An interview with Yukio Tono

In memoriam Geoffrey Leech

Yukio Tono is Professor of Corpus Linguistics at Tokyo University of Foreign Studies, Japan and a long-standing researcher in Learner Corpus Research. He holds a PhD in Corpus Linguistics from Lancaster University, UK, with a dissertation on “The Role of Learner Corpora in Second Language Acquisition Research and Foreign Language Learning: The Multiple Comparison Approach”. His research interests are in Corpus Linguistics (especially learner corpora and applications of corpora for English language teaching), second language vocabulary acquisition, and pedagogical lexicography. He has been involved in corpus projects such as the Japanese EFL Learner Corpus (JEFLL; http://jefll.corpuscobo.net/) and the International Corpus of Crosslinguistic Interlanguage (ICCI; http://cblle.tufs.ac.jp/llc/icci/)

IJLCR editors: Thank you very much, Prof. Yukio Tono, for agreeing to answer our questions in this interview for the inaugural issue of the International Journal of Learner Corpus Research. Learner English on Computer (Granger, ed. 1998) was the first book dedicated to Learner Corpus Research (LCR). In his preface to this landmark volume, Geoffrey Leech, one of the pioneers in corpus linguistics, already identified several key issues and future challenges for the field. In this interview, we would like to take stock of LCR and reconsider some of Leech’s comments with you to discuss how LCR has developed vis-a-vis the neighbouring fields of Corpus Linguistics and Second Language Acquisition (SLA). Geoff sadly passed away on 19 August 2014. We would like to dedicate the interview to his memory, thereby paying tribute to his immense contribution to Corpus Linguistics.

In his preface Leech wrote that “[l]ike any healthily active and developing field of inquiry, learner corpus research has to continue to face challenges both material and intellectual before it wins a secure and accepted place in the discipline of applied linguistics” (Leech 1998: xx). How do you judge the development of LCR as a discipline since the publication of that preface? Has it already gained its accepted place in applied linguistics?
Yukio Tono: LCR is buzzing with activity: the field now has its own association (Learner Corpus Association, http://www.learnercorpusassociation.org/), its biannual conference (i.e. the Learner Corpus Research series) and, with the publication of IJLCR, its own scientific journal. In my view, however, LCR has become an important subdiscipline of Corpus Linguistics, but it has not yet gained widespread recognition in the general field of Applied Linguistics. In Second Language Acquisition, for example, learner corpora have only recently started to be recognized as one data type among many others and the impact of LCR (and perhaps more particularly of its methods and tools) is yet to be seen (see Question 2).

IJLCR editors: LCR is sometimes — probably in a simplistic way — conceived of as ‘SLA meets corpus linguistics’, which, however, “is not likely to be a meeting of unalloyed joy and goodwill. Rather, it may well be an encounter marked by some suspicion and non-understanding” (Leech 1998:xvi). And indeed, it seems that some mutual theoretical and methodological scepticism still poses challenges to a firmer establishment of LCR as a methodological key player in SLA research. What is your position on this issue? How can LCR form closer links and gain more recognition in the SLA camp?

Yukio Tono: One of the ironies of LCR is that even though Sylviane Granger proposed LCR as a new paradigm to investigate SLA in the early 1990s, most learner corpus-based studies to date have not been met with great enthusiasm from SLA researchers. There are several explanations for this. First, SLA researchers are usually interested in testing specific hypotheses within a particular theoretical framework such as Universal Grammar (UG) and the large majority of learner corpora that have been compiled so far represent too open-ended types of tasks for this purpose. To investigate phenomena such as the relationship between verb raising and subject-verb agreement (Lardiere, 1998) or the preference of full clauses over small clause complements (Chen 2005), for example, it is often necessary to use more experimental data collection techniques designed to elicit particular linguistic features. Recent publications such as Dominguez et al. (2013) and Tracey-Ventura & Myles (this issue), however, show that it is possible to combine the respective advantages of learner corpus data and more experimental data and compile “peripheral learner corpora” that include more constrained data types (Nesselhauf 2005:128). Secondly, many SLA researchers, especially those in UG-based SLA research, want to investigate the optionality of linguistic competence. In other words, they are interested in whether L2 learners allow certain linguistic constructions as acceptable in their interim grammar. To do this, SLA researchers usually focus on a series of sentences containing a target construction and rely on grammaticality/acceptability judgement tasks. They are not interested in whether learners actually use the construction. The primary concern is whether learners
have the knowledge of what is possible and what is not possible. Learner corpora are evidence of how learners actually perform, but corpus data does not show the entire picture of what is possible in learners’ interlanguage.

