CHALLENGES IN BUILDING WRITTEN AND SPOKEN COMPARABLE CORPORA IN CORPUS LINGUISTICS

Turakulova Dilafruz Muxammadsoli qizi
Uzbekistan State World Languages University
E-mail: dilimturaulova1309@gmail.com

Abstract: This article covers challenges while creating written and spoken comparable corpora. Moreover, it suggests possible solutions to some problems on the creation and use of various linguistic corpora.

Keywords: ICC corpus; contrastive linguistics; comparable corpus; ICE corpus; data sustainability; copyright.

Even though translation (parallel) corpora make up the majority of corpus-based contrastive studies, similar data is also being used more and more frequently (see, e.g., Mauranen 1998; Aijmer and Altenberg 2013). The International Comparable Corpus (ICC), a collaborative project of currently twelve national teams, aims to provide highly comparable datasets of spoken and written registers across a range of carefully matched text categories. This contrasts with extensive comparable corpora mined from the web which are used in natural language processing for the development of machine translation and crosslingual information retrieval systems (Sharoff et al. 2013). The ICC begins with the notion of linguistic data reuse and so contributes to the topic of data sustainability on the one hand, and the current dearth of comparable datasets for contrastive investigations on the other. The contemporary landscape of contrastive studies is largely centered on linguistic comparisons between pairs, with English being used as one of those languages quite frequently. A brief scan of the last five volumes (15–19) of Languages in Contrast, the premier publication in contrastive linguistics, swiftly confirms this trend. With the exception of two special issues, 39 of the 47

research articles that were published contained comparisons of two languages, and 38 of them involved English. There is no question that a lack of appropriate linguistic resources is one of the reasons for this two-language, English-focused study. Another interesting finding is that, with a few significant exceptions, all the study is virtually restricted to written language.

The ICE family corpora project was established in the early 1990s, when corpus linguistics research was only beginning to intensely examine issues of data sampling and data comparability and when huge corpora like the British National Corpus started to be generated (McEnery and Hardie 2013). The ICE sampling frame is composed of 15 spoken discourse scenarios and 17 written text kinds, and it is based on identical-length extracts (2,000 words) arranged according to text type categories (for more details see Greenbaum 1996: 3). The proportion of spoken to written languages represented in the ICE corpus has been maintained for the ICC, but a few text categories have been changed to allow for cross-linguistic comparison. Comparability of texts between languages is a challenging problem (see, e.g., Granger 2010). Contrastive cross-linguistic comparisons are based on the idea of "comparability," a "background of sameness" (James 1980: 169) that allows for a comparison of language differences. Therefore, it follows that comparison is always a matter of degree and, as James (1980: 168) notes, "does not imply absolute identity, but merely a degree of shared similarity." Practically speaking, the ICC attempts to create data comparability by matching different text properties, such as time of generation or text type, with varying degrees of success. The matching of text kinds across languages is far more difficult, even when parameters like the year of publication may be rather straightforward. Other corpus projects have demonstrated that certain text kinds may be very culturally distinctive. For instance, it proved impossible to discover science fiction texts in the Nepali National Corpus (Yadava et al. 2008), while McEnery and Xiao (2004) explain how to match FLOB corpus text genres to the Lancaster Corpus of Mandarin Chinese. This also applied to the ICC; for instance, the national teams

elected not to include the two text categories contained in the spoken component of the ICE corpora, legal cross-examinations and legal presentations.

The ICC employs the written text types of the ICE-Ireland corpus for its English component (Kallen and Kirk 2007, 2008). In addition to these written writings from 1990 to 1994 (Kallen and Kirk 2008: 65–79 provides a bibliography), it was deemed desirable to include texts that are mostly contemporaneous, i.e., texts produced after 2000 whenever possible (see Section 3.1). It was also determined that an element of online texts should be incorporated to reflect the evolving nature of contemporary communication (see, for example, Crystal 2004). In light of this, ICC corpora will no longer include the non-printed texts category (existing in ICE) in favor of including blogs, which will be gathered for all relevant languages, including English.

