JYX
JYVÄSKYLÄN YLIOPISTO
UNIVERSITY OF JYVÄSKYLÄ

This is a self-archived version of an original article. This version may differ from the original in pagination and typographic details.

Author(s): Čermáková, Ann; Jantunen, Jarmo; Jauhiainen, Tommi; Kirk, John; Křen, Michal; Kupietz, Marc; Uí Dhonnchadha, Elaine

Title: The International Comparable Corpus : Challenges in building multilingual spoken and written comparable corpora

Year: 2021

Version: Published version

Copyright: © 2021 Research in Corpus Linguistics

Rights: CC BY 4.0

Rights url: https://creativecommons.org/licenses/by/4.0/

Please cite the original version:

Čermáková, A., Jantunen, J., Jauhiainen, T., Kirk, J., Křen, M., Kupietz, M., & Uí Dhonnchadha, E. (2021). The International Comparable Corpus : Challenges in building multilingual spoken and written comparable corpora. Research in Corpus Linguistics, 9(1), 89-103. https://doi.org/10.32714/ricl.09.01.06

RiCL Research in Corpus Linguistics

The *International Comparable Corpus*: Challenges in building multilingual spoken and written comparable corpora

Anna Čermáková[a] – Jarmo Jantunen[b] – Tommi Jauhiainen[c] – John Kirk[d] –
Michal Křen[a] – Marc Kupietz[e] – Elaine Uí Dhonnchadha[f]
Charles University[a] / Prague
University of Jyväskylä[b] / Finland
University of Helsinki[c] / Finland
University of Vienna[d] / Austria
Institut für Deutsche Sprache, Mannheim[e] / Germany
Trinity College Dublin[f] / Ireland

Abstract – This paper reports on the efforts of twelve national teams in building the *International Comparable Corpus* (ICC; https://korpus.cz/icc) that will contain highly comparable datasets of spoken, written and electronic registers. The languages currently covered are Czech, Finnish, French, German, Irish, Italian, Norwegian, Polish, Slovak, Swedish and, more recently, Chinese, as well as English, which is considered to be the pivot language. The goal of the project is to provide much-needed data for contrastive corpus-based linguistics. The ICC corpus is committed to the idea of re-using existing multilingual resources as much as possible and the design is modelled, with various adjustments, on the *International Corpus of English* (ICE). As such, ICC will contain approximately the same balance of forty percent of written language and 60 percent of spoken language distributed across 27 different text types and contexts. A number of issues encountered by the project teams are discussed, ranging from copyright and data sustainability to technical advances in data distribution.

Keywords – ICC corpus; contrastive linguistics; comparable corpus; ICE corpus; data sustainability; copyright

1. INTRODUCTION

While corpus-based contrastive studies largely rely on translation (parallel) corpora, they also increasingly draw on comparable data (see, e.g., Mauranen 1998; Aijmer and Altenberg 2013). Unlike extensive comparable corpora mined from the web which are used in natural language processing for the development of machine translation and cross-lingual information retrieval systems (Sharoff *et al*. 2013), the ultimate goal of the *International Comparable Corpus* (ICC), a collaborative project of currently twelve

national teams,[1] is to provide highly comparable datasets of spoken and written registers across a range of carefully matched text categories.

The ICC starts with the idea of linguistic data reusability, and thus contributes to a discussion of data sustainability, on the one hand, and the current lack of comparable datasets for contrastive studies, on the other. A substantial proportion of the current landscape in contrastive studies is based on comparisons of pairs of languages, very often one of those languages being English. This trend is quickly confirmed by a quick survey of the last five volumes (15 to 19) of *Languages in Contrast*,[2] the leading journal in contrastive linguistics. Two special issues aside, out of the 47 published research articles, 39 involved two-language comparisons and 38 articles involved English. There is no doubt that one of the contributing factors to this two-language English-centered research is a lack of suitable linguistic resources. Another notable observation is that all the research (with a few exceptions) is essentially focused on written language only.

The aim of the ICC is, therefore, to provide a highly comparable, multilingual dataset of both spoken and written language to support contrastive and cross-linguistic research.[3] It was decided that the design of the ICC will be modelled on the *International Corpus of English* (ICE)[4] (see Greenbaum 1996), where each ICE corpus comprises one million words made up of 40 percent written samples and 60 percent spoken samples. The provision of comparable spoken datasets across several languages will be unique and will also allow the much-needed contrastive comparisons of spoken language. In addition to English, the languages currently involved in the ICC compilation, and in various stages of completion, are Czech, Finnish, French, German, Irish, Italian, Norwegian, Polish, Slovak, Swedish and, the most recent acquisition, Chinese.

