Learner language

Gaëtanelle Gilquin & Sylviane Granger
Centre for English Corpus Linguistics
Université catholique de Louvain

1. Learner corpus research

1.1. State-of-the art survey

While a large range of language varieties had been explored by corpus linguists from the emergence of the field, it was only in the late 1980s that corpus linguists began to show an interest in learner language and started collecting learner corpora, i.e. electronic collections of writing or speech produced by foreign or second language learners. The new research strand, which is commonly referred to as learner corpus research (LCR), encroached on a field which until then had been the sole remit of second language acquisition (SLA). The two fields have the same object of study, i.e. learner language, but they differ markedly in their objectives and methods of analysis. One of the main differences is that SLA studies focus on competence: "[t]he main goal of SLA research is to characterize learners' underlying knowledge of the L2, i.e. to describe and explain their competence" (Ellis 1994: 13). LCR studies, on the other hand, focus on performance; their main objective is to describe the use of language by learners in actual production. To do so, they apply the tools and techniques of corpus linguistics, which, thanks to the degree of automation they involve, allow for the study of whole learner populations. This is to be contrasted with the more manual methods of analysis of traditional SLA studies, which are better suited for the investigation of a small number of individual learners. The field of LCR also has links with English as a Lingua Franca (ELF), although here again the perspective is different. ELF is not so much interested in the learning process itself as in the use of an L2 for communication purposes (see Mauranen 2011).

LEARNER CORPUS DATA

There are many different types of performance data and only some of them qualify as learner corpus data. Unlike the more experimental data types often used in SLA, where learners are forced to produce a particular form (as in fill-in-the-blanks exercises or reading-aloud tasks), the focus in learner corpus data is on message conveyance and the possibility for learners to use their own wording. In principle, like any other corpora, learner corpora need to be authentic, i.e. "gathered from the genuine communications of people going about their normal business" (Sinclair 1996). However, learners, especially those who learn the target language in a country where that language is not spoken, rarely – if ever – use the target language in their normal everyday activities. The criterion of authenticity therefore needs to be relaxed in the case of learner corpora so as to include data, such as free compositions or informal interviews, "that are elicited for the corpus but that use procedures exerting very little control" (Nesselhauf 2004: 128). The notion of control is fluid, however, and some learner corpus researchers make use of more "peripheral types of learner corpora" (ibid.), such as picture descriptions or student translations that involve a higher degree of control (cf. the *Giessen-Long Beach Chaplin Corpus*, a corpus of retellings of a silent Charlie Chaplin movie).

Learner corpora can be of different types, including general or specific, written or spoken, synchronic or longitudinal, mono-L1 or multi-L1 data, which can be produced by learners of different origins and different proficiency levels. A survey of the learner corpora currently

available (see http://www.uclouvain.be/en-cecl-lcworld.html), however, reveals that some types of learner corpora are more common than others. Thus, while in principle learner corpus data can be collected from learners at all proficiency levels, most corpora to date represent the more advanced stages. This initially stemmed from a wish to fill a double-sided gap, i.e. a general neglect of the more advanced proficiency levels by both SLA researchers and designers of teaching resources, but it would now be desirable to revisit the other stages on the basis of corpus data. Another feature is that the number of written corpora by far outnumbers that of spoken corpora. This imbalance results from the difficulty of collecting and transcribing oral data produced by learners. To some extent, this focus on writing can be seen as a positive shift from SLA studies that generally prioritize L2 oral production and indeed, as will be shown below, LCR studies have greatly contributed to the analysis of learner writing. However, this balance needs to be redressed and projects such as the *Role Play Learner Corpus* and the *Louvain International Database of Spoken English Interlanguage* are particularly welcome. To take full advantage of these spoken learner corpora, which are usually released as transcriptions, it would also be advisable to have access to the audio files, so as to allow the investigation of learner pronunciation and prosody.

The majority of learner corpus studies are based on raw data, i.e. data devoid of any linguistic annotation. This is changing, however, and an increasing number of studies make use of annotated data, usually in the form of part-of-speech (POS) tagged or error-tagged data. Annotation of learner data comes with potential problems. Having been developed on the basis of native corpus data, POS-taggers may not perform as well when applied to learner texts. Studies have shown the success rate to be sensitive to errors, especially spelling errors, although for higher proficiency level texts which contain relatively few formal errors, the accuracy rate remains quite high (Granger et al. 2009: 16). Error tagging poses challenges of a different order (see also Section 1.3), having to do mainly with the manual nature of the annotation task and the variety of error typologies, which makes the results difficult to compare across LCR studies. However, annotated learner corpus data are worth the effort they involve, since they expand the possibilities of linguistic analysis (cf. Section 1.2, Granger & Rayson 1998).

LINGUISTIC PHENOMENA

Learner corpus research has tackled aspects of learner language that had been neglected until then. As against SLA studies which have traditionally prioritized morphology and grammar, LCR is characterized by a strong focus on lexis, lexico-grammar and a range of discourse phenomena. This has been made possible by corpus software tools such as *AntConc* (Anthony 2004) or *WordSmith Tools* (Scott 2012), with which it is easy to draw up frequency lists of single words or sequences of words of a given length (bigrams, trigrams, etc.), extract all occurrences of a particular linguistic item and identify its typical lexico-grammatical patterning. Studies using these methods have uncovered a facet of learner language which had hitherto been largely left untouched, i.e. the learner phrasicon, in particular learners' use of collocations (Nesselhauf 2005) and lexical bundles (Chen & Baker 2010). Discourse features are also strongly in evidence in learner corpus studies, in particular the use of connectors (Leńko-Szymańska 2008), discourse markers (Müller 2005), stance markers (Hasselgård 2009) and involvement features (Ädel 2008). Grammatical features, on the other hand, have been under-researched in LCR. The reason is that the majority of researchers rely on raw learner corpus data and investigations of grammar ideally require the use of POS-tagged or parsed data. Studies of learner grammar that make use of learner corpora tend to deal with aspects of grammar that can be studied on the basis of raw data, either because they involve one specific linguistic item, such as causative *make* (Gilquin 2012) or *what*-clefts (Callies

2009), or a closed class, such as modal auxiliaries (Aijmer 2002), articles (Díez-Bedmar & Papp 2008) or demonstrative pronouns (Petch-Tyson 2000). Studies relying on POS-tagged data are less common (Granger & Rayson 1998, Borin & Prütz 2004).

