RiCL Research in Corpus Linguistics

Challenges of combining structured and unstructured data in corpus development

Tanja Säily[a] – Jukka Tyrkkö[b]
University of Helsinki[a] / Finland
Linnaeus University[b] / Sweden

Abstract – Recent advances in the availability of ever larger and more varied electronic datasets, both historical and modern, provide unprecedented opportunities for corpus linguistics and the digital humanities. However, combining unstructured text with images, video, audio as well as structured metadata poses a variety of challenges to corpus compilers. This paper presents an overview of the topic to contextualise this special issue of *Research in Corpus Linguistics*. The aim of the special issue is to highlight some of the challenges faced and solutions developed in several recent and ongoing corpus projects. Rather than providing overall descriptions of corpora, each contributor discusses specific challenges they faced in the corpus development process, summarised in this paper. We hope that the special issue will benefit future corpus projects by providing solutions to common problems and by paving the way for new best practices for the compilation and development of rich-data corpora. We also hope that this collection of articles will help keep the conversation going on the theoretical and methodological challenges of corpus compilation.

Keywords – structured data; unstructured data; metadata; rich data; corpus annotation; corpus design

As an evidence-based and empirical discipline, corpus-linguistic research relies on the quality and composition of the primary data. Consequently, the principles and methods of compiling corpora and concepts such as representativeness and sample size have been central concerns in corpus linguistics since the discipline first emerged in the 1960s (cf. Francis and Kučera 1964; Biber 1993; McEnery and Hardie 2012). Even today, after more than half a century of theoretical and technological advances, many questions related to corpus compiling remain current and relevant, and new multimodal and linked data types present entirely new challenges to corpus developers.

Over the last twenty years, increasing attention has understandably been paid to so-called mega-corpora, which differ from traditional corpora in several ways, most especially in the much more cursory approach that is by necessity taken to strict

sampling and inclusion criteria (see, for example, Davies 2012; Hundt and Leech 2012). Nevertheless, linguistic datasets comprising billions of words that would have been fantastical dreams only a decade or two ago are now everyday research tools, and the new opportunities they afford have revolutionised many aspects of linguistic inquiry (cf. Tichý 2018; Tyrkkö 2020). In addition to datasets specifically compiled for linguistic research, the newfound availability of social media data, repositories of born-digital documents, and digitised archives of heritage data make it possible to apply corpus-linguistic methods to vast collections of texts that, in some cases, approach the threshold between sample and population.

At the same time, however, small- and medium-sized corpora that match the original definitions of linguistic corpora more closely also continue to be used and developed. Exciting and attractive as mega-corpora of hundreds of millions or billions of words are, they are usually also messy and unpredictable, lacking in metadata, and difficult to study from sociolinguistic or philological perspectives (see, for example, Koplenig 2017). Smaller corpora, on the other hand, can provide valuable insights into these and other areas of inquiry where more data is needed at the linguistic, metalinguistic and metatextual levels. Not only can layers of automatic and semi-automatic annotation be applied more reliably to the language in smaller corpora, but other analytical features can also be made searchable. Multimodal features such as paratextual devices, phonetic and prosodic characteristics, gestures and facial expressions can be annotated into the corpora and be provided as linked data, such as hyperlinks to online repositories of facsimile images, audio and video data, etc. As a consequence of technological developments, linguistic corpora comprising these kinds of 'rich' data have become increasingly realistic to compile, but that does not mean that all the related challenges are already solved (cf. Hiltunen *et al*. 2017).

The contributors to this special issue address a variety of issues that arise from the complexities of linguistic phenomena and their associated metadata. In digital humanities and data science, the terms 'structured data' and 'unstructured data' refer to the way in which data is stored in a computer system (cf. Schöch 2013). When data is described as structured, it is made up of clearly defined and mutually exclusive variables, which can be stored as a database and queried with great efficiency, accuracy, and speed. Structure can be added to linguistic data by, for example, tokenising the text into lexical units and assigning each token linguistic information, such as a word class

or a semantic category. Likewise, metadata describing the texts or authors included in a corpus can be broken down into systematic variables, such as year of publication, genre, or level of education, which facilitate focused queries or the comparison of search results between subsections of the dataset. Importantly, whenever unstructured data is transformed into structured data, many theoretical, analytical, and practical decisions have to be made. The compilers will have to decide on the most appropriate way of selecting and defining independent variables, the appropriate level of granularity that is both sufficiently descriptive but also practically and theoretically feasible to implement, and striking the right balance between description and analysis (cf. Meurman-Solin and Nurmi 2007).

