# Rethinking Applied Corpus Linguistics from a Language-pedagogical Perspective: New Departures in Learner Corpus Research

*Joybrato Mukherjee & Jan-Marc Rohrbach*

**Abstract:**

*Corpus linguists have been right in demanding to rethink language pedagogy from a corpus perspective. However, as the present paper argues, it is also necessary for corpus linguists who are interested in the language-pedagogical applications and implications of corpus research to constantly redefine and evaluate their work from the point of view of actual teachers and learners in the EFL classroom. Therefore, the present paper starts off from a brief discussion of the complementary roles of the corpus-linguistic and the language-pedagogical perspective in the corpus-informed classroom. Focusing on learner corpus research, we then sketch out how the relevance of applied corpus linguistics to actual teachers and learners can be increased by making them work with learner output themselves. To this end, it is suggested that traditional analyses of large reference learner corpora be complemented, firstly, by analysing individual learners' output (and comparing the results) and, secondly, by making teachers and learners compile their own local learner corpora which would then be used for corpus-based activities. In this context, the present paper will report on an on-going learner corpus project at the University of Giessen, i.e. the German component of the Louvain International Database of Spoken English Interlanguage (LINDSEI-Ger), and on the beginnings of a more experimental classroom project, i.e. the Giessen-Göttingen Local Learner Corpus of English (GGLLC).*

There is no doubt that corpus-linguistic research has exerted an enormous influence on the teaching of English as a Foreign Language (EFL) around the world. This is a welcome development, but – as already pointed out elsewhere (cf. Mukherjee 2004:242 f.) – we have the impression that in EFL countries like Germany there is a widening gap and a growing lag

between on-going and intensive corpus-linguistic research on the one hand and classroom teaching on the other. As Granger (2004:136) reports, research into the use of corpora for language teaching is almost entirely done by linguists; the contribution of SLA researchers to – and the participation of EFL teachers in – what happens in corpus linguistics is still relatively low. It is, however, crucial to involve SLA researchers and the ELT community, including teachers and learners, to a much larger extent in the actual work on corpora and in the systematic evaluation and improvement of corpus-based activities in the classroom. It seems to us in particular that so far many applied corpus linguists have been reluctant to adopt an average EFL teacher's perspective while many EFL teachers very often do not see the relevance of the corpus-linguistic perspective to their teaching.

The very gap between the applied corpus linguist's and the average EFL teacher's perspectives serves as a starting-point for the present paper. In the first part, we will thus pick out and briefly discuss some important corpus-linguistic perspectives on language teaching and combine them with specific language-pedagogical perspectives that corpus linguists very often tend to neglect. In the second part, we will sketch out how the corpus-linguistic and language-pedagogical perspectives may complement each other and discuss benefits that may accrue from this. To this end, we will focus on learner corpus research and present an on-going learner corpus project at Giessen University and the beginnings of a more experimental classroom project, which is also based on learner data. In this context, we will argue in particular that research into large reference learner corpora can be fruitfully complemented with the 'quick-and-dirty' compilation and the learner-centred analysis of local learner corpora, as already envisaged by Seidlhofer (2002).[1]

[1] We would like to thank the students at Justus Liebig University Giessen and at Hainberg-Gymnasium Göttingen whose data have become part of the German component of the Louvain International Database of Spoken English Interlanguage (LINDSEI-Ger) and the Giessen-Göttingen Local Learner Corpus of English (GGLLC) respectively. For the actual transcription and compilation work we are indebted to Rosemary Bock, Christiane Brand, Susanne Kämmerer, Simone Müller, Simone Tausend and Christian Woppowa.

## 1. The corpus-informed classroom: corpus-linguistic and language-pedagogical perspectives

Applied corpus linguistics, i.e. research into the use of corpora in the EFL classroom, finds itself at the crossroads of corpus-based descriptive linguistics, SLA research and language pedagogy. In the following, we will focus on some general differences in perspective between corpus linguists and English language teachers which need to be overcome in order to involve teachers and learners in corpus-based activities and in order to establish a genuinely corpus-informed classroom. As shown somewhat simplistically in Figure 1, we would contend that at present the corpus-linguistic perspective on what should happen in the EFL classroom is markedly different from what the ELT community and EFL teachers in particular focus on in their daily teaching practice.

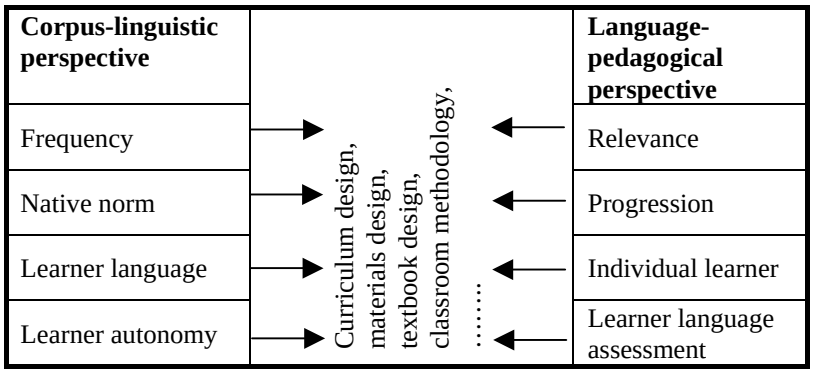


**Fig. 1:** Corpus linguistics vs. language pedagogy in the EFL classroom.

Fig. 1 focuses on four particularly relevant aspects of the corpus-linguistic and the language-pedagogical perspectives on curriculum design, materials design, textbook design, classroom methodology and so on. To some extent, there is a tension between those concepts that can be found in the same line in the right-hand and in the left-hand columns, and we will briefly comment on the four dichotomies in Fig. 1 in the following.

To begin with, in applied corpus linguistics, frequency is very often considered to be the most central criterion for the selection and sequencing of forms and structures to be taught and learned. Schlüter's (2002) recent corpus-based suggestion as to how to revamp the teaching

of the present perfect in the EFL classroom in Germany, for instance, is a very good example of corpus-linguistic approaches to language-pedagogy that are firmly based on frequencies. In the ELT community the issue of relevance is of equal importance. The notion of relevance here refers to the decisions on what kind of English in general and which specific forms, structures and meanings in particular are relevant to the foreign-language learner of English. This includes, for example, specifying a standard native variety of English as the target norm and specifying the range of registers that should be taught. At times, it may thus be the case that frequent forms are not considered to be relevant from a language-pedagogical perspective, for example the teaching of high-frequency swearwords.