The following sections will discuss some of the issues being faced in the compilation of the corpus. Section 2 will discuss the design of the ICC corpus and legacy issues arising from the ICE design, including comparability of text categories. Section 3 will discuss, in more detail, some of the issues being faced by the individual national teams, such as the questions of formatting and annotation, while Section 4 looks into

[1] https://korpus.cz/icc/languages
[2] https://benjamins.com/catalog/lic
[3] For discussion of terminology, see e.g. Ebeling and Ebeling (2013: 4).
[4] https://www.ice-corpora.uzh.ch/

possibilities and problems concerning the ICC data release, as well as the dissemination of the corpus to the wider research community.

2. Designing the ICC

The ICE family corpora project was initiated in the early 1990s, at a time when questions of data sampling and data comparability were only beginning to be intensively discussed within corpus linguistics research, and when large corpora such as the *British National Corpus* started to be built (McEnery and Hardie 2013). The ICE sampling frame is based on same-length extracts (2,000 words) organized around text type categories and involves 15 spoken discourse situations and 17 written text types (for more details see Greenbaum 1996: 3). For the ICC, the ratio of written to spoken language represented in the ICE corpus has been kept, but a few text categories have been revised for comparability across the languages involved. Cross-linguistic text comparability is a thorny issue (see, e.g., Granger 2010). Contrastive cross-linguistic comparisons rely on the notion of 'comparability', a "background of sameness" (James 1980: 169) against which the differences between languages can be contrasted. Comparability is, therefore, always a matter of degree and, as James (1980: 168) points out, it "does not presuppose absolute identity, but merely a degree of shared similarity." In practical terms, data comparability is being achieved by the ICC, with various degrees of success, through matching various text parameters, such as time of production or text type. While parameters such as the year of publication may be relatively easy to match, matching text types across languages is far more challenging. As other corpus projects show, some text types may be highly culturally specific. For example, in the case of the *Nepali National Corpus* (Yadava *et al*. 2008), it was not possible to find science fiction texts, and see McEnery and Xiao (2004) for discussion on matching FLOB corpus text types to *Lancaster Corpus of Mandarin Chinese*. This was also the case with the ICC; for example, it was decided among the national teams not to include legal cross-examinations and legal presentations, two text types present in the spoken component of the ICE corpora.

As its English component, the ICC uses the written text types of the ICE-Ireland corpus (Kallen and Kirk 2007, 2008). Apart from these written texts which date from 1990–1994 (a bibliography is provided in Kallen and Kirk 2008: 65–79), it was felt also desirable to include texts that are largely contemporary —that is, wherever possible, texts

published after 2000 (see Section 3.1). To reflect the changing nature of current communication (e.g. Crystal 2004), it was also decided that a component of on-line texts should be included. Accordingly, ICC corpora will drop the category of non-printed texts (present in ICE) and, instead, include blogs which will be collected for all the languages involved, including English. For the final set of categories in the ICC design, see Table 1 (for other design criteria see also Kirk and Čermáková 2017: 10).

Spoken	**Words**	**Written**	**Words**
Dialogue/conversation		**Printed**	
Direct, face-to-face conversation	180,000	Humanities (academic)	20,000
Telephone conversation	20,000	Social sciences (academic)	20,000
Classroom lessons	40,000	Natural sciences (academic)	20,000
Broadcast discussions	40,000	Technical (academic)	20,000
Parliamentary debates	20,000	Humanities (popular)	20,000
Business transactions	20,000	Social sciences (popular)	20,000
Monologue		Natural sciences (popular)	20,000
Spontaneous commentaries	40,000	Technical (popular)	20,000
Unscripted speeches	60,000	Reportage	40,000
Demonstrations	20,000	Administrative/regulatory prose	20,000
Broadcast interviews	20,000	Skills & Hobbies	20,000
Broadcast news	40,000	Press editorials	20,000
Broadcast talks	40,000	Fiction	40,000
Scripted speeches (not broadcast)	20,000	**Web/Internet**	
Total	560,00	Blogs	100,000
		Total	400,000
Grand total 960,000			