The special issue focuses on three main types of challenge: multimodality, principles and practices of corpus annotation, and the complexities of historical data. **Marie-Louise Brunner** and **Stefan Diemer** address the challenges of annotating nonverbal elements into conversational corpora, which the authors argue is crucially important. In order to transform multimodal and unstructured elements such as gestures, facial expressions, and physical stance into useful structured annotations, it is necessary first to develop a robust transcription system that can be accessed using standard query tools and does not require excessive prior familiarity from end-users. Using their work on the *Corpus of Video-mediated English as a Lingua Franca Conversations* (ViMELF 2018) as an example, the authors show that many existing transcription schemes are not readily usable in corpus-based research due to their complexity and lack of transparency. The authors describe the feature selection process that focuses on salient features and show how the elements are annotated into the corpus. Finally, examples are given of studies making use of the annotated corpus.

Camille Debras discusses the annotating of gestures and other visual features in video recordings of political speeches included in the *Diachronic Corpus of Political Speeches* (DCPS), currently being compiled by an international team at Linnaeus University, the University of Paris Nanterre, and Tampere University. Introducing the open-source video editing tool ELAN (cf. Wittenburg *et al.* 2006), Debras discusses the wide variety of multimodal features that could be annotated for the benefit of multimodal political discourse analysis, such as camera framing and camera angle, continuity of filming, interpausal and intonation units, and gestures. The author focuses on revealing the rich data associated with gestures made with different parts of the body

and the many functions that they may serve in performative discourse. A short repertoire of gestures commonly used by politicians is also provided to show how the data could be used. The article ends with a set of practical recommendations for researchers working on similar data.

The contribution by **Nele Põldvere, Johan Frid, Victoria Johansson** and **Carita Paradis** draws our attention to one of the key challenges of compiling multimodal corpora, namely, how to release the multimodal primary data to the research community. Focusing on the *London-Lund Corpus* 2 (LLC2), compiled at Lund University (cf. Põldvere *et al.* in press), the authors discuss both the technical and legal challenges of releasing the audio recordings. Starting with a very useful overview of transcribed spoken language in corpora and a survey of British English corpora with audio data, the article focuses on the technical aspects of aligning audio and text using timestamps, and the anonymisation of the audio files in accordance with the *European Union's General Data Protection Regulation* (GDPR). Noting that previously used techniques, such as muting personal names, have the effect of removing potentially important prosodic information, the authors opted to replace tagged segments of the original audio with a non-lexical noise that nonetheless retains the pitch and intensity of the original. The article concludes with discussion of the technique's scalability to larger corpora and a brief overview of the next steps for the LLC2 corpus.

Anna Čermáková, Jarmo Jantunen, Tommi Jauhiainen, John Kirk, Michal Křen, Marc Kupietz and **Elaine Uí Dhonnchadha** discuss the principles and practices of compiling the *International Comparable Corpus* (ICC), modelled after the widely known *International Corpus of English* (ICE) family of corpora. The authors draw attention to a range of issues that reflect the changing of times, such as the need to include linguistic data representative of online use, the pros and cons of reusing pre-existing data as sources, and challenges to do with compiling a multilingual corpus, such as the selection of schema for part-of-speech tagging of multiple languages when the existing language-specific models may reflect different underlying linguistic theories. Another important consideration discussed is the dissemination of the corpus. The initial plan of the project was to make ICC available on one online query platform but, for reasons of copyright restrictions and the lack of a robust interface for contrastive multilingual analysis, the dissemination strategy was changed and now involves multiple query platforms hosted by various project members.

Continuing on the theme of dissemination, **Katrin Menzel, Jörg Knappen** and **Elke Teich** tackle the problem of generating and managing different types of metadata for diachronic corpora according to the FAIR principles (Findable, Accessible, Interoperable, Reusable; Wilkinson *et al*. 2016). The *Royal Society Corpus* (RSC; cf. Kermes *et al.* 2016), which consists of scientific journal articles published by the Royal Society of London in 1665–1996, comes with descriptive and structural metadata inherited from the two databases from which the corpus was compiled, hosted by JSTOR and the Royal Society itself. The authors describe the process of matching and integrating the metadata from these two sources into a cohesive whole. They also illustrate how they enriched the RSC by generating contextual metadata on the fields of discourse of each text, based on topic modelling. Together, these metadata facilitate both (socio)linguistic research and biographical studies of the writers. The authors stress the importance of the FAIR principles in generating metadata that enables reuse of the corpus by a wide variety of researchers.