A second key issue in applied corpus linguistics, based on native corpora, is the conception of native usage as a given target norm. In language pedagogy, on the other hand, the process of approximation to this target, i.e. progression, is of prime importance. Curriculum designers and teachers have to decide on when to introduce which forms, structures and meanings in the learning process. This of course raises some general questions about the suitability of specific native control corpora. For example, does it make sense to compare young learners' language at a specific stage with the language of native speakers of the same age? Many language-educational professionals argue that this kind of comparison may lead to the teaching of a variety of English which at some point will turn out to be fossilised when compared to the adult native norm (Michael K. Legutke, personal communication).

Thirdly, corpus-based studies of learner language, still a relatively recent development in corpus linguistics (cf. Granger 2002, Nesselhauf 2004), tends to focus on quantitatively feasible trends in learner language in general at a given stage in the learning process and compared to a specific native control corpus. In language pedagogy, on the other hand, the individual learner's language and his/her progression is of far greater importance than the general and supraindividual description of learner language. This, by the way, also holds true for the language-pedagogical need for differentiation in classroom methodology. For example, it is worth discussing whether the highly cognitive approach to language learning, on which the analysis of concordances in data-driven learning methods is based, is equally suitable for all types of learners.

Fourthly, in most corpus-based approaches to language teaching, the notion of learner autonomy is a key concept:

> Perhaps the greatest attraction of corpora in language pedagogy is their potential for autonomous learning: as Leech has put it, 'the main rationale of corpora in teaching is their immediate availability for students' use' (1997:7).
>
> (Aston 2001:41)

Note in this context that in the ELT community, the systematic assessment of learner language has recently turned into a major objective of EFL teaching; in Germany, which ranked low in various international surveys including PISA, the issue of assessment has become almost a matter of national concern. Again, there is a kind of tension between the plea for increasingly autonomous and individualised language learning on the one hand and, on the other, the demand to assess learners' proficiency levels by referring to standardised schemes like the *Common European Framework of Reference for Languages* (Council of Europe 2001).

What does all of this mean? In our view, corpus linguists have certainly been right in demanding to *rethink language pedagogy from a corpus perspective*, as echoed, for example, in the title of the proceedings of the third TALC Conference (cf. Burnard/McEnery 2000). However, it is also necessary for us to rethink our field of research from a language-pedagogical perspective. From our experience in teacher-training programmes, in teacher education and in the classroom in secondary schools, we know, for example, that very often teachers are confronted with suggestions of corpus-based activities which, at least in Germany, are difficult (if not to say impossible) to put into practice or which contradict mainstream language-pedagogical positions in the ELT community. A good case in point is the recent suggestion to base the teaching of English as a foreign language not on the native target norm, but to establish a target norm that is based on English-as-a-lingua-franca communication in conversations between non-native speakers because "this is the most likely situation for the majority of learners in the 21st century" (Jenkins 2004:65). As Jenkins (2000, 2004) and Seidlhofer (2001, 2004) argue, only the forms, structures and meanings that are necessary for successful communication in non-native international contexts should be taught. Deviations from native-like language use that do not impede successful communication should thus no longer be treated as errors. Without going

into details about the English-as-a-lingua-franca teaching model, it needs to be pointed out that – at least in the German context – this model ignores many language-pedagogical counterarguments, three of which we would like to briefly mention.

Firstly, in Germany all curricula place special emphasis on the need for learners to be able to communicate with native speakers. Most curricula directly refer to the *Common European Framework of Reference for Languages*, in which, e.g., the following self-assessment descriptor for spoken interaction at the level B2 ("Vantage") can be found:

> I can interact with a degree of fluency and spontaneity that makes regular interaction with native speakers quite possible. I can take an active part in discussion in familiar contexts, accounting for and sustaining my views.
>
> (*Common European Framework of Reference for Languages* 2001:27)

Secondly, many language-educational professionals are very critical of Seidlhofer's (2001) and Jenkins's (2000) attempt to dissociate the English language as a communicative device from its sociocultural context and to reduce its complexity at the level of pronunciation, lexis, grammar and pragmatics. At a very early stage, Vielau (1991) has already criticised the inherent danger of 'pidginising' the English language in the EFL classroom if the only goal is to get the message across to one's interlocutor and if the sociocultural context of a language community is left out of consideration:

> All dies bedeutet, daß man eine Fremdsprache zwar relativ schnell 'so einigermaßen' erlernen kann, daß diese Art des Lernens aber recht bald an Grenzen stößt, da sie der Eigenart einer natürlichen Fremdsprache streng genommen äußerlich bleibt. Ohne praktische Erfahrungen mit den kulturspezifischen Denk- und Handlungmustern, ohne Verständnis der Kultur, Geschichte und Mentalität eines Volkes ist letztlich auch dessen Sprache nicht angemessen lernbar.
>
> (Vielau 1991:24)

Thirdly, the kind of non-native, English-as-a-lingua-franca target norm that proponents of the English-as-a-lingua-franca norm suggest is also clearly rejected by teachers and learners. Figure 2 shows what the students – many of them future English teachers – in the lecture on World Englishes at Justus Liebig University in the summer semester 2004 viewed as the norm that they would like to approximate to. It is only a

small minority that would accept non-native English in international contexts as their target norm.

