learner corpora as a "fuzzy variable", claiming that in learner corpus research proficiency levels are often not adequately defined. As was stated above, LINDSEI uses an institutional definition of proficiency, which cannot guarantee comparability of results. Listening to the recordings indeed confirmed just how "fuzzy" this approach was; proficiency levels in them ranged from B2 to C2. Whilst LINDSEI sets out to be a corpus of advanced English, this proficiency definition problem would appear to make the whole corpus into a multi-level rather than exclusively advanced corpus.

In 2016 an international team comprising Taiwanese researcher Lanfen Huang, co-ordinator of LINDSEI_TW, and the present author received a Taiwanese government grant[14] for one of its joint projects: the carrying out of a post-hoc, perceptive proficiency rating in our two LINDSEI subcorpora (_TW, _CZ). To this end we engaged three professional IELTS examiners who had been previously trained to rate proficiency in accordance with CEFR levels. They gave ratings for lexical range, accuracy, fluency, phonological control, coherence and overall impression. As shown in Table 3, the Czech subcorpus appears to be a level ahead of the Taiwanese, with the majority of the Czech speakers rated as C1 and of the Taiwanese speakers as B2. In both corpora the figures for the respective highest levels (C2 for Czechs and C1 Taiwanese) are of low representativeness. It might also be argued that the high ratio within each corpus of speakers with lower proficiency somewhat undermines the corpus's aims to be one of advanced English.

14 Ministry of Science and Technology, Taiwan, grant number MOST105-2628-H-158-001.

Proficiency level	LINDSEI_CZ	LINDSEI_TW
B1	0	9 (18%)
B2	12 (24%)	39 (78%)
C1	36 (72%)	2 (4%)
C2	2 (4%)	0
Total	50	50

TABLE 4: Comparison of proficiency levels in the Czech and Taiwanese subcorpora of LINDSEI.

4.3 CONCERNS REGARDING TRANSCRIPTION

Transcription systems are rarely ideal for all purposes. Graddol, Cheshire and Swann (1994, p. 185) note that a compromise must be sought between validity and ease of reading; if too much detail is recorded, the resulting transcriptions are less readable. As this corpus, however, is distributed only as a body of transcriptions (i.e. without recordings), transcription quality is actually of crucial importance. Whilst LINDSEI strikes a good balance between granularity and readability, some of the included features are so difficult to transcribe with accuracy that their inclusion in the transcriptions may be viewed as problematic.

The first such feature is the length of unfilled pauses. LINDSEI distinguishes between pauses shorter than one second, pauses between one and two seconds, and longer pauses (marked using ., .. and ... respectively), but it does not define how to identify a pause nor, indeed, what the minimum length of a pause actually is. Entering pauses in the transcriptions is thus rather haphazard and impressionistic.[15] Neither is there yet available any reliable software for pause detection,[16], and even identification of pauses visually through waveform inspection is unreliable and extremely tedious.

Similarly, the length of filled pauses may be difficult to ascertain. LINDSEI distinguishes between "short" pauses (*eh*, *em*) and "longer" ones (*er*, *erm*) but no guidelines are given as to how these ought to be identified. Neither is it always possible to distinguish between "nasalised" (*eh*, *er*) and "unnasalised pauses" (*em*, *erm*).

Another somewhat problematic feature is the marking of vowel/syllable lengthening, as it is again very difficult to know where to draw the line between "short" and "long" sounds. This is also the case with the unreduced pronunciation of the definite article (*the[i:]*).

Besides phonetic features, the transcriptions may also contain inaccuracies owing to the transcriber's not having understood correctly what was being said. As has been

15 A short test was carried out in which the present author's seminar participants were asked to listen to an excerpt from the corpus and insert pause marks in the transcription provided. The level of agreement was only moderate ($\kappa = 0.55$).

