From ICE to ICC: A proposal for an International Comparable Corpus

John Kirk (Dresden University of Technology, Germany)
and Anna Čermáková (Charles University Prague, Czech Republic)

There is broad agreement that the *International Corpus of English* project has been highly successful because it has facilitated numerous comparisons of L1 and L2 national varieties of English worldwide. Those comparisons encompass the lexical and morpho-syntactic structural levels, as well as comparisons of discourse types and written registers (cf. e.g. Greenbaum 1996; Hundt & Gut 2012; Aarts et al. 2013; and the papers in the Special Issues of *World Englishes* vol. 15(1) (1996) and vol. 36(3) (2017), to mention but a few key studies). No small part of this success rests with the fact that for each national variety there has been chosen a set of spoken and written text categories which are deemed to be representative of each national variety: 15 discourse situations (totalling 60%) and 17 written registers (totalling 40%). A major review of the ICE project has been undertaken and its results and outcomes are to be agreed upon at ICAME in Prague in May 2017. It seems likely that the text categories will be expanded to include electronic texts and some flexibility in text category choice will become possible.

At the same time, spoken and/or written corpora have been compiled for other languages (cf. list of non-English corpora in e.g. O'Keeffe et al. (2007: 294-296) or the non-English corpora discussed in Xiao (2008) or Ostler (2008)). Xiao makes comparisons with corpora of English: for instance, the *Polish National Corpus* replicates the structure of the *British National Corpus* (Xiao 2008: 387), as does the *Czech National Corpus* (Čermák 1997), which contains spoken texts similar to those of demographically sampled component of BNC (Xiao 2008: 388-389, Čermák 2009). However, no corpus of another language appears to be composed with the range and balance of text categories and quantities of texts as contained within an ICE corpus. The existing corpora in various languages are generally compiled on very different principles and do not allow direct cross-linguistic contrastive comparisons.

Corpus-based contrastive studies are a growing research area and researchers have voiced need for more rigorous analytical framework (e.g. Aijmer et al. 1996, Altenberg & Granger 2002, Marzo et al. 2012, Aijmer & Altenberg 2013, Altenberg & Aijmer 2013). The majority of contrastive studies are being carried out on two languages only, one of the reasons being the lack of comparable data. Contrastive analysis relies on two types of data (Granger 2003): translation (parallel) corpora and comparable corpora (cf. McEnery & Xiao 2007). While translation corpora contain original texts and their translations, comparable corpora contain original texts in two or more languages that have been selected on comparable criteria for text categories and quantities for each category, such as the *Lancaster Corpus of Mandarin Chinese*, which uses the same sampling frame of the *Lancaster/Oslo-Bergen Corpus*, or the *Aarhus Corpus of Contract Law* (both cited in McEnery & Hardie 2012: 19; cf. also e.g. Sharoff et al. 2014). Comparable corpora are an essential data source to support contrastive analyses, since the translation corpora are usually limited as far as text types are concerned (Johansson 2007).

What we are introducing is not a parallel translation corpus such as the *English-Swedish Parallel Corpus*, the *English-Norwegian Parallel Corpus* (ENPC), or the *InterCorp* corpus; rather, it is the creation of an International Comparable Corpus (ICC – pronounced to rhyme with *lick*) with as many languages as wish to come on board. Phase I will start with national, standard(ised) European languages. An expression of interest to collaborate on this project has been expressed for the following languages: German, French, Czech, Slovak, Polish, Finnish, Norwegian, Swedish, and Scottish Gaelic. The first collaborative meeting is to be held in June 2017 in Prague.

The ultimate goal of this project is the facilitation of contrastive studies between English and other languages involving highly comparable datasets of spoken, written and probably electronic registers. A striking and unique feature of each new corpus will be its substantial spoken component, at present comprising 600,000 words (or 60% of the current total). The revised ICE format, to be adopted here, is likely to safeguard this large amount of spoken texts but will include electronic texts as well. Such provision of spoken data across 15 or so discourse situations for contrastive analysis will be unprecedented and invaluable for future research. This will then also allow the much-needed cross-linguistic comparisons of spoken language, further investigations may include the area of pragmatics, such as pragmatic discourse markers (cf. e.g. Aijmer & Vandenbergen 2006).

The proposed comparable corpus ICC will allow substantially to add to existing contrastive corpus-based research (e.g. studies of English-German contrasts, such as König & Gast (2012), or English-Norwegian contrasts, such as Ebeling & Ebeling (2013)), and will allow replicability and comparisons with other languages, i.e. a corpus-based empirical approach to each pair of contrasts, with spin-offs for the others, would all become possible. A further application will almost certainly be possible in bilingual lexicography (as shown by the papers in Sharoff et al. 2013).