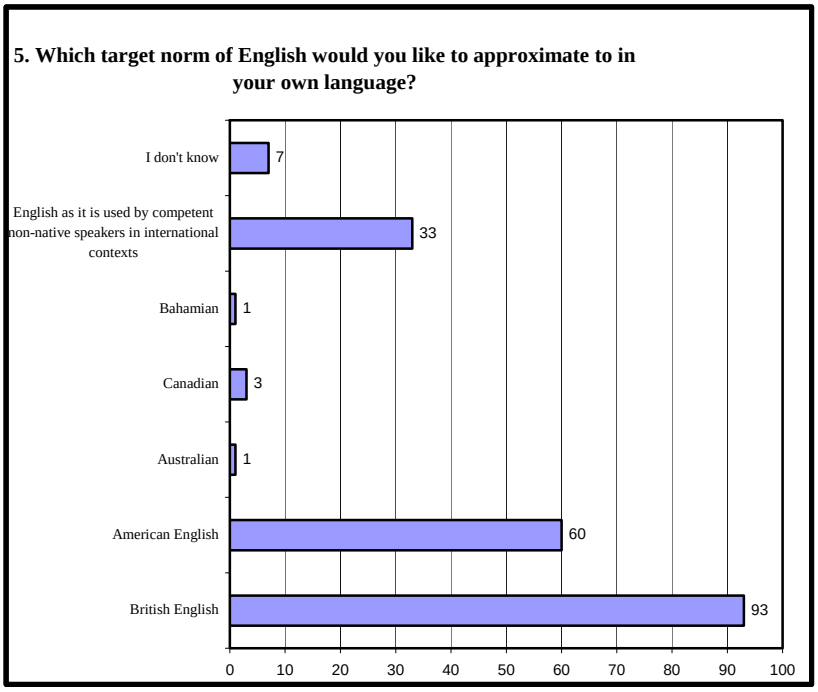


**Fig. 2:** Questionnaire V-S04-5c, lecture on "English in the world and world Englishes" (University of Giessen, summer semester 2004), corrected version (n = 197).[2]

It thus seems that the concept of an English-as-a-lingua-franca norm, which is based on the analysis of the Vienna Oxford International Corpus of English (VOICE, cf. Seidlhofer 2002), is not in line with (1) what curricula and language reference frameworks demand, (2) what the overall goal of intercultural competence requires and (3) what future teachers and learners actually want to learn.

In the light of the discussion so far, we would thus suggest that applied corpus linguists be more aware of the language-pedagogical side

[2] One student gave two answers, which explains the discrepancy between the number of participants (197) and the number of replies (198).

of things in the EFL classroom. We firmly believe that corpora will be used more routinely by teachers and learners in the EFL context of countries like Germany only if the use of corpora has a surplus value within a given language-pedagogical framework. In other words: corpus technology is not a language-pedagogical framework itself, but should best be seen as an "added value," as Bernardini (2004:33) calls it, and as a problem-solving resource.

In the following, we will focus exclusively on learner corpus research. Specifically, we will sketch out three steps in the process of bringing learner corpus research to the classroom so that by using learner-corpus resources and methods language-pedagogically relevant questions can be answered.

## 2. Three steps in bringing learner corpora to the classroom

As already mentioned, the *Common European Framework of Reference for Languages* with its detailed taxonomic description of six proficiency levels in learner language has had substantial repercussions on the ELT community, especially with regard to the standardised assessment of learner language. The problem, of course, is how the general and very often abstract descriptions in the *Reference Framework* can be put in concrete terms so that teachers are enabled to relate actual learner language behaviour to the taxonomy of the *Reference Framework*. There are various ways in which this problem can be solved. At the University of Fribourg (Switzerland), typical examples that illustrate the various stages at the levels of vocabulary, grammar, pragmatics, etc. are being collected and put together in reference materials and video tapes (Günther Schneider, personal communication). Schneider/North (2000) is one of the first pilot studies to try to concretise the learner language levels of the *Reference Framework* for the purpose of assessment. While Schneider's research group collects examples of reference output from randomly collected classroom data, another approach would be the in-depth analysis of a learner corpus, which would provide quantitative and more systematic data on the forms, structures and meanings that learners at a specific stage of their learning process have at their disposal. It is this learner corpus approach to which we will now turn.

### *2.1 Learner corpus analysis: the case of discourse markers*

One of the areas that learner corpus research in Giessen has concentrated on is the use of discourse markers in learner language. The native-like use of discourse markers certainly belongs to the most difficult areas in learning a foreign language and they are therefore acquired at a relatively late stage. The Giessen Long Beach Chaplin Corpus (GLBCC), which was compiled in the late 1990s by Andreas Jucker, Simone Müller and Sara Smith, provides a suitable database for research into discourse markers in advanced learners' language (cf. Jucker et al. 2003, Müller 2004). GLBCC is a spoken corpus which comprises 136 recordings in English and 25 additional recordings in German; all speakers are university students. The English section of the corpus includes 53,028 words produced by 34 American native speakers and 95,555 words produced by 77 German learners of English as a foreign language. The data include oral narratives and conversations between two students, the subject of which is the silent Chaplin movie *The Immigrant* (hence the name of the corpus). The transcription follows, by and large, DuBois's (1991) transcription design principles.

Discourse markers are language-pedagogically relevant at an advanced level because it is by using discourse markers that advanced learners like German university students of English language and literature and of Teaching English as a Foreign Language can progress towards the reference levels C2 ("Mastery") in the *Common European Framework of Reference for Languages*. With respect to the "functional competence" of "spoken fluency," the *Reference Framework* specifies level C2 as follows:

> Can express him/herself at length with a natural, effortless, unhesitating flow. Pauses only to reflect on precisely the right words to express his/her thoughts or to find an appropriate example or explanation.
>
> (*Common European Framework of Reference for Languages* 2001:129)

Like many other descriptions offered by the *Reference Framework*, the entire scale for spoken fluency can also be found in most modern ELT curricula in Germany. The target level for learners in German secondary schools, however, is usually C1 or B2. Hasselgren (2002) shows that in the learner language of Norwegian learners of English there is a strong

correlation between the use of discourse markers – which she subsumes under the notion of smallwords – and the overall fluency of learners:

> The findings from this analysis of smallword use seem to support the hypothesis that the more fluent pupils used smallwords in a more native-like way than the less fluent, as far as quantity and distribution across turns are concerned.
>
> (Hasselgren 2002:154)

For advanced learner language of German speakers of English, Müller's (2004) PhD thesis, which is based on the GLBCC, provides the first in-depth analysis of the pragmatic functions of four discourse markers, namely *well*, *like*, *you know* and *so*. For the ELT community, her study provides a wealth of useful data because her findings reveal fields in which the use of these four discourse markers is already similar to the native norm and fields in which there still is a marked discrepancy. For example, Müller (2004) finds that there are only two interactional functions of *you know* for which there is no significant difference between advanced learners in the GLBCC and native speakers (see Figure 3). These two functions are labelled by her "see the implication" (SIM) and "appeal for understanding" (AFU). Some typical examples are given in (1) and (2).