Table 1: The ICC corpus composition across text categories[5]

3. Compiling the ICC

The ICC compilation relies largely on the idea of reusability. The data to be included in the ICC are meant to be selected primarily from already existing linguistic resources. While some of the languages involved may draw on large depositories of their national corpora (Czech,[6] German,[7] Polish,[8] Slovak[9]) and others are able to collect data from various sources (Finnish,[10] French, Italian, Norwegian,[11] Chinese), all languages will

[5] Whereas the ICC is based on ICE, we are aware that a total of 960,000 words falls short of the ICE's one-million words total. This shortfall is due solely to the ICC's dropping of spoken legal texts. We are currently discussing in what ways this shortfall may be rectified, in order for the grand total to become the rounded one-million words. However, we are also aware that not all ICE corpora have indeed completed every text category or provided one-million words, and that Kirk and Nelson (2018) envisage that second-generation ICE corpora may come to have variable word totals.
[6] http://korpus.cz/
[7] http://www.dereko.de/, https://dgd.ids-mannheim.de/
[8] http://nkjp.pl/
[9] https://korpus.sk/
[10] https://www.kielipankki.fi/language-bank/
[11] https://www.hf.uio.no/ilos/english/services/knowledge-resources/icc-no/

need to collect new data for some of the categories. Some languages (e.g. Swedish and Irish) will need to start essentially from scratch, especially for the collection of most of the spoken categories. The need for collecting new data does not always arise from the fact that a particular text type has not been collected before. The idea of data re-usability has proved extremely difficult to pursue due to complex copyright reasons. More often than not, corpora compiled in the past have usage agreements tied to those specific corpora, specific research purposes or institutions, so that the re-use of the texts has not always proven possible.

This section will discuss in more detail various issues encountered while compiling the written (Section 3.1) and spoken (Section 3.2) ICC resources. Section 3.3, in turn, will discuss the technical issues related to formatting and annotating the corpora.

3.1. The ICC written component

In order to compile the ICC written components, languages with large national corpora are in a relatively more comfortable situation as they already have data to draw from. The SYN-series corpora of contemporary written Czech being compiled at the *Czech National Corpus* (CNC)[12] can be described as traditional (as opposed to the web-crawled corpora), featuring well-defined composition, reliability of annotation and high-quality text processing. The SYN series also includes SYN2015, a representative reference corpus that contains a good mix of fiction, non-fiction, newspapers and magazines. It has been compiled with diversity in mind, so that it not only contains all registers common for written (printed) Czech but, within each register, it also comprises a large variety of texts by various authors, from various publishers, etc. (Křen *et al*. 2016). Based on SYN2015, the Czech written component of the ICC (ICC-CZ) has been selected and made internally available in June 2019 through the institute's corpus query engine *Kontext* (see Section 4).

For German, the situation is almost as good as for Czech. Drawing on resources in the *German Reference Corpus* (DeReKo),[13] the first draft version of the ICC-DE was completed in July 2019. However, some domains still need to be sampled more broadly

[12] The Czech ICC component and the preparation of this publication has been supported within the Czech National Corpus project (LM2018137) funded by the Ministry of Education, Youth and Sports of the Czech Republic within the framework of *Large Research, Development and Innovation Infrastructures*.
[13] https://www.ids-mannheim.de/digspra/kl/projekte/korpora

before the corpus release. Fortunately, in this case, some licensees were willing to release texts for the ICC under a Creative Commons license (CC), so that in the future the German ICC part may be available for download (see Section 4 for further discussion).

The compilation of the Finnish component of the ICC presents one of the examples where it is difficult, in some cases impossible, to re-use already existing resources. The investigation of existing and matching data in Finland was done in 2017.[14] The corpora distributed through the *Language Bank of Finland* were identified as the most promising source of material for the ICC corpus. During the last ten years, the *Language Bank of Finland*, maintained by the FIN-CLARIN consortium, has aimed to collect and give centralized access to various corpora compiled by the consortium members, which include most of the Finnish academic institutions dealing with linguistic data. The initial driving idea behind the ICC corpus was to collect a separate collection under a CC-BY or CC-BY-NC license. Some of the identified corpora from the *Language Bank of Finland* were indeed readily available for download and redistribution with such licenses. However, the remainder of the texts identified as suitable for inclusion in the ICC are available under a variety of more restrictive licenses issued by the different rights-holding universities, research institutes, private companies, or even individuals. The attempts to renegotiate the more restrictive licenses with their rights-holders were mostly unsuccessful. Consequently, due to these strict licenses and distribution limitations, it has not been possible to re-use many of the existing suitable corpus resources. As a similar situation has occurred also with other languages, the ICC corpus distribution will need to be reconsidered (see Section 4). One of the proposed solutions is to make the data available through the respective institutional corpus query interfaces such as the *Korp*[15] offered by the *Language Bank of Finland*.

