Published in: Bański, Piotr/Kupietz, Marc/Lüngen, Harald/Rayson, Paul/Biber, Hanno/Breiteneder, Evelyn/Clematide, Simon/Mariani, John/stevenson, Mark/Sick, Theresa (eds.): Proceedings of the Workshop on Challenges in the Management of Large Corpora and Big Data and Natural Language Processing (CMLC-5+BigNLP) 2017 including the papers from the Web-as-Corpus (WAC-XI) guest section. Birmingham, 24 July 2017. - Mannheim: Institut für Deutsche Sprache, 2017. Pp. 7-12

From ICE to ICC: The new *International Comparable Corpus*

John Kirk
4 Parkvue Manor
Belfast, BT5 7TJ
Northern Ireland
jk@etinu.com

Anna Čermáková
Charles University Prague
Institute of the Czech National Corpus
nám. J. Palacha 2
116 38, Praha 1
anna.cermakova@ff.cuni.cz

Abstract

This paper outlines the broad research context and rationale for a new international comparable corpus (ICC). The ICC is to be largely modelled on the text categories and their quantities the *International Corpus of English* with only a few changes. The corpus will initially begin with nine European languages but others may join in due course. The paper reports on those and other agreements made at the inaugural planning meeting in Prague on 22-23 June 2017. It also sets out the project's goals for its first two years.

1 International Corpus of English (ICE) project

There is broad agreement that the *International Corpus of English* (ICE) project has been highly successful because it has facilitated numerous systematic 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 and 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).

ICE does not sample populations, nor does it relate national component sizes proportionately to the size of the population. Rather ICE is entirely text-based, being organized around text categories and the quantity which has been designated for each one, with each corpus following the same pattern regardless of population size. Each corpus thus amounts to its text collection, no matter whether it is the USA, with a population of 321.4 million, or Malta, with its population of 419,000. It's the identical nature and quantities of that collection which allow for the comparability of the component corpora. No small part of the success of ICE 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%) (see Table 1) and 17 written registers (totalling 40%) (see Table 2). The importance of retaining those categories for any second generation ICE corpus for comparability was confirmed in a major review of the ICE project in May 2017.

SPOKEN TEXTS			
DIALOGUE (180)	**Private (100)**	direct conversations	90
		distanced (telephone) conversations	10
	Public (80)	class lessons or seminars	20
		broadcast discussions	20
		broadcast interviews	10
		parliamentary debates	10
		legal cross-examinations	10
		business transactions	10
MONOLOGUE (120)	**Unscripted (70)**	spontaneous commentaries	20
		unscripted speeches	30
		demonstrations	10

		legal presentations	10
	Scripted (50)	broadcast news	20
		broadcast talks	20
		speeches (not broadcast)	10

Table 1: Spoken text categories in ICE.

The spoken texts are categorized by a principled, top-down approach with regard to the speech situation: whether there is one speaker or more than one; whether the speech is public or private; and whether the speech is scripted or spontaneous. The final choice is based largely on functional domain, such as broadcasting, parliament, education, or the law courts. As Table 1 shows, most categories are collected in similar quantities, except private, face-to-face conversation, which predominate, not least because they are regarded as the quintessential form of spoken interaction. However, public speech accounts for two-thirds of all spoken texts (200/300 texts).

In these ways, although they are not without criticism, ICE has come to represent a fair sampling of all the major spoken and written varieties of English in the present day and throughout the world, particularly in L1 countries.

WRITTEN TEXTS			
NON-PRINTED (50)	**Non-professional writing (20)**	student untimed essays	10
		student examination	10
	Correspondences (30)	social letters	15
		business letters	15
PRINTED (150)	**Informational (learned) (40)**	humanities	10
		social sciences	10
		natural sciences	10
		technology	10
	Informational (popular) (40)	humanities	10
		social sciences	10
		natural sciences	10
		technology	10
	Informational (reportage) (20)	press news reports	20
	Instructional (20)	administrative/regulatory prose	10
		skills/hobbies	10
	Persuasive (10)	press editorials	10
	Creative (20)	novels/short stories	20

Table 2: Written text categories in ICE.