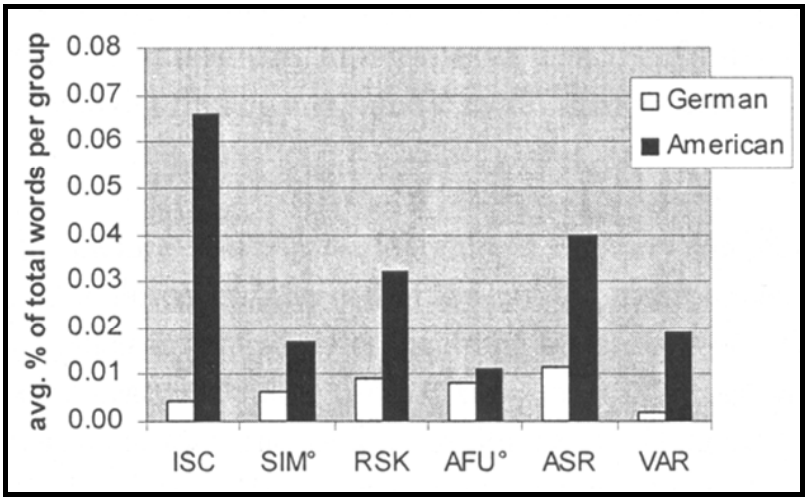


**Fig. 3:** Distribution of the interactional functions of *you know* in the GLBCC (Müller 2004:196).

(1) *you know* – "see the implication" (SIM, cf. Müller 2004:180 f.):

(a)

**B:** and he goes in the in the restaurant, .. thinking that he's got money, but he hasn't got money. **you know?** um, and he orders some food,

(non-native)

(b)

**B:** and he goes to pay the bill, .. and the waiter .. bends the
coin it's no good. [**you know**],
**A:** [o=h],
**B:** (H) so then he doesn't know what he's gonna do.

(native)

(2) *you know* – "appeal for understanding" (AFU, cf. Müller 2004:186 f.):

(a)

… remember [this] dress and this &
**B:** [yeah],
**A:** & um …(1.1) <L2 Kopftuch L2>?
[[<L2 Kopftuch L2>]]?
**B:** [[yeah I don't know]]. [yeah]
**A:** [this]
**you know**, and he he was falling down,

(non-native)

(b)

**B:** and he paid for it, but it was fake, … cause it could bend,
so then he goes <Q oh wai' wai' Q> & &I want .. some coffee
.. for her, so to- .. ***you know***.
[<X final X>]—
**A:** [he knows] he knows it's faked?

(native)

On the other hand, as shown in Figure 3, the remaining three interactional functions are used considerably less frequently by advanced learners than

by native speakers. These functions are labelled "imagine the scene" (ISC), "reference to shared knowledge" (RSK), and "acknowledge that the speaker is right" (ASR). Some typical native examples are given in (3) to (5).

(3) *you know* – "imagine the scene" (ISC, cf. Müller 2004:177):

…(1.6) made it look like he wasn't too & & obvious <X in X> like
he's picking up—picking up coins up off the ground for fun, ***you know***,

(native)

(4) *you know* – "reference to shared knowledge" (RSK, cf. Müller 2004:184):

the= … the guy the thief was really & & convincing. .. ***you know***
the big guy.
**B:** [yea=h]
**A:** [<SV at the] beginning SV>.

(native)

(5) *you know* – "acknowledge that the speaker is right" (ASR, cf. Müller 2004:188):

.. did you like it.
**B:** … I thought it was pretty good, .. ***you know***, for no words, …
it was amusing,

(native)

It is obvious that such findings provide a good starting point for an improvement of ELT textbooks and materials, given that – apart from *well* – all other discourse markers are notoriously underrepresented even in modern materials. In future materials, the various functions of frequently used discourse markers should be given more prominence because mastery of discourse markers is part and parcel of the pragmatic competence and spoken fluency that is necessary for achieving an overall proficiency in line with the levels C1 or C2 in the *Common European Framework of Reference for Languages*. This kind of analysis of a large

reference learner corpus is thus a good example of the first important step in bringing learner corpus data to the classroom. The next two steps to be taken are individualisation and localisation.

### *2.2 Individualisation of the analysis: the case of discourse markers in LINDSEI-Ger*

Like most other corpus-based studies of learner language, Hasselgren (2002) and Müller (2004) describe learner language at a specific stage as a system, abstracting from the corpus general trends across individual learners' usages. Also, the learners are taken to be representative of a specific stage in the learning process; they are not any single teacher's own students. In order to involve teachers in learner corpus research and to increase the relevance of learner data analysis to the teaching practice, it is necessary in our view to do two things: (1.) to individualise the analysis; (2.) to localise the database.

In this context, individualisation of the analysis is intended to mean that for the purpose of individual assessment and analysis of the variation between learners, it would be useful to complement the learner-language-as-a-whole perspective by also taking into account the differences between learners. For illustrative purposes, we would like to briefly refer to the case of discourse markers again. At the University of Giessen, the German component of LINDSEI (cf. De Cock et al. 2003) – the spoken counterpart of the International Corpus of Learner English (ICLE, cf. Granger et al. 2002), including 50 interviews with monologic and dialogic parts – has been compiled and transcribed (and is scheduled for publication on CD-ROM in 2006, cf. Brand/Kämmerer 2006). The design of the interviews is broadly comparable to the GLBCC corpus. However, the German component of LINDSEI (LINDSEI-Ger) will include only learner language of German university students of English. While it is no doubt useful to analyse the final corpus in its entirety, it is also insightful to compare individual learners' use of discourse markers to unveil the wide range of individual proficiencies within what we usually believe to be a relatively homogeneous group.

Consider the case of two individual learners in the German component of LINDSEI, namely speakers 047 and 050. Their use of discourse markers is markedly different, although their learner profiles (e.g.

with regard to age, number of years of English at school and university, stays abroad) are nearly identical. In Table 1, the overall frequencies of the four discourse markers *like*, *so*, *well* and *you know* in the interviews of speakers 047 and 050 are given. The right-hand columns give the corresponding frequencies in the learner language component and in the native component of GLBCC.

| | **LINDSEI GER 047 Speaker B** | | **LINDSEI GER 050 Speaker B** | | **GLBCC learner language** | | **GLBCC US native component** | |
|---|---|---|---|---|---|---|---|---|
| Number of words | **1,866** | | **1,698** | | **95,555** | | **53,028** | |
| *like* (all) | 10 | 0.53% | 17 | 1.00% | | | | |
| *like* (as DM) | **1** | **0.05%** | **12** | **0.71%** | **212** | **0.22%** | **754** | **1.42%** |
| *so* (all) | 17 | 0.91% | 22 | 1.30% | | | | |
| *so* (as DM) | **12** | **0.64%** | **17** | **1.00%** | **634** | **0.67%** | **639** | **1.21%** |
| *well* (all) | 10 | 0.53% | 4 | 0.24% | | | | |
| *well* (as DM) | **5** | **0.27%** | **1** | **0.06%** | **315** | **0.33%** | **73** | **0.14%** |
| *you know* (all) | 0 | 0.00% | 0 | 0.00% | | | | |
| *you know* (as DM) | **0** | **0.00%** | **0** | **0.00%** | **68** | **0.07%** | **197** | **0.37%** |