Thirdly, there is the issue of the nature of learner corpus data. The original idea behind the *International Corpus of Learner English* (ICLE; Granger et al. 2002), for example, was to compile a corpus of ‘learner English’ that could be compared with other varieties of English in the framework of the *International Corpus of English* (ICE; Greenbaum 1996). The concept of ‘learner English’, however, is not easy to define. The ICLE team decided to collect data from higher-intermediate to advanced learners majoring in English at university, but this focus on higher proficiency levels made the ICLE too narrow in scope to allow for SLA-oriented studies of L2 developmental aspects. Another issue relates to the type of learner data collected: argumentative essays were probably a good type of data to compare against other varieties of English, but they did not fit into the mainstream of SLA research, whose interest was on spontaneous spoken utterances as a reflection of L2 competence.

The compilation of a learner corpus as a representative sample of a particular L2 learner group nevertheless remains a meaningful enterprise. Back in the late 1960s and the early 1970s, many ‘applied’ linguists became interested in the notion of ‘interlanguage’. If electronic learner corpora had been available then, SLA researchers might have seized the opportunity to produce detailed descriptions of learners’ interlanguages. Unfortunately, however, they quickly shifted their attention towards the “explanation” of the acquisition process and did not spend much time on its description. This is partly due to the influence of the claim made by Chomsky that explanatory adequacy is superior to descriptive adequacy when it comes to develop a theory of grammar (1964: 24). I would argue that a thorough description of L2 learners’ interlanguage systems across proficiency levels must go hand in hand with an explanation of the transition between the different levels.

**IJLRC editors**: In a recent paper in *Language Learning*, Lourdes Ortega has appraised SLA’s disciplinary progress over the last 15 years and reflects on transdisciplinary relevance as the field has completed 40 years of existence. She advocates a disciplinary reorientation in SLA, arguing that “[p]erhaps a more profound and complete reframing for SLA in the 21st century is needed, one that targets a deeper renewal of disciplinary goals and harnesses the unique benefits of studying bi/multilingualism and late learning, all while making transdisciplinary connections of relevance” (2013: 16). What could be the role of LCR in this reframing of SLA?

**Yukio Tono**: In my opinion, one of the key methodological breakthroughs in SLA in the next few decades will be the use of big data. As exemplified in Alexopoulou et al. (this issue), large datasets of learner performance can be exploited by using
various natural language processing and statistical methods. Educational big data can be collected from a variety of sources including testing companies, conversation schools, e-learning systems, and massive open online courses. LCR has a major role to play in this development in terms of data collection tools and techniques. To be useful for SLA modelling, educational big data should be accompanied with the type of detailed learner profiles that are being used in LCR. This being said, SLA specialists have always been interested in the socio-cultural aspects of second language acquisition. I think that learner corpora would appeal much more to those researchers if they included more context-rich learner profiles describing, for example, the type of L2 input received inside and outside the classroom (textbooks used, teaching practices, classroom interaction, etc.).

I also think that future SLA research will show increasing interest in a wider range of genres and text types. Corpus linguistic research has shown that genres and text types are key factors influencing language use (e.g. Biber 1988) and these findings have already been taken on board in language assessment and testing. In order to attain C-levels, for example, the Common European Framework of Reference for Languages (CEFR; Council of Europe 2001) states that language learners need to know the particular use of lexical and grammatical items in academic or genre-specific contexts. These factors should become more prominent in SLA research (especially when the focus is on advanced learners) and I believe LCR could make a valuable contribution in this area.

**IJLCR editors:** LCR studies often focus on the variable of the learners’ first language and its potential effects on their learning of a second/foreign language. Leech wrote that to him it appeared odd that SLA research had not yet found clear answers to questions concerning the influence of the L1 and that “the study of learner corpora for the first time provides for a research programme which will lead to its being answered” (Leech 1998: xv). And clearly L1 influence has been the focus of much LCR. But Leech also warned us that “it is conceivable that particular observed effects may be due to other factors, such as differences in cultural context, or in educational institutions and practices. The conclusion is that we have to be cautious in drawing general inferences from the results of a corpus, and alert to the influence of hidden variables implicit in the way we collected or sampled the data” (Leech 1998: xix). How do you value the achievements that LCR has made in the study of transfer? Do you think that the field has taken Leech’s warnings about other potential influential factors seriously enough?