The written texts are similarly categorized by a principled, top-down approach with regard to the register situation: whether the text has been printed or not; and what its primary function is. Cutting across two of the main informational functions are domain choices. There are also two types of writing from newspapers: reporting as a further instance of informational writing; and editorials as an instance of persuasive writing. Printed texts account for three-quarters of all written texts (150/200).

2 Contrastive (corpus) linguistics

While ICE has been developing over the last thirty years or so, 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, to an extent, 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 spoken and written text categories and quantities of texts as contained within the ICE corpus. The existing corpora in various languages are generally compiled on very different principles and thus 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 and Granger, 2002; Marzo et al., 2012; Aijmer and Altenberg, 2013; Altenberg and Aijmer,

2013; Ebeling and Ebeling, 2013). The majority of contrastive studies are being carried out on two languages only (and very often one of the compared languages is English), 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 and Xiao, 2007). While translation corpora contain original (source) texts with their aligned translations, comparable corpora[1] 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 and 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 (e.g. Johansson, 2007; Mauranen, 1999).

3 The International Comparable Corpus (ICC)

3.1 Rationale for ICC

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 electronic registers. What we are introducing is not a parallel translation corpus;[2] but rather, it is the creation of an *International Comparable Corpus* (ICC – pronounced to rhyme with *lick*), with as many languages as may 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: Czech, Finnish, French, German, Norwegian, Polish, Slovak, and Swedish. The first collaborative meeting was held on 22–23 June 2017 in Prague[3].

The ICC corpus is based, on the one hand, on the idea that there are plenty of various language data for many languages that could be reused if carefully selected and, on the other, that contrastive analysis very often relies on comparisons with English. Thus the ICC corpus will largely rely on re-using existing language resources and will be modelled for comparability with the ICE family corpora. For the field of contrastive linguistics, a striking and unique feature of each new corpus in ICC will be its substantial spoken component. Such provision of spoken data across 13 or so discourse situations for contrastive analysis among several languages is entirely unique as it will allow the much-needed and unprecedented cross-linguistic corpus-based comparisons of spoken language. Together with balanced data across written registers, ICC will become invaluable for future research[4]. The approach will also allow replicability and comparisons with and between other languages.

3.2 Composition of ICC

Let us now turn to some specifics about the new ICC. Following agreement in Prague, the ICC will broadly follow the composition of the ICE corpus, see Tables 3 and 4, the rationale for those text categories as briefly outlined above being taken largely for granted. Individual texts will comprise approximately 2,000 tokens each, ending with sentences or paragraphs completed (if possible); many texts will be excerpts, derived from a good spread of beginnings, middles and endings of their source texts. If texts are shorter than 2,000 words, composite texts are to be created, to make up the desired total. Within categories, the texts are to be chosen on the basis of the range, spread and diversity of the category or the function which the texts represent. Texts are to post-date 2000, and there are to be no

[1] Terminology may differ, but here we mean by parallel corpora, source language texts aligned to their translations. Comparable corpora may be multilingual as referred to in this article but also monolingual, containing comparable datasets in one language, e.g. non-translated language and translated language such as the *The Translation English Corpus* (TEC) (Baker, 1995), available at http://www.monabaker.com/tsresources/TranslationalEnglishCorpus.htm).

[2] Such as the *English-Swedish Parallel Corpus*, the *English-Norwegian Parallel Corpus* (ENPC), or the *InterCorp* corpus.

[3] We would like to thank the following participants in the ICC planning meeting: Michal Křen (Czech), Oliver Wicher (French), Marc Kupietz (German), Signe Oksefjell Ebeling (Norwegian), Jarle Ebeling (Norwegian), Rafal Gorski (Polish), Radovan Garabík (Slovak), Vladimir Benko (Slovak), and the following for their input and support of the ICC idea: Jarmo Jantunen (Finnish), Dirk Siepmann (French), Christoph Bürgel (French), Sascha Diwersy (French), Thomas Schmidt (German), Mária Šimková (Slovak), and Karin Aijmer (Swedish).