GENERAL RESEARCH ORIENTATIONS

Unlike non-corpus-based SLA studies which are typically hypothesis-testing, LCR studies tend to be exploratory or descriptive. This general research orientation is consistent with the very nature of learner corpora which are usually collected as generic resources to be used to answer a wide range of research questions not identified at the time of collection. However, some learner corpus researchers have adopted a more explanatory research design, which uses learner corpora to revisit important SLA findings. For instance, Tono (2000) and Housen (2002) revisit, on the basis of learner corpus data, the order of acquisition of morphemes established in SLA. Studies of this type are important as they can help bridge the gap between SLA and LCR, two fields which are still distant from each other – although recent research shows signs of a rapprochement, with learner corpus data slowly earning their rightful place among standard research methods in SLA (Mackey & Gass 2012).

Another important orientation of LCR resides in its strong links with teaching, which were in evidence from the start. Several studies demonstrate how learner corpora can be used to develop pedagogical tools and methods which target more accurately the needs of the learner. Wible et al. (2001), for example, describe a web-based writing environment that allows for the collection and pedagogical implementation of learner corpus data. Learner corpora are also used to inform pedagogical lexicography and language testing. However, as Flowerdew (2012: 207) rightly points out, they "still seem to be 'the poor relation' as far as pedagogic applications are concerned". Most studies include a section on the pedagogical implications of the findings, but up-and-running pedagogical resources that make full use of learner corpus data are still relatively rare.

1.2. Representative studies

1) Nesselhauf, N. (2003). The use of collocations by advanced learners of English and some implications for teaching. *Applied Linguistics* 24(2): 223-242.

This study looks into German learners' use of verb-noun collocations such as *take a break* or *shake one's head*, which were extracted manually from raw texts taken from the German component of the *International Corpus of Learner English* (ICLE) (Granger et al. 2009). The collocations were evaluated according to their degree of restriction (high as in *fail an exam* or low as in *exert influence*) and their degree of acceptability (correct, wrong or dubious). The results show that about a quarter of the collocations contain one or several mistakes (mainly due to a wrong choice of verb) and that the difficulty of the collocations for the learners appears to be less affected by their degree of restriction than by the influence of the L1. The L1 seems to be responsible for over half of the incorrect collocations, and its role is visible in all types of collocational mistakes. This study is interesting because it demonstrates that collocations still represent a major problem for advanced learners and that the L1 has a significant influence on the (mis)use of collocations. It is also interesting in terms of methodology and pedagogical implications. From a methodological point of view, the article includes a thorough and honest discussion of the problems that may be involved in classifying authentic corpus data (for example, to distinguish between restricted collocations and free combinations). From a pedagogical point of view, it addresses some implications of the

findings for teaching, such as the necessity of explicitly teaching some of the most difficult collocations and the desirability of adopting an L1-based approach which underlines differences between L1 and L2.

2) Osborne, J. (2008). Adverb placement in post-intermediate learner English: A contrastive study of learner corpora. In G. Gilquin, S. Papp & M. B. Díez-Bedmar (eds.) *Linking Up Contrastive and Learner Corpus Research* (pp. 127-146). Amsterdam: Rodopi.

This grammatical study on the placement of the adverb in learner English relies on two learner corpora, one containing essays produced by French learners only (*Chambéry Corpus*) and the other one including essays produced by learners from different mother tongue backgrounds (ICLE). These corpora were compared with two corpora of essays written by native speakers, viz. the *Louvain Corpus of Native English Essays* (LOCNESS) and the *Essay Bank*. All four corpora were POS-tagged so as to permit the automatic extraction of adverbs. The main objective of the study is to find out whether the learners' L1 may have an influence on the use of the Verb-Adverb-Object order (e.g. *to see clearly the contrast*), which is normally not allowed in English. The large proportion of this structure among French, Italian and Spanish learners might be explained by syntactic transfer from the L1, since French, Italian and Spanish all allow this structure. However, the finding that learners with other L1s also produce Verb-Adverb-Object sequences suggests that transfer is not the only explanation. Besides the information it provides about adverb placement in learner English, this article is important because it clearly demonstrates that problems seemingly due to L1 transfer could in fact be shared by other learner populations and thus be attributed to other factors. This should serve as a warning against studies that hastily conclude that transfer is at work without considering further evidence such as data from other learner populations or data from the learners' L1.

3) Flowerdew, L. (1998). Integrating 'expert' and 'interlanguage' computer corpora findings on causality: Discoveries for teachers and students. *English for Specific Purposes* 17(4): 329-345.

The objective of this study is to investigate the rhetorical function of causality in scientific text on the basis of an expert corpus (taken from the *MicroConcord Academic Corpus Collection*) and a learner corpus (taken from the *Hong Kong University of Science and Technology Learner Corpus*). The careful examination of concordances of 52 devices expressing reason-result, means-result or grounds-conclusion reveals a number of features that are distinctive for the learners (as compared to the expert writers). These differences include the overuse of logical connectors as markers of local coherence, their predominantly sentence-initial position, the reliance on a small set of devices, the use of idiosyncratic phrases (e.g. *it is because*), the absence of certain grammatical patternings (like reduced relative clauses with causative verbs) and the lack of mitigating markers such as modal verbs or adverbs in the direct environment of the causal devices. The article is especially interesting for its discourse-oriented perspective, starting from a rhetorical function rather than from one or two isolated items, as well as for its focus on scientific prose and its plea for more corpus-based English for Academic/Specific Purposes textbooks. It is also a good illustration of how useful and enlightening it is to consider the functionality of items in context (qualitative approach) in addition to their overall frequency (quantitative approach).

4) Granger, S. & Rayson, P. (1998). Automatic lexical profiling of learner texts. In S. Granger (ed.) *Learner English on Computer* (pp. 119-131). London & New York: Addison Wesley Longman.

The objective of this study is to establish whether it is possible to identify distinctive stylistic characteristics of learner writing fully automatically on the basis of POS-tagged corpora. The frequencies of POS categories – both major categories such as pronouns and subcategories like personal or indefinite pronouns – were compared in two similar-sized corpora, a corpus of native novice writing, LOCNESS, and the French component of ICLE. The comparison generated a number of distinctive POS configurations which highlight the speech-like nature of learner writing. The learners tend to underuse the categories typical of academic writing (e.g. nouns and prepositions) and overuse those typical of speech (e.g. indefinite determiners and pronouns, first and second personal pronouns). The study is a good example of a fully corpus-driven analysis which allows generalizations to emerge bottom-up from the learner data. It demonstrates the interest of using POS-tagged data, an approach that is still more the exception than the rule in LCR. The study paves the way for Biber-inspired multidimensional approaches to the analysis of learner language (e.g. Asención-Delaney & Collentine 2011), typology-driven studies (Szmrecsanyi & Kortmann 2011) and natural language processing applications, in particular automatic L1 identification (Jarvis & Paquot forthcoming).

5) Thewissen, J. (2013). Capturing L2 accuracy developmental patterns: Insights from an error-tagged EFL learner corpus. *The Modern Language Journal* 97(S1): 77-101.