As discussed in Section 2, the ICC preference is to include contemporary data (post-2000). Search for the potential data for the inclusion in the ICC-FI has revealed that this requirement is challenging. For example, a major source of written data, the *Finnish Text Collection*,[16] consists of newspapers, journals and fiction texts dating back to the 1990s. One reason for a limited number of corpora that contain current language is that they have

[14] We wish to thank the Department of Language and Communication Studies at the University of Jyväskylä for providing financial support for this project.
[15] https://www.kielipankki.fi/support/korp/
[16] https://www.kielipankki.fi/news/ftc-in-korp/

been compiled within projects that ended before or around 2000, and data compilation ceased thereafter.

The compilation of the Norwegian ICC (ICC-NO) written component has been finished as well.[17] The texts were selected from various digital archives or from sources in the public domain. Again, most effort went into obtaining copyright clearance from the archive owners.

Another case in point is the French component.[18] Even though the extensive French corpus FRANTEXT,[19] spanning texts from the twelfth to twentieth centuries, amounts to 250 million words, the majority of its texts are literary, with many of the text types needed for the ICC simply not covered. The copyright licenses vary across the French corpora; for instance, FRANTEXT limits access to its online interface. Text samples for the ICC-FR have had to become selected manually and, as with all the other corpora, this involves a laborious process of requesting permissions for further distribution.

The case of Irish (ICC-GA) is different in that it is a minority language with limited written and spoken corpora. Although Irish is constitutionally the first language of Ireland (with English being the second language), in practice, English is the first language of discourse and business for much of the population. This means that many domains of Irish language usage are under pressure from English in terms of lexicon and language structure. Therefore, a balanced corpus design such as the ICC is of immense importance for inspiring the collection of data for spoken and written domains, which are not only difficult to obtain but do not yet feature in existing Irish corpora. However, it is envisaged that the Irish written component will draw on texts from existing sources, such as the *The New Corpus for Ireland*[20] (Kilgarriff *et al*. 2006) and the *Corpus of Contemporary Irish*.[21]

As discussed in Section 2, as an additional new component that is not present in the ICE corpora, it has been decided to include texts that display some of the characteristics of internet language. The ICC corpora will therefore include various blog posts that will

[17] The Norwegian team would like to thank the Department of Literature, Area Studies and European Languages at the University of Oslo (further acknowledgments to be found at https://www.hf.uio.no/ilos/english/services/knowledge-resources/icc-no/acknowledgements.html).
[18] Personal communication with Oliver Wicher, the compiler of the ICC-FR component.
[19] https://www.frantext.fr/
[20] http://corpas.focloir.ie/
[21] https://www.gaois.ie/g3m/en/

be specifically collected for the project amounting to about 100,000 words per each language.

3.2. The ICC spoken component

Obviously, the ICC spoken categories pose many more challenges for data collection than the written ones (see Table 1 in Section 2). Current state-of-the-art spoken corpora have sound-aligned transcripts; however, our pivot language corpus, the ICE-Ireland, unfortunately contains only transcriptions with no aligned sound files. Therefore, for maximum efficiency and re-use of data, the spoken component of the ICC-English is to comprise data from the new *London-Lund Corpus* 2 (LLC-2),[22] with any gaps to be filled by fresh recordings and transcriptions.

Generally, spoken language is often underrepresented in language resource collections and some categories are not available even in the large national corpora, and will need to be collected and transcribed. In transcribing spoken data, the usual practice is to protect the anonymity of participants by anonymizing personal and identifying references in the transcriptions, and also by bleeping the relevant sections of the audio files where necessary. Under the new European Union General Data Protection Regulation (GDPR), this is now a strict requirement, and care must be taken not to hold any unnecessary personal or identifying data. In a spoken corpus, the human voice itself can be considered an identifying feature. Therefore, new consent agreements with participants for the newly collected data need to make reference to this issue, which may also need to be considered in the case of pre-existing recordings.