**Table 1:** Discourse markers (DM) in two advanced learners' speech.

In some regards, the two learners show the same deviation from the native norm: what is most striking is perhaps the total absence of *you know* as a discourse marker. On the other hand, GER speaker 050 is much closer to the native norm in her use of discourse-marker *so* and, to a lesser extent, *like*. It does not come as a surprise that also with regard to the various functions of the discourse markers *so* and *like*, speaker 050 is closer to the native norm than speaker 047. For example, while speaker 047 uses *like* as a discourse marker only once to introduce an example, speaker 050 uses *like* for many more native-like functions, including "exemplification," "explanation" and "focusing." Some examples are given in (6); note also her use of *kind of like* in (6c).

(6) Examples of *like* as a discourse marker in speaker GER 050's speech:

(a) *like* – "exemplification":

<B> really . erm .. but erm yeah and it wasn't allowed to drink . at all . even in the families **like** a glass of wine during the meal it wasn't . was really strict <\B>

(b) *like* – "explanation":

<B> yes . uhu . I had a exchange year erm . **like** last semester in England <\B>

(c) *like* – "focusing":

<B> . her friends: . I only can see the face of the one . girl . and I think she: erm is .. kind of **like** smiling <\B>

In the light of these differences between speakers 047 and 050, it would seem that learner-corpus-as-a-whole figures provide useful average numbers, but that individual learners may clearly deviate from this average. For example, while the frequency of *well* in the speech of speaker 047 is relatively close to the frequency in GLBCC, speaker 050 uses *well* less frequently than in GLBCC; but in this regard, speaker 050 again approximates more clearly to the native speaker norm. From the perspective of, say, a lecturer who happens to have speakers 047 and 050 in his or her class, this kind of individual analysis of learners' L2 idiolects would provide relevant data both for the individual assessment of and for the individual feedback on learner language.

In general, then, it is quite obvious that from the perspective of a specific classroom context – or, for that matter, from a specific teacher's perspective – learner corpora are needed that include the language of the learners that are present in that very classroom. While the analysis of large reference learner corpora like ICLE, GLBCC and LINDSEI provide an abstract description of what Granger (1998:7) calls the "archetypal learner." teachers in the classroom are also interested in the output of the learners in their own classrooms. Seidlhofer (2002) is thus absolutely right in arguing the case for "local learner corpora," compiled and analysed in a specific classroom context:

> FL pedagogy, and presumably any pedagogy, has to be local, designed for specific learners and settings.
>
> (Seidlhofer 2002:220)

This brings us to the third step, i.e. the localisation of the database.

### *2.3 Localisation of the database: the case of error analysis in a local learner corpus*

In order to illustrate the principles and the potential benefits not only of an individualisation, but also of a localisation of learner corpus research, we would like to report on a more experimental classroom project which has been initiated at Hainberg-Gymnasium Göttingen. The starting point for our idea was the trivial yet significant observation that every single teacher constantly receives an enormous amount of learner texts, corrects them and hands them back without ever storing – let alone, analysing – the sheer amount of learner data systematically. Our main aim, thus, is to find ways of how this waste of learner data can be overcome and how teachers – and learners – can compile and analyse their own local learner corpora with minimal effort.

What comes to mind immediately are written examinations.[3] We collected the entirety of learner texts that were produced by the students of one specific class, the *Leistungskurs Englisch EN 1/3 L1 ROR* (*ROR* standing for the English teacher of the class, Jan-Marc Rohrbach), at Hainberg-Gymnasium Göttingen in the context of two written examinations.[4] The resulting corpus, the Giessen-Göttingen Local Learner Corpus of English (GGLLC), thus, at present consists of two subcorpora with

[3] We could have collected *Facharbeiten*, i.e. proto-scientific student papers on specific topics, much more easily, given that these papers are usually produced electronically. However, we regard written examinations as more spontaneous writings and as more reliable representations of students' language competence because it can be safely ruled out that the texts have been edited by the student or proof-read by another person.

[4] In Germany, the *Leistungskurs* is a course for advanced students who choose the subject at hand as a field of specialisation in classes 12 and 13 and in their final examinations, the *Abitur*. In a *Leistungskurs*, students usually have 5 to 6 lessons per week.

data from one class: all learner texts produced in a written examination in class 12 (EN1L1), and all texts of the same learners produced in a written examination one year later, i.e. in class 13 (EN3L1). The general design of GGLLC is sketched out in Table 2.[5]

| | **Subcorpus EN1L1 ROR (Class 12)** | **Subcorpus EN3L1 ROR (Class 13)** |
|---|---|---|
| **Date** | 6 Oct 2003 | 4 Oct 2004 |
| **Size of subcorpus** | 13,076 words | 19,297 words |
| **Overall size** | 32,373 words | |
| **Number of texts**[6] | 12 | 16 |
| **Mean text size** | 1,090 words per text | 1,206 words per texts |
| **Topic of the Examination** | "Capitalism and the Atlantic divide rich nations of the West" (Guided analysis of a text about US and European economics ) | "Melting pot or tossed salad?" (Guided analysis of a text about the current debate in Britain over multicultural education) |

**Table 2:** The Giessen-Göttingen Local Learner Corpus of English (GGLLC).

It is obvious that computerising the exams must be 'quick and dirty' in style, without sophisticated markup and the like, if the learners are to work with the data themselves soon after the examinations have been corrected. Thus, it is only the plain learner texts (i.e. the handwritten examinations) with the teacher's correction marks that have been typed in. All learner texts were then stored in three different formats:

- **Plain format:** the learner's own output only, see example (7a);

[5] In its current version, GGLLC serves as a pilot corpus in order to try out and evaluate various teacher-led and learner-centred corpus-based activities. At a later stage, GGLLC could easily be extended by including data from other classes at Hainberg-Gymnasium Göttingen, potentially ranging from the early secondary level to the *Abitur*.