This study makes use of data from ICLE to trace second language developmental trajectories in terms of accuracy. It is based on data sampled at the same point in time from learners at different proficiency levels. The data were submitted to two independent processes: they were rated by testing experts according to the *Common European Framework of Reference for Languages* (Council of Europe 2001) and fully annotated for errors on the basis of the Louvain error-tagging system (Granger 2003). Errors were counted using a new method, called potential occasion analysis, which quantifies errors in relation to the number of times they could be produced rather than in relation to the total number of words in the corpus. The study investigates 45 error types across four proficiency levels: lower intermediate, upper intermediate, advanced and near-native. The results show that the developmental patterns vary in function of the error category. While many error types (e.g. spelling errors or verb dependent preposition errors) are characterized by progress followed by stabilization, others (e.g. noun number errors) show steady linear development and others still (e.g. verb tense errors) show no sign of development. The findings generally bring support to SLA claims about the nonlinear nature of language development. They also have major pedagogical implications, in particular for language testing research which has started to extract error patterns together with other features of interlanguage (accurate use, over- and underuse) from learner corpora to achieve a more accurate description of language proficiency levels.

6) Belz, J. & Vyatkina, N. (2005). Learner corpus research and the development of L2 pragmatic competence in networked intercultural language study: The case of German modal particles. *The Canadian Modern Language Review* 62(1): 17-48.

This study demonstrates that learner corpus data can be used to develop L2 pragmatic competence thanks to a methodology that fully integrates learner corpus data into normal pedagogical activities. The data used are part of *Telekorp*, a bilingual corpus that results from a telecollaborative partnership involving email and synchronous chat between German-

speaking learners of English and English-speaking learners of German. The corpus contains native and non-native data collected longitudinally in the two languages over a period of three weeks. The study focuses on pragmatic competence in German, and more particularly the use of modal particles such as *ja* or *doch*. Data-driven learning is at the heart of the pedagogical intervention: learners are made to reflect on L1 and L2 data directly extracted from their own telecollaborative interactions. A comparison of learners' use of modal particles before and after the experiment shows a marked increase in the number of modal particles used, improved accuracy of use and a heightened awareness of their importance in German. One of the main advantages of using data from computer-mediated communication between native and non-native speakers is that the data are collected in the same interactions. This ensures that the native control data are fully comparable to the learner data with respect to learner and task variables.

1.3. Methodological issues for learner corpus research

While the methods generally applied in corpus linguistics can be applied in LCR too, there are a number of methods that have been specifically designed to deal with learner corpus data. They include contrastive interlanguage analysis, the integrated contrastive model and computer-aided error analysis. New directions can (and should) also be envisaged to complement these three methods.

CONTRASTIVE INTERLANGUAGE ANALYSIS

Many learner corpus studies rely (explicitly or implicitly) on a method called contrastive interlanguage analysis (CIA) (see Granger 1996). This method involves two types of comparison: one between a learner corpus and a normative reference corpus, and another one between different learner (sub)corpora.

The first type of comparison makes it possible to highlight features that are distinctive for learner language. This includes errors (which were the focus of pre-corpus interlanguage studies), but also cases of under- or overuse, i.e. the use of significantly fewer or more instances of a particular item as compared to the reference corpus. Particularly important in this type of comparison is the use of corpora that are as comparable as possible in terms of genre, subjects, etc. De Cock's (2002) analysis of pragmatic prefabs in several learner and native corpora suggests that a comparison between the *Louvain International Database of Spoken English Interlanguage* (LINDSEI) and a corpus of spontaneous conversations like (the demographic component of) the *British National Corpus* would point to an overuse of *and then* in the learner corpus. Yet, a comparison with the *Louvain Corpus of Native English Conversation* (LOCNEC), which is made up of informal interviews like LINDSEI, shows that learners actually underuse *and then*. Similarly, using data produced by expert or novice writers as a control corpus may have an impact on the results of the comparison. It should also be emphasized that the reference for comparison need not necessarily be native in the strict sense (e.g. it could represent a new developing variety of World Englishes) nor need it be unique/monolithic (cf. Mukherjee 2005) – the most important being that this corpus-based reference is made explicit.

The second type of comparison involved in CIA is between (varieties of) interlanguages. Most commonly, this comparison is made between several learner populations with different L1s, which helps identify the possible source of certain non-standard features. Features that are limited to one L1-group (or several groups with L1s from the same language family) are

likely to be due to transfer from the mother tongue (interlingual features). On the other hand, features that are shared by a large number of L1-groups are more probably linked to inherent difficulties in acquiring the target language (developmental features). This type of comparison is crucial to avoid attributing to transfer problems that are common to learners from different L1s (cf. Section 1.2, Osborne 2008). Here again, it is important to select comparable corpora for the analysis. Multi-L1 learner corpora, which contain data produced by learners from different L1s, are usually good resources for this, as they contain subcorpora collected according to the same design criteria. However, even these may be open to several interpretations. Ädel (2008) shows that differences in the use of involvement features in argumentative essays from ICLE, previously attributed to the learners' different mother tongue backgrounds, may in fact result from differences in the composition of the L1-subcorpora, and more precisely the proportion of timed vs. untimed essays. This pleads in favour of going beyond the L1 investigation and taking account of other variables that have largely been neglected in LCR up to now. In ICLE no less than 21 variables are recorded per learner text (e.g. learner's age, gender, knowledge of other languages, access to reference tools), all of which can be used to compile tailor-made subsets of data. One variable that cannot accurately be controlled for in ICLE is proficiency: it is imperfectly measured in terms of number of years of English at university – a rough estimation of the learners' exact proficiency levels. Other learner corpora rely on more objective measures, like a standardized test, to establish proficiency (e.g. the *NICT JLE Corpus*). Taking such variables into consideration also opens up new ways of implementing the second type of comparison in the CIA model. Instead of comparing learners with different mother tongues, one could compare learners with different proficiency levels, learners with different degrees of exposure to the target language, learners expressing themselves in speech vs. writing, or indeed different individual learners (see Section 1.4).