While collections of direct conversation are less well represented for other languages (see below), there are two Czech corpus series on which the ICC component will draw: the older ORAL (5.4 million words in total) and the newer ORTOFON (currently one million words), which features a manual, two-tier transcription. Each of the series includes samples from the entire Czech Republic and the latter is fully balanced for the main sociolinguistic categories (Komrsková *et al*. 2017). In addition to the category of direct conversation (see Table 1), the *Czech National Corpus* has recently added to its spoken resources a collection of more formal and prepared speeches

[22] https://www.sol.lu.se/en/subjects/engelska/research/llc2/. We would like to express our gratitude to Nele Põldvere and Carita Paradis for their willingness to collaborate with the provision of these data.

(monologues): the ORATOR corpus (0.58 mil. words), which was released in 2019 (Kopřivová *et al.* 2019). ORATOR includes, for example, lectures, instructions, guided tours, welcome addresses and sermons. However, even with these rich resources of spoken data, many of the remaining text types will still need to be collected.

The German ICC component will draw on data from the *Archive for Spoken German*.[23] Although the transcriptions are richly annotated with metadata, some sub-domains will need to be added. Furthermore, legal issues concerning restrictions in the use of public broadcast media data have arisen. In this respect, legal expertise has been sought and we have been advised that under current copyright regulations, the use and distribution for research purposes needs to be limited to small excerpts only.

The Norwegian spoken component is currently under construction, with recordings of conversations to be made. Other text types need to be transcribed and consent forms conforming to the current GDPR legislation are being issued. For Irish, the compilation of the ICC spoken component will virtually need to be started from scratch. The *Comhrá Corpus of Spoken Irish* (Uí Dhonnchadha *et al.* 2012) (250,000 words approx.) consists mainly of transcribed broadcast discussions, news and interviews, as well as a small number of personal conversations. Broadcast dialogues and news make up approximately 20 percent of the ICC spoken part, therefore, at least 80 percent of the Irish spoken sub-corpus will need to be recorded and transcribed specifically for the ICC, in accordance with GDPR regulations.

3.3. Formatting and annotating the ICC

The most challenging aspect of the ICC compilation relates to general issues of corpus design and comparability across languages. In comparison, the technical issues, though some are laborious, are not particularly challenging. Some of the legacy corpora being used, including ICE-Ireland, needed to be converted to XML format. As the ICE design uses 2,000-word extracts, these needed to be selected and annotated with appropriate metadata.

The ICC uses TEI P5 XML as a common data format, and it will also attempt to harmonize the mark-up of the individual national components. One of the still open

[23] http://agd.ids-mannheim.de/index_en.shtml

questions concerns the part-of-speech (POS) tagging scheme. There are many national tagging systems that could be used to tag the individual ICC languages. However, the national tagsets reflect various linguistic theories, and they also differ formally, so that the tagsets render individual linguistic categories to some extent differently. This is why Universal Dependencies (UD; Nivre *et al.* 2016) was introduced, as a standard for consistent annotation of morphology and syntax across many languages. UD are becoming widely accepted by the community, so that they present an obvious solution for the ICC in the long run. However, currently, the size and quality of UD training data for the individual languages vary considerably, which means that, for some languages, the accuracy of UD tagging could prove significantly lower than that of their national taggers. However, there is the possibility of using the national taggers and converting the tagged output to UD format.

4. Making the ICC available

As discussed above, the central idea of collecting data for the ICC was to re-use as much as possible already existing linguistic resources. In terms of the ICC accessibility and distribution, we were initially hoping to be able to gather all the ICC components centrally with CC licenses and make them accessible through an online interface suitable for contrastive research. We were also hoping to offer the data for download to researchers, in order to be processed with their own tools and methods. However, in the course of the project (our first meeting took place in 2017), both of these options have become major stumbling blocks.

Given the fact that the copyright issues are still not resolved satisfactorily across the ICC languages, and that there is currently no frontend that would support contrastive language research, we plan to make the ICC available to the community through several corpus query interfaces on various project sites. The user interfaces being currently considered are *KorAP*, *KonText* and *Korp*.