Lassi Saario, Tanja Säily, Samuli Kaislaniemi and **Terttu Nevalainen** discuss challenges to do with updating legacy corpora. Originally developed decades ago, these corpora are small but carefully compiled and continue to be useful for linguistic research. However, their format is often outdated and ill suited for modern concordancing software. Moreover, enriching them with new linguistic annotation or other metadata would extend their use to new kinds of research questions. The authors illustrate the issues involved by describing the production process of the *Tagged Corpus of Early English Correspondence Extension* (TCEECE). The untagged legacy corpus consists of personal letters written in the long eighteenth century, sampled and digitised from previously published letter editions. Producing the TCEECE involved updating the format of the untagged corpus from COCOA to TEI-XML, normalising historical spellings to improve the output of the tagger developed for Present-day English, tokenisation and part-of-speech tagging by the CLAWS software, and evaluating the accuracy of the tagging. The authors discuss their decisions and come up with solutions for streamlining the process in future projects.

Finally, **Mikko Tolonen, Eetu Mäkelä, Ali Ijaz** and **Leo Lahti** assess the potential for linguistic research of massive historical text databases not compiled according to the corpus-linguistic principles of balance and representativeness. More specifically, they discuss the database of *Eighteenth Century Collections Online*

(ECCO), which is the most comprehensive machine-readable source available for eighteenth-century English printed texts. Unlike the pre-eighteenth century *Early English Books Online* (EEBO), no significant portion of ECCO has been keyed in manually, meaning that researchers need to rely on text automatically recognised through Optical Character Recognition (OCR), the variable quality of which is illustrated by the authors. By comparing ECCO with a harmonised and enriched version of the *English Short-Title Catalogue* (ESTC), which is the most comprehensive collection of metadata on eighteenth-century publications, and by utilising the scant metadata that comes with ECCO itself, the authors are able to quantify the biases of ECCO with respect to, for instance, geography, writers, genres, and reprints (which linguists would often prefer to exclude from their studies). The verdict is promising: despite its biases, ECCO —especially when complemented with ESTC metadata— is a potentially valuable data source, as long as researchers pay close attention to historical source criticism.

As has long been the case in corpus linguistics, knowing their corpus will help scholars account for biases when designing their research but, with big data in particular, that knowledge needs to be quantitative as well as qualitative, and the work may benefit from interdisciplinary collaboration between linguists, other humanities scholars, and data scientists. Arguably, one of the particular domains of the corpus linguist is corpus design, that is, understanding the process of compiling a corpus and knowing the best practices that turn unstructured linguistic data into structured data. The contributions in this special issue each highlight one or more areas of corpus design that require the insights of scholars who have practical hands-on experience of working with corpora. We hope that these articles shed light on timely and relevant issues, raise new questions, and inspire fellow corpus linguists to continue the long tradition of looking for the best practices in our field.

References

Biber, Douglas. 1993. Representativeness in corpus design. *Literary and Linguistic Computing* 8/4: 243–257.

CLAWS. Computer program. Developed by UCREL at Lancaster University. http://ucrel.lancs.ac.uk/claws/ (25 June, 2021.)

Davies, Mark. 2012. Some methodological issues related to corpus-based investigations of recent syntactic changes in English. In Terttu Nevalainen and Elizabeth C. Traugott eds., 157–174.