INTEGRATED CONTRASTIVE MODEL

While, as noted above, comparisons between different L1 populations help make more reliable claims about transfer, learner corpus researchers have sought to develop even more rigorous methods to identify transfer, given its central place in SLA. One such method is the integrated contrastive model (see Granger 1996, Gilquin 2000/2001), which combines CIA with a (corpus-based) contrastive analysis between the target language and the learner's mother tongue. Lack of reliable contrastive data and analysis has been shown to sometimes lead to misinterpretation of the results and misattribution of certain phenomena to transfer (see Kamimoto et al. 1992 on Schachter 1974). This risk can be limited by conducting a careful contrastive analysis on bilingual corpora, which allows the researcher to get a clear picture of the characteristics of the learner's L1 and the differences it presents with the target language. On the basis of such information, the researcher can predict possible occurrences of transfer or provide plausible explanations for certain characteristics of the interlanguage. This methodology is used by Demol & Hadermann (2008) to study discourse organisation in learner French and Dutch. By combining the learner corpus analysis with a contrastive analysis of native French and Dutch, they are able to show, among other things, that present participles in secondary predication are more common in French than in Dutch, which could account for Dutch learners' tendency to underuse this type of structure in French. (See also the other contributions in Gilquin et al. 2008).

COMPUTER-AIDED ERROR ANALYSIS

The methods described up to now do not specifically require that the learner corpora be annotated – even though annotation such as POS-tagging or parsing may enhance the possibilities of analysis (cf. Section 1.1). There is one method, however, that requires a

specific type of annotation, namely computer-aided error analysis (Dagneaux et al. 1998), which relies on error-tagged corpus data, i.e. data in which learners' errors have been identified, tagged and, usually, corrected (see Díaz-Negrillo & Fernández-Domínguez 2006 for a survey of error-tagging systems). This method is less popular than CIA, probably because it entails a number of problems at each of its stages. The stage of identification is very time-consuming because it is normally carried out manually (although there have been some attempts at automatic identification of errors). It involves making a distinction between real errors and infelicitous (but acceptable) forms. Different correctors may have different opinions on what counts as an error and it is therefore advisable to have several correctors and high inter-rater reliability for this stage of identification. Each identified error is then tagged, using a system of error tags which should be very thoroughly documented so as to ensure consistency in the assignation of the tags. The correction stage is problematic too, as it may not always be easy to reconstruct what the learner meant to say and several corrections may be equally plausible. In this respect, a multi-level error annotation system like that proposed by Lüdeling et al. (2005) may be useful as it allows for the multiple correction of errors. Despite these difficulties, computer-aided error analysis represents a major improvement over traditional error analysis, not least because instead of considering errors in isolation, it sees them as part of a system which includes both correct and incorrect uses.

FUTURE DIRECTIONS

The methods used in LCR have gradually become more sophisticated since the first learner corpus studies conducted over twenty years ago. Yet, a number of directions should be further pursued before LCR can be said to meet the methodological requirements that are expected of corpus research and empirical research in general. A case in point is the use of statistics in LCR. As opposed to earlier LCR studies that did not include any statistics, most current studies now follow the general trend in corpus linguistics by providing some sort of statistical analysis. Most of the time, however, it merely consists in testing the statistical significance of the results (e.g. the frequency of a word in a native and learner corpus). Often, this is done through the use of the well-known chi-square test – a test which, incidentally, may not even be adequate to compare word frequencies between corpora (see Bestgen 2013). More refined statistical techniques would be desirable, though, such as a multifactorial method, which allows the analyst to take into consideration the numerous factors that may affect learners' choices (cf. Gries & Deshors forthcoming). Another aspect that would benefit from further exploration is the use of multi-method approaches. In contrast to the methodological monism that tends to characterise current LCR, these makes it possible to consider learner language from different (but often complementary) perspectives. The exclusive reliance on corpus data, for instance, gives a very limited access to competence, since performance can sometimes be a very imperfect window on competence. A more experimental or elicitation-based approach may therefore be necessary to bring to light what learners really know (cf. Granger 1998 on prefabricated patterns or Gilquin 2007 on *make*-collocations). By triangulating findings generated by different methods including learner corpus-based methods, researchers may hope to gain new insights into the nature of interlanguage.

1.4. Advances and challenges

Since its emergence in the late 1980s LCR has established itself as a vibrant research strand at the crossroads between corpus linguistics, second language acquisition and foreign language teaching. The advances in data collection, methodology, linguistic analysis and applications

are impressive but each of these areas still has a number of challenges to overcome before the field can truly come into its own.

DATA

One of the main contributions of LCR, if not *the* main contribution, is the amount and variety of learner data it has made available to the community of researchers interested in L2 acquisition for theoretical or applied purposes. The empirical basis on which researchers can now rely, especially for writing, is more reliable than previous data collections which, in the eyes of SLA specialists themselves, suffered from a lack of representativeness. At first limited to English, the coverage in terms of L2s now includes a large number of different languages. On the downside, however, is the fact that very few learner corpora contain truly longitudinal data, with the same learners followed for an extended period of time. This dearth of longitudinal data, which is also deplored by mainstream SLA researchers (Ortega & Byrnes 2008), has begun to be addressed by the LCR community, as evidenced by some recent corpus initiatives like the LONGDALE project or the *Corpus of Advanced Learner Finnish*. Another area in which efforts should be directed is that of spoken corpora. Besides compiling new and larger spoken learner corpora with a view to countering the dominance of writing, it is time to make full use of the existing ones, i.e. analyse the sound files (when they are available) instead of relying solely on the transcripts. A last point concerns annotation. POS-tagging, though undoubtedly useful, is a very crude way to access syntax. Parsing of learner corpora is still extremely rare. This is a pity as preliminary studies that have applied shallow parsing to learner corpora (e.g. Nagata et al. 2011) have had encouraging results.

METHODOLOGY

Contrastive interlanguage analysis (CIA) and computer-aided error analysis (CEA) have been used successfully to analyse a wide range of linguistic phenomena and have contributed to a number of important debates in language acquisition. CIA studies raised the issue of the norm (native vs. non-native; novice vs. expert) and pointed to the benefit of relying on an explicit corpus-based norm rather than the implicit and intuitive norm that underlies many SLA studies. CEA, on the other hand, was the opportunity to ponder on the notion of error and introduce a higher degree of standardization at each level of the error analysis process: from error identification to error interpretation through error annotation and counting methods. Another particularly positive development is the integrated contrastive model which establishes a close link between learner corpus studies and contrastive studies, thereby paving the way for more rigorous investigations of transfer. One of the major methodological weaknesses of LCR to date is the exclusive use of aggregate data. While there are undeniable benefits to be gained from pooling data from a large number of learners, the reliability of the results is not guaranteed if possible differences between individual learners are disregarded. This methodological weakness can be addressed by the use of statistical techniques like the analysis of variance (ANOVA), which compares the between-group variation to the within-group variation. The combination of learner corpus data and experimental data is another area in need of further exploration.