[4] Cf. e.g. studies of English-German contrasts, such as König and Gast (2012), or English-Norwegian contrasts, such as Ebeling and Ebeling (2013).

translations. Ideally, no source text is to be used more than once. Moreover, ICC has decided to drop all non-printed texts (undergraduate essays and letters, totalling 100 texts) and the two spoken categories of legal texts (each 10 texts). However, it has been decided to add a category of (electronic) blogs (50 texts each of 2,000 words) equivalent to the non-printed texts now dropped. The total for the spoken component will be 560,000 words. As far as possible texts are to be selected from existing national resources, to maximise their re-usability and to minimise the effort. ICC is not intended to replicate or compete with national corpora; rather the emphasis is on systematic comparability between and across languages. As with ICE, so will it be for ICC: identical types and quantities of texts will neutralize any population differences between participating countries, whether 81 million for Germany, or 5.2 million for Norway. As ICE is a corpus of English, so ICC is to be a corpus of languages.

	CZE	FIN	FRE	GER	NOR	POL	SLO	SWE
Hum. (acad.)	√		√	√	√	√	√	
Soc. sci. (acad.)	√		√		√	√	√	
Nat. sci. (acad.)	√		√		√	√	√	
Technol. (acad.)	√		√	√	√	√	√	
Hum. (pop.)	√	√	√	√	√	√	√	
Soc. sci. (pop.)	√	√	√	√	√	√	√	
Nat. sci. (pop.)	√	√	√	√	√	√	√	
Technol. (pop.)	√		√	√	√	√	√	
Reportage	√	√	√	√	√	√	√	
Instruct. (admin.)	√	√	√		√	√	√	
Instruct. (hobbies)	√		√		√	√	√	
Press editorials	√			√	√	√	√	
Fiction	√		√	√	√	√	√	
Blogs	√	√	√	√	√	√	√	√

Table 3: Written categories agreed for ICC and their availability from currently identified resources (as of June 2017).

	CZE	FIN	FRE	GER	NOR	POL	SLO	SWE
Direct convers.	√	√	√	√		√	√	√
Telephone convers.			√	√				√
Class (uni) lessons								
Broadcast discussions	√	√	√		√	√	√	
Broadcast interviews	√		√		√	√	√	
Parliament debates			√		√		√	
Business transact.			√	√				√
Spontan. comment.			√					
Unscripted speeches			√				√	
Demonst. (broadc.)								
Broadcast News		√	√		√	√		
Broadcast Talks		√			√	√		
Speeches (not broadcast)			√					

Table 4: Spoken categories agreed for ICC and their availability from currently identified resources (as of June 2017).

Both written and spoken texts are to be marked up in a format conforming to TEI P5 XML,[5] keeping the original characters (multiple dashes, apostrophes etc.). Each text is to be accompanied by metadata in an accompanying header. As, for the spoken component, sound alignment is strongly desired wherever possible, a multi-layer environment will be needed, such as ELAN, in which one-layer will contain the orthographic transcription. Transcription details are to be language-dependent. However, overlaps and pauses are to be included and marked according to TEI. A minimum markup scheme is being drawn up.

Texts are to be annotated with regard to the part of speech (POS) status with POS taggers representing state-of-the-art for each language. As, among the languages, considerable morphological variation exists, another simultaneous tagging layer was considered for mapping language-specific POS annotation schemes onto higher level "universal" schemes

[5] http://www.tei-c.org/Guidelines/P5/

(e.g. 'universal dependencies' or simplified tagset used for the *Aranea* corpora series)[6] to support cross-linguistic comparisons. A further aspiration for the future is for cross-linguistic syntactic tagging and parsing.

The ICC is to be made available through a common search interface with distributed indexes (KorAP).[7] However, there is a preference for ICC components to be downloadable, at least partially[8], and with non-destructive annotation, but that will depend on copyright permissions being cleared in the first instance. As a plan of action, it was decided to re-negotiate licensing of written texts (CC BY-NC), and to choose and attempt to transform the spoken texts into TEI P5 XML format by the end of the first year. By the end of two years, missing spoken texts are to be collected and the pilot written corpus should have been completed.