16 The author has experimented with various packages which enable the user to search for areas with low dB levels, but encountered the problem that the software either would only select longer areas than those typically occupied by pauses, or else would not select anything at all because of the presence of background noise during the speaker's pauses.

noted in some of the national subcorpora, this affects mainly the transcriptions of native-speaker interlocutors.

As the above-stated examples show, for a variety of reasons transcriptions are a compromise, and certain features in particular — as has been observed in some of the national subcorpora[17] — may be transcribed differently by different research teams.

4.4 METHODOLOGICAL CONCLUSIONS

Many lessons regarding learner corpus construction may be learnt through analysis of the weaknesses outlined in the preceding sections. In particular the quality of the instructions provided to coordinators seems to be of paramount importance. If homogeneity and consistency are to be guaranteed — and this is especially desirable for large, international corpora with the participation of larger numbers of self-contained research teams and coordinators — instructions ought to be as detailed as possible, describing the purpose of the corpus and its individual components and offering possible solutions to problems which might arise. These instructions should specify whether attempts should be made to elicit particular linguistic features, and should exemplify how this might be achieved. The purpose of the individual tasks should be carefully considered and explained; this is important not only with regard to the replicability of future studies, but also because task design — as has been shown (cf. e.g. Foster — Skehan, 1996; Götz, 2013; Gráf, 2015; Levkina — Gilabert, 2012; Robinson, 2001; Skehan, 2001; Skehan — Foster, 1999; Tracy-Ventura — Myles, 2015) — has a significant effect on language production.

Consideration should also be given to recording environment. Whilst recording in a sound-proof studio provides ideal conditions soundwise, the authenticity of the situation might be decreased by the surroundings, so that speakers might perhaps underperform owing to increased anxiety. If recording studios are to be used it might be worthwhile experimenting with the use of an independent microphone for each of the two participants, which would also simplify the separation of the two tracks as well as some subsequent analysis, such as the measuring of speech rate. On the other hand, such a specification might also further decrease the authenticity of the situation.

The results of the post-hoc proficiency rating reveal that recruiting speakers on the basis of an institutional definition of proficiency may result in a rather heterogeneous mixture of levels. A simple solution to this issue is not easy to find. As pre-recording proficiency tests are hardly time- and cost-efficient, an easier solution might be to select participants based on the coordinator's own or locally reported knowledge of their performance. Alternatively, the interviewer might make a decision after each interview as to whether the given student's performance merits inclusion in the relevant corpus. In the case of LINDSEI_CZ, a second version is being prepared in which all transcriptions of speakers below C1/C2 level will be replaced with transcriptions of more advanced students. This might also be a solution for certain problems of an

17 Differences may be found especially in the marking of the length of pauses, overlapping speech, the length and type of filled pauses, truncations and syllable lengthening, and in the level of detail provided with regard to prosodic information.

ethical nature, e.g. in cases where students have disclosed personal details they might later regret revealing.

Finally, in corpus design careful attention should be paid to the transcription system, and to deciding which aspects can realistically be expected to be transcribed with any degree of reliability. As is shown above, some features are so problematic to define and distinguish that they perhaps ought not to become the focus of subsequent analyses, or at least not unless the researcher has access to the corresponding recordings and can verify that the given features were reliably identified and transcribed.

5 LINDSEI RESEARCH AND FUTURE PLANS

Despite some of the methodological weaknesses outlined here, LINDSEI offers a wealth of valuable data for analyses of learner language. The LINDSEI bibliography published and regularly updated on the website[18] of the project currently lists the titles of 98 studies and MA and PhD theses. The topics covered in these works include lexis (27%), discourse (27%), methodology (10%), pragmatics (8%), pronunciation (7%), grammar (6%), fluency (5%), and learner language in general (5%). As the majority of these studies deal with lexis and discourse markers it would appear that researchers prefer topics which allow easier retrieval of data (lexical items and discourse markers) using concordancers to those involving the more laborious study of such complex phenomena as fluency and pronunciation.