LINGUISTIC ANALYSIS

Learner corpus research has put in evidence an aspect of interlanguage which had hitherto been intractable, i.e. frequency of use. It has amply demonstrated that interlanguages are characterized by patterns of under- and overuse which are as distinctive as, if not more than, downright errors. The shift of focus from morphosyntax to lexis and discourse has proved to be particularly fruitful for the analysis of advanced interlanguage. A number of features related to lexical expansion and genre diversification have been identified and found to be

systematic across different L1 populations (Cobb 2003). A question that still remains largely open, however, is how these features develop across time. To answer this question we need more longitudinal studies covering the full interlanguage continuum – from beginner to advanced – to complement the cross-sectional studies that have been the backbone of LCR to date. The corpus-based study of interlanguage should also be expanded to include less computationally tractable phenomena, including grammatical aspects such as tense and aspect or the complexity of NPs or VPs, as well as phonetic aspects like pronunciation and prosody – two types of aspects that have received relatively little attention in LCR up to now. Another criticism that can be levelled at LCR is that it has so far been relatively weak on interpretation. Myles (2005: 380), for example, deplores that learner corpus studies "often remain rather descriptive, documenting differences between learner and native language rather than attempting to explain them" and that studies "are also often not sufficiently informed by SLA theory". The few SLA-informed studies relying on learner corpora have shown that these provide a solid empirical basis to revisit some of the main findings of SLA research that were established on limited (often non-electronic) samples of language and/or experimental data. The influence of the full range of factors that affect learners' behaviour should also be investigated. To date, the influence of the L1 has often been discussed, although there has been a lack of baseline L1 corpora to confirm the presence of transfer. However, other factors have been largely neglected. The collection of a wide variety of metadata, together with the learner data themselves, should therefore be encouraged, and so should the inclusion of these metadata in the analysis, for example in the form of a multifactorial approach that seeks to assess the relative effect of each of a number of variables.

APPLICATIONS

Potential applications of LCR are extremely varied: they include materials and syllabus design, language testing, lexicography, data-driven learning, as well as a number of natural language processing (NLP) applications. One of the earliest applications is the integration of error warnings into monolingual learners' dictionaries (Gillard & Gadsby 1998). Language testing was also prompt to use learner corpus data for development and validation purposes (Hawkey & Barker 2004). Automatic error detection, which had been identified as a promising application very early on (Granger & Meunier 1994), is making increased use of learner corpus data to develop more efficient algorithms (Leacock et al. forthcoming). Other NLP applications involve automated scoring (Briscoe et al. 2010) and automatic L1 identification (Jarvis & Paquot forthcoming). Sadly, however, progress has been very slow, if at all visible, in the area where advances had been most ardently expected by early learner corpus researchers, that of textbooks and other pedagogical tools, in particular grammars. One reason for this is that the LCR field is still very young and that one should not expect too much too soon. That said, the field has already generated a wealth of relevant findings, but they are scattered and therefore not easy to integrate into pedagogical resources. There is a pressing need for a synthesis of current findings, which could result in a learner corpus-informed follow-up to Swan & Smith's (1987) intuition-based volume on the characteristic difficulties of learners from several mother tongue backgrounds.

2. Empirical study: two-word discourse markers in native and non-native speech

Some of the aspects discussed above will now be illustrated by means of a study of two-word discourse markers (DMs) in native and non-native speech. Through this study, we were interested in finding out how learners of English from different mother tongue backgrounds use DMs, and how their use compares to that of native speakers (quantitatively and

qualitatively), but we also wanted to demonstrate that when doing learner corpus research one should consider individual data in addition to pooled data. The role of DMs in efficient communication has often been underlined, both in native speech (e.g. Erman 1987) and non-native speech (e.g. Hasselgren 2002), and the problems foreign learners encounter when using (or not using) them have been brought to light repeatedly, often on the basis of corpora. However, studies of DMs in LCR tend to be limited in two ways: first they are usually restricted to one learner population (e.g. German learners in Müller 2005, French-speaking learners in Gilquin 2008, Swedish learners in Aijmer 2011) and second they adopt a global approach that does not take account of individual variation (Götz 2013 is a major exception to this). Our study departs from these in that it examines data produced by learners from eleven L1s and considers them both as groups and as individual learners. The investigation of several learner populations (rather than just one) makes it possible to determine the influence of the mother tongue on the use of DMs. As for the analysis at the level of individual speakers, it follows a recent trend in corpus linguistics that recognizes that "corpora are inherently variable internally" (Gries 2006: 110) and that comparing "native and non-native corpora as wholes (…) runs the risk of disguising differences between individual texts, and may therefore potentially produce misleading results" (Durrant & Schmitt 2009: 168).

Two corpora lie at the basis of our study, namely LINDSEI, which is made up of informal interviews with higher intermediate to advanced foreign learners of English (Gilquin et al. 2010), and LOCNEC, which is an exact replica of LINDSEI but with native speakers of English (De Cock 2004). In total, twelve L1 groups are represented: Bulgarian, Chinese, Dutch, French, German, Greek, Italian, Japanese, Polish, Spanish and Swedish (each corresponding to a subcorpus of LINDSEI consisting of some 50 interviews), as well as British English (corresponding to the 50 interviews of LOCNEC). We applied a corpus-driven method to select the DMs. After extracting all bigrams from LINDSEI and LOCNEC, we selected those bigrams occurring with a minimum frequency of 200 occurrences in at least one of the L1 subcorpora. We then identified the DMs among these, for a total of six different DMs, viz. *and so, and then, I mean, in fact, sort of* and *you know*. Finally, we manually excluded the concordance lines where the bigram did not behave as a syntactically non-obligatory string, keeping a sentence like *You can't start doing things like that* ***you know*** but excluding one like ***you know*** *everything about them*. The data were analysed by means of a contrastive interlanguage analysis, of which we exploited the two types of comparison: a comparison between learners and native speakers and a within-learner comparison (comparison between different learner populations and between different learners).

Considering the corpora as aggregates reveals a general underuse of DMs among learners: *you know*, *sort of*, *I mean* and *and then* are significantly more frequent in LOCNEC than in LINDSEI, while the frequency of *and so* is not significantly different between the two corpora; *in fact* is the only exception, being significantly overused by the learners (see Table 1).

	LOCNEC	LINDSEI
you know	514.9	266.8
sort of	487.2	56.1
I mean	372.7	159.2
and then	291.3	184.3
and so	65.9	79.4
in fact	5.2	55.3

Table 1. Relative frequency per 100,000 words of DMs in LOCNEC and LINDSEI

There is however a great deal of variation between the subcorpora, depending on the learners' L1. For instance, *sort of* (Figure 1) is rarely used among most learner groups, but the Swedish (SW) group displays a very high frequency which brings it close to the level of the native speakers (EN). On the other hand, a DM like *in fact* (Figure 2) shows three main peaks of frequency for French (FR), Bulgarian (BU) and Italian (IT) learners, whereas all the other populations (including the native speakers) display a rather low frequency.

