E ISSN 1512-3146 (online)
ISSN 1987-9601 (print)
International Journal
of Multilingual Education
www.multilingualeducation.org

Zakharia Pourtskhvandze

Johan Wolfgang Goethe University of Frankfurt, Germany

Learner Corpora and Their Potential for Multilingual Teaching

Abstract

This article primarily deals with building and using a type of language corpus - the learner corpus - for multilingual teaching. It describes all relevant aspects of the conceptualization, motivation and construction of learner corpora including the case example of the German learner corpus FALKO (*Fehlerannotiertes Lernerkorpus* 'error annotated learner corpus'). In addition we discuss the possibility of a learner corpus for the Georgian language using examples from real Georgian language courses at Goethe University Frankfurt and Tbilisi State University. The article stresses the potential of learner corpora for multilingual teaching and multilingual teacher education.

***Keywords**: Learner Corpus, Multilingual Teaching.*

1. Learner Corpus - getting to know

According to the common definition, a learner corpus is an electronic collection of authentic texts (language material) produced by foreign or second language learners stored in an electronic database (Anna O´Keeffe, 2007, S.23.). Additionally, computer learner corpora are electronic collections of authentic FL/SL textual data assembled according to explicit design criteria for a particular SLA/FLT purpose. The corpora are encoded in a standardized and homogeneous way and documented as to their origin and provenance (Granger et al. 2002: 7).

The crucial determination for the learner corpus is the idea of language error, which can be recognized as „*a linguistic form or combination of forms which, in the same context and under similar conditions of production, would, in all likelihood, not be produced by the speakers' native speaker counterparts*" (Corder, 1983; corder, tephan,1986).

The language materials can be analyzed by a software and edited. The analyse serves different purposes. A learner corpus is a new

type of language corpus that started appearing in the early 1990s.[1] Since then many learner corpora have been developed for different languages. The Catholic University of Louvain list 138 different learner corpora[2]. The list is not complete but contains the main learner corpora and gives a good overview. The learner corpora are classified there by different attributes, for example target language, medium and text type. As expected, most of them have English as their target language. 87 of the 138 are for English, 10 for French, 9 for German, 8 for Spanish, 3 for Italian and so on. The corpora work with different media. 87 of them use written media (e.g. The Advanced Learner English Corpus (ALEC), Uppsala University - texts composed/written by students), 33 of them use spoken media (e.g. The ANGLISH corpus, University of Provence - readings, oral language), 11 of them use written and spoken media and 3 of them use multimedia. Generally these corpora have only one target language, but beside the 127 monolingual corpora the list also contains 11 multilingual corpora (e.g. The Corpus of Young Learner Interlanguage (CYLIL) Vrije Universiteit Brussel or The Eastern European English learner corpus Eberhard Karls University of Tübingen).

1.1.The case of FALKO

Now we will look at one learner corpus in detail – The FALKO corpus (Fehlerannotiertes Lernerkorpus ‘error annotated learner corpus’) (https:// www.linguistik.hu-berlin.de/institut/ professuren/korpuslinguistik/forschung/falko.; https: //korpling.german.hu-berlin.de/falko-suche/ search.html.). The FALKO was developed at Humboldt University in Berlin by Anke Lüdeling and Maik Walter in 2004. The main FALKO corpus can be divided into five smaller corpora - FalkoSummaryVL, FalkoSummaryL1 V1.2, FalkoSummaryL2 V1.2, FalkoEssayL1 V1.2, FalkoEssayL2 V2.0.

1. Learner texts (Fal*koSummaryL2*) (Reznicek, 2012, S.8ff.): Collection of summaries (linguistic texts and literary studies), made by advanced German learners (C1-C2). They were written in the framework of an exam, which is obligatory for foreign students who have German philology as their main subject. The examination took place at the Free University of Berlin.
2. Native speaker texts (*FalkoSummary L1*): Collection of the same texts as in FSL2, written by native German speakers.
3. Original texts (*FalkoSummaryVL*): Collections of the linguistic and philological texts which served as templates for the

summaries. Altogether contains 197 texts written by 98 learners.