KorAP[24] is an open-source corpus analysis platform that has been developed at IDS Mannheim since 2012 as successor of the COSMAS II system, which is used by over 45,000 German linguists (Bański *et al*. 2013). Apart from the support of unlimited, multi-level annotations and dynamically definable virtual corpora, *KorAP* has some features

[24] https://github.com/KorAP, https://korap.ids-mannheim.de/

that make it particularly suitable for use within the ICC. *KorAP* has been designed to be able to query corpora distributed over different locations, so that it will be able to handle the expected complicated license conditions in an optimal way. Furthermore, *KorAP* is already used for contrastive research within the EuReCo project (Kupietz *et al*. 2020) and, in this context, is being further developed together with the Romanian and Hungarian academies (Cosma and Kupietz 2019; Diewald *et al*. 2019). *KorAP* supports various search query languages, such as *Poliqarp*,[25] the CQP variant developed for the *Polish National Corpus*, and can thus be easily adopted by experienced users from different communities, but also by inexperienced users via the so called 'query by match' mechanism, which allows constructing and learning complex annotation queries by selecting (i.e. clicking on) annotation elements of query hits.

KonText (Machálek 2020) is an advanced, highly customizable open-source corpus query interface that supports various corpus types; for instance, detailed views of spoken corpora can be rendered as dialogues with clear indication of speaker turns and overlaps, as well as audio playback. *KonText* is a mature software developed at the *Czech National Corpus* and deployed also by other centers. The development of *KonText* takes place on GitHub,[26] where developers and users are welcome to contribute in different ways —fixing/improving code, reporting bugs or discussing new features. Among the recently implemented functionalities, there is a UD tagset support in the Tag Builder widget and support for displaying the UD syntactic trees. We believe that the additional functionality will provide a user-friendly experience for working with the ICC corpora in *KonText*.

The *Korp* search engine, used by the *Language Bank of Finland*, in addition to providing access to the ICC-FI, may also provide hosting services for other ICC components. *Korp* is an MIT licensed corpus search tool which is developed by the Swedish Språkbanken.[27] The software includes a user-friendly frontend; its backend is based on IMS Open Corpus Workbench.[28] *Korp* is currently in active production use in Sweden, Finland, Estonia, Norway, Iceland and Denmark.[29]

---

[25] nkjp.pl/poliqarp/
[26] https://github.com/czcorpus/kontext
[27] https://spraakbanken.gu.se/en
[28] http://cwb.sourceforge.net/download.php
[29] https://korp.keeleressursid.ee/#?stats_reduce=word&cqp=%5B%5D (Tartu, Estonia); http://gtweb.uit.no/korp/#?cqp=%5B%5D&stats_reduce=word (Tromsø, Norway); https://malheildir.arnastofnun.is/?mode=rmh2018#?stats_reduce=word&isCaseInsensitive&searchBy=word&cqp=%5B%5D (Reykjavík, Iceland);

Other options as possible corpus management systems are being explored as well, for example, TEITOK[30] (Janssen 2016). This web-based platform allows viewing, creating and editing corpora with structural mark-up and linguistic annotation. It has a modular design, which supports both text and audio and has an attractive and flexible query interface.

The individual national ICC components are being finished at a different pace: some of the written components are finished and ready to be released very soon, some are only in initial stages. The written and spoken components are collected separately, the blogs are planned to be collected centrally for each language. Therefore, the individual parts will be released separately as they become available.

5. Conclusions and future work

The ICC is, in a way, a unique 'grassroots' collaborative effort of national teams and individuals. The simple idea around data sustainability, with which the ICC started, has proved much more complex than anticipated. Although there is a vast amount of various linguistic resources that were collected at various times and places, often funded from public resources, their wider use often clashes with their restrictive user licenses. Even though the ICC sub-corpora with one million words per language are in today's terms small in size and the text samples are short, it is proving, in many cases, that this is not a sufficient case for exemption. As collecting linguistic data, other than harvesting the web, is a costly and time-consuming activity, the sustainability and accessibility of those data should ideally be ensured beyond the existence of the individual projects they have been collected for. Efforts in this direction have certainly greatly advanced. Sophisticated linguistic infrastructures, such as CLARIN,[31] provide easy and sustainable access to digital language data. However, coordinated creation of language resources is not a part of their mission. A complex task, such as compilation of a carefully sampled comparable corpus, is therefore beyond the reach of individual researchers or even teams.