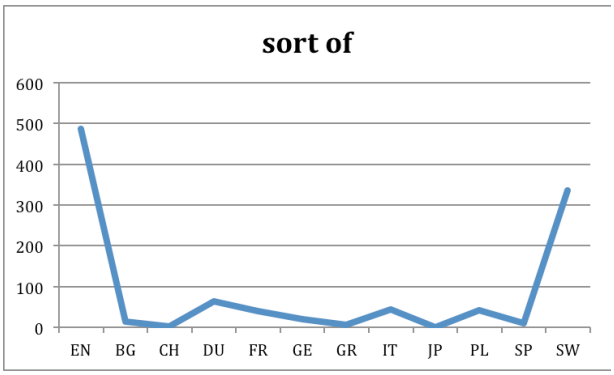

Figure 1. Relative frequency per 100,000 words of *sort of* in the different subcorpora

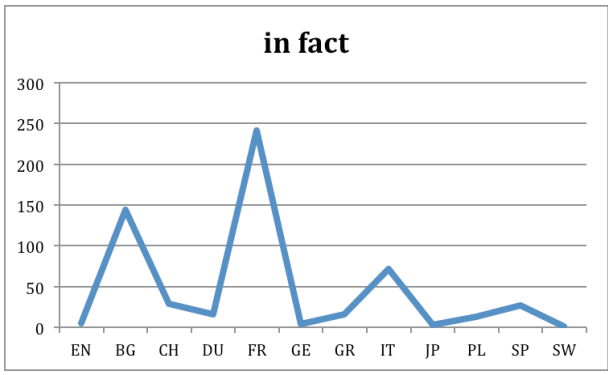

Figure 2. Relative frequency per 100,000 words of *in fact* in the different subcorpora

An ANOVA test reveals the presence of interesting clusters of L1 populations for each of the six DMs, which demonstrates the relevance of the L1 variable. It should be emphasized, however, that the variation is not necessarily a direct reflection of the L1 influence. The overuse of *in fact* among the French-speaking learners, for example, is very probably a consequence of the frequent use of the equivalent French DM *en fait*, but the high frequency of *sort of* among the Swedish learners may simply be due to these learners' high proficiency in speaking skills.

At the last level of analysis, that of the individual interviews, we also found important variation, in the form of speakers' idiolectal preferences for certain DMs. Interestingly, this variability also appears in the native corpus, where it is actually the most marked of all (sub)corpora. While some native speakers have a strong preference for *sort of*, for instance, others clearly prefer *you know*. The same is true of the learners, who may have strong preferences for certain DMs. The main difference between the two groups, however, is that the learners' preferences tend to be exclusive: they show a preference for one or two DMs to the exclusion of the others, which they do not (or at least hardly ever) use; the native speakers, by contrast, may display certain preferences but usually still use the whole range of DMs (except for *and so* and *in fact*, which are less common in LOCNEC). This corpus-internal variability should be taken into account when doing LCR. More generally, one should bear in mind that the native norm is not unique, nor is the learner behaviour.

It may be tempting, in corpus linguistics in general and in LCR in particular, to limit the analysis to a quantitative approach. While quantitative results provide interesting insights into certain aspects of interlanguage, they may also hide some qualitative trends that are worthy of interest. In the case of two-word DMs, we wanted to go beyond the simple equation between high frequency of DMs and fluency (as established, e.g., in Hasselgren 2002) to see how native-like learners' use of DMs was. With this aim in mind, we selected two groups of learners whose use of *you know* showed different tendencies, namely the French-speaking learners who use relatively few occurrences of *you know* (228 per 100,000 words) and the

Polish learners who use many occurrences of it (715 per 100,000 words), and we compared these two populations with the native group (relative frequency of *you know*: 515 per 100,000 words). By making a distinction between cases where *you know* does not interrupt any structure (e.g. *it's just like a curtain **you know** so you've gotta change it*) and cases where it interrupts phrases or closely-knit structures (e.g. *if we **you know** make some= something legal*), we noticed that the French-speaking learners' behaviour was closer to the native norm than that of the Polish learners. While the French-speaking learners use *you know* in an interrupted structure in 37% of the cases (to be compared with 35% in LOCNEC), the Polish learners interrupt closely-knit structures in 60% of the cases. In other words, it turns out that the French-speaking learners underuse *you know* but when they do use it, the breakdown of interrupted vs. non-interrupted structures is very similar to that found among the native speakers. On the other hand, the Polish learners use *you know* very often (more, in fact, than the native speakers) but they tend to use it within strings that should in principle have been uttered as uninterrupted chunks. Especially striking are the interruptions between a copula (usually *be*) and the nominal part of the predicate (as in *he stopped being **you know** humorous*) and the interruptions between a preposition and the rest of the prepositional phrase (as in *that was some guy from **you know** the upper classes*). Such uses, rather than contributing to the fluency of the Polish learners' speech, mark it as particularly disfluent and non-native-like, contrary to what the quantitative analysis would have suggested.

While both the quantitative and qualitative analyses should be expanded to provide a full picture of the use of two-word DMs by foreign learners of English, this study shows the potential of LCR to bring to light linguistic features typical of certain (groups of) learners. The exploitation of a native corpus, used in combination with the learner corpora, makes it possible to see how the learner data are situated in relation to a certain reference norm (without being limited by it), whereas the inclusion of the L1 variable gives a glimpse of the possible influence of the mother tongue. This 'classic' application of the CIA model is furthermore refined by also adopting a more individual approach, which considers the speakers individually and underlines the idiolectal variation that can be found within corpora. Like many other learner corpus-based studies, however, this one does not examine the wide range of variables that are available in learner corpora like LINDSEI and that may have an effect on the learners' use of DMs (such as the time spent in an English-speaking country or the knowledge of other languages), nor does it consider the general context in which the learners acquire English (e.g. input-rich vs. input-poor environment). It also fails to fully exploit the spoken corpus at hand, since we did not use the audio files, even though these would have been useful for example to disambiguate between the DM and non-DM uses of the six bigrams under study. In addition, since we only worked on the basis of the (transcribed) utterances produced by the interviewees, without looking at the interviewers' turns, we could not adopt a more pragmatic perspective, which would have consisted in investigating how interaction is created between speakers by means of DMs, nor a sociolinguistic perspective, which could for instance have studied whether the interviewer's profile (e.g. male or female, native or non-native speaker of English) may have an impact or not on the presence of DMs in the learner's language. Finally, this study is mainly descriptive and does not seek to establish links with two fields which can be seen as close (and complementary) to LCR, namely second language acquisition (SLA) and foreign language teaching (FLT). An SLA approach would for example favour a more interpretative perspective and a more pronounced focus on competence than is the case with learner corpora (which may give an indication of what learners know or fail to know about the target language but strictly speaking only provide access to performance). An FLT approach would look into the ways in which the corpus findings can be applied for pedagogical purposes; in

the case of our study, this could start with the question of whether (and how) DMs should be taught in the classroom. A combined LCR-SLA-FLT approach would thus result in an integrated descriptive/interpretative/applied perspective which is still largely lacking in the field of interlanguage studies.