4. *FalkoSummaryL1* 1.1 (Reznicek, 2012, S.17.): For this corpus, 4 data collections have been carried out. They took place at the Free University of Berlin and at Humboldt University of Berlin. Again only students with German philology as their main subject took part in these data collections. The conditions of these data collections were all the same.
5. *Falko Essay Corpus (*Reznicek, 2012, S.19f.*)*: This corpus contains two sub-corpora.

a. *FalkoEssayL2*: contains a collection of essays written by advanced German learners. 4 different topics were given for the essays and the participants had to achieve at least 60 from 100 points in a C-test.

b. *FalkoEssayL1*: contains a collection of essays written by native speakers. The participants were graduating class pupils of three different secondary schools. The topics were the same as in FalkoEssayL2 as were the conditions of the exam.

All sub-corpora have different levels of annotation and FALKO's architecture allows the addition of more annotations levels (multi-layer stand-off annotation). In general FALKO contains written texts of advanced German learners. The most annotated sub-corpus is a collection of summaries (Siemen et al., FALKO S.1.). The lemmata were automatically annotated by Treetagger (Mark Reznicek et al. Das Falko-Handbuch. Korpusaufbau und Annotationen, Version 2.01, 2012, S.4.). The database also contains explicit information about the authors, e.g. level of education, level of language ability and much else (Reznicek, 2012, S.6.).

Table 1. Falko Annotation Levels (Karin Schmidt, 2015). ((word) – Learner utterance, (kpos) – Part of speech, (target-hypothesis) - Assumption about proposed utterance, (ref) – evidence reference.)

word	Dabei	Ist	es	zu	beachten
kpos	PAV	VAFIN	PPER	PTKZU	VVINF
lemma	dabei	Sein	es	zu	beachten
target_hypothesis	*Dabei ist zu beachten*				
ref	70	71	72	73	74

The annotation level contains a target hypothesis to allow the reconstructing of the error made by the learner. The errors are identified by comparing original utterances with so-called reconstructed utterances, that is, correct utterances having the meaning intended by the learner.

Table 2. Error analysis in FALKO (Falko-Handbuch S.39.)

word	Fraue n	konnten		sol- chen		geselschaftlichen	Zust and	verändern
target_hyp_ 1	Fraue n	konnten	eine n	sol- chen		gesellschaftlichen	Zust and	verändern
¬target_hyp _1	Fraue n	konnten		solch	eine n	gesellschaftlichen	Zust and	verändern

The use of FALKO has shown which aspects of the German language are more difficult for learners (for example proper use of articles and prepositions) and hence, which aspects need to be prioritised in teaching.

2. About the motivation, construction and function of learner corpora

The main task of a learner corpus is the annotation of errors. Therefore texts which are written by learners have to be compared to those of native speakers. This assumes that there is, compared to the mistakes made by the learners, a correct version given by the native speakers. This seems to be easy, but in reality there is no right way to express yourself in the first language (Siemen FALKO - S. 2.). Language is something very flexible, so there are a lot of different ways to say the exact same sentence. Additionally, language is in a constant process of development, so it changes constantly. What may seem correct nowadays can be completely wrong in the future. Nonetheless learner corpora are an important instrument for didactical studies and didactics themselves.

The motivations to build a learner corpus may be various. For example, in foreign language teaching some verb constructions can be very complex for beginners. Some constructions are almost completely neglected in teaching materials. This would be a chance to prove that corpora are useful for cases like this. Learner corpus analyses are prone to a criticism similar to what recommendations for teaching based on native speaker corpora have been subjected to for a while: that they only take into account one criterion that is important for teaching, and disregard others. In the case of teaching recommendations based on native

speaker corpora, it has often been objected that the only criterion considered is frequency in native speaker usage. But the learner corpus would definitely motivate the learner and promote language awareness. They stimulate the student to work actively and independently, and in this way, they probably increase both the motivation of the student and the learning effect. In summary a corpus will be used in the education of teachers of a foreign language, as a source of examples usable in the classroom and for educational tools, and will help tailor instructions and teaching materials to specific groups of learners.