**Yukio Tono:** Transfer phenomena are quite complex. As the history of SLA research clearly shows, identifying the causes of errors is often tricky, and the way errors are analyzed often depends on the theoretical framework adopted for the study. Dulay, Burt and Krashen (1982), for instance, believed in the active role of
the Language Acquisition Device (LAD) and as a result, minimized the role of L1 transfer in their study of grammatical morphemes as developmental features. Several recent studies, however, have emphasized the active role of cross-linguistic influence and reinterpreted what used to be treated as developmental features as L1-induced phenomena (see Jarvis & Pavlenko 2010: 6).

In 1996, Granger published a seminal article in which she put forward Contrastive Interlanguage Analysis (CIA). This model is important in that it relies explicitly on a comparison of different interlanguages of the same language (e.g. the English of French learners, Spanish learners, Chinese learners, etc.) to “differentiate between features which are shared by several learner populations and are therefore likely to be developmental and those which are peculiar to one national group and therefore possibly L1-dependent” (Granger 2002: 13). It is however unfortunate that CIA has sometimes been used rather carelessly without considering the many other factors involved in foreign language learning and teaching. Many studies have reported patterns of L1 influence without controlling for other variables that may have distinguished the different L1 learner populations under study (e.g. proficiency levels, teaching practices). Similarly, studies have often not taken into account the contextual factors that may influence the use of a specific linguistic feature. In my PhD (Tono 2002), I investigated the acquisition of verb argument structures in L2 English and compared the use of argument structures for ten lexical verbs. I specified the factors affecting the use of argument structures such as L1 effects (frequency of similar/different argument structure properties in L1 corpora), L2 input factors (frequency of subcategorization patterns in ELT textbook corpora) and L2 internal factors (frequency of different verb classes and alternations defined by COMLEX Lexicon and Levin 1993’s list), as well as developmental factors (frequency of use/misuse of argument structures from the learner corpus), thus explaining the use of verb argument structures as a complex interplay of these competing factors using log-linear analysis. Recently a similar approach has been proposed by Gries (2013), who employed binary logistic regression to compare the use of of- and s-genitives by L2 learners and native speakers. This type of approach has the advantage of examining L1 effects in conjunction with other related variables.

IJLCR editors: Leech projected that “through the coding of speaker/writer identity in the corpus header information, the productions of individuals can be retrieved from a corpus and compared. As corpora are enlarged in the future, it is likely that this aspect will be developed further, particularly in a longitudinal dimension” (1998: xviii–xix). Until very recently, however, learner corpora were typically used to abstract away from individual learners to provide a corpus-based description of a learner group (e.g. Chinese learners of English, advanced learners). In line with
standard practice in Corpus Linguistics at large, researchers usually aggregate and analyse data drawn from all learners that meet a chosen selection criterion such as first language or global L2 proficiency level. Current research trends seem to indicate that more attention is now devoted to inter-learner variability. Why do you think Leech’s projection took fifteen years or so to start being realised? What is your judgment on longitudinal learner studies in LCR?

**Yukio Tono:** Personally, I do not see any problem in looking at learners as a group as long as the groups are clearly defined and comparable and the analysis takes into account all the variables that may affect learner production. As Granger (1998:4) commented, previous SLA research has tended to rely on small samples of elicited data, and the importance of looking at the acquisition process by analysing a large quantity of natural language usage data was neglected until learner corpora became available. So I think it is still meaningful to look at learner corpora as aggregate datasets and analyse some recurrent patterns across a learner population in terms of overuse, underuse, or misuse. More serious problems lie in the way some researchers treat learner corpus data. Sinclair (2004)’s third guiding principle for corpus building states that “only those components of corpora which have been designed to be independently contrastive should be contrasted.” Despite this warning, however, it has recently been suggested that the rich learner profiles available in learner corpora such as the ICLE should be used to compile sub-corpora representing other learner variables than L1 background (e.g. learners’ stay abroad, particular essay topics). Unlike the criterion of L1 background, however, these variables were often not originally part of the learner corpus design (i.e. there was no attempt to compile a balanced corpus as regards those variables) but were recorded as part of the learner profiles. Therefore, as Sinclair suggested, the size of the sub-corpora that could potentially be created on the basis of these variables will often be too small or skewed to make any meaningful generalization.

Up to now, we have also tended to focus on patterns of use common to specific learner populations because that is what corpus linguistics has been focusing on in general. If we take a general corpus like the *British National Corpus* (BNC), 4,000 independent files could show individual variation across writers/speakers in a particular genre or register. However, very few corpus-based studies have so far focused on individual differences, most probably because we believe that individual differences will be minimized as we accumulate more data and capture some general pictures of patterns of use. This is ultimately the strength of corpus-based analyses.