[6] Four students who took part in the 2004 exam had not taken part in the 2003 exam, which explains the difference in the overall number of texts between the two subcorpora.

- **Marked format:** the learner's output and the teacher's correction marks, see example (7b);
- **Corrected format:** the learner's output, the teacher's correction marks and the teacher's actual suggestions for correction, see example (7c).[7]

(7) (a) plain format:

The headmistress Daphne Gould support the attitude that anti-racism should be treated as serious as other subjects Talking about racism is one measure to reduce xenophobic attitudes Moreover she explains that ethnic minorities and their white counterparts are both responsible of racism and that teachers need to support anti-racist campaigns and multicultural approaches

(2004-EN3L1-10-plain-text)

(b) marked format:

The headmistress, Daphne Gould, support <Gr> the attitude that anti-racism should be treated as serious <Gr> as other subjects. Talking about racism is one measure to reduce xenophobic attitudes. Moreover, she explains that ethnic minorities and their white counterparts are both responsible of <Pr> racism and that teachers need to support anti-racist campaigns and multicultural approaches.

(2004-EN3L1-10-marked-text)

(c) corrected format:

The headmistress, Daphne Gould, support <Gr supports> the attitude that anti-racism should

[7] The correction marks that are used in the excerpts in (7) are *Gr* (*Grammatik*) for grammar mistakes and *Pr* (*Präposition*) for wrong choice of prepositions. Other standard correction marks include *W* (*Wortschatz*, lexical choice), *A* (*Ausdruck*, idiomaticity), *Sb* (*Satzbau*, sentence structure), *St* (*Stellung*, word order), *T* (*Tempus*, tense) and *R* (*Rechtschreibung*, orthography). Combinations of the codes are possible, e.g. *Gr Sb*.

> be treated as serious <Gr seriously> as other subjects. Talking about racism is one measure to reduce xenophobic attitudes. Moreover, she explains that ethnic minorities and their white counterparts are both responsible of <Pr for> racism and that teachers need to support anti-racist campaigns and multicultural approaches.
>
> (2004-EN3L1-10-corrected-text)

The data presently included in GGLLC allow for various activities on the part of the teacher. First and foremost, the teacher is now able to analyse both quantitatively and qualitatively his own students' language at a specific point in the learning process. For example, the teacher could – by using standard corpus-linguistic software like *WordSmith Tools* – generate wordlists to check the range of general and topic-related vocabulary that students in general or individual students have used. Secondly, the teacher is enabled not only to analyse the corpus in its entirety, but also to focus on individual learners. For example, the teacher can provide specific feedback to an individual learner by providing him/her with concordance lines that highlight frequently occurring kinds of mistakes in that particular student's learner language. Thirdly, GGLLC makes it possible for the teacher to also evaluate the progression in learner language by comparing the 2003 data and the 2004 data; again, by applying this longitudinal perspective, we may wish to focus either on the class as a whole or on specific learners in particular. For example, the teacher may be interested in finding out whether specific kinds of errors occur more frequently or less frequently after one year in an individual learner's output.

Consider in this context the case of learners' use of the noun *conclusion* (which certainly is a central word in argumentative writing) in the 2003 and 2004 written examinations. Figure 4 gives all instances of *conclusion* in the 2003 subcorpus EN1L1 ROR (Class 12) of GGLLC. Note that there is a clear prevalence of the phrase *in conclusion*. While some frequent lexicogrammatical patterns are used (*come to a conclusion, lead to the conclusion that*), others are missing (e.g. *draw a conclusion, reach a conclusion*).

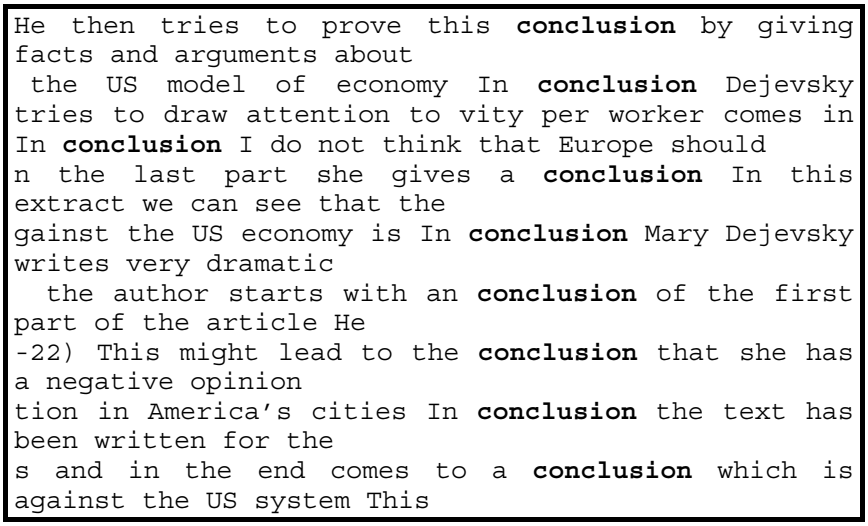


**Fig. 4:** Concordance for conclusion in GGLLC, subcorpus EN1L1 ROR (Class 12).

Figure 5 gives all instances of *conclusion* in the 2004 written examination in the same class. Interestingly, the most frequent phrase is no longer *in conclusion*, but *as a conclusion*. This certainly is a problematical development because *in conclusion* is much more frequent and idiomatic than *as a conclusion*, the latter being notoriously overused by German learners of English at university level as well.[8] It is here that the teacher should try to make learners aware of their overuse of a phrase which rarely occurs in native-like language use. Also, there are non-idiomatic patterns that are used in connection with *conclusion*: **This draws one to the conclusion* is an inadmissible form, which should also be pointed out to the learners. In general, then, Figures 4 and 5 show that there is not always a clear progression in learner language at an advanced level in secondary schools.