Linguists have different motivations for constructing a learner corpus. The main purpose may be to improve didactical methods. Learner corpora can identify specific problems learners have with a certain language. These perceptions can help improve learning methods for these learners. Hence, it is an important tool for foreign language didactics and allows the analysis of the mistake/error typology of certain learner groups. Therefore, it is a win-win-situation for both the learners and the teachers. By comparing the learner texts with those of native speakers, the learners themselves can learn from it and improve their language skills, and the teachers can adapt their methods to specific learner groups. In general, the main target groups of learning corpora are learners of a foreign language and teachers teaching foreign languages. Besides them, linguists and those who research didactical methods also benefit from this type of corpora.

Although learner corpora open up new possibilities for foreign language didactics they are still seldom seen in schools and language classes. One of the main reasons for this may be the lack of information and the fact that corpora are seen as a scientific tool, not a teaching tool (Karin Aijmer, 2009 S.47f.). Therefore, it is important to instruct the teachers and train the student so they can learn how to use learner corpora. At this point, schools and universities have to show initiative and start workshops. To help the students learn a new language, the teacher can include learner corpora in their lesson. They can, for example, give exercises which can only be solved by using the learner corpora. Many words have a wide range of meanings and are therefore used in a wide range of contexts. With the aid of the learner corpora, students can compare the usage of these words in the native text and identify the different lexical categories (Aijmer, 2009 JBPC, S.50f.). Or, if the students have a certain question, they can answer it by searching in the learner corpora potentially turning students into language researchers (John McHary Sinclair, 2004, S.16.). Learner corpora can serve as a supplement for grammar studies by exemplifying the grammar rules.

3. How is a learner corpus built?

To build a learner corpus it is important to collect a great amount of written and/or spoken materials. Written corpora are easier to create than spoken corpora, because a written corpus can use the internet as a source. They may contain recorded speech, interviews, essays, exams and so on. These must be written by learners. For comparison the same materials must be available written and spoken by native speakers (Anna O´Keeffe, 2007).

A basic language corpus can be assembled from spoken or written texts and can be used with commercially available corpus software, which any average home computer user can manipulate with relative ease. Of course, a spoken corpus takes considerably longer to build, because the speech, for example in videos, has to be transcribed and possibly coded for some of its non-verbal features. By comparison, building a written corpus is very quick using the internet as a source. Every corpus needs design principles. You have to consider not only the design, but also the feasibility, because there are struggles with what is available, what is ethical or what is legal. This could be a leading factor. Also deciding what to represent and how to represent the best for the general purpose is very important. In that case, you have to decide on the amount of data you want to collect and use.

In the case of spoken corpora, the next step is recording the data. There are a number of options for recording including analogue cassettes, digital media and audiovisual digital recorders. Traditional analogue, though they are inexpensive, have a number of drawbacks. They are cumbersome to store and unlike digital recordings, they cannot easily be computerized and aligned with the transcription later. Using digital devices leaves open the option of aligning sound (and image if you use an audiovisual recorder) with your transcription.

An important aspect is permission. Permission to record should be cleared in advance with the speakers and consent forms should be signed authorizing the use of the recordings for research or commercial pedagogical materials, etc. It may be necessary to specify how the recordings will be used when obtaining permission. After that, the main task is the transcription, because spoken data needs to be manually transcribed and this is what makes corpora of spoken language such a challenge. They are best stored as ‘plain text’ files, as this offers the maximum flexibility of use with different software suites. One hour of recorded speech may take days to transcribe, depending on the complexity of the language. In most cases, every word, vocalization, truncation, hesitation, overlap, and so on, is transcribed, as opposed to a cleaned-up version of what the speakers said. The level of detail of the transcription is relative to the purpose of your corpus. If you

have no requirement to know where overlapping utterances and interruptions occur, then there is no point in spending time transcribing to that level of detail. This hard work includes *pattern matching* (1), *collocations* (2), *lemma and part of speech* (3), *synonyms and antonyms* (4), more complex searches using combinations of the preceding *types of searches* (5), queries based on the *frequency of the construction in different historical periods and registers of the language* (6), and queries involving *customized, user-defined lists* (7).

Transcription files need to be organized so that source information can be traced. For example, it may be useful to be able to retrieve information such as gender, age, number of speakers, place of birth, occupation, level of education, where the recording took place, relationship of speakers and so on. This information can be stored at the beginning of each transcript as an information 'header', or in a separate database, where the information is logged with the file name. In short, the corpus should be richly annotated and should allow searches for many types of linguistic phenomena. The content of every corpus is a collection of texts and expressions in a language. Of course, we have to differentiate between written and spoken corpora. The materials for written corpora are comparatively easy to collect, because everything is physically available. The content of spoken corpora, as mentioned above, is more difficult to collect and to edit. The basic materials for spoken corpora are generally given through audio or audiovisual recorders.