Following the launch of *ICE Phase II* at the ICAME conference in Prague in May 2017 and the first ICC meeting in June 2017, *Corpus Linguistics 2017* seems an ideal and opportune moment to introduce this exciting new international, multi-lingual corpus project and to present at the outset some of the issues and challenges it raises as well as the solutions being adopted.

References

Aarts, B., Close, J, Leech, G. & Wallis, S. (Eds.). (2013). *The Verb Phrase in English*. Cambridge: Cambridge University Press.

Aijmer, K. & Altenberg, B. (Eds.). (2013). *Advances in Corpus-based Contrastive Linguistics*. Amsterdam: John Benjamins.

Aijmer, K., Altenberg, B. & Johansson, M. (Eds.). (1996*). Languages in Contrast. Papers from a Symposium on Text-based Cross-Linguistic Studies*, *Lund, 4–5 March 1994*. Lund: Lund University Press.

Aijmer, K. & Vandenbergen, A.-M. (Eds.). (2006). *Pragmatic Markers in Contrast*. Amsterdam: Elsevier.

Altenberg, B. & Aijmer, K. (Eds.). 2013. *Text-based Contrastive Linguistics*. Special Issue of *Languages in Contrast* 13(2).

Altenberg, B. & Granger, S. (Eds.). (2002). *Lexis in Contrast: Corpus-based Approaches*. Amsterdam: John Benjamins.

Čermák, F. (1997). Czech National Corpus: A Case in Many Contexts. *International Journal of Corpus Linguistics*, *2*(2), 181–197.
Čermák, F. (2009). Spoken Corpora Design: Their Constitutive Parameters. *International Journal of Corpus Linguistics, 14*(1), 113–123.
Ebeling, J. & Ebeling, S. O. (2013). *Patterns in Contrast*. Amsterdam: John Benjamins.
Granger, S. (2003). The Corpus Approach: A common way forward for contrastive linguistics and translation studies? In S. Granger, J. Lerot & S. Petch-Tyson (Eds.), *Corpus-based Approaches to Contrastive Linguistics* (pp. 17–29). Amsterdam: Rodopi.
Greenbaum, S. (1996). *Comparing English World-Wide*. Oxford: Clarendon Press.
Hundt, M. & Gut, U. (Eds.). (2012). *Mapping Unity and Diversity World-Wide.* Amsterdam: John Benjamins.
Johansson, S. (2007). *Seeing through Multilingual Corpora: On the use of corpora in contrastive studies.* Amsterdam: John Benjamins.
König, E. & Gast, V. (2012). *Understanding English-German Contrasts*. Berlin: Erich Schmidt Verlag.
Marzo, S., Heylen, Kr. & De Sutter, G. (Eds.). (2012). *Corpus Studies in Contrastive Linguistics*. Amsterdam: John Benjamins.
McEnery, T. & Hardie, A. (2012). *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press.
McEnery, T. & Xiao, R. (2007). Parallel and Comparable Corpora: What is Happening? In G. M. Anderman & M. Rogers (Eds.), *Incorporating Corpora: The Linguist and the Translator* (pp. 18–31). Clevedon: Multilingual Matters.
O'Keeffe, A. & McCarthy, M. (2010). *The Routledge Handbook of Corpus Linguistics*. Abingdon: Routledge.
O'Keeffe, A., McCarthy, M. & Carter, R. (2007). *From Corpus to Classroom: Language Use and Language Teaching*. Cambridge: Cambridge University Press.
Ostler, N. (2008). Corpora of less studied languages. In A. Lüdeling & M. Kytö (Eds.), *Corpus Linguistics: An International Handbook* (pp. 457–483). Berlin: Mouton de Gruyter.
Sharoff, S., Rapp, R., Zweigenbaum, P. & Fung, P. (Eds.). (2013). *Building and Using Comparable Corpora*. Heidelberg: Springer.
World Englishes (1996) vol. 15(1), special issue on the International Corpus of English, guest-edited by S. Greenbaum & G. Nelson.
World Englishes (2017) vol. 36(3), special issue on the International Corpus of English, guest-edited by G. Nelson, R. Fuchs & U. Gut.
Xiao, R. (2008). Existing Corpora. In A. Lüdeling & M. Kytö (Eds.), *Corpus Linguistics: An International Handbook* (pp. 383–456). Berlin: Mouton de Gruyter.