References

Ädel, A. (2008). Involvement features in writing: Do time and interaction trump register awareness? In G. Gilquin, S. Papp & M. B. Díez-Bedmar (eds.) *Linking up Contrastive and Learner Corpus Research* (pp. 35-53). Amsterdam & Atlanta: Rodopi.*Metadiscourse in L1 and L2 English*. Studies in Corpus Linguistics 24. Amsterdam & Philadelphia: John Benjamins.

Aijmer, K. (2002). Modality in advanced Swedish learners' written interlanguage. In S. Granger, J. Hung & S. Petch-Tyson (eds.) *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching* (pp. 55-76). Amsterdam & Philadelphia: Benjamins.

Aijmer, K. (2011). *Well I'm not sure I think…* The use of *well* by non-native speakers. *International Journal of Corpus Linguistics* 16(2): 231-254.

Anthony, L. (2004). AntConc: A learner and classroom friendly, multi-platform corpus analysis toolkit. *Proceedings of IWLeL 2004: An Interactive Workshop on Language e-Learning*, 7-13.

Asención-Delaney, U. & Collentine, J. (2011). A multidimensional analysis of a written L2 Spanish corpus. *Applied Linguistics* 32(3): 299-322.

Bestgen, Y. (2013). Inadequacy of the chi-squared test to examine vocabulary differences between corpora. *Literary and Linguistic Computing* (Advance access).

Borin, L. & Prütz, K. (2004). New wine in old skins? A corpus investigation of L1 syntactic transfer in learner language. In G. Aston, S. Bernardini & D. Stewart (eds.) *Corpora and Language Learners* (pp. 67-87). Amsterdam & Philadelphia: John Benjamins.

Briscoe, T., Medlock, B. & Andersen, Ø. (2010). Automated assessment of ESOL free text examinations. Cambridge ESOL. Cambridge: University of Cambridge Computer Laboratory. http://www.cl.cam.ac.uk/techreports/UCAM-CL-TR-790.pdf.

Callies, M. (2009). 'What is even more alarming is...' - A contrastive learner-corpus study of *what*-clefts in advanced German and Polish L2 writing. In M. Wysocka (ed.) *On Language Structure, Acquisition and Teaching. Studies in honour of Janusz Arabski on the occasion of his 70th birthday* (pp. 283-292). Katowice: Wydawnictwo Uniwersyetu Slaskiego.

Chen, Y.-H. & Baker, P. (2010). Lexical bundles in L1 and L2 academic writing. *Language Learning & Technology* 14(2): 30-49.

Cobb, T. (2003). Analyzing late interlanguage with learner corpora: Québec replications of three European studies. *The Canadian Modern Language Review – La Revue canadienne des langues vivantes* 59(3): 393-423.

Council of Europe (2001). *Common European Framework of Reference for Languages*. Cambridge: Cambridge University Press.

Dagneaux, E., Denness, S. & Granger, S. (1998). Computer-aided error analysis. *System* 26: 163-174.

De Cock, S. (2002). Pragmatic prefabs in learners' dictionaries. In A. Braasch & C. Povlsen (eds.) *Proceedings of the Tenth EURALEX International Congress, EURALEX 2002, Copenhagen, Denmark, August 13-17 2002, Volume II* (pp. 471-481). Center for Sprogteknologi.

De Cock, S. (2004). Preferred sequences of words in NS and NNS speech. *Belgian Journal of English Language and Literatures (BELL)*, New Series 2: 225-246.

Demol, A. & Hadermann, P. (2008). An exploratory study of discourse organisation in French L1, Dutch L1, French L2 and Dutch L2 written narratives. In G. Gilquin, S. Papp & M. B. Díez-Bedmar (eds.) *Linking up Contrastive and Learner Corpus Research* (pp. 255-282). Amsterdam & Atlanta: Rodopi.

Díaz-Negrillo, A. & Fernández-Domínguez, J. (2006). Error tagging systems for learner corpora. *RESLA* 19: 83-102.

Díez-Bedmar, M. B. & Papp, S. (2008). The use of the English article system by Chinese and Spanish learners. In G. Gilquin, S. Papp & M. B. Díez-Bedmar (eds.) *Linking up Contrastive and Learner Corpus Research* (pp. 147-175). Amsterdam & Atlanta: Rodopi.

Durrant, P. & Schmitt, N. (2009). To what extent do native and non-native writers make use of collocations? *IRAL* 47: 157-177.

Ellis, R. (1994). *The Study of Second Language Acquisition.* Oxford: Oxford University Press.

Flowerdew, L. (2012). *Corpora and Language Education*. London: Palgrave Macmillan.

Erman, B. 1987. *Pragmatic Expressions in English. A study of* you know, you see *and* I mean *in face-to-face conversation.* Stockholm: Almqvist & Wiksell.

Gillard, P. & Gadsby, A. (1998). Using a learners' corpus in compiling ELT dictionaries. In S. Granger (ed.) *Learner English on Computer* (pp. 159-171). Addison Wesley Longman: London & New York.

Gilquin, G. (2000/2001). The Integrated Contrastive Model. Spicing up your data. *Languages in Contrast* 3(1): 95-123.

Gilquin, G. (2007). To err is not all. What corpus and elicitation can reveal about the use of collocations by learners. *Zeitschrift für Anglistik und Amerikanistik* 55(3): 273-291.

Gilquin, G. (2008). Hesitation markers among EFL learners: Pragmatic deficiency or difference? In J. Romero-Trillo (ed.) *Pragmatics and Corpus Linguistics: A Mutualistic Entente* (pp. 119-149). Berlin, Heidelberg & New York: Mouton de Gruyter.

Gilquin, G. (2012). Lexical infelicity in English causative constructions. Comparing native and learner collostructions. In J. Leino & R. von Waldenfels (eds.) *Analytical Causatives. From 'give' and 'come' to 'let' and 'make'* (pp. 41-63). München: Lincom Europa.

Gilquin, G., De Cock, S. & Granger, S. (2010). *Louvain International Database of Spoken English Interlanguage. Handbook and CD-ROM*. Louvain-la-Neuve: Presses universitaires de Louvain.

Gilquin, G., Papp, S. & Díez-Bedmar, M. B. (eds.) (2008). *Linking up Contrastive and Learner Corpus Research*. Amsterdam & Atlanta: Rodopi.