Possible draft for a Georgian learner Corpus

The sociolinguistic situation of Georgia can characterized as multilingual. In border areas of Georgia to Azerbaijan and Armenia, but also in central regions, classroom settings are multilingual. It is an educational challenge to develop suitable language learning contents, which uses pointedly the spread errors of Georgian language learners. One of the first steps in that direction is the collection and unified documentation of all available errors in both the written and the spoken register.

As a first source of material, learner groups at the high schools of Georgia can be tapped. According to the official statistics (GeoStat. http://www.geostat.ge/?action=page&p_id=205&lang=geo, 25.11.14, 13.00.), about 2000 non-native speakers of Georgian enter higher education in Georgia every year.

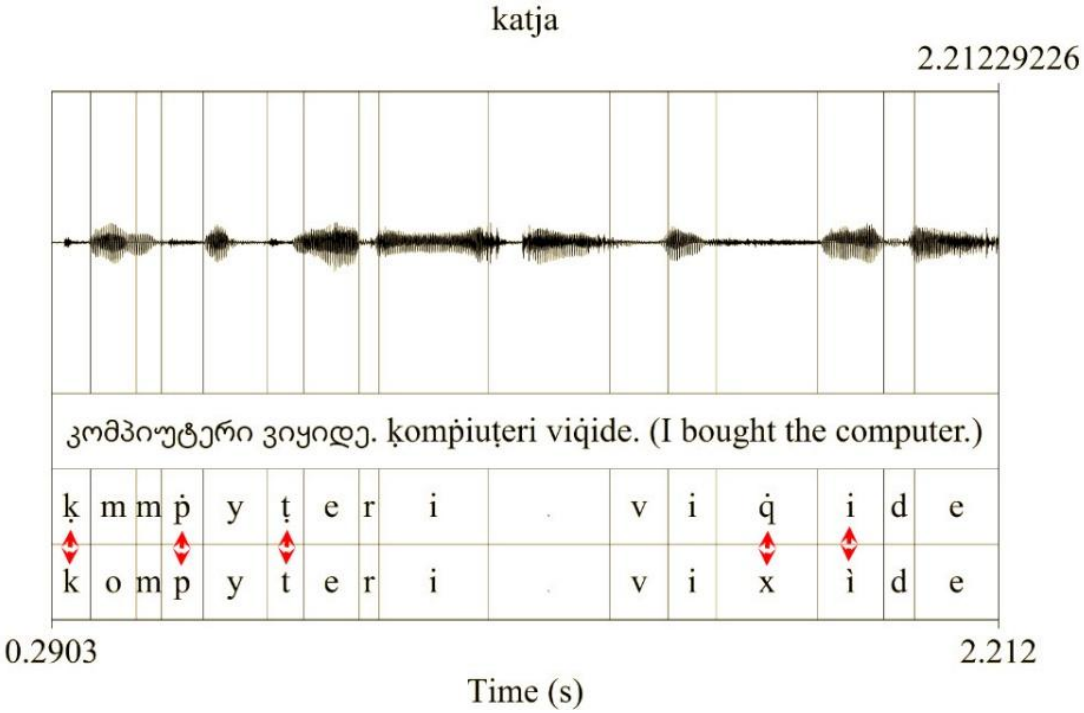
The teachers and language trainers can be constrained to notify the multilingual teaching experiences and systemize the recurring errors. These observations act as groundwork for the further development of the database containing error patterns. As we saw with FALKO

(Ch.1.1), such a-database is need for the establishing of hypothesis as kind of the correction preliminary.

There are a few Georgian language-learning programs provided currently in Georgia (http://www.ice.ge/web/elearning_geo.html) and abroad (http://195.178.225.22/DiasporaGeo/Georgianonline.html) (online distance learning course offered by Malmö University, Sweden). A target learner group are Georgian citizens who speak Azerbaijani or Armenian as their first language. The topic structure of the program syllabi represents the program creators' presumptions about possible difficulties of the learner. The topics are not confirmed based on empirical evidence, despite the fact that the emphasis of any specific subject matter must be strengthened oriented on the errors made by learners in the real learning process.