Languages with extensive national corpora are in a comparably better position to construct the ICC written components as they already have data to draw from. As opposed to web-crawled corpora, the SYN-series corpora of modern written Czech being created at the Czech National Corpus (CNC) can be classified as traditional, with clearly specified composition, reliable annotation, and high-quality text processing. SYN2015, a representative reference corpus with a solid balance of fiction, non-fiction, newspapers, and periodicals, is another entry in the SYN series. It was put together with diversity in mind, thus in addition to including all the registers used frequently in written (printed) Czech, each register also includes a wide range of texts from different authors, publishers, etc. In June 2019, the institute's corpus query engine Kontext made the Czech written component of the ICC (ICC-CZ) internally accessible based on SYN2015 .

One instance where it is difficult, or even impossible, to reuse already existent resources is the compilation of the Finnish component of the ICC. In Finland, a 2017 study into the data that already existed and matched was conducted. The Language Bank of Finland corpora were determined to be the most promising source of data for the ICC corpus. The Language Bank of Finland, run by the FIN-CLARIN consortium, has been collecting and providing centralized access to various corpora compiled by

consortium members for the past ten years. These members include the majority of Finnish academic institutions dealing with linguistic data. A CC-BY or CC-BY-NC license was intended to be used to collect a separate collection for the ICC corpus. With such licenses, some of the recognized corpora from the Language Bank of Finland were in fact easily downloadable and distributable. However, the remaining texts that were determined to be appropriate for the ICC are only accessible through a variety of more stringent permissions that have been granted by the many universities, research institutions, private businesses, or even private persons who control the relevant rights. Most attempts to renegotiate the stricter licensing with the owners of those rights were fruitless. As a result, many of the current appropriate corpus materials cannot be reused because of the severe permissions and distribution restrictions. The distribution of the ICC corpus will need to be reviewed because a similar scenario has occurred with other languages as well. Making the data accessible through the relevant institutional corpus query interfaces, such as the Korp15 provided by the Language Bank of Finland, is one of the alternatives suggested. In general, spoken language is frequently underrepresented in collections of linguistic resources, and some categories are not even present in the major national corpora, necessitating collection and transcription. The standard procedure while transcribing spoken data is to protect the anonymity of participants by obscuring personal and identifiable references in the transcriptions and, when necessary, bleeping the pertinent audio files. This is now a compulsory requirement under the new General Data Protection Regulation (GDPR) of the European Union, and care must be made to not hold any unneeded personal or identifiable information. The human voice itself can be regarded as a distinguishing characteristic in a spoken corpus.. Therefore, this matter must be addressed in new consent agreements with participants for the newly acquired data, and it may also need to be taken into account in the case of earlier recordings.

In a way, the ICC represents a special "grassroots" collaboration of national teams and individuals. The ICC's original, straightforward notion of data sustainability has proven to be much more difficult to implement than anticipated. Although there are

many different language resources that were amassed over time and space and frequently paid for using public funds, their limited user licenses frequently prevent their wider use. Even while the one million words per language ICC sub-corpora are little by today's standards and the text samples are brief, this is frequently shown that this is not a strong enough argument for exemption. Since gathering linguistic data outside of web harvesting is an expensive and time-consuming process, it is important to ensure that the data are accessible and sustainable beyond the lifespan of the specific projects for which they were gathered. There is little doubt that efforts in this area have come a long way. Modern linguistic infrastructures like CLARIN make it simple and enduring to access digital language data. Their aim does not include the coordinated creation of linguistic materials, nevertheless. Therefore, an intricate undertaking like compiling a thoroughly sampled comparable corpus is beyond the capabilities of lone researchers or even teams.

REFERENCE

1. Aijmer, Karin and Bengt Altenberg eds. 2013. Advances in Corpus-based Contrastive Linguistics: Studies in Honour of Stig Johansson. Amsterdam: John Benjamins.
2. 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.
3. 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.