Whilst the need for more longitudinal data in SLA has often been mentioned, it is difficult to collect data from the same learners over a long period of time. Most learner corpora built in the last two decades have tended to focus on advanced
learners only, probably due to the ease of access to data from such learner groups. I have stressed the importance of developmental patterns from the beginning of my LCR work, and built corpora covering a range of proficiency levels (e.g. JEFLL, ICCI), although the design was still cross-sectional. Despite the difficulties, it is now time for LCR to compile truly longitudinal learner corpora as a priority and focus on the process of learning/acquisition. Recent learner corpus compilation projects (e.g. Longitudinal Database of Learner English, http://www.uclouvain.be/en-cecl-longdale.html, Meunier & Littré, 2013) seem to show that the field is shifting in this direction.

**IJLRCR editors:** Recently, LCR has been criticised for paying too much attention to errors and deficiencies rather than being interested in what learners do right from a communication and learning perspective, and how they manage to improve. Guy Aston suggests that “such interest has been missing from much work on learner corpora. Those focusing on error have paid too little attention to success — what are the things that learners get right, or that allow them to solve their communication problems even while making mistakes? Those who turn their attention to frequency of use have rarely asked why learners underuse/overuse particular features, and whether these tendencies may not result from the particular contexts in which they find themselves as learners” (Aston 2011: 11). Do you agree with this criticism and if so, what are ways out of this dilemma?

**Yukio Tono:** I partly agree with the criticism, but it is perfectly natural to focus on errors in learner language because that is the most noticeable gap between learners and native speakers. Error analysis (EA) that originally emerged in the 1960s and 1970s did not deal with the frequencies of errors very well, so today’s renewed interest in error frequencies seems reasonable to me. Having said that, we are now witnessing a paradigm shift from mere error analysis to the analysis of performance in its entirety, which is exactly what happened after severe criticism of EA in the past. So we are on the right track. Now there is a new thread of research called ‘L2 learner profiling’, in which researchers try to profile what learners can do with language at each proficiency level as specified by the CEFR on the basis of learner corpora (Hawkins and Filipovic 2012). The focus is on both positive and negative linguistic properties as well as usage distributions, so the analysis covers pretty much everything.

Aston’s question about “why” is closely related to the context of use that I mentioned earlier. Maybe we should incorporate contextual factors, such as the amount of input from the textbook, the amount of exposure in the classroom interactions, the amount of support learners can get from their teacher or peers, which affect the underuse, overuse, and misuse of particular linguistic features and build a model explaining the cause-effect relationship. Some factors external
to linguistic issues such as motivation or anxiety cannot be dealt with straightforwardly in a corpus study, but these factors could be researched by combining different data elicitation methods (e.g. diaries, questionnaires or interviews) alongside learner corpora.

**IJLCR editors:** Much LCR literature is largely descriptive. Despite its assumed pedagogical basis, there is yet comparatively little in terms of practical use of learner corpora for teaching and learning purposes. However, learner corpora are at the core of recent NLP-oriented projects on automatic error detection and correction, automated scoring, native language identification, etc. Do you agree with this statement and if so, why do you think pedagogical material developers/publishers do not make more use of learner corpus data? Do you think there is a role for learner corpus data in the classroom?

**Yukio Tono:** Compared to the recent upsurge in popularity of learner corpora in the area of automated scoring in language testing, for example, the pedagogical applications of learner corpora seem to be underdeveloped. So far, insights from learner corpus data have mainly been used to provide usage information about common learner errors in monolingual learners’ dictionaries (e.g. the *Longman Essential Activator*, *Longman Dictionary of Contemporary English*, *Macmillan English Dictionary*, *Cambridge Learner’s Dictionary*, and *Cambridge Advanced Learner’s Dictionary*). I have not seen any learner corpus-informed reference grammar yet, but this is a very promising direction. One difficulty is that, in order to provide information in a dictionary or a reference grammar, we need to examine every grammar item exhaustively in terms of learner performance, which is particularly time-consuming. Perhaps a more serious problem is that there is great variability in learners’ problems (depending on proficiency levels, L1 background and L1 influence) and that it might be particularly challenging to write a grammar appropriate for all learners (which is what a publishing house would require for obvious marketing reasons). Moreover, the data at hand sometimes do not seem big enough for such purposes. Only a sufficiently large learner corpus such as the *Cambridge Learner Corpus* can do the job.