Despite the many challenges, the ICC will provide valuable material for contrastive languages studies and many other kinds of linguistic research. It has a greater breadth and

http://alf.hum.ku.dk/korp/#?stats_reduce=word&corpus=lspconstructioneb1,lspconstructioneb2&cqp=%5B%5D (Denmark).

[30] https://wiki.tei-c.org/index.php/TEITOK

[31] https://www.clarin.eu/

variety of written and spoken genres than found in many large modern web-sourced corpora. Its focus on spoken data differentiates it from any other comparable corpora. For some languages, the ICC provides the impetus for spoken corpus collection. Even though the focus of the ICC is on European languages, from a typological point of view, it represents all the major varieties. With the recent addition of Chinese, the ICC will face new challenges but at the same time open up new avenues in contrastive linguistic research, including linguistic annotation. This will, hopefully, be an impetus for a development of new state-of-the-art query interfaces for this type of research.

REFERENCES

Aijmer, Karin and Bengt Altenberg eds. 2013. *Advances in Corpus-based Contrastive Linguistics: Studies in Honour of Stig Johansson.* Amsterdam: John Benjamins.

Bański, Piotr, Joachim Bingel, Nils Diewald, Elena Frick, Michael Hanl, Marc Kupietz, Piotr Pęzik, Carsten Schnober and Andreas Witt. 2013. KorAP: The new corpus analysis platform at IDS Mannheim. In Zygmunt Vetulani and Hans Uszkoreit eds. *Human Language Technologies as a Challenge for Computer Science and Linguistics. Proceedings of the 6th Language and Technology Conference*. Poznan: Uniwersytet im. Adama Mickiewicza w Poznaniu, 586–587.

Calzolari, Nicoletta, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asunción Moreno, Jan Odijk and Stelios Piperidis eds. 2016. *Proceedings of the 10th International Conference on Language Resources and Evaluation*, LREC 2016. Portorož: European Language Resources Association.

Cosma, Ruxandra and Marc Kupietz. 2019. On design, creation and use of the Reference Corpus of Contemporary Romanian and its analysis tools. CoRoLa, KorAP, DRuKoLA and EuReCo. *Revue Roumaine de Linguistique*, 64/3. Editura Academiei Române.

Crystal, David. 2004. *The Language Revolution*. London: John Wiley & Sons.

Diewald, Nils, Verginica Barbu Mititelu and Marc Kupietz. 2019. The KorAP user interface. Accessing CoRoLa via KorAP. *Revue Roumaine de Linguistique* 64/3: 265–277. http://www.lingv.ro/images/RRL%203%202019%2006-%20Diewald.pdf

Ebeling, Jarle and Signe Oksefjell Ebeling. 2013. *Patterns in Contrast.* Amsterdam: John Benjamins.

Granger, Sylviane. 2010. Comparable and translation corpora in cross-linguistic research. Design, analysis and applications. *Journal of Shanghai Jiaotong University* 2: 4–21.

Greenbaum, Sidney ed. 1996. *Comparing English Worldwide*. Oxford: Clarendon Press.

James, Carl. 1980. *Contrastive Analysis.* London: Longman.

Janssen, Maarten. 2016. TEITOK: text-faithful annotated corpora. In Calzolari *et al*. eds, 4037–4043.

Kallen, Jeffrey L. and John Kirk. 2007. ICE-Ireland: Local variations on global standards. In Joan C. Beal, Karen P. Corrigan and Hermann L. Moisl eds. *Creating and Digitizing Language Corpora*. London: Palgrave Macmillan, 121–162.

Kallen, Jeffrey L. and John Kirk. 2008. *ICE-Ireland: A User's Guide.* Belfast: Cló Ollscoil na Banríona.

Kilgarriff, Adam, Michael Rundell and Elaine Uí Dhonnchadha. 2006. Efficient corpus development for lexicography: Building the *New Corpus for Ireland*. *Language Resources & Evaluation* 40/2: 127–152.