The most common difficulty in learning Georgian (like other Caucasian languages) was and still is the canonical consonant pronunciation. There are single consonants or consonant clusters, which are characteristic phonetic features of Caucasian languages. Hence, it is a significant intellectual and physical challenge for the learner to acquire and use these sounds.

Table 3.Pronunciation example by Georgian learner. 5 pronunciation errors in 2 words.



The targeted recordings of the audio material with L2 learners act as crucial database for closely exploring frequent errors in the phonetic acquisition and allow the focusing of teaching process on these errors. Even super-

ficial observation of frequent errors highlights problematic areas, which should be a central point of attention. Below are some examples of prototypical language errors noted by Georgian language trainers (Prof. Ketevan Gochitashvili. Tbilisi State University).

Table 4. Word order error.

wording	sad	šen	iqavi?
lemma	sad_*wh*	šen_*PPron.2Sg*	qopna.*be*
hypothesis	šen	sad	iqavi?
Eng.	*Where have you been?*		

Table 5. Agreement error.

wording	ḳargi	var.
lemma	ḳargi_*good*	qopna.*be*
hypothesis	ḳargad	var.
Eng.	*I am fine.*	

Table 6. Lexical error.

wording	didi	gemrieli-a
lemma	didi_*big*	gemrieli - qopna.*be. Encl*
hypothesis	ʒ alian	gemrielia
Eng.	*It is very tasty.*	

Table 7. Syntax error, unused word order

wording	saxli	romeli	dgas	kalakši
lemma	saxli_*house*	romel_*wh*	dogma_*stand*	kalaki_*city*
hypothesis_1	saxli	romlic	dgas	kalakši
hypothesis_2	saxli	romlic	kalakši	dgas
Eng.	*The house, which is (standing) in the city.*			

wording	reṣtorani	sad	viqavi	gušin
lemma	reṣtorani *_restau-rant*	sad_*wh*	qopna_*be*	gušin *_yes-terday*
hypothesis_1	reṣtorani	sadac	viqavi	gušin
hypothesis_2	reṣtorani	sadac	gušin	viqavi
Eng.	*The restaurant I was in yesterday.*			

References:

Aijmer Karin, Corpora and Language Teaching, 2009 John Benjamins Publishing Company.

Biber, Douglas; Jones, James K. (2009): Quantitative methods in corpus linguistics. In: Lüdeling, Corder, Steven Pit (1981): Error Analysis and Interlanguage. Oxford; Oxford University Press.

Corder, Stephen Pit (1986): The role of interpretation in the study. In: Corder, Stephen P. (Hrsg.): Error analysis and interlanguage. 4. impr. Oxford: Oxford University Press.

Diaz-Negrillo, Ana; Fernandez-Dominguez, Jesus (2006): Error tagging systems for learner corpora. In: RESLA 19.

Granger, Sylviane. (2002): A Bird's-eye View of Computer Learner Corpus Research. In: Granger S., Kytö, Merja Anke (Hg.): Corpus Linguistics. An International Handbook. Vol. 2. Berlin: Mouton de Gruyter.

O´Keeffe Anna et al: From corpus to classroom. Cambridge University Press 2007.

Reznicek et al. Das Falko-Handbuch. Korpusaufbau und Annotationen, Version 2.01, 2012.

Schmidt Karin, Lernerkorpora: Ressourcen für die Deutsch-als-Fremdsprache-Forschung (publikationen.ub.uni-frankfurt.de/files/.../Schmidt_Lernerkorpora.pdf).

Siemen et al., FALKO - Ein fehlerannotiertes Lernerkorpus des Deutschen.

Sinclair John McHary, How to use Corpora in Language Teaching Studies in Corpus Linguistics, Silvia Bernardini, 2004 John Benjamins Publishing Company.

http://www.uclouvain.be/en-cecl-lcworld.html.

https://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung/falko.

https://korpling.german.hu-berlin.de/falko-suche/search.html

http://www.ice.ge/web/elearning_geo.html

http://195.178.225.22/DiasporaGeo/Georgianonline.html