[8] Note that *in conclusion* occurs 315 times in the 100-million-word British National Corpus (BNC World Edition, Burnard 2000), while *as a conclusion* can be found only 8 times.

ominantly single culture In **conclusion** as I said before I would think that
ll will depelop itself As a **conclusion** I would say that maybe we should
ass full of immigrants As a **conclusion** it becomes obvious that an approach
ir own school (l 13) In her **conclusion** Mrs Balinska refers back to
a multicultural society In **conclusion** one can say that Maria Balinska
none This draws one to the **conclusion** that she wants to draw the reader's
into this murder led to the **conclusion** that moral anti-racism failes its
t with Hanif and Aruna As a **conclusion** we can say that the author uses many
support ratio To come to a **conclusion**: The concepts of a multicultural

**Fig. 5:** Concordance for *conclusion* in GGLLC, subcorpus EN3L1 ROR (Class 13).

From a methodological point of view, Figures 4 and 5 illustrate one of the key advantages of a computerised local learner corpus for the teacher: he/she is now able to immediately identify patterns of use in learner language across all students in his/her class. The identification of 'typical mistakes' can now be put on an empirical footing and be based on the analysis of all learners' actual output.

Learner-centred activities using local learner corpus data are of course also possible. By giving learners access to the entirety of data that accumulate in a written examination, they not only profit from the correction of their own mistakes but also from the analysis of their fellow-students' errors and their corrections.[9] A very useful activity is the learner-centred mutual correction of examinations. Learners will use the plain format of fellow-students' texts and insert corrections auto-

[9] Needless to say, learners need to be familiarised with the basic functions of corpus-linguistic software like *WordSmith Tools*, i.e. learners must become more or less 'corpus literate' (cf. Mukherjee 2002:179 f.) before being able to work with corpora.

nomously; at the same time their own texts will be corrected by fellow-students. After discussing the mutual corrections in smaller groups, the students' corrections should be compared and contrasted with the teacher's correction marks in the marked format and, in a second step, with the teacher's actual corrections in the corrected format. This is a very effective way of increasing learners' language awareness by making them work with their own output.

Since corpus-linguistic software allows learners not only to look for particular words and patterns but also for particular categories of errors (in the marked and corrected formats of the corpus), learners may also find it useful to review their errors in terms of error categories. For example, student 1 will find that in the 2004 examination she made only one orthographical mistake (category *R*), see Figure 6.

```
obvious that this problem is definitly <R> not
solved <Gr T> at all.
```

**Fig. 6:** Concordance for *R* errors in the 2004 written examination of student 1.

On the other hand, the same student, when looking for all of her grammar mistakes (*Gr*), will find quite a few instances, see Figure 7. The concordance lines in Figure 7 should be taken by the student as a starting-point for a revision of some general rules in English grammar from which she deviated, e.g. in the field of article usage (cf. lines 2 and 4) and with regard to the distinction between past tense and present perfect (cf. lines 5,7 and 8).

```
ridicule the difficult situations <Gr> at
multiracial schools like <W>
anti-racism should be part of the <Gr> education. As
for her not doing so
ary to <A> these two opinions for <Gr> Mr. Ray
Honeyford, a former
Their aim is to fight against the <Gr> multiracial-
schools-racism <A>.
      school for over 10 years is <Gr T> dominated
by anti-racism and
anti-racism and multi-culturalism <Gr Sb>. The
opinion of this headmistress
```

```
problem is definitly R not solved <Gr T> at all.
Equality which is already adopted <Gr T> by many
Local Education
```

**Fig. 7:** Concordance for *Gr* errors in the 2004 written examination of student 1.

There is a wide range of options for the use of local leaner corpus data like GGLLC. For example, various types of data-driven learning material could be designed on the basis of learner data (in addition to native data, cf. Mukherjee 2003), and the analysis of learner data could also be fruitfully combined with recent approaches to the teaching of genre competence (cf. Rohrbach 2003). In all of these actvities, the learner's analysis of his/her own output serves as a 'bottom-up' complement to the 'top-down' teaching of the native-speaker target norm (cf. Osborne 2004).

In the long run, the particular attraction of the kind of local learner corpus that has been described in the present section will also lie in its monitoring quality since it will document learner language progression both at the level of the class in its entirety and at the level of individual learners. Each individual learner will also be able to trace his/her progression in the various error categories. A local monitor learner corpus, which – if corpus compilation starts early enough – could well include learner data from several years of learning English as a foreign language, would be an invaluable resource both for the teacher and his/her students in getting access to data on language learning success and failure.[10] It goes without saying that other kinds of learner texts are also suitable for inclusion in such local learner corpora, especially long, term-paper-like essays that all students at advanced secondary level have to write on a specific topic (i.e. the previously mentioned *Facharbeiten*); they are usually written on the computer and would just have to be changed into plain texts. In fact, it would be insightful to actually compare the same learners' output in various communication situations by including data not only from written examinations, but also from, say, edited and proof-read student essays, written homework assignments and – if possible – from

[10] Note that this kind of long-term learner language documentation ties in very well with the recent emphasis on language portfolios in foreign language teaching (cf. Schärer 2001).

spoken classroom discourse. This would make it possible to describe both the diachronic progression of learner language and the synchronic variation of learner language across diverse communication situations.

## 3. Concluding remarks

In the present paper, we started out from a theoretical discussion of differences in perspective between applied corpus linguistics and language pedagogy. We then focused on learner corpus research and argued that in order to increase the relevance of learner corpora to the EFL classroom it would be useful to complement research into reference learner language corpora like ICLE and GLBCC with an analysis of individual learners' data and with the compilation and analysis of local learner corpora. The compilation of a local learner corpus like GGLLC, which we sketched out in the last part, has just been tried out in the first pilot study in a class in years 12/13 at Hainberg-Gymnasium Göttingen. The data – and the work done so far on their basis – lead us to believe that this kind of localisation of learner corpus compilation and analysis will no doubt prove to be a promising and enticing new avenue in learner corpus research. While it is true that the use of learner data in the EFL classroom as envisaged in the present paper remains a controversial issue in the ELT community (cf. Meunier 2002), we do think that, firstly, the focus on their own students' output will involve many more teachers in corpus-based activities and that, secondly, the exploration of learner data by the learners themselves will motivate many more learners to reflect on their language use and thus raise their foreign language awareness.

Coming full circle, we would argue that the gap between corpus-linguistic research and classroom teaching, which was bemoaned at the beginning of the present paper, can best be bridged if special emphasis is placed on the teacher's perspective.

> […] the most constructive way forward is to recognise and act upon the need for empirical classroom-based action research conducted by teachers who are aware of the potential as well as the limitations of corpus linguistics.
>
> (Seidlhofer 2002:215)

We couldn't agree more with Seidlhofer with respect to her plea for more classroom-based corpus-linguistic action research. We hope that more teachers will start compiling their own local mini-corpora and make the best possible use of the vast amount of learner output produced in their own EFL classrooms.