Kirk, John and Anna Čermáková. 2017. From ICE to ICC: The new *International Comparable Corpus.* In Piotr Bański, Marc Kupietz, Harald Lüngen, Paul Rayson, Hanno Biber, Evelyn Breiteneder, Simon Clematide, John Mariani, Mark Stevenson and Theresa Sick eds. *Proceedings of the Workshop on Challenges in the Management of Large Corpora and Big Data and Natural Language Processing* (CMLC-5+BigNLP). Mannheim: Institut für DeutscheSprache, 7–12. https://ids-pub.bsz-bw.de/frontdoor/deliver/index/docId/6243/file/2.+Kirk_Cermakova_From_ICE_to_ICC_2017.pdf

Kirk, John and Gerald Nelson. 2018. The *International Corpus of English* project: A progress report. *World Englishes* 37/4: 697–716.

Komrsková, Zuzana, Marie Kopřivová, David Lukeš, Petra Poukarová and Hana Goláňová. 2017. New spoken corpora of Czech: ORTOFON and DIALEKT. *Jazykovedný časopis* 68/2: 219–228.

Kopřivová, Marie, Zuzana Laubeová, David Lukeš and Petr Poukarová. 2019. ORATOR v1: Korpus monologů. Ústav Českého národního korpus FF UK, Praha. https://www.korpus.cz

Křen, Michal, Václav Cvrček, Tomáš Čapka, Anna Čermáková, Milena Hnátková, Lucie Chlumská, Tomáš Jelínek, Dominika Kováříková, Vladimir Petkevič, Pavel Procházka, Hana Skoumalová, Michal Škrabal, Petr Truneček, Pavel Vondřička and Adrian Jan Zasina. 2016. SYN2015: *Representative Corpus of Contemporary Written Czech*. In Calzolari *et al.* eds., 2522–2528.

Kupietz, Marc, Nils Diewald, Beata Trawiński, Ruxandra Cosma, Dan Cristea, Dan Tufiş, Tamás Váradi and Angelika Wöllstein. 2020. Recent developments in the European Reference Corpus EuReCo. In Sylviane Granger and Marie-Aude Lefer eds. *Translating and Comparing Languages: Corpus-based Insights*. *Selected Proceedings of the Fifth Using Corpora in Contrastive and Translation Studies Conference*. Louvain-la-Neuve: Presses universitaires de Louvain, 257–273.

Machálek, Tomáš. 2020. KonText: Advanced and Flexible Corpus Query Interface. In Calzolari, Nicoletta, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asución Moreno, Jan Odijk and Stelios Piperidis eds. *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille: European Language Resources Association, 7003–7008.

Mauranen, Anna. 1998. Will 'translationese' ruin a contrastive study? *Languages in Contrast* 2/2: 161–185.

McEnery, Tony and Andrew Hardie. 2013. The history of corpus linguistics. In Keith Allan ed. *The Oxford Handbook of the History of Linguistics*. Oxford: Oxford University Press, 727–746.

McEnery, Tony and Richard Xiao. 2004. *The Lancaster Corpus of Mandarin Chinese.* https://www.lancaster.ac.uk/fass/projects/corpus/LCMC/lcmc/lcmc_info.htm

Nivre, Joakim, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty and Daniel Zeman. 2016. Universal Dependencies v1: A Multilingual Treebank Collection. In Calzolari *et al*. eds, 1659–1666.

Sharoff, Serge, Reinhard Rapp, Pierre Zweigenbaum and Pascale Fung eds. 2013. *Building and Using Comparable Corpora*. Berlin: Springer.

Uí Dhonnchadha, Elaine, Alessio Frenda and Brian Vaughan. 2012. Issues in designing a *Corpus of Spoken Irish*. In Calzolari, Nicoletta, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asunción Moreno, Jan Odijk and Stelios Piperidis eds. *Proceedings of the 8th International Conference on Language Resources and Evaluation*, LREC 2012. Istanbul: European Language Resources Association.

Yadava, Yogendra, Andrew Hardie, Ram Lohani, Bhim N. Regmi, Srishtee Gurung, Amar Gurung, Tony McEnery, Jens Allwood and Pat Hall. 2008. Construction and annotation of a corpus of contemporary Nepali. *Corpora* 3/2: 213–225.

Corresponding author
Anna Čermáková
Institute of the *Czech National Corpus*
Charles University Prague
nám. J. Palacha 2
116 38, Prague
Czech Republic
e-mail: anna.cermakova@ff.cuni.cz

received: February 2020
accepted: June 2021