As regards LINDSEI_CZ, three MA theses and one PhD thesis on the subcorpus have so far been successfully defended. Gillová (2014) discussed the applicability of current tagging systems to the tagging of spoken corpora; Štěpánová (2015) carried out a pilot study of disfluency features in advanced learner English; Zvěřinová (2016) compared the use of 4-grams in learner and native English; and Gráf (2015) analysed accuracy and fluency in LINDSEI and LOCNEC. Other theses (e.g on the use of articles, on the use of tenses and on L1 transfer of fluency) are currently being written.

As far as the future of the project is concerned, the Louvain team plan to introduce a ver. 2 in 2017, which should be POS-tagged and include several new national subcorpora. Unlike ver. 1, it is to be distributed not on a CD-ROM but through a web interface. Plans are being drawn up for the compiling of a corpus of intermediate learner English based on the same structure and working towards a pseudo-longitudinal corpus. And LINDSEI itself might grow even further as other national teams are still being invited to make contributions.

6 CONCLUSION

Just as ICLE paved the way for learner corpus linguistics, LINDSEI has undoubtedly paved the way for spoken learner corpora and serves to this day as a source of inspiration for data collection and research in this rapidly expanding field. Despite cer-

18 See <https://www.uclouvain.be/en-cecl-lindsei-biblio.html> (last accessed on 29 August, 2017).

tain drawbacks, mentioned above, the LINDSEI corpus is a truly invaluable source of highly authentic learner data. Its key strength is its size in terms not only of the number of speakers per country and the number of participating countries, but also of the length of the interviews it contains. Another strong point is the existence of the parallel native-speaker corpus LOCNEC. Particular mention must also be made of the fact that, due to the informal character of the interview format, the collected data come across as highly authentic and natural.

Lessons are to be learnt from both its strong and its weak points, some of which have been outlined above. Learner corpus research has shown LINDSEI to have tremendous potential in many respects, although the ways in which that potential might be best exploited are still to be explored.

Pedagogical implications stereotypically outlined at the end of most LCR studies still remain mostly on paper and lack empirical validation in classroom practice; they should stretch beyond learner language descriptions and move into such areas as testing and assessment, proficiency definitions, and alignment with the Common European Framework of Reference for Languages. Interdisciplinary links between LCR, SLA, NLP and psycholinguistics ought to be sought and encouraged. Methodological implications should include not only corpus design but also research and publication standards, as only well-defined and well-described phenomena can be analysed to lead to results that are reliable and studies that are replicable. Much more attention ought to be paid to the collection of a variety of metadata, especially contextual metadata such as the role of learning and teaching materials, teaching contexts, contact with the L2 outside the classroom, and motivation. Much of the knowledge available to us as we work to achieve these ends has been obtained thanks to the pioneering role of ICLE, LINDSEI and the whole Louvain family of learner corpora. After more than twenty-five years in the field, LINDSEI can now afford to take the time to reflect on the period of its existence and give consideration to the formulation of future priorities and guidelines newly informed by what has been achieved to date.

ACKNOWLEDGMENTS

The author would like to thank the Institute of the Czech National Corpus, Faculty of Arts, Charles University, for providing financial and methodological support in the compilation of LINDSEI_CZ and for hosting the subcorpus in their collection of corpora available through the KonText interface. This research was supported by the Charles University, project Progres 4, Language in the shiftings of time, space, and culture.

REFERENCES:

Adolphs, Svenja — Knight, Dawn (2012): Building a spoken corpus. In: Anne O'Keeffe — Michael McCarthy (eds.), *The Routledge Handbook of Corpus Linguistics*. Abingdon — New York, NY: Routledge, pp. 38–52.

Carlsen, Cecilie (2012): Proficiency level — a fuzzy variable in computer learner corpora. *Applied Linguistics*, 33(2), pp. 161–183.

Foster, Pauline — Skehan, Peter (1996): The influence of planning and task type on second language performance. *Studies*

in Second Language Acquisition, 18(3), pp. 299–323.