Learner corpus researchers should actively promote the use of learner corpora in the classroom by developing user-friendly techniques and tools enabling both teachers and learners to benefit. In Japan, the JEFLL corpus can be accessed via a web query system for free and is used by many English teachers to observe common errors in student writing. We prepared six writing topics (20-minute in-class essays), so teachers can ask their students to carry out the free composition task using one of these topics and compare the results against the data of more than 10,000 other users. This is a good exercise for reviewing some grammar points
or teaching writing beyond the sentence level; it has turned out to be a useful resource for teachers.

**IJLCC editors:** LCR probably originated in Europe but it also seems to be a vibrant research field in Asia today. What do you think of LCR in Asian countries? How does it compare with what is happening in Europe? Would you like to pinpoint one or two innovative or promising learner corpus research projects in Asia?

**Yukio Tono:** LCR is becoming an increasingly popular area of research in Asian countries. In Europe, the initiative of the ICLE team has paved the way for current developments in the field, while in Asia there are various national research niches with different interests and foci. In Korea and Taiwan, major learner corpus projects are closely tied to language testing projects and institutions. Institutions such as the Korean Institute of Curriculum and Evaluation (KICE) or the Language Training & Testing Center (LTTC) in Taiwan have compiled in-house learner corpora to improve language tests. In China, despite being latecomers compared to Japan, they built resources very quickly: the *Spoken and Written English Corpus of Chinese Learners* (SWECCCL, Wen et al. 2008), the *Chinese Learner English Corpus* (CLEC; Gui & Yang 2003), and the *Parallel Corpus of Chinese EFL Learners* (PACCEL) are just a few examples. The CLEC has been fully error-tagged. This represents a major achievement compared to other learner corpus projects: many learner corpora are only available in raw format and, even when error-tagged versions exist, they are often not released. In Singapore, the *Singapore Corpus of Research in Education* (SCoRE) is extremely ambitious, consisting of video recordings from more than 350 classrooms at primary and secondary schools in Singapore, thus producing a corpus of classroom discourse (Hong 2005). Unfortunately the data is not available to the public, but this is exactly the type of corpus that Ellis considers necessary to verify the input and interaction process in the classroom (Ellis and Barkhuizen 2005).

In Japan, I have been involved in three major learner corpus projects, i.e. the *Japanese EFL Learner Corpus*, the *International Corpus of Crosslinguistic Interlanguage* (ICCI), and the *National Institute of Communications Technology Japanese Learner English Corpus* (NICT JLE Corpus). Uniquely, all of them are collections of learners at different proficiency levels, with a particular focus on beginning to upper-intermediate learners. Each corpus was designed to enable the user to compare sub-corpora against each other in order to examine features across different proficiency levels. New European learner corpus projects such as the *Cambridge English Profile Corpus* have adopted the CEFR framework to compile sub-corpora representing different proficiency levels. So I have recently been working on the preparation of a new dataset that includes the JEFLL or the ICCI,
re-organised by CEFR levels. This will make it possible to compare our corpora against European ones.

**IJLCR editors:** As editors of the *International Journal of Learner Corpus Research*, we strongly believe that the publication of an international peer-reviewed journal dedicated to LCR represents an important step in the maturity of the field. Do you agree? What do you think of the initiative? What topics would you like to see on the LCR research agenda? What types of articles would you like to read (or not read...) in IJLCR?

**Yukio Tono:** I fully agree! I thought it would take more time for this kind of dedicated journal to appear. I was pleasantly surprised to learn about this new publication, and I really admire the vision of the IJLCR editors. I think the LCR research agenda is still in its infancy and needs to address a number of methodological issues. I would very much like to read proposals for new types of learner corpora, new types of annotations for learner performance, and new types of query systems for learner data. I would also like to see more research papers with solid multifactorial designs. Overuse/underuse and use/misuse should be judged under comparable contexts. To do so, we need to describe linguistic as well as learning environments where the given linguistic feature is observed. I’d be interested to read studies that compare results from corpus data and experimental data. Finally, I would particularly welcome large-scale studies using big data and sophisticated statistics (it would be interesting to collect learners’ errors and their correction candidates in this way) and also more individualized, qualitative approaches to the details of L2 acquisition processes. This will allow for more detailed analyses of how a particular learner behaves under particular learning conditions for a period of time. So I am very much looking forward to seeing the contents of forthcoming issues of IJLCR!

**IJLCR editors:** Thank you again, Professor Tono, for sharing your thoughts on LCR with us and our readers.

Marcus Callies & Magali Paquot  
IJLCR editors

**References**