## References

Aston, G. (2001). "Learning with corpora: an overview." In: Guy Aston (ed.). *Learning with Corpora*. Houston, TX: Athelstan. 7-45.

Bernardini, S. (2004). "Corpora in the classroom: an overview and some reflections on future developments." In: J. Sinclair (ed.). *How to Use Corpora in Language Teaching*. Amsterdam: John Benjamins. 15-36.

Brand, C./S. Kämmerer (2006). "The Louvain international database of spoken English interlanguage: compiling the German component." In: S. Braun/K. Kohn/J. Mukherjee (eds.). *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*. Frankfurt am Main: Peter Lang. In press.

Burnard, L. (2000). *Reference Guide for the British National Corpus (World Edition)*. Available at <http://www.hcu.ox.ac.uk/BNC/World/html/urg.html> [accessed 1 Aug 2004].

Burnard, L./T. McEnery (eds.) (2000). *Rethinking Language Pedagogy from a Corpus Perspective: Papers from the Third International Conference on Teaching and Language Corpora*. Frankfurt/M.: Peter Lang.

Council of Europe (2001). *Common European Framework of Reference for Languages*: *Learning, Teaching, Assessment.* Cambridge: CUP.

De Cock, S./S. Granger/S. Petch-Tyson (2003). "The Louvain International Database of Spoken English Interlanguage – LINDSEI." <http://www.lftr.ucl.ac.be/fltr/germ/etan/cecl/Cecl-Projects/Lindsei/lindsei.htm> [accessed 13 Jul 2003].

DuBois, J. (1991). "Transcription design principles for spoken discourse research." *Pragmatics* 1:1. 71-106.

Granger, S. (1998). "The computer learner corpus: a versatile new source of data for SLA research." S. Granger (ed.). *Learner English on Computer*. London: Longman. 3-18.

Granger, S. (2002). "A bird's eye view of learner corpus research." In: S. Granger/J. Hung/S. Petch-Tyson (eds.). *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Amsterdam: John Benjamins. 3-33.

Granger, S./E. Dagneaux/F. Meunier (eds.) (2002). *International Corpus of Learner English*. Louvain-la-Neuve: Presses Universitaires de Louvain.

Granger, S. (2004). "Computer learner corpus research: current state and future prospects." In: U. Connor/T. Upton (eds.). *Applied Corpus Linguistics: A Multidimensional Perspective*. Amsterdam: Rodopi. 123-145.

Hasselgren, A. (2002). "Learner corpora and language testing: small-words as markers of learner fluency." In: S. Granger/J. Hung/S. Petch-Tyson (eds.). *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Amsterdam: John Benjamins. 143-173.

Jenkins, J. (2000). *The Phonology of English as an International Language*. Oxford: OUP.

Jenkins, J. (2004). "ELF at the gate: the position of English as a lingua franca." *The European English Messenger* 13:2. 63-69.

Jucker, A.H./S.W. Smith/T. Lüdge (2003). "Interactive aspects of vagueness in conversation." *Journal of Pragmatics* 35. 1737-1769.

Leech, G. (1997). "Teaching and language corpora: a convergence." In: A. Wichmann/S. Fligelstone/T. McEnery/G. Knowles (eds.). *Teaching and Language Corpora*. London: Longman. 1-23.

Meunier, F. (2002). "The pedagogical value of native and learner corpora in EFL grammar teaching." In: S. Granger/J. Hung/S. Petch-Tyson (eds.). *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Amsterdam: John Benjamins. 119-141.

Müller, S. (2004). *Discourse Markers in Native and Non-native English Discourse*. University of Giessen: PhD thesis.

Mukherjee, J. (2002). *Korpuslinguistik und Englischunterricht: Eine Einführung*. Frankfurt a.M.: Peter Lang.

Mukherjee, J. (2003). "Korpusbasierte Aktivitäten im Englischunterricht: Konzepte und Vorschläge für die Unterrichtspraxis." In: G. Fehrmann/ E. Klein (eds.). *Schüleraktivierung im Fremdsprachenunterricht: Beiträge zur Tagung des FMF-Nordrhein am 10. September 2002 in Aachen*. Bonn: Romanistischer Verlag. 41-53.

Mukherjee, J. (2004). "Bridging the gap between applied corpus linguistics and the reality of English language teaching in Germany." In: U. Connor/T. Upton (eds.). *Applied Corpus Linguistics: A Multidimensional Perspective*. Amsterdam: Rodopi. 239-250.

Nesselhauf, N. (2004). "Learner corpora and their potential for language teaching." In: John Sinclair (ed.). *How to Use Corpora in Language Teaching*. Amsterdam: John Benjamins. 125-152.

Osborne, J. (2004). "Top-down and bottom-up approaches to corpora in language teaching." In: U. Connor/T. Upton (eds.). *Applied Corpus Linguistics: A Multidimensional Perspective*. Amsterdam: Rodopi. 251-265.

Rohrbach, J. (2003). "'Don't miss out on Göttingen's nightlife': Genreproduktion im Englischunterricht." *Praxis des Neusprachlichen Unterrichts* 50. 381-389.

Schärer, R. (2001). *A European Language Portfolio: Final Report on the Pilot Project (1998-2000)*. Strasbourg: Council of Europe.

Schlüter, N. (2002). *Present Perfect: Eine korpuslinguistische Analyse des englischen Perfekts mit Vermittlungsvorschlägen für den Sprachunterricht*. Tübingen: Gunter Narr.

Schneider, G./B. North (2000). *"Dans d'autres langues, je suis capable de …": Echelles pour la description, l'évaluation et l'auto-évaluation des compétences en langues étrangères*. Berne: Direction du Programme national de recherche 33, en collaboration avec le Centre suisse de coordination pour la recherche en éducation.

Seidlhofer, B. (2001). "Closing a conceptual gap: the case for a description of English as a lingua franca." *International Journal of Applied Linguistics* 11:2. 133-158.

Seidlhofer, B. (2002). "Pedagogy and local learner corpora: working with learning-driven data." In: S. Granger/J. Hung/S. Petch-Tyson (eds.). *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Amsterdam: John Benjamins. 213-234.

Seidlhofer, B. (2004). "Research perspectives on teaching English as a lingua franca." *Annual Review of Applied Linguistics* 24. 209-239.

Vielau, A. (1991). "Sprachlos in vielen Sprachen." *Praxis des Neusprachlichen Unterrichts* 38. 20-28.