EEBO = *Early English Books Online*. https://quod.lib.umich.edu/e/eebodemo/
ECCO = *Eighteenth Century Collections Online*. https://www.gale.com/intl/primary-sources/eighteenth-century-collections-online
ESTC = *English Short Title Catalogue*. http://estc.bl.uk
Francis, W. Nelson and Henry Kučera. 1964. *Manual of Information to Accompany a Standard Corpus of Present-Day Edited American English, for Use with Digital Computers*. Providence, Rhode Island: Brown University.
Hiltunen, Turo, Joseph McVeigh and Tanja Säily. 2017. How to turn linguistic data into evidence? In Turo Hiltunen, Joseph McVeigh and Tanja Säily eds. *Big and Rich Data in English Corpus Linguistics: Methods and Explorations*. Helsinki: VARIENG. https://varieng.helsinki.fi/series/volumes/19/introduction.html (24 June, 2021.)
Hundt, Marianne and Geoffrey Leech. 2012. "Small is beautiful": On the value of standard reference corpora for observing recent grammatical change. In Terttu Nevalainen and Elizabeth C. Traugott eds., 175–188.
ICC = *International Comparable Corpus*. https://korpus.cz/icc/languages
ICE = *International Corpus of English*. https://www.ice-corpora.uzh.ch/en.html
Kermes, Hannah, Stefania Degaetano-Ortlieb, Ashraf Khamis, Jörg Knappen and Elke Teich. 2016. The *Royal Society Corpus*: From uncharted data to corpus. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Hèléne Mazo, Asunción Moreno, Jan Odijk and Sterlios Piperidis eds. *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (LREC 2016). Portorož, Slovenia: European Language Resources Association, 1928–1931.
Koplenig, Alexander. 2017. The impact of lacking metadata for the measurement of cultural and linguistic change using the Google Ngram data sets – reconstructing the composition of the German corpus in times of WWII. *Digital Scholarship in the Humanities* 32/1: 169–188.
McEnery, Tony and Andrew Hardie. 2012. *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press.
Meurman-Solin, Anneli and Arja Nurmi eds. 2007. *Annotating Variation and Change*. Helsinki: VARIENG. https://varieng.helsinki.fi/series/volumes/01/ (24 June, 2021.)
Nevalainen, Terttu and Elizabeth C. Traugott eds. 2012. *The Oxford Handbook of the History of English*. Oxford: Oxford University Press.
Põldvere, Nele, Victoria Johansson and Carita Paradis. In press. On the *London-Lund Corpus 2*: Design, challenges and innovations. *English Language and Linguistics* 25/3.
Schöch, Christof. 2013. Big? Smart? Clean? Messy? Data in the humanities. *Journal of Digital Humanities* 2/3. http://journalofdigitalhumanities.org/2-3/big-smart-clean-messy-data-in-the-humanities/ (24 June, 2021.)
TCEECE = *Tagged Corpus of Early English Correspondence Extension*. 2020. Annotated by Lassi Saario and Tanja Säily. Spelling standardised by Mikko Hakala, Minna Palander-Collin, Minna Nevala, Emanuela Costea, Anne Kingma and Anna-Lina Wallraff. Compiled by Terttu Nevalainen, Helena Raumolin-Brunberg, Samuli Kaislaniemi, Mikko Laitinen, Minna Nevala, Arja Nurmi, Minna Palander-Collin, Tanja Säily and Anni Sairio at the Department of Modern Languages, University of Helsinki. https://varieng.helsinki.fi/CoRD/corpora/CEEC/

TEI Consortium, eds. 2020. *Guidelines for Electronic Text Encoding and Interchange*. http://www.tei-c.org/P5/ (24 June, 2021.)

Tichý, Ondřej. 2018. Lexical obsolescence and loss in English: 1700–2000. In Joanna Kopaczyk and Jukka Tyrkkö eds. *Applications of Pattern-driven Methods in Corpus Linguistics*. Amsterdam: John Benjamins, 81–103.

Tyrkkö, Jukka. 2020. The war years: Distant reading British parliamentary debates. In Joacim Hansson and Jonas Svensson eds. *Doing Digital Humanities: Concepts, Approaches, Cases*. Växjö: Linnaeus University Press, 169–199.

ViMELF. 2018. *Corpus of Video-Mediated English as a Lingua Franca Conversations.* Birkenfeld: Trier University of Applied Sciences. http://umwelt-campus.de/case

Wilkinson, Mark D., Michel Dumontier *et al*. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3: 160018.

Wittenburg, Peter, Hennie Brugman, Albert Russel, Alex Klassmann and Han Sloetjes. 2006. ELAN: A professional framework for multimodality research. In Nicoletta Calzolari, Khalid Choukri, Aldo Gangemi, Bente Maegaard, Joseph Mariani, Jan Odijk and Daniel Tapias eds. *Proceedings of the Fifth International Conference on Language Resources and Evaluation* (LREC 2006). Genoa, Italy: European Language Resources Association, 1556–1559.

Corresponding author
Tanja Säily
P.O. Box 24
FI-00014
University of Helsinki
Finland
tanja.saily@helsinki.fi

Helsinki and Växjö, 9 July 2021