These, then, are the parameters in terms of which the ICC is to come into being. We are pleased to introduce this exciting, new international corpus. The project welcomes further participation.

References

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

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

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

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

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

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

Baker M. 1995. Corpora in Translation Studies. An Overview and Suggestions for Future Research. *Target*, 7(2): 223-243.

Benko V. 2014a. Aranea: Yet Another Family of (Comparable) Web Corpora. In P. Sojka, A. Horák, I. Kopeček and K. Pala (Eds.), *Text, Speech and Dialogue. 17th International Conference, TSD 2014*, Brno, Czech Republic, September 8-12, 2014. Proceedings. LNCS 8655, 257-264. Springer International Publishing Switzerland.

Benko V. 2014b. Compatible Sketch Grammars for Comparable Corpora. In A. Abel, C. Vettori and N. Ralli (Eds.), *Proceedings of the XVI EURALEX International Congress: The User In Focus*. 15–19 July 2014, 417-430. Bolzano/Bozen: Eurac Research.

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*. John Benjamins, Amsterdam.

Granger S. 2003. The Corpus Approach: A common way forward for contrastive linguistics and translation studies? In S. Granger, J. Lerot and S. Petch-Tyson (Eds.), *Corpus-based Approaches to Contrastive Linguistics*, 17–29. Rodopi, Amsterdam.

Greenbaum S. 1996. *Comparing English World-Wide*. Clarendon Press, Oxford.

Hundt M. and Gut U. (Eds.). 2012. *Mapping Unity and Diversity World-Wide*. John Benjamins, Amsterdam.

Johansson S. 2007. *Seeing through Multilingual Corpora: On the use of corpora in contrastive studies.* John Benjamins, Amsterdam.

König E. and Gast V. 2012. *Understanding English-German Contrasts*. Erich Schmidt Verlag, Berlin.

Mauranen A. 1999. Will 'translationese' ruin a contrastive study? *Languages in Contrast*, 2 (2): 161-185.

[6] http://universaldependencies.org/ and http://unesco.uniba.sk/aranea_about/index.html (Benko, 2014a, b)

[7] Korpusanalyseplattform der nächsten Generation; cf. http://www1.ids-mannheim.de/kl/projekte/korap.html

[8] By CC BY-NC licensing; f. https://creativecommons.org/licenses/by-nc/2.0/

Marzo S., Heylen K. and De Sutter G. (Eds.). 2012. *Corpus Studies in Contrastive Linguistics*. John Benjamins, Amsterdam.

McEnery T. and Hardie A. 2012. *Corpus Linguistics: Method, Theory and Practice*. Cambridge University Press, Cambridge.

McEnery T. and Xiao R. 2007. Parallel and Comparable Corpora: What is Happening? In G. M. Anderman and M. Rogers (Eds.), *Incorporating Corpora: The Linguist and the Translator*, 18–31. Multilingual Matters, Clevedon.

O'Keeffe A. and McCarthy M. 2010. *The Routledge Handbook of Corpus Linguistics*. Routledge, Abingdon.

O'Keeffe A., McCarthy M. and Carter R. 2007. *From Corpus to Classroom: Language Use and Language Teaching*. Cambridge University Press , Cambridge.

Ostler N. 2008. Corpora of less studied languages. In A. Lüdeling and M. Kytö (Eds.), *Corpus Linguistics: An International Handbook*, 457–483. Mouton de Gruyter, Berlin.

Sharoff S., Rapp R., Zweigenbaum P. and Fung P. (Eds.). 2013. *Building and Using Comparable Corpora*. Springer, Heidelberg.

World Englishes (1996) vol. 15(1), special issue on the International Corpus of English, guest-edited by S. Greenbaum and G. Nelson.

World Englishes (2017) vol. 36(3), special issue on the International Corpus of English, guest-edited by G. Nelson, R. Fuchs and U. Gut.

Xiao R. 2008. Existing Corpora. In A. Lüdeling and M. Kytö (Eds.), *Corpus Linguistics: An International Handbook*, 383–456. Mouton de Gruyter, Berlin.