Gillová, Lucie (2014): *Tagging a Spoken Learner Corpus* [MA Thesis; online]. Prague: Charles University. Cit. 10. 11. 2017. Retrieved from WWW: <https://is.cuni.cz/webapps/zzp/detail/148736>.

Götz, Sandra (2013): *Fluency in Native and Nonnative English Speech*. Amsterdam — Philadelphia, PA: John Benjamins.

Graddol, David — Cheshire, Jenny — Swann, Joann (1994): *Describing Language*. Buckingham: Open University Press.

Gráf, Tomáš (2015): *Accuracy and Fluency in the Speech of the Advanced Learner of English* [Ph.D. Thesis; online]. Prague: Charles University. Cit. 10. 11. 2017. Retrieved from WWW: <https://is.cuni.cz/webapps/zzp/detail/151663>.

Gráf, Tomáš (2017): Repeats in advanced spoken English of learners with Czech as L1. *Acta Universitatis Carolinae, Philologica 3, Phonetica Pragensia*, pp. 65–78.

Granger, Sylviane (1996): From CA to CIA and back: an integrated approach to computerized bilingual and learner corpora. In: Karin Aijmer — Bengt Altenberg — Mats Johansson (eds.), *Languages in Contrast: Text-Based Cross-Linguistic Studies*. Lund: Lund University Press, pp. 37–51.

Granger, Sylviane (2004): Computer learner corpus research: current status and future prospects. In: Ulla Connor — Thomas A. Upton (eds.), *Applied Corpus Linguistics: A Multidimensional Perspective*. Amsterdam — Atlanta, GA: Rodopi, pp. 123–145.

Granger, Sylviane — Gilquin, Gaëtanelle — Meunier, Fanny (eds.) (2015): *The Cambridge Handbook of Learner Corpus Research* [online]. Cit. 10. 11. 2017. Retrieved from WWW: <http://www.cambridge.org/us/academic/subjects/languages-linguistics/applied-linguistics-and-second-language-acquisition/cambridge-handbook-learner-corpus-research?format=HB>.

Levkina, Mayya — Gilabert, Roger (2012): The effects of cognitive task complexity on L2 oral production. In: Alex Housen — Folkert Kuiken — Ineke Vedder (eds.), *Dimensions of L2 Performance and Proficiency: Complexity, Accuracy and Fluency in SLA*. Amsterdam — Philadelphia, PA: John Benjamins.

Robinson, Peter (2001): Task complexity, task difficulty, and task production: exploring interactions in a componential framework. *Applied Linguistics*, 22(1), pp. 27–57.

Skehan, Peter (2001): Tasks and language performance assessment. In: Martin Bygate — Peter Skehan — Merrill Swain (eds.), *Researching Pedagogic Tasks, Second Language Learning, Teaching and Testing*. Harlow: Longman.

Skehan, Peter — Foster, Pauline (1999): The influence of task structure and processing conditions on narrative retellings. *Language Learning*, 49(1), pp. 93–120.

Štěpánová, Tereza (2015): *"I I Erm I Thought": Selected Performance Phenomena of Czech Advanced Speakers of English in Comparison with the Native Speaker Norm* [MA Thesis; online]. Prague: Charles University. Cit. 10. 11. 2017. Retrieved from WWW: <https://is.cuni.cz/webapps/zzp/detail/147232>.

Tracy-Ventura, Nicole — Myles, Florence (2015): The importance of task variability in the design of learner corpora for SLA research. *International Journal of Learner Corpus Research*, 1(1), pp. 58–95.

Zvěřinová, Simona (2016): *N-Grams in the Speech of Czech and Native Speakers of English* [MA Thesis; online]. Prague: Charles University. Cit. 10. 11. 2017. Retrieved from WWW: <https://is.cuni.cz/webapps/zzp/detail/167908>.

Tomáš Gráf | Department of English Language and ELT Methodology, Faculty of Arts, Charles University
<tomas.graf@ff.cuni.cz>