Götz, S. (2013). *Fluency in Native and Nonnative English Speech*. Amsterdam & Philadelphia: John Benjamins.

Granger, S. (1996). From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora. In K. Aijmer, B. Altenberg & M. Johansson (eds.) *Languages in Contrast. Papers from a Symposium on Text-based Cross-linguistic Studies* (pp. 37-51). Lund: Lund University Press.

Granger, S. (1998). Prefabricated patterns in advanced EFL writing: Collocations and formulae. In A. P. Cowie (ed.) *Phraseology. Theory, Analysis, and Applications* (pp. 145-160). Oxford: Oxford University Press.

Granger, S. (2003). Error-tagged learner corpora and CALL: A promising synergy. *CALICO* 20(3): 465-480.

Granger, S., Dagneaux, E., Meunier, F. & Paquot, M. (2009). *The International Corpus of Learner English. Handbook and CD-ROM. Version 2*. Louvain-la-Neuve: Presses universitaires de Louvain.

Granger, S. & Meunier, F. (1994). Towards a grammar checker for learners of English. In U. Fries & G. Tottie (eds.) *Creating and using English language corpora* (pp. 79-91). Rodopi: Amsterdam & Atlanta.

Granger, S. & Rayson, P. (1998). Automatic lexical profiling of learner texts. In S. Granger (ed.) *Learner English on Computer* (pp. 119-131). London & New York: Addison Wesley Longman.

Gries, S. Th. (2006). Exploring variability within and between corpora: Some methodological considerations. *Corpora* 1(2): 109-151.

Gries, S. Th. & Deshors, S. C. (Forthcoming). Using regressions to explore deviations between corpus data and a standard/target: Two suggestions. *Corpora*.

Hasselgård, H. (2009). Thematic choice and expressions of stance in English argumentative texts by Norwegian learners. In K. Aijmer (ed.) *Corpora and Language Teaching* (pp. 120-39). Amsterdam & Philadelphia: John Benjamins.

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

Hawkey, R. & Barker, F. (2004). Developing a common scale for the assessment of writing. *Assessing Writing* 9: 122-159.

Housen, A. (2002). A corpus-based study of the L2-acquisition of the English verb system. In S. Granger, J. Hung & S. Petch-Tyson (eds.) *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching* (pp. 77-116). Amsterdam & Philadelphia: John Benjamins.

Jarvis, S. & Paquot, M. (Forthcoming). Native language identification. In S. Granger, G. Gilquin & F. Meunier (eds.) *The Cambridge Handbook of Learner Corpus Research*. Cambridge: Cambridge University Press.

Kamimoto T., Shimura, A. & Kellerman, E. (1992). A second language classic reconsidered. The case of Schachter's avoidance. *Second Language Research* 8(3): 251-277.

Leacock, C., Chodorow, M. & Tetreault, J. (Forthcoming). Automatic spell- and grammar-checking. In S. Granger, G. Gilquin & F. Meunier (eds.) *The Cambridge Handbook of Learner Corpus Research*. Cambridge: Cambridge University Press.

Leńko-Szymańska, A. (2008). Non-native or non-expert? The use of connectors in native and foreign language learners' texts. *Acquisition et interaction en langue étrangère* 27: 91-108. http://aile.revues.org/4213.

Lüdeling, A., Walter, M., Kroymann, E. & Adolphs, P. (2005). Multi-level error annotation in learner corpora. In *Proceedings from the Corpus Linguistics Series* 1(1). http://www.birmingham.ac.uk/research/activity/corpus/publications/conference-archives/2005-conf-e-journal.aspx.

Mackey, A. & Gass, S. (eds.) (2012). *Research Methods in Second Language Acquisition: A Practical Guide.* Chichester: Wiley-Blackwell.

Mauranen, A. (2011). Learners and users – Who do we want corpus data from? In F. Meunier, S. De Cock, G. Gilquin & M. Paquot (eds.) *A Taste for Corpora: In honour of Sylviane Granger* (pp. 155-171). Amsterdam: John Benjamins.

Mukherjee, J. (2005). The native speaker is alive and kicking: Linguistic and language-pedagogical perspectives. *Anglistik* 16(2): 7-23.

Müller, S. (2005). *Discourse Markers in Native and Non-native English Discourse.* Amsterdam & Philadelphia: John Benjamins.

Myles, F. (2005). Interlanguage corpora and second language acquisition research. *Second Language Research* 21(4): 373-391.

Nagata, R., Whittaker, E. & Sheinman, V. (2011). Creating a manually error-tagged and shallow-parsed learner corpus. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, Portland, Oregon, June 19-24 2011* (pp. 1210-1219).

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

Nesselhauf, N. (2005). Collocations in a Learner Corpus. Amsterdam & Philadelphia: John Benjamins.

Ortega, L. & Byrnes, H. (eds.) (2008). *The Longitudinal Study of Advanced L2 Capacities*. New York & London: Routledge.

Petch-Tyson, S. (2000). Demonstrative expressions in argumentative discourse. A computer-based comparison of non-native and native English. In S. Botley & A. M. McEnery (eds.) *Corpus-based and Computational Approaches to Discourse Anaphora* (pp. 43-64). Amsterdam & Philadelphia: John Benjamins.

Schachter, J. (1974). An error in error analysis. *Language Learning* 24(2): 205-214.

Scott, M. (2012). *WordSmith Tools version 6*. Liverpool: Lexical Analysis Software.

Sinclair, J. (1996). Preliminary recommendations on corpus typology, Technical report, EAGLES (Expert Advisory Group on Language Engineering Standards). http://www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html.

Swan, M. & Smith, B. (1987). *Learner English. A Teacher's Guide to Interference and Other Problems*. Cambridge: Cambridge University Press.

Szmrecsanyi, B. & Kortmann, B. (2011). Typological profiling: Learner Englishes versus indigenized L2 varieties of English. In J. Mukherjee & M. Hundt (eds.) *Exploring Second-Language Varieties of English and Learner Englishes: Bridging a Paradigm Gap* (pp. 167-187). Amsterdam & Philadelphia: John Benjamins.

Tono, Y. (2000). A computer learner corpus-based analysis of the acquisition order of English grammatical morphemes. In L. Burnard & T. McEnery (eds.) *Rethinking Language Pedagogy from a Corpus Perspective. Papers from the Third International Conference on Teaching and Language Corpora* (pp. 123-132). Frankfurt am Main: Peter Lang.

Wible, D., Kuo, C. H., Chien, F.-Y., Liu, A. & Tsao, N.-L. (2001). A web-based EFL writing environment: Integrating information for learners, teachers, and researchers. *Computers and Education* 37(3